Statistics, computation, and adaptation in high dimensions

Ashwin Pananjady Martin

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2020-159
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2020/EECS-2020-159.html

August 13, 2020

Acknowledgement

Research in this dissertation was partly funded by grants from the National Science Foundation and Office of Naval Research.

Statistics, Computation, and Adaptation in High Dimensions

by

Ashwin Pananjady Martin

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Engineering – Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Martin J. Wainwright, Co-chair
Professor Thomas A. Courtade, Co-chair
Professor Adityanand Guntuboyina
Professor Michael I. Jordan

Summer 2020

Statistics, Computation, and Adaptation in High Dimensions

Abstract

Statistics, Computation, and Adaptation in High Dimensions

by

Ashwin Pananjady Martin

Doctor of Philosophy in Engineering – Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Martin J. Wainwright, Co-chair

Professor Thomas A. Courtade, Co-chair

With a focus on designing flexible, tractable, and adaptive methodology for some canonical machine learning tasks, we establish several results for the class of permutation-based models, index models, and Markov reward processes. First, we study permutation-based models in the vector, matrix, and tensor settings, which provide robust representations in "broken-sample" problems and of human-generated data. We design tractable and adaptive methodological solutions for fitting these models that, among other things, narrow statistical-computational gaps conjectured in the literature. Second, we study a subclass of index models—widely used in dimensionality reduction and exploratory data analysis—through a computational lens, focusing on avoiding the (statistical) curse of dimensionality and on achieving automatic adaptation to the noise level in the problem. Our perspective yields efficient algorithms for solving these non-convex fitting problems that come with provable guarantees of sample efficiency and adaptation. Finally, we turn to studying some statistical questions in reinforcement learning, focusing in particular on instance-dependent guarantees for the policy evaluation problem. We show that while some algorithms attain the optimal, "local" performance for this problem, other popular methods fall short and must be modified in order to achieve the desired levels of adaptation.

To Amma, Appa, Papa, and Vidya

Contents

Acknowledgments

I must begin by thanking my advisors Martin Wainwright and Thomas Courtade for tolerating me over all these years.

Martin has been a tremendous intellectual force in my fledgling career, and has molded almost all facets of my work, from the way I think and what I choose to think about, to the way I write and present my research. Right from our chance encounter[1] during the Berkeley visit days, it was clear to me that he was both personable and brilliant in equal measure. Many of my research interests were also shaped by taking his class on theoretical statistics in my first year, and my initial enthusiasm for the subject was largely stoked by the extraordinary clarity of his teaching. Martin's knack for giving just the right amount of guidance is well known, and the breadth of the group's research interests always challenged me to learn new things. Martin was also influential in other ways: he bears sole responsibility for my (over-)dependence on espresso coffee[2], and he has taught me to stand on principle about my email and parking rights. I have strived to learn as much as possible from him during my time at Berkeley, and hope to guide and advise my own students with the generosity and time that Martin invested in me.

Being Tom's first student has been a privilege. It is safe to say that I have met very few people who are anywhere near as sharp, technically creative, and gifted at problem solving as Tom. Tom was the one who convinced me that coming to Berkeley to do a PhD was a "no-brainer", and my interactions with him have consistently shown that this was indeed the right environment for me. I will always treasure the hours that we spent at his (extremely expansive) whiteboard, and the "Aha!" moments that often accompanied those discussions. Tom was instrumental in building both my confidence and skillset during the first few years of my PhD, but also extremely open to giving me independence to explore my own interests later on. I worked with Tom on several topics that do not appear in this thesis [5, 70, 71, 242], and I thank him for giving me the opportunity to explore and learn about many beautiful mathematical areas. Tom is also a gifted writer and communicator, and I have learned a lot from him on these counts.

I must also thank the two other members of my thesis committee. Aditya "Mozart" Guntuboyina has repeatedly amazed me with his clarity of thought and humility, and I am grateful for his time and patience over the course of our collaborations. Thank you, Aditya, for trying your best to fill the many gaps in my statistical knowledge, and I hope that I can continue to badger you in the future! Michael Jordan has also been generous with his time and thoughts; I consider myself very lucky to have witnessed from close quarters his awe-inspiring breadth and vision for the field.

I was also fortunate to work with some exceptional people over the course of internships and (sometimes virtual) visits. Thank you to all the members of the Seminar für Statistik at ETH Zürich for their warmth during my brief visit in October 2015. Thank you to Denny Zhou and Lihong Li for hosting me for an internship at Microsoft Research in summer 2017 that kindled my interest in reinforcement learning. Thank you to Lee Dicker and Dean Foster for an awesome summer 2018 at Amazon New York that exposed me to a variety of interesting, real-world research problems. I must

[1] Prof. Dr. Mr. "Swami" Wainwright has an often-repeated story about how I was "disrespectful" during this interaction, and also during every subsequent interaction to date. I maintain that sass $\neq$ disrespect.

[2] Martin's latte art skills, I am sorry to say, have hovered around "needs improvement" for years now.

particularly call out Dean, who has been incredibly generous with his time and praise over the last couple of years. It has been a privilege to work with someone who has such a broad variety of deep interests, and I have learned an immense amount from you (usually after spending hours thinking about the things that you found obvious!). I would also like to thank Richard Samworth, who has been an extremely fun collaborator through the lockdown of summer 2020. I admire his attention to detail, appetite for long and technical papers, and enthusiasm for English cricket.

While I learned how to do research during my PhD, I also learned how to teach here. I was a graduate student instructor (GSI) for two classes: EECS189 (Introduction to machine learning) taught by Anant Sahai and Stella Yu, and EE229A (Information theory) taught by Tom Courtade. I would like to thank them for giving me the opportunity to witness firsthand their approaches to lecturing, grading, and managing a class. In particular, I am grateful in this regard to Anant Sahai, whose passion, drive, and conviction as a teacher are an inspiration. I also thank my fellow GSIs in EECS189 for a fun semester, and Anant, Jennifer Listgarten, and Jiantao Jiao for inviting me, later on in my PhD, to deliver guest lectures in their classes. On a related note, I myself took many classes at Berkeley; I thank Tom Courtade, Laurent El Ghaoui, Jean Walrand, Martin Wainwright, Venkat Anantharam, Michael Jordan, Steve Evans, Ben Recht, John Canny, Christos Papadimitriou, and Luca Trevisan for being terrific lecturers who brought great energy into the classroom.

I also thank the many other faculty members both in Berkeley and the broader academic community who contributed immensely to my experience in graduate school. Kannan Ramchandran has always exemplified the intellectual energy of Berkeley, and I thank him for our collaborations. Peter Bartlett has influenced many of my interests, and I look forward to continuing our interactions at the Simons Institute. Venkat Anantharam has always been ready to lend an ear when I have had technical questions. Jiantao Jiao has been very generous with his time over the course of the academic job market season. I also thank Pramod Viswanath and Po-Ling Loh for their encouragement and advice over the years, and David Tse for making it possible for me to attend the visit days as an international student. Finally, I thank Rahul Vaze, Andrew Thangaraj, and Sounaka Mishra for introducing me to research as an undergraduate student and for setting me on this path.

It is fair to say that I have learned almost everything I know about research during my PhD from my phenomenal co-authors, whom I list here in (roughly) chronological order: Thomas Courtade, Martin Wainwright, Dong Yin, Dimitris Papailiopoulos, Max Lam, Kannan Ramchandran, Peter Bartlett, Max Fathi, Cheng Mao, Vidya Muthukumar, Dean Foster, Dhruv Malik, Kush Bhatia, Koulik Khamaru, Efe Aras, Kuan-Yun Lee, Avishek Ghosh, Adityanand Guntuboyina, Feng Ruan, Michael Jordan, Anca Dragan, and Richard Samworth. Among this list of amazing people, I would like to particularly call out two of my peers: Cheng Mao, who has been an amazingly fun person to work and hang out with, and I look forward to much more of that going forward; and Feng Ruan, who has spent the last few months trying his best to teach me some Le Cam theory. I should also mention the students who trusted me with "mentoring" them in some capacity during our interactions——Efe Aras, Kush Bhatia, Avishek Ghosh, Koulik Khamaru, Kuan-Yun Lee, and Dhruv Malik——I hope you have learned as much from me as I have from you. I would also like to thank Vivek Bagaria and N Safina Devi, partners-in-crime on my undergraduate research projects.

A big thank you is also owed to the BLISS community and broader EECS and Statistics communities. I would like to thank Gireeja Ranade, Kate Harrison, Vijay Kamble, Varun Jog, Kangwook Lee, Guilia Fanti, Mert Pilanci, Ramtin Pedarsani, Nihar Shah, Rashmi Vinayak, Vasuki Narasimha Swamy, Fanny Yang, Yuting Wei, Reza Abbasi-Asl, Payam Delgosha, Dong Yin, Orhan Ocal, Vidya Muthukumar, Raaz Dwivedi, Soham Phade, Billy Fang, Joseph Borja, Jianbo Chen, Sang Min Han, Mark Velednitsky, Niladri Chatterji, Kush Bhatia, Avishek Ghosh, Vipul Gupta, Laura Brink, Aviva Bulow, Koulik Khamaru, Nick Bhattacharya, Jake Soloff, Efe Aras, Kuan-Yun Lee, Vignesh Subramanian, Wenlong Mou, Banghua Zhou, Jichan Chung, Josh Sanz, Wenshuo Guo, Nived Rajaraman, Chinmay Maheshwari, Kabir Chandrasekher, Leah Dickstein, Tavor Baharav, Alvin Wan, Dimitris Papailiopoulos, Simon Li, Sivaraman Balakrishnan, Yudong Chen, Aaditya Ramdas, Leila Wehbe, Yun Yang, Nhat Ho, Swanand Kadhe, Reinhard Heckel, Ilan Shomorony, Ludwig Schmidt, and Feng Ruan. In particular, thanks to Payam, Orhan, Dong, and Vidya for taking this journey together; it's been a fun ride! Thanks to Fanny for the hikes, concerts, and picture(s), and for your unparalleled intensity. Thanks to Mert, Yuting, Siva, Ilan, Nihar, Rashmi, and Aadi for the career (and life) advice. Thank you Sang Min, Swanand, Reinhard, Ludwig, Kabir, Vipul, Avishek, and Kush for the much-needed levity and breaks from work. On that note, Aymeric Dieuleveut also deserves mention, and I also thank the broader community of ultimate frisbee enthusiasts that joined us every Friday. Finally, I owe thanks to Shirley Salanio for her perennial smile and readiness to help with all things, Kim Kail for entertaining my requests surrounding travel reimbursements and coffee orders, and Eric Arvai for electronic support and barista lessons.

I would also like to thank a few other friends who have made this journey enjoyable. Thanks to the Saras 9th wing for the frequent visits and reminiscence, especially to Pavan, Nath, and Smit. Thanks to Chengtao and Jianbo for the table tennis in Redmond. Thanks to Mayur, Siva, Aadi, Vish, Nadia, and Despina for your company during the AIDS LifeCycle (ALC) ride. Thanks to the broader ALC community for its cheer and optimism, and to the many who donated to a noble cause. I thank Amod, Vaishnavi, Sanjay, and Ramya for the boardgames. Thanks are also owed to Akshay, Liam, Noel, and Jason for being fun roommates, and to Shromona for helping me find my first Berkeley apartment.

Finally, and most importantly, I thank my family for their support through all these years. Let me begin by thanking Muthu uncle, Viji aunty, and Gayathri for taking me in as one of their own; I am honored to be part of this most functional family. Thanks in particular to aunty for setting up a robust support system complete with mouthwatering meals. Next, I thank Swathi and Indira, my own parents, for their love, encouragement, and sacrifices through all these years. Thank you Appa, for raising me with puzzles, games and laughter, and for making your dreams my own. Thank you Amma, for instilling in me a sense of purpose, duty, and discipline, and for that all-important statistical gene. Thanks to Kasturi, my sister, for being there through all our childhood adventures. I am extremely proud of her mental fortitude and conviction, and also of her (ever growing) list of Ivy League degrees. And finally, thank you to Vidya, my spectacularly gifted and beautiful wife who, through her own example, challenges me to be a better person every day. Your kindness, passion, and generosity of spirit are an inspiration, and I am incredibly lucky to have spent all these years with you. Thank you for four wonderful years of college, for teaching me to ski, for biking with me to LA, for your baking prowess, for your selfless care (with Amma and Viji aunty) after my ACL reconstruction surgery, and thank you for being my enthusiastic partner in all things. May this next journey with you be just as exciting and fun as our time in Berkeley!

Chapter 1

Introduction and high-level background

Many rapidly developing fields of scientific inquiry use massive amounts of data to find patterns in natural phenomena that are then further explored and leveraged in order to guide discovery, decision making, and policy. We now have intuitive and easily accessible tools to collect, organize, and visualize data, and consequently, a quantitative approach to data science has pervaded many disciplines. The *high dimensional* nature of many of these datasets raises a host of important and exciting questions that statisticians, engineers, and computer scientists are in a prime position to address, such as:

- How does one model *structure* in natural phenomena?
- Once a model has been chosen, how much *information*, or data, is needed in order to perform a certain estimation or inference task?
- Can this task be performed by a low-complexity algorithm that has the capacity to *scale* to massive data sets?
- Can we use structure in the underlying problem to *benchmark* our algorithms, and design algorithms that take advantage of additional structure if and when it exists?

This thesis is guided by such fundamental questions, and centered around the theoretical and methodological aspects of drawing principled conclusions from large, noisy data sets. In order to distill concrete research directions from the broad questions outlined above, we will set down three criteria according to which procedures will be evaluated. The first is *statistical*: we will focus on developing solutions to inference problems that come with sharp theoretical bounds on their sample efficiency, and on providing information-theoretic lower bounds that guarantee the optimality of procedures for the task at hand. The second is *computational*: we will demand that our procedures are not just statistically optimal but also efficient, or tractable, in that they can be implemented at scale on large datasets. Finally, we will adopt a more fine-grained view of the properties of a procedure by assessing its ability to *adapt* to underlying structure when it exists. In particular, while multiple procedures may be statistically and computationally efficient in a “worst-case” sense over a class of problem instances, the data scientist should prefer the one that enjoys improved efficiency whenever the problem is “simpler” in a specific sense.

We showcase this perspective on three concrete and canonical models in machine learning. The first is the class of permutation-based models, which, while having classical roots [82, 314], has emerged as a robust modeling framework for some modern regression and ranking tasks [59, 275, 307]. The second is the class of index models, which is very popular in statistics [195], econometrics [197] and statistical signal processing [253] as a vehicle of dimensionality reduction and exploratory data analysis. Finally, we consider the class of Markov reward processes, classical representations of stochastic phenomena arising in operations research, communications engineering, robotics, and artificial intelligence [27, 28, 292]. We focus here on estimating the long-term value function of the process, which has widespread utility particularly in applications of reinforcement learning [77]. In an overall sense, our approaches to these problems are firmly grounded within the framework of statistical learning theory, but we will move back and forth between optimization, convex geometry, information theory, statistical signal processing, and control theory.

The rest of this chapter is organized as follows. In Section 1.1, we briefly introduce the general mathematical framework that underlies our development. A reader who is interested more in the applications that we study should feel free to skip to Section 1.2. Here, we guide the reader through the specific models considered in this thesis, focusing in particular on how our perspective asks (and answers) fundamental questions from both the statistical and computational points of view. We also include, for the mathematically inclined reader, brief descriptions of what we see as being the major technical takeaways from the three parts of the thesis. Except for the current chapter, all the chapters of the thesis are essentially self-contained and presented (almost) independently, allowing the reader to browse the contents of a particular chapter solely with the context provided by reading the introduction. In Section 1.3, we include brief descriptions of closely related work that does not appear in this thesis, and most of our notation is detailed in Section 1.4.

1.1 A brief glimpse into the mathematical framework

The paradigm of statistical decision theory has had a preponderant influence on both the mathematics and practice of statistics, and deals with problems of the following type. Suppose that we observe a sample Y from an underlying distribution P, and are interested in the value of some unknown[1] “functional” $\phi(P)$. Any estimator $T(Y)$ can then be evaluated by measuring its *risk*

$$\mathbb{E}\left[\ell(T(Y), \phi(P))\right],$$

where $\ell(\phi, \phi^*)$ measures the loss incurred if the prediction ϕ is made when the underlying “true” functional is ϕ^*. For instance, in this thesis alone, we will use the zero-one loss, squared loss, and ℓ_∞ loss to evaluate our procedures, but many other alternatives exist in the literature. The flexibility to choose the tuple (P, ϕ, ℓ) endows the framework with significant expressive power (see, e.g. the

[1] Our notation in this preliminary section is intentionally non-standard; we will use more specialized notation in the various chapters.

books [167, 192] for particular examples). For the purposes of this thesis, it will be convenient to specialize the framework to the *sequence model* [157], in which for each $i = 1, \dots, N$, we have the scalar observation model

$$Y_i = \phi_i^* + \epsilon_i. \tag{1.1}$$

Here, the sequence $\{\epsilon_i\}_{i=1}^N$ represents zero-mean "noise" in the model, and for now, we will assume that the entries of this sequence are independent and use $\sigma > 0$ to (informally) denote the *noise level*[2]. Thus, the vector $\phi^* \in \mathbb{R}^N$ is the unknown mean of the observations, and we are typically interested in using the sample $Y \in \mathbb{R}^N$ to estimate a functional $g(\phi^*)$. Abusing notation slightly, let $T(Y)$ denote such an estimator; then its risk at ϕ^* now takes the form

$$\mathcal{R}_{N,\sigma}(T(Y), g(\phi^*)) := \mathbb{E}\left[\ell(T(Y), g(\phi^*))\right], \tag{1.2}$$

where the loss function ℓ is the same as before, and we have chosen to be explicit in our notation about both the sample size N and noise level σ.

Statistical modeling and estimation: Consider the case where g is the identity function, in which case the number of parameters of interest grows with the sample size N. In order to produce estimators that have non-trivial power, it is of interest to use prior knowledge about the problem in order to posit the existence of some set $\Omega \subseteq \mathbb{R}^N$, so that $\phi^* \in \Omega$. The tuple (Ω, σ) then specifies our *statistical model*. Typically, fewer than N parameters are necessary to describe Ω. For example, when Ω is equal to the range of a known, d-dimensional subspace, we only need d parameters to describe it, and when Ω is the class of all bounded, "monotone" parameters with non-decreasing entries, the effective number of parameters grows with (but is still less than) the sample size N. In rough terms, the former case is called a *parametric* model, and the latter case a *nonparametric* model since it typically places fewer assumptions on the quantity of interest; see Wasserman [329] for a more detailed discussion.

While the risk $\mathcal{R}_{N,\sigma}(T(Y), g(\phi^*))$ provides information about the behavior of our estimator at a particular ϕ^*, a "good" estimator T is one that has acceptable performance across the entire model class of interest. The *minimax* principle, originally put forth by Wald [326], suggests that estimators be compared according to their worst-case risk

$$\sup_{\phi^* \in \Omega} \mathcal{R}_{N,\sigma}(T(Y), g(\phi^*)). \tag{1.3}$$

A minimax procedure is thus one that has the smallest possible worst-case risk, and while Bayesian approaches are equally popular [25], we will generally use the minimax principle in this thesis. Since we evaluate our estimators in the *non-asymptotic* regime where N is finite, we will be satisfied with *rate-optimality*, i.e., with attaining the smallest possible worst-case risk up to a constant, or even polylogarithmic, factor.

[2] The reader may find it convenient to think of the noise variables as Gaussian, with standard deviation σ.

Information-theoretic limits: In order to evaluate the optimality of procedures, the first step is to produce lower bounds on the worst-case risk (1.3) that depend solely on our observation model (1.1), and to ask if there exists any procedure that is able to attain these lower bounds—without any further restrictions. The development of these information-theoretic lower bounds guides the framing of many of our questions about computation and adaptation. In particular, with the exception of Chapter 7, all the other chapters of this thesis derive the information-theoretic limits for their corresponding problems. Indeed, establishing these lower bounds is not solely a pessimistic exercise; the process of constructing "difficult" problem instances exposes structural properties of the problem at hand, and guides the development of many of our algorithms. We note that the use of information theory in statistical estimation problems is classical, and goes back at least to the 1950s [172, 199].

Demanding computational efficiency: The minimax principle places no restrictions on the underlying procedure T. In particular, it allows the statistician access to infinite computational resources, which is often unrealistic, particularly with the scale of modern, nonparametric problems. How do the fundamental limits of the problem change under the additional restriction that T be computable efficiently? As we will see in many portions of this thesis—particularly Chapters 3 through 6—it is possible that the most natural estimators for the problem are in fact challenging to compute efficiently, and furthermore, that there is a significant gap between the information-theoretic estimation limits and the performance of the best-known, tractable estimators. Can we develop better, efficient estimators that take advantage of structure in the optimization problem? Can we prove complexity-theoretic lower bounds? Our work in these chapters complements many modern investigations of related phenomena; e.g. [26, 44, 207, 277].

Quantifying adaptation: While the worst-case risk (1.3) provides a convenient measure of an estimator's performance, there could be multiple procedures that achieve the same worst-case risk. In such cases, which procedure should one prefer? A common-sense approach is given by the following consideration: If the particular problem at hand—specified by the unknown tuple (ϕ^*, σ) is "easy"—then the statistician ought to prefer the procedure that exhibits improved risk properties. While in classical, parametric statistics, the asymptotic behavior of the maximum likelihood estimator in sufficiently "benign" problems is fully characterized by the (inverse) Fisher information at the unknown ϕ^* (see, e.g., [192]), assessing adaptation in the nonparametric case is often more subtle (see, e.g., [48, 122]). Accordingly, our exploration of adaptation in this thesis is multi-pronged. In Chapter 4, we consider the *adaptation factor* [193, 194]—or index—of a procedure, by evaluating the risk of an estimator on a sequence of simpler model classes $\Omega_1, \ldots, \Omega_M \subseteq \Omega$, and asking for the ratio

$$\max_{1 \le i \le M} \frac{\sup_{\phi^* \in \Omega_i} \mathcal{R}_{N,\sigma}(T(Y), g(\phi^*))}{\inf_T \sup_{\phi^* \in \Omega_i} \mathcal{R}_{N,\sigma}(T(Y), g(\phi^*))} \tag{1.4}$$

to be small. In other words, such an adaptation factor measures how much worse the estimator T is at exploiting the additional structure guaranteed by the inclusion $\phi^* \in \Omega_i$ than the minimax-optimal estimator that knows of this inclusion in advance. An exploration of this factor exposes statistical-computational trade-offs of its own. In Chapter 6, we demand that our procedures adapt to the noise level of the problem, i.e., in the case where $\sigma \downarrow 0$ and the problem gets progressively easier (and eventually, noise-free), we ask if the worst-case risk (1.3) gets progressively smaller as a function of σ. Does this occur at the optimal rate? Finally, in Chapters 8 and 9, we consider the local minimax framework [49, 135, 186] in order to evaluate *instance-specific* adaptation properties, both asymptotically and non-asymptotically.

1.2 Model classes covered in this thesis

We now introduce, at a high level, the model classes that are touched upon in the three parts of this thesis. Our focus here is partly on applications, and partly on introducing the mathematical formulation of the problem within the general framework of Section 1.1. The first sections of the specific chapters provide more detailed introductions to the particular problems.

1.2.1 Permutation-based models

We first consider a class of models in which parametric assumptions can be coupled with *combinatorial* ones in order to produce more flexible models. In particular, we consider the class of permutation-based models, defined informally below.

Definition 1.2.1. *(Informal) A permutation-based model is one in which the set Ω is specified in part by a tuple of (unknown) permutations.*

The incorporation of permutations into our model could be the consequence of our prior knowledge about the problem at hand, or alternatively, could capture a lack of prior knowledge. Indeed, in the specific examples to follow, we will see cases where the permutations model inherent "error" in the observation process, and other cases in which they allow us to incorporate flexibility within our modeling assumptions.

"Broken-sample" regression problems: Traditional statistical procedures make the tacit assumption that the correspondence between the covariates and responses is fully known. But what if this is not actually the case? The lack of correspondence could be implicit in the data (for example, in archaeology or genomic data [149, 267]), or correspondence information may have been intentionally removed (for example, in anonymized data sets [228]). This "broken-sample" problem has classical roots, with applications to record linkage and correspondence estimation in image processing [82, 213]. In Chapter 2, we take an information-theoretic viewpoint on the linear regression problem without correspondence information, and sharply characterize the fundamental limits of estimating the unknown correspondence. In particular, given a covariate (or design) matrix $A \in \mathbb{R}^{N \times d}$, we consider the sequence model (1.1) in which the set Ω can be written as

$$\Omega = \{\phi \in \mathbb{R}^N : \phi \in \mathsf{range}(\Pi A) \text{ for some permutation matrix } \Pi\}.$$

We establish a *phase transition* in the fundamental properties of permutation recovery—using the loss $\ell(\widehat{\Pi}, \Pi) = \mathbf{1}\left\{\widehat{\Pi} \neq \Pi\right\}$ in order to evaluate the permutation $\widehat{\Pi}$ as an estimator—depending on an appropriate notion of signal-to-noise ratio in the problem. Our analysis establishes a statistical baseline for this problem that has since been built upon to study many interesting computational questions in related models (e.g. [1, 133, 147, 247, 285]). Work in this chapter is joint with M. J. Wainwright and T. A. Courtade, and based on the paper [248].

Flexible models for human-generated data: The problem of ranking from pairwise comparison data arises in voting and recommendation systems, where large data sets consisting of noisy and often inconsistent human choices must be consistently aggregated into a ranking in order to inform future recommendations [17, 20, 60]. The crowd labeling problem involves aggregating data labeled by humans—often workers on a crowdsourcing platform such as Amazon Mechanical Turk—which contain inconsistent information due to factors such as heterogeneous levels of worker expertise, varying difficulty levels of the questions being asked, or spammers on these platforms [80, 241, 332]. As an illustrative example of pairwise comparisons, consider modeling the $n \times n$ matrix of pairwise comparison probabilities between n items, illustrated in Figure 1.1(a); the *strong stochastic transitivity*, or SST, model makes the following assumption: the items can be ordered from "best" to worst such that for any triple (p, q, r) respecting this order, the probability that p beats r in a comparison must exceed both the probability that p beats q and the probability that q beats r. Such an assumption is known to be both flexible and representative in several applications [16, 103, 214], and Chapter 3 also collects other applications (spanning psychometrics [314] and crowd-labeling [275]) in which a similar perspective yields a flexible permutation-based model.

Mathematically, our observation model can then be described using the sequence model (1.1) as follows: The entries of the underlying mean ϕ^* can be rearranged into an $n \times n$ matrix of pairwise comparison probabilities, such that this matrix is *bivariate isotonic* with unknown permutations. In particular, we have

$$\Omega = \{\phi \in \mathbb{R}^{n \times n} : \phi = \Pi M \Pi^\top \text{ for a permutation matrix } \Pi \text{ and matrix } M \\ \text{having entries that increase along each row and along each column}\}.$$

Here Π represents the unknown ranking between items, and endows the model with significant flexibility; indeed, the permutation-based SST assumption is significantly more robust to misspecification than its parametric counterparts [16, 103, 214]. In addition, and perhaps surprisingly, it is known that the minimax rate of matrix estimation (under squared ℓ_2 loss) over the class of SST matrices is essentially the same as that over its parametric counterparts [59, 274]. Thus, these permutation-based models occupy a nice "sweet-spot" on the bias-variance tradeoff.

However, existing computationally efficient estimators suffer from sub-optimal rates. Consequently, a "statistical-computational" gap was conjectured in a series of papers on this topic [55, 59, 274, 277]. The focus of Chapter 3 is on the computational question, and we produce an estimator that uniformly outperforms existing, tractable procedures. It obtains minimax-optimal estimation rates in a certain regime of the problem, and in other regimes, it narrows the statistical-computational gap. Work in this chapter is joint with C. Mao and M. J. Wainwright, and based on the paper [211].

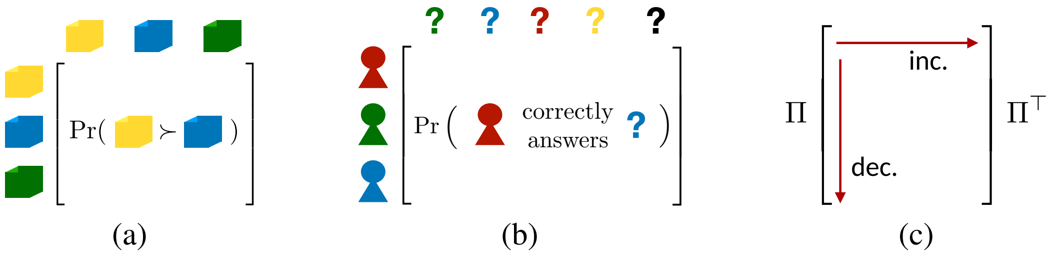


Figure 1.1: Matrix models for human-generated data. (a) Pairwise comparisons, in which entry i, j of the matrix represents the probability with which item i beats item j in a comparison. (b) Psychometrics (and crowd-labeling), in which entry i, j represents the probability with which student (or crowd-worker) i answers question j correctly, and (c) The strong stochastic transitivity, or SST, assumption for pairwise comparisons, which gives rise to a permutation-based model.

Higher-order models for discrete choice data: The SST assumption from Chapter 3 has a higher-order analogue for multiple-comparison data, where a user is presented with a set of d options and must choose one of them[3]. The analogous parametric model in this case is the multinomial logit, or Plackett-Luce model [204, 219, 252], which has found applications in a multitude of disciplines [24, 54]. The parametric assumption in this class of models is also known to be limiting [97], and is generalized by the *simple scalability* assumption from the economics literature [177, 305]. This gives rise to a "tensor" version of the SST assumption; we refer the reader to Section 4.1 for a more comprehensive introduction.

Once again, we study the estimation problem under (squared) ℓ_2 loss, starting by establishing its information-theoretic limits. Surprisingly, whenever $d \geq 3$, we show that there is no longer a statistical-computational gap, and that a simple estimator attains the optimal worst-case risk. The question of adaptation is more closely studied in this chapter, where we show that our estimator enjoys improved performance from both the *statistical and computational* standpoints when there is underlying parametric structure in the tensor. We use the adaptivity index (1.4) in order to evaluate the statistical adaptation properties, showing that interesting statistical-computational gaps in this index manifest under an average-case complexity assumption. Finally, in order to drive home the utility of our estimator for this problem, we show that multiple minimax-optimal estimators in this problem have necessarily sub-optimal adaptation behavior. Work in this chapter is joint with R.J. Samworth, and based on the paper [245].

Technical takeaways: We conclude our discussion by highlighting, for the mathematically inclined reader, a few technical takeaways from this part of the thesis. From Chapter 2, we highlight our application of the *strong converse* to the channel coding problem, which yields sharper results than the often-used Fano's method [302]. While the strong converse is classical in information theory [278, 343], it does not seem to have gained as much popularity in the statistical estimation literature. A second technical contribution, spanning Chapters 3 and 4, is our analysis of

[3] In the pairwise comparison model, we have $d = 2$.

least squares estimators over non-convex sets. Indeed, we present five distinct proof techniques for obtaining such results depending on the structure of the underlying set, and the associated noise process. Finally, we highlight our proof of Theorem 3.4.2 from Chapter 3, which contains many interesting approximation-theoretic properties of multivariate isotonic functions that are of independent interest.

1.2.2 Index models

While the previous part of the thesis was concerned with endowing parametric models with the virtues of their nonparametric counterparts, the index model provides a happy medium between these approaches. On the one hand, it accommodates non-linearity in linear models, and is thus robust to misspecification in these models. On the other hand, it allows the statistician to impose parametric structure in high-dimensional nonparametric models and perform non-linear dimensionality reduction. Let us introduce the index-model within our framework from Section 1.1.

Definition 1.2.2. *(Informal) In an index model, we have access to set of d-dimensional covariate vectors $x_1, \ldots, x_N$, and the set Ω takes the form*

$$\Omega = \{\textit{There exist vectors} \quad \theta_1, \ldots, \theta_k \in \mathbb{R}^d : \phi_i = g(\langle \theta_1, x_i \rangle, \ldots, \langle \theta_k, x_i \rangle) \quad \textit{for} \quad i = 1, \ldots, N\}.$$

Here, the function $g : \mathbb{R}^k \to \mathbb{R}$ is either known, or unknown but belongs to some function class $\mathcal{G}$.

Index models are important dimensionality reduction and exploratory data analysis tools; see the book [197] for a sample of their wide-ranging applications. In this part of the thesis, we study two specific instantiations of index models.

Fitting convex functions in high dimensions: How do humans value different quantities of various commodities? How does one model the long term value of a certain inventory position? What is a good model for a value function in reinforcement learning? Convex regression is a flexible framework within which one can study many such questions that abound in econometrics and operations research [15, 139, 198]. However, being nonparametric, it suffers from the curse of dimensionality, in that the number of samples required to fit a convex function to within a prescribed error tolerance scales exponentially in the ambient dimension of the problem [46, 115, 131]. On the other hand, a natural index model, with the function g chosen to be the coordinate-wise maximum, provides piece-wise linear approximations to convex functions, where the number of linear components k must be chosen by the statistician. In Chapter 5, we show that these "max-affine"

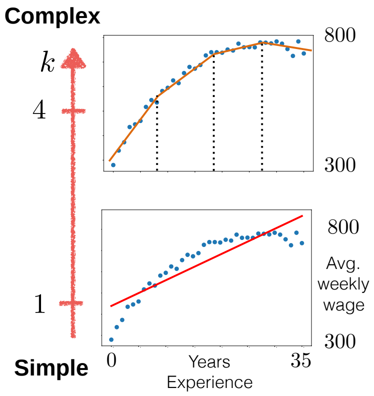


Figure 1.2: Piece-wise linear approximation to two variables in the "Wage and Race" dataset from the Sleuth3 package [304]. The number of affine pieces k controls the complexity of the model, and trades off interpretability with flexibility.

models are natural and interpretable for many applications involving convex regression while admitting provably tractable estimation with a non-convex optimization procedure. An illustration in dimension 1 is provided in Figure 1.2. These results also have implications for fitting convex sets to support function measurements, which is a problem that arises in computational tomography [116, 125, 257], and for the phase retrieval problem [102, 117]. Work in this chapter is joint with A. Ghosh, A. Guntuboyina, and K. Ramchandran, and based on the paper [119].

A noise-adaptive framework for single-index models: The single-index model is the specific case of the index model where $k = 1$, and we study the semiparametric setting in which the function g in unknown. This model has been classically studied [33, 45, 145, 195], but existing algorithms are not adaptive to the noise level of the model. As alluded to before, this is a fundamental property that one should desire from any method: in particular, we would like to be able to perform accurate estimation with significantly fewer samples as the noise level goes to zero. In Chapter 6, we provide a natural, iterative, and computationally tractable procedure that exhibits *automatic adaptation* to the noise level of the problem. The methodology is flexible, and reduces the problem of parameter estimation to estimating well-chosen "inverse", nonparametric functions in low dimensions; for the latter, the statistician can choose any nonparametric method. We present particular consequences for the monotone single-index model, in which $\mathcal{G}$ is the class of all monotonic functions. Work in this chapter is joint with D. P. Foster, and based on the paper [243].

Technical takeaways: We now highlight two technical themes from this part of the thesis. The first is optimization-theoretic: Both chapters in this part of the thesis make use of alternating projections as an algorithm to solve the associated non-convex fitting problem. This algorithm is shown to have automatic noise-adaptation properties for single-index models, and also converges very quickly in a local neighborhood of the optimal solution. In that sense, the two chapters in this part of the thesis are also unified under an algorithmic lens. The second technical component that we highlight is from Chapter 5: Our analysis of the alternating projection algorithm for the max-affine regression problem requires a fine-grained understanding of Gaussian random vectors under truncations to convex sets, and more importantly, involves random matrix theory for these truncated Gaussian random vectors. We expect our techniques to be more broadly applicable to the analysis of iterative algorithms in statistical settings.

1.2.3 Policy evaluation in reinforcement learning

With the proliferation of reinforcement learning (RL) algorithms in a variety of applications such as robotics and competitive gaming, an important consideration is that of algorithm choice. There are currently many procedures that are available to train these RL agents; given a task at hand, which algorithm should one prefer based on the amount of data and computation available? Which of these procedures come with formal theoretical guarantees? Can a model for structure in a dynamical system be used to better delineate the pros and cons of various methods, and can adaptation to such structure be used as a meaningful yardstick to compare these methods?

We take steps towards answering these questions by considering the problem of policy evaluation in finite dimensional (also known as tabular) Markov decision processes (MDPs), which is an important sub-routine used within approximate dynamic programming. The problem setting is very simple: for a "reward" vector $r \in \mathbb{R}^D$ and a matrix $\mathbf{P} \in \mathbb{R}^{D \times D}$ representing the transition matrix of a Markov chain, we are interested in the solution θ^* to the *Bellman equation*

$$\theta^* = r + \gamma \mathbf{P} \theta^*, \tag{1.5}$$

where $\theta^* \in \mathbb{R}^D$ collects the long-term, discounted value of the D states in the MDP. In the generative (or simulator-based) model for RL [166], our observations consist of independent and noisy observations of the pair $(\mathbf{P}, r)$, and our goal is to estimate θ^* from these observations. In that respect, this problem bears resemblance to the classical problem of errors-in-variables regression studied widely in statistics [22]. In contrast to classical work, however, we demand *non-asymptotic* guarantees in the ℓ_∞-norm, and use the simple linear model (1.5) in order to ask and answer some refined adaptation questions. In particular, we use the benchmark of *instance-dependent* performance, which is the strictest form of adaptation considered in this thesis, in order to evaluate various procedures. Chapter 7 introduces this problem in more detail, and provides background on the setting that we consider. Chapter 8 considers the plug-in estimator for the problem, which goes by the names of least-squares temporal difference learning [41], and the "model-based" estimator in the RL literature [294]. We establish refined, instance-specific guarantees and evaluate their optimality from a local minimax standpoint over subclasses of parameter space. Work in Chapter 8 is joint with M. J. Wainwright, and based on the paper [246]. In Chapter 9, we further refine our lower bounds to produce an instance-specific local minimax characterization, both in the asymptotic regime, where the sample size goes to infinity, and in the non-asymptotic, finite-sample regime. Finally, we use this local understanding to evaluate the popular, "model-free" *temporal difference (TD)* learning update [293] and its variants [255, 270]. Overall, our results demonstrate that while the plug-in estimator always adapts to the instance-specific complexity given by the lower bound, the TD update has to be modified with a variance-reduction device in order to attain instance-optimality. Note that the perspective of *non-asymptotic adaptation* is crucial here; indeed, the TD update (for a certain choice of step-size) with Polyak-Ruppert averaging is optimal in the worst-case over a natural, bounded model class, and also asymptotically instance-optimal [255]. Work in Chapter 9 is joint with K. Khamaru, F. Ruan, M. J. Wainwright, and M. I. Jordan, and is based on the paper [169].

Technical takeaways: We conclude by highlighting two techniques from this part of the thesis that are likely to be more broadly useful. In Chapter 8, we introduce a leave-one-out decoupling technique to establish non-asymptotic performance guarantees for the plug-in estimator. Variants of such techniques have also been used in the literature [3, 63], and are likely to be useful in generating non-asymptotic guarantees in ℓ_∞ norm. A second takeaway is in Chapter 9, in which we carry out extensive numerical experiments showing that the finite-sample performance of variance-reduction as applied to stochastic approximation improves upon that of Polyak-Ruppert averaging. We expect this observation to have broader optimization-theoretic consequences.

1.3 Related work not appearing in this thesis

In this subsection, we briefly mention some other papers on closely related topics that the author contributed to during his PhD, along with references and background for the interested reader.

1. Permutation-based modeling

   a. A treatment of the statistical and algorithmic issues at play when denoising the permutation-based, multiple linear regression model, and algorithms for recovering the unknown linear transformation both in the noiseless and noisy settings [247].

   b. A curious dichotomy between "worst-case" and "average-case" assumptions on rankings in the permutation-based matrix regression model, with an emphasis on SST matrices and the subclass of "noisy sorting" matrices [244].

   c. SST matrix estimation in a row-based metric that has applications to learning mixtures of rankings [211]. Part of this paper appears in Chapter 3.

2. Index models

   a. An analysis of the alternating minimization methodology under "small-ball" design assumptions, and consequences for the phase retrieval problem [118].

   b. A new algorithm for phase retrieval [243] based on the idea of a labeling oracle introduced in Chapter 6.

3. Reinforcement learning

   a. Sharp rates for zero-order optimization when applied as a "model-free" method for learning optimal policies from data generated by a linear quadratic control system [210].

   b. Learning from multi-criteria preferences in the context of human-in-the-loop reinforcement learning [31].

1.4 Notation

As alluded to before, the notation used in Chapter 1 was mainly to set up the themes of this dissertation at an abstract level; the chapters to follow will use more specific notation that is tailored to the problems at hand. In this section, we collect some of our notational conventions; some chapter-specific notation will be introduced as and when needed.

For a positive integer n, let $[n] := \{1, 2, \ldots, n\}$. For a finite set S, we use $|S|$ to denote its cardinality. For two scalars a and b, we use the convenient shorthand $a \vee b := \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$. For two sequences $\{a_n\}_{n=1}^{\infty}$ and $\{b_n\}_{n=1}^{\infty}$, we write $a_n \lesssim b_n$ if there is a universal positive constant C such that $a_n \leq C b_n$ for all $n \geq 1$. The relation $a_n \gtrsim b_n$ is defined analogously. We also use standard order-wise notation $f_n = \mathcal{O}(g_n)$ to indicate that $f_n \lesssim g_n$ and $f_n = \widetilde{\mathcal{O}}(g_n)$ to indicate that $f_n \lesssim g_n \log^c n$, for a universal positive constant c. The notation $f_n = o(g_n)$ is used when $\lim_{n\to\infty} \frac{f_n}{g_n} = 0$, and $f_n = \omega(g_n)$ when $g_n = o(f_n)$. We use $c, C, c_1, c_2, \ldots$ to denote universal constants that may change from line to line. The notation v_i denotes the i-th entry of a vector v. We let $v_{(i)}$ denote the i-th order statistic of a vector v, i.e., the i-th largest entry of v. For a pair of vectors (u, v) of equal dimensions, we use the notation $u \preceq v$ to indicate that the difference vector $v - u$ is entry-wise non-negative. The relation $u \succeq v$ is defined analogously. When used for (PSD) matrices, the notation $\succeq$ will denote ordering with respect to the positive (semi-)definite cone. We let e_j denote the j-th standard basis vector; the dimension of this vector can usually be inferred from context.

We use $\text{Ber}(p)$ to denote the Bernoulli distribution with success probability p, the notation $\mathsf{Bin}(n, p)$ to denote the binomial distribution with n trials and success probability p, the notation $\mathsf{Poi}(\lambda)$ to denote the Poisson distribution with mean λ, and $\mathcal{N}(\mu, \Sigma)$ to denote a Gaussian distribution with mean μ and covariance matrix $\Sigma \succeq 0$.

For a (semi-)normed space $(\mathcal{F}, \|\cdot\|)$ and $\epsilon > 0$ let $N(\epsilon; \mathcal{F}, \|\cdot\|)$ denote its ϵ-covering number, i.e., the minimum cardinality of any set $U \subseteq \mathcal{F}$ such that $\inf_{u \in U} \|x - u\| \leq \epsilon$ for all $x \in \mathcal{F}$. Let $\mathbf{1}\left\{\cdot\right\}$ denote the indicator function; we sometimes abuse notation slightly and also use $\mathbf{1}$ to denote the all-ones vector, whose dimension can typically be inferred from context. Let $\mathsf{sgn}(t)$ denote the sign of a scalar t, with the convention that $\mathsf{sgn}(0) = 1$. All logarithms are to the natural base unless otherwise stated.

Specific notation for Part I: Permutations are denoted by small Greek letters (e.g. π) and permutation matrices by capital Greek letters (e.g. Π). Let $\mathfrak{S}_n$ denote the set of all permutations $\pi : [n] \to [n]$, although we sometimes abuse notation to let it also denote the set of all $n \times n$ permutation matrices. Let id denote the identity permutation, where the dimension can be inferred from context. We use $\pi(i)$ to denote the image of an element i under the permutation π. We sometimes use the compact notation y_π (or y_Π) to denote the vector y with entries permuted according to the permutation π (or Π).

Specific notation for Part II: In this part of the thesis, we work exclusively with the ℓ_2 norm, and so throughout this part, we use the notation $\|v\|$ to denote the ℓ_2 norm of a vector v. We also denote the d-dimensional unit shell by $\mathbb{S}^{d-1} = \{v \in \mathbb{R}^d : \|v\| = 1\}$, and $\mathbb{B}^d := \left\{v \in \mathbb{R}^d : \|v\| \leq 1\right\}$ to denote the d-dimensional unit ball.

Specific notation for Part III: This portion of the thesis deals largely with ℓ_∞ norm guarantees, and so we require some notation for operations on vectors. We let $|u|$ denote the entry-wise absolute value of a vector u; squares and square-roots of vectors are, analogously, taken entrywise. Note that for a positive scalar λ, the statements $|u| \preceq \lambda \cdot \mathbf{1}$ and $\|u\|_\infty \leq \lambda$ are equivalent.

Part I

Permutation-based models

Robust methodology for regression and ranking problems

Chapter 2

Linear regression with a broken sample

In this chapter, we study a permutation-based model for regression problems involving vector-valued data. This "broken-sample" setting is one where the unknown permutation serves to model a specific type of uncertainty in correspondence between our samples.

2.1 Introduction

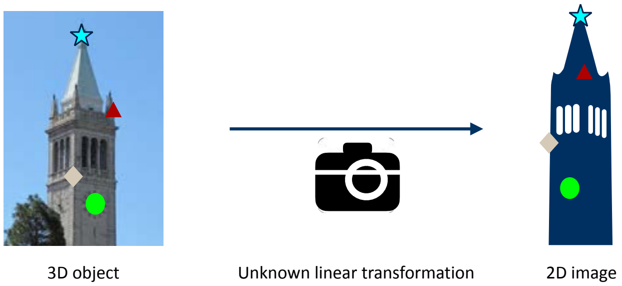
Recovery of a vector based on noisy linear measurements is the classical problem of linear regression, and is arguably the most basic form of statistical estimator. A variant, the "errors-in-variables" model [200], allows for errors in the measurement matrix; classical examples include additive or multiplicative noise [203]. In this chapter, we study a form of errors-in-variables in which the measurement matrix is perturbed by an unknown permutation of its rows.

More concretely, we study an observation model of the form

$$y = \Pi^* A x^* + w, \tag{2.1}$$

where $x^* \in \mathbb{R}^d$ is an unknown vector, $A \in \mathbb{R}^{n \times d}$ is a measurement (or design) matrix, Π^* is an unknown $n \times n$ permutation matrix, and $w \in \mathbb{R}^n$ is observation noise. We refer to the setting where $w = 0$ as the *noiseless case*. As with linear regression, there are two settings of interest, corresponding to whether the design matrix is **(i)** deterministic (the fixed design case), or **(ii)** random (the random design case). There are also two complementary problems of interest: recovery of the unknown Π^*, and recovery of the unknown x^*. In this chapter, we focus on the former problem; the latter problem is also known as unlabeled sensing [307].

The observation model (2.1) is frequently encountered in scenarios where there is uncertainty in the order in which measurements are taken. The model has classical roots, going back to record linkage problems [82] wherein it went by the name of "regression with a broken sample". Other illustrative example include sampling in the presence of jitter [13], timing and molecular channels [268], and flow cytometry [1]. Another such scenario arises in multi-target tracking problems [256]. For example, in the robotics problem of simultaneous localization and mapping [298], the environment in which measurements are made is unknown, and part of the problem is to estimate



$$\begin{pmatrix} p_1 & q_1 & r_1 \\ \vdots & \vdots & \vdots \\ p_n & q_n & r_n \end{pmatrix} \times \begin{pmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{pmatrix} = \Pi \begin{pmatrix} u_1 & u_2 \\ \vdots & \vdots \\ u_n & v_n \end{pmatrix}$$

Figure 2.1: Example of pose and correspondence estimation. The camera introduces an unknown linear transformation, or pose. The unknown permutation represents the correspondence between points, which is shown in the picture via colored shapes, and needs to be estimated.

relative permutations between measurements. Archaeological measurements [267] also suffer from an inherent lack of ordering, and another compelling example of such an observation model is in data anonymization, in which the order, or "labels", of measurements are intentionally deleted to preserve privacy. The inverse problem of data de-anonymization [228] is to estimate these labels from the observations.

Let us also mention two other applications. First, in large sensor networks, it is often the case that the number of bits of information that each sensor records and transmits to the server is exceeded by the number of bits it transmits in order to identify itself to the server [168]. In applications where sensor measurements are linear, model (2.1) corresponds to the case where each sensor only sends its measurement but not its identity. The server is then tasked with recovering sensor identities, or equivalently, with determining the unknown permutation. A second application is to the pose and correspondence estimation problem in image processing [213], illustrated in Figure 2.1. The capture of a 3D object by a 2D image can be modeled by an unknown linear transformation called the "pose", and an unknown permutation representing the "correspondence" between points in the two spaces. One of the central goals in image processing is to identify this correspondence information, which in this case is equivalent to permutation estimation in the linear model.

It is worth noting that both the permutation recovery and vector recovery problems have an operational interpretation in applications. Permutation recovery is equivalent to "correspondence estimation" in vision tasks [213], and vector recovery is equivalent to "pose estimation". In sensor network examples, permutation recovery corresponds to sensor identification [168], while vector recovery corresponds to signal estimation. Clearly, accurate permutation estimation allows for recovery of the regression vector, while the reverse may not be true. From a theoretical standpoint, such a distinction is similar to the difference between the problems of subset selection [322] and sparse vector recovery [50] in high dimensional linear regression, where studying the model selection and parameter estimation problems together helped advance our understanding of the statistical model in its entirety.

2.1.1 Related work

Prior to our own work on the subject, related work had largely focused on x^* recovery. The observation model (2.1) appears in the context of compressed sensing with an unknown sensor permutation [94]. The authors consider the matrix-based observation model $Y = \Pi^* A X^* + W$, where X^* is a matrix whose columns are composed of multiple, unknown, sparse vectors. Their contributions include a branch and bound algorithm to recover the underlying X^*, which they show to perform well empirically for small instances under the setting in which the entries of the matrix A are drawn i.i.d. from a Gaussian distribution. In the context of pose and correspondence estimation, the paper [213] considers the noiseless observation model (2.1), and shows that if the permutation matrix maps a sufficiently large number of positions to themselves, then x^* can be recovered reliably. In the context of molecular channels, the model (2.1) has been analyzed for the case when x^* is some random vector, $A = I$, and w represents non-negative noise that models delays introduced between emitter and receptor. Rose et al. [268] provide lower bounds on the capacity of such channels. In particular, their results yield closed-form lower bounds for some special noise distributions, e.g., exponentially random noise.

Most closely related to our model is the paper by Unnikrishnan et al. [307], which considers the question of when the equation (2.1) has a unique solution x^*, i.e., the identifiability of the noiseless model. The authors show that if the entries of A are sampled i.i.d. from any continuous distribution with $n \geq 2d$, then equation (2.1) has a unique solution x^* with probability 1. They also provide a converse showing that if $n < 2d$, any matrix A whose entries are sampled i.i.d. from a continuous distribution does not (with probability 1) have a unique solution x^* to equation (2.1). While the paper shows uniqueness, the question of designing an efficient algorithm to recover a solution, unique or not, is left open. The paper also analyzes the stability of the noiseless solution, and establishes that x^* can be recovered exactly when the SNR goes to infinity.

Since the publication of the paper on which this chapter is based [248], a line of recent work has considered variants of the observation model (2.1); we mention a few representative papers. Elhami et al. [93] show that there is a careful choice of the measurement matrix A such that it is possible to recover the vector x^* in time $\mathcal{O}(dn^{d+1})$ in the noiseless case. Hsu, Shi, and Sun [147] show that the vector x^* can be recovered efficiently in the noiseless setting (2.1) when the design matrix A is i.i.d. Gaussian. They also demonstrate that in the noisy setting, it is not possible to recover the vector x^* reliably unless the signal-to-noise ratio is sufficiently high. See Section 4 of their paper [147] for a detailed comparison of their results with our own. Our own follow-up work [247] establishes the minimax rate of prediction for the more general multivariate setting, and proposes an efficient algorithm for that setting with guaranteed recovery provided some technical conditions are satisfied. Haghighatshoar and Caire [133] consider a variant of the observation model (2.1) in which the permutation matrix is replaced by a row selection matrix, and provide an alternating minimization algorithm with theoretical guarantees.

2.1.2 Contributions

Our primary contribution addresses permutation recovery in the noisy version of observation model (2.1), with a random design matrix A. In particular, when the entries of A are drawn i.i.d. from a standard Gaussian matrix, we show sharp conditions—in Theorems 2.3.1 and 2.3.2—on the SNR under which exact permutation recovery is possible. We also derive necessary conditions for approximate permutation recovery to within a prescribed Hamming distortion.

Chapter-specific notation: Recall the notational convention introduced in Section 1.4. We complement this notation with a few other definitions that are used solely in this chapter and the corresponding technical proof section in Appendix A.1. We let $\mathsf{d_H}(\pi, \pi')$ denote the Hamming distance between two permutations. More formally, we have $\mathsf{d_H}(\pi, \pi') := \#\{i \mid \pi(i) \neq \pi'(i)\}$. Additionally, we let $\mathsf{d_H}(\Pi, \Pi')$ denote the Hamming distance between two permutation matrices, which is to be interpreted as the Hamming distance between the corresponding permutations. We use the notation $a_i^\top$ to refer to the i-th row of A.

2.2 Background and problem setting

In this section, we set up notation, state the formal problem, and provide concrete examples of the noiseless version of our observation model by considering some fixed design matrices.

2.2.1 Formal problem setting and permutation recovery

As mentioned in the introduction, we focus exclusively on the noisy observation model in the random design setting. In other words, we obtain an n-vector of observations y from the model (2.1) with $n \geq d$ to ensure identifiability, and with the following assumptions:

Signal model The vector $x^* \in \mathbb{R}^d$ is fixed, but unknown. We note that this is different from the *adversarial* signal model of Unnikrishnan et al. [307], and we provide clarifying examples in Section 2.2.2.

Measurement matrix The measurement matrix $A \in \mathbb{R}^{n \times d}$ is a random matrix of i.i.d. standard Gaussian variables chosen without knowledge of x^*. Our assumption on i.i.d. standard Gaussian designs easily extends to accommodate the more general case when rows of A are drawn i.i.d. from the distribution $\mathcal{N}(0, \Sigma)$. In particular, writing $A = W\sqrt{\Sigma}$, where W in an $n \times d$ standard Gaussian matrix and $\sqrt{\Sigma}$ denotes the symmetric square root of the (non-singular) covariance matrix Σ, our observation model takes the form

$$y = \Pi^* W \sqrt{\Sigma} x^* + w,$$

and the unknown vector is now $\sqrt{\Sigma} x^*$ in the model (2.1).

Noise variables The vector $w \sim \mathcal{N}(0, \sigma^2 I_n)$ represents uncorrelated noise variables, each of (possibly unknown) variance σ^2. As will be made clear in the analysis, our assumption that the noise is Gaussian also readily extends to accommodate i.i.d. σ-sub-Gaussian noise. Additionally, the permutation noise represented by the unknown permutation matrix Π^* is arbitrary.

The main recovery criterion addressed in this chapter is that of exact permutation recovery, which is formally described below. Following that, we also discuss two other relevant recovery criteria.

Exact permutation recovery The problem of exact permutation recovery is to recover Π^*, and the risk of an estimator is evaluated on the 0-1 loss. More formally, given an estimator of Π^* denoted by $\widehat{\Pi} : (y, A) \to \mathfrak{S}_n$, we evaluate its risk by

$$\Pr\{\widehat{\Pi} \neq \Pi^*\} = \mathbb{E}\left[\mathbf{1}\{\widehat{\Pi} \neq \Pi^*\}\right], \tag{2.2}$$

where the probability in the LHS is taken over the randomness in y induced by both A and w.

Approximate permutation recovery It is reasonable to think that recovering Π^* up to some distortion is sufficient for many applications. Such a relaxation of exact permutation recovery allows the estimator to output a $\widehat{\Pi}$ such that $\mathsf{d_H}(\widehat{\Pi}, \Pi^*) \leq D$, for some distortion D to be specified. The risk of such an estimator is again evaluated on the 0-1 loss of this error metric, given by $\Pr\{\mathsf{d_H}(\widehat{\Pi}, \Pi^*) \geq D\}$, with the probability again taken over both A and w. While our results are derived mainly in the context of exact permutation recovery, they can be suitably modified to also yield results for approximate permutation recovery.

Recovery with side information In this variation, the unknown permutation matrix is not arbitrary, but known to be in some Hamming ball around the identity matrix. In other words, the estimator is provided with side information that $\mathsf{d_H}(\Pi^*, I) \leq \bar{h}$, for some $\bar{h} < n$. In many applications, this may constitute a prior that leads us to believe that the permutation matrix is not arbitrary. In multi-target tracking, for example, we may be sure that at any given time, a certain number of measurements correspond to the true sensors that made them (that are close to the target, perhaps). Our results also address the exact permutation recovery problem with side information.

We now provide some examples in which the noiseless version of the observation model (2.1) is identifiable.

2.2.2 Illustrative examples of the noiseless model

In this section, we present two examples to illustrate the problem of permutation recovery and highlight the difference between our signal model and that of Unnikrishnan et al. [307].

Example 1 Consider the noiseless case of the observation model (2.1). Let $\nu_i, \nu_i' \quad (i = 1, 2, \ldots, d)$ represent i.i.d. continuous random variables, and form the design matrix A by choosing

$$a_{2i-1}^\top := \nu_i e_i^\top \text{ and } a_{2i}^\top = \nu_i' e_i^\top, \quad i = 1, 2, \ldots, d.$$

Note that $n = 2d$. Now consider our fixed but unknown signal model for x^*. Since the permutation is arbitrary, our observations can be thought of as the unordered set $\{\nu_i x_i^*, \nu_i' x_i^* \mid i \in [d]\}$. With probability 1, the ratios $r_i := \nu_i/\nu_i'$ are distinct for each i, and also that $\nu_i x_i^* \neq \nu_j x_j^*$ with probability 1, by assumption of a fixed x^*. Therefore, there is a one to one correspondence between the ratios r_i and x_i^*. All ratios are computable in time $\mathcal{O}(n^2)$, and x^* can be exactly recovered. Using this information, we can also exactly recover Π^*.

Example 2 A particular case of this example was already observed by Unnikrishnan et al. [307], but we include it to illustrate the difference between our signal model and the adversarial signal model. Form the fixed design matrix A by including 2^{i-1} copies of the vector e_i among its rows. We therefore[1] have $n = \sum_{i=1}^{d} 2^{i-1} = 2^d - 1$.

Our observations therefore consist of 2^{i-1} repetitions of x_i^* for each $i \in [d]$. The value of x_i^* can therefore be recovered by simply counting the number of times it is repeated, with our choice of the number of repetitions also accounting for cases when $x_i^* = x_j^*$ for some $i \neq j$. Notice that we can now recover *any* vector x^*, even those chosen adversarially with knowledge of the A matrix. Therefore, such a design matrix allows for an *adversarial* signal model, in the flavor of compressive sensing [50].

Having provided examples of the noiseless observation model, we now return to the noisy setting of Section 2.2.1, and state our main results.

2.3 Fundamental limits of permutation estimation

In this section, we state our main theorems and discuss their consequences. Proofs of the theorems can be found in Section 2.4.

2.3.1 Statistical limits of exact permutation recovery

Our main theorems in this section provide necessary and sufficient conditions under which the probability of error in exactly recovering the true permutation goes to zero.

In brief, provided that d is sufficiently small, we establish a threshold phenomenon that characterizes how the signal-to-noise ratio $\mathsf{snr} := \frac{\|x^*\|_2^2}{\sigma^2}$ must scale relative to n in order to ensure identifiability. More specifically, defining the ratio

$$\Gamma(n, \mathsf{snr}) := \frac{\log(1 + \mathsf{snr})}{\log n},$$

[1] Unnikrishnan et al. [307] proposed that e_i be repeated i times, but it is easy to see that this does not ensure recovery of an adversarially chosen x^*.

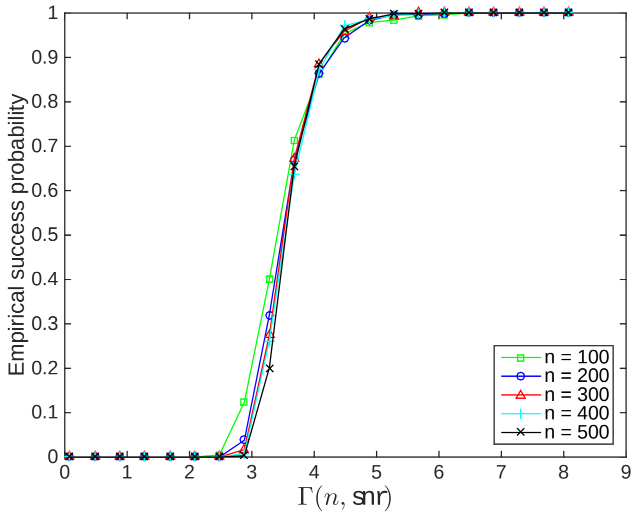


Figure 2.2: Empirical frequency of the event $\{\widehat{\Pi}_{\mathsf{ML}} = \Pi^*\}$ over 1000 independent trials with $d = 1$, plotted against $\Gamma\left(n, \mathsf{snr}\right)$ for different values of n. The probability of successful permutation recovery undergoes a phase transition as $\Gamma\left(n, \mathsf{snr}\right)$ varies from 3 to 5. This is consistent with the prediction of Theorems 2.3.1 and 2.3.2.

we show that the maximum likelihood estimator recovers the true permutation with high probability provided $\Gamma(n, \mathsf{snr}) \gg c$, where c denotes an absolute constant. Conversely, if $\Gamma(n, \mathsf{snr}) \ll c$, then exact permutation recovery is impossible. For illustration, we have plotted the behavior of the maximum likelihood estimator for the case when $d = 1$ in Figure 2.2. Evidently, there is a sharp phase transition between error and exact recovery as the ratio $\Gamma(n, \mathsf{snr})$ varies from 3 to 5.

Let us now turn to more precise statements of our results. We first define the maximum likelihood estimator (MLE) as

$$(\widehat{\Pi}_{\mathsf{ML}}, \widehat{x}_{\mathsf{ML}}) = \arg \min_{\substack{\Pi \in \mathfrak{S}_n \\ x \in \mathbb{R}^d}} \|y - \Pi A x\|_2^2. \tag{2.3}$$

The following theorem provides an upper bound on the probability of error of $\widehat{\Pi}_{\mathsf{ML}}$, with (c_1, c_2) denoting absolute constants.

Theorem 2.3.1. *For any $d < n$ and $\epsilon < \sqrt{n}$, if*

$$\log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right) \geq \left(c_1 \frac{n}{n-d} + \epsilon\right) \log n, \tag{2.4}$$

then $\Pr\{\widehat{\Pi}_{\mathsf{ML}} \neq \Pi^*\} \leq c_2 n^{-2\epsilon}$.

Theorem 2.3.1 provides conditions on the signal-to-noise ratio $\mathsf{snr} = \frac{\|x^*\|_2^2}{\sigma^2}$ that are sufficient for permutation recovery in the non-asymptotic, noisy regime. In contrast, the results of Unnikrishnan et al. [307] are stated in the limit $\mathsf{snr} \to \infty$, without an explicit characterization of the scaling behavior.

We also note that Theorem 2.3.1 holds for all values of $d < n$, whereas the results of Unnikrishnan et al. [307] require $n \geq 2d$ for identifiability of x^* in the noiseless case. Although the recovery of Π^* and x^* are not directly comparable, it is worth pointing out that the discrepancy also arises due to the difference between our fixed and unknown signal model, and the adversarial signal model assumed in the paper [307].

We now turn to the following converse result, which complements Theorem 2.3.1.

Theorem 2.3.2. *For any $\delta \in (0, 2)$, if*

$$2 + \log\left(1 + \frac{\|x^*\|_2^2}{\sigma^2}\right) \leq (2 - \delta)\log n, \tag{2.5}$$

then $\Pr\{\widehat{\Pi} \neq \Pi^*\} \geq 1 - c_3 e^{-c_4 n \delta}$ for any estimator $\widehat{\Pi}$.

Theorem 2.3.2 serves as a "strong converse" for our problem, since it guarantees that if condition (2.5) is satisfied, then the probability of error of any estimator goes to 1 as n goes to infinity. Indeed, it is proved using the strong converse argument for the Gaussian channel [278], which yields a converse result for any *fixed* design matrix A (see (2.15)). In fact, we are also able to show the following "weak converse" for Gaussian designs in the presence of side information.

Proposition 2.3.1. *If $n \geq 9$ and*

$$\log\left(1 + \frac{\|x^*\|_2^2}{\sigma^2}\right) \leq \frac{8}{9}\log\left(\frac{n}{8}\right),$$

then $\Pr\{\widehat{\Pi} \neq \Pi^*\} \geq 1/2$ for any estimator $\widehat{\Pi}$, even if it is known a-priori that $\mathsf{d_H}(\Pi^*, I) \leq 2$.

As mentioned earlier, restriction of Π^* constitutes some application-dependent prior; the strongest such prior restricts it to a Hamming ball of radius 2 around the identity. Proposition 2.3.1 asserts that even this side information does not substantially change the statistical limits of permutation recovery. It is also worth noting that the converse results of Theorem 2.3.2 and Proposition 2.3.1 hold uniformly over d.

Taken together, Theorems 2.3.1 and 2.3.2 provide a crisp characterization of the problem when $d \leq pn$ for some fixed $p < 1$. In particular, setting ϵ and δ in Theorems 2.3.1 and 2.3.2 to be small constants and letting n grow, we recover the threshold behavior of identifiability in terms of $\Gamma(n, \mathsf{snr})$ that was discussed above and illustrated in Figure 2.2. In the next section, we find that a similar phenomenon occurs even with approximate permutation recovery.

When d can be arbitrarily close to n, the characterization obtained using these bounds is no longer sharp. In this regime, we conjecture that Theorem 2.3.1 provides the correct characterization of the limits of the problem, and that Theorem 2.3.2 can be sharpened.

2.3.2 Limits of approximate permutation recovery

The techniques we used to prove results for exact permutation recovery can be suitably modified to obtain results for approximate permutation recovery to within a Hamming distortion D. In particular, we show the following converse result for approximate recovery.

Theorem 2.3.3. *For any* $2 < D \leq n-1$, *if*

$$\log\left(1 + \frac{\|x^*\|_2^2}{\sigma^2}\right) \leq \frac{n-D+1}{n} \log\left(\frac{n-D+1}{2e}\right), \tag{2.6}$$

then $\Pr\{\mathsf{d_H}(\widehat{\Pi}, \Pi^*) \geq D\} \geq 1/2$ *for any estimator* $\widehat{\Pi}$.

Note that for any $D \leq pn$ with $p \in (0,1)$, Theorems 2.3.1 and 2.3.3 provide a set of sufficient and necessary conditions for approximate permutation recovery that match up to constant factors. In particular, the necessary condition resembles that for exact permutation recovery, and the same SNR threshold behavior is seen even here. We remark that a corresponding converse with side information can also be proved for approximate permutation recovery using techniques similar to the proof of Proposition 2.3.1. It is also worth mentioning the following:

Remark 2.3.1. *The converse results given by Theorem 2.3.2, Proposition 2.3.1, and Theorem 2.3.3 hold even when the estimator has exact knowledge of* x^*.

2.4 Proofs of main results

In this section, we prove our main results. Technical details are deferred to the appendices. Throughout the proofs, we assume that n is larger than some universal constant. The case where n is smaller can be handled by changing the constants in our proofs appropriately. We also use the notation c, c' to denote absolute constants that can change from line to line. Technical lemmas used in our proofs are deferred to the appendix.

We begin with the proof of Theorem 2.3.1. At a high level, it involves bounding the probability that any fixed permutation is preferred to Π^* by the estimator. The analysis requires precise control on the lower tails of χ^2-random variables, and tight bounds on the norms of random projections, for which we use results derived in the context of dimensionality reduction by Dasgupta and Gupta [78].

In order to simplify the exposition, we first consider the case when $d = 1$ in Section 2.4.1, and later make the necessary modifications for the general case in Section 2.4.2. In order to understand the technical subtleties, we recommend that the reader fully understand the $d = 1$ case along with the technical lemmas before moving on to the proof of the general case.

2.4.1 Proof of Theorem 2.3.1: $d = 1$ case

Recall the definition of the maximum likelihood estimator

$$(\widehat{\Pi}_{\mathsf{ML}}, \widehat{x}_{\mathsf{ML}}) = \arg \min_{\Pi \in \mathfrak{S}_n} \min_{x \in \mathbb{R}^d} \|y - \Pi A x\|_2^2.$$

For a fixed permutation matrix Π, assuming that A has full column rank[2], the minimizing argument x is simply $(\Pi A)^\dagger y$, where $X^\dagger = (X^\top X)^{-1} X^\top$ represents the pseudoinverse of a matrix X. By computing the minimum over $x \in \mathbb{R}^d$ in the above equation, we find that the maximum likelihood estimate of the permutation is given by

$$\widehat{\Pi}_{\mathsf{ML}} = \arg \min_{\Pi \in \mathfrak{S}_n} \|P_\Pi^\perp y\|_2^2, \tag{2.7}$$

where $P_\Pi^\perp = I - \Pi A (A^\top A)^{-1} (\Pi A)^\top$ denotes the projection onto the orthogonal complement of the column space of ΠA.

For a fixed $\Pi \in \mathfrak{S}_n$, define the random variable

$$\Delta(\Pi, \Pi^*) := \|P_\Pi^\perp y\|_2^2 - \|P_{\Pi^*}^\perp y\|_2^2. \tag{2.8}$$

For any permutation Π, the estimator (2.7) prefers the permutation Π to Π^* if $\Delta(\Pi, \Pi^*) \leq 0$. The overall error event occurs when $\Delta(\Pi, \Pi^*) \leq 0$ for some Π, meaning that

$$\{\widehat{\Pi}_{\mathsf{ML}} \neq \Pi^*\} = \bigcup_{\Pi \in \mathfrak{S}_n \setminus \Pi^*} \{\Delta(\Pi, \Pi^*) \leq 0\}. \tag{2.9}$$

Equation (2.9) holds for any value of d. We shortly specialize to the $d = 1$ case. Our strategy for proving Theorem 2.3.1 boils down to bounding the probability of each error event in the RHS of equation (2.9) using the following key lemma, proved in Section A.1.1. Technically speaking, the proof of this lemma contains the meat of the proof of Theorem 2.3.1, and the interested reader is encouraged to understand these details before embarking on the proof of the general case. Recall the definition of $\mathsf{d_H}(\Pi, \Pi')$, the Hamming distance between two permutation matrices.

Lemma 2.4.1. *For $d = 1$ and any two permutation matrices Π and Π^*, and provided $\frac{\|x^*\|_2^2}{\sigma^2} > 1$, we have*

$$\Pr\{\Delta(\Pi, \Pi^*) \leq 0\} \leq c' \exp\left(-c \, \mathsf{d_H}(\Pi, \Pi^*) \log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right)\right).$$

We are now ready to prove Theorem 2.3.1.

Proof of Theorem 2.3.1 for $d = 1$. Fix $\epsilon > 0$ and assume that the following consequence of condition (2.4) holds:

$$c \log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right) \geq (1 + \epsilon) \log n, \tag{2.10}$$

[2] An $n \times d$ i.i.d. Gaussian random matrix has full column rank with probability 1 as long as $d \leq n$

where c is the same as in Lemma 2.4.1. Now, observe that

$$\begin{aligned}
\Pr\{\widehat{\Pi}_{\mathsf{ML}} \neq \Pi^*\} &\leq \sum_{\Pi \in \mathfrak{S}_n \setminus \Pi^*} \Pr\{\Delta(\Pi, \Pi^*) \leq 0\} \\
&\overset{\text{(i)}}{\leq} \sum_{\Pi \in \mathfrak{S}_n \setminus \Pi^*} c' \exp\left(-c\, \mathsf{d_H}(\Pi, \Pi^*) \log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right)\right) \\
&\leq c' \sum_{2 \leq k \leq n} n^k \exp\left(-c\, k \log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right)\right) \\
&\overset{\text{(ii)}}{\leq} c' \sum_{2 \leq k \leq n} n^{-\epsilon k} \\
&\leq c' \frac{1}{n^\epsilon(n^\epsilon - 1)},
\end{aligned}$$

where step (i) follows since $\#\{\Pi : \mathsf{d_H}(\Pi, \Pi^*) = k\} \leq n^k$, and step (ii) follows from condition (2.10). Relabeling the constants in condition (2.10) proves the theorem. □

In the next section, we prove Theorem 2.3.1 for the general case.

2.4.2 Proof of Theorem 2.3.1: Case $d \in \{2, 3, \ldots, n-1\}$

In order to be consistent, we follow the same proof structure as for the $d = 1$ case. Recall the definition of $\Delta(\Pi, \Pi^*)$ from equation (2.8). We begin with an equivalent of the key lemma to bound the probability of the event $\{\Delta(\Pi, \Pi^*) \leq 0\}$. As in the $d = 1$ case, this constitutes the technical core of the result.

Lemma 2.4.2. *For any $1 < d < n$, any two permutation matrices Π and Π^* at Hamming distance h, and provided $\left(\frac{\|x^*\|_2^2}{\sigma^2}\right) n^{-\frac{2n}{n-d}} > \frac{5}{4}$, we have*

$$\Pr\{\Delta(\Pi, \Pi^*) \leq 0\} \leq c' \max\left[\exp\left(-n \log\frac{n}{2}\right), \exp\left(ch\left(\log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right) - \frac{2n}{n-d}\log n\right)\right)\right]. \tag{2.11}$$

We prove Lemma 2.4.2 in Section A.1.2. Taking it as given, we are ready to prove Theorem 2.3.1 for the general case.

Proof of Theorem 2.3.1, general case. As before, we use the union bound to prove the theorem. We begin by fixing some $\epsilon \in (0, \sqrt{n})$ and assuming that the following consequence of condition (2.4) holds:

$$c \log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right) \geq \left(1 + \epsilon + c\frac{2n}{n-d}\right) \log n. \tag{2.12}$$

Now define $b(k) := \sum_{\Pi:\mathsf{d}_\mathsf{H}(\Pi,\Pi^*)=k} \Pr\{\Delta(\Pi, \Pi^*) \leq 0\}$. Applying Lemma 2.4.2 then yields

$$\frac{(n-k)!}{n!} b(k) \leq c' \max \left\{ \exp\left(-n \log \frac{n}{2}\right), \exp\left(-ck \left(\log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right) - \frac{2n}{n-d}\log n\right)\right)\right\}. \tag{2.13}$$

We upper bound $b(k)$ by splitting the analysis into two cases.

Case 1 If the first term attains the maximum in the RHS of inequality (2.13), then for all $2 \leq k \leq n$, we have

$$\begin{aligned} b(k) &\leq c' n! \exp(-n \log n + n \log 2) \\ &\overset{\text{(i)}}{\leq} c' e\sqrt{n} \exp(-n \log n + n \log 2 + -n + n \log n) \\ &\overset{\text{(ii)}}{\leq} \frac{c'}{n^{2\epsilon+1}}, \end{aligned}$$

where inequality (i) follows from the well-known upper bound $n! \leq e\sqrt{n}\left(\frac{n}{e}\right)^n$, and inequality (ii) holds since $\epsilon \in (0, \sqrt{n})$.

Case 2 Alternatively, if the maximum is attained by the second term in the RHS of inequality (2.13), then we have

$$b(k) \leq n^k c' \exp\left(-ck\left(\log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right) - \frac{2n}{n-d}\log n\right)\right) \overset{\text{(iii)}}{\leq} c' n^{-\epsilon k},$$

where step (iii) follows from condition (2.12).

Combining the two cases, we have

$$b(k) \leq \max\{c' n^{-\epsilon h}, c n^{-2\epsilon - 1}\} \leq \left(c' n^{-\epsilon h} + c n^{-2\epsilon-1}\right).$$

The last step is to use the union bound to obtain

$$\begin{aligned} \Pr\{\widehat{\Pi}_{\mathsf{ML}} \neq \Pi^*\} &\leq \sum_{2\leq k\leq n} b(k) \\ &\leq \sum_{2\leq k \leq n} \left(c' n^{-\epsilon h} + c n^{-2\epsilon-1}\right) \\ &\overset{\text{(iv)}}{\leq} c n^{-2\epsilon}, \end{aligned} \tag{2.14}$$

where step (iv) follows by a calculation similar to the one carried out for the $d = 1$ case. Relabeling the constants in condition (2.12) completes the proof. □

2.4.3 Information theoretic lower bounds on $\Pr\{\widehat{\Pi} \neq \Pi^*\}$

We prove Theorem 2.3.2 in Section 2.4.3 via the strong converse for the Gaussian channel. We prove the weak converse of Proposition 2.3.1 in Section 2.4.3 by employing Fano's method. We note that the latter is a standard technique for proving minimax bounds in statistical estimation; see the historical overview in [323, Chapter 15].

Proof of Theorem 2.3.2

We begin by assuming that the design matrix A is fixed, and that the estimator has knowledge of x^* a-priori. Note that the latter cannot make the estimation task any easier. In proving this lower bound, we can also assume that the entries of Ax^* are distinct, since otherwise, perfect permutation recovery is impossible.

Given this setup, we now cast the problem as one of coding over a Gaussian channel. Toward this end, consider the codebook

$$\mathcal{C} = \{\Pi A x^* \mid \Pi \in \mathfrak{S}_n\}.$$

We may view $\Pi A x^*$ as the codeword corresponding to the permutation Π, where each permutation is associated to one of $n!$ equally likely messages. Note that each codeword has power $\|Ax^*\|_2^2$.

The codeword is then sent over a Gaussian channel with noise power equal to $\sum_{i=1}^n \sigma^2 = n\sigma^2$. The decoding problem is to ascertain from the noisy observations which message was sent, or in other words, to identify the correct permutation.

We now use the non-asymptotic strong converse for the Gaussian channel [343]. In particular, using Lemma A.2.4 (see Appendix A.2.3) with $R = \frac{\log n!}{n}$ then yields that for any $\delta' > 0$, if

$$\frac{\log n!}{n} > \frac{1+\delta'}{2} \log\left(1 + \frac{\|Ax^*\|_2^2}{n\sigma^2}\right),$$

then for any estimator $\widehat{\Pi}$, we have $\Pr\{\widehat{\Pi} \neq \Pi\} \geq 1 - 2 \cdot 2^{-n\delta'}$. For the choice $\delta' = \delta/(2-\delta)$, we have that if

$$(2-\delta)\log\left(\frac{n}{e}\right) > \log\left(1 + \frac{\|Ax^*\|_2^2}{n\sigma^2}\right), \tag{2.15}$$

then $\Pr\{\widehat{\Pi} \neq \Pi\} \geq 1 - 2 \cdot 2^{-n\delta/2}$. Note that the only randomness assumed so far was in the noise w and the random choice of Π.

We now specialize the result for the case when A is Gaussian. Toward that end, define the event

$$\mathcal{E}(\delta) = \left\{1 + \delta \geq \frac{\|Ax^*\|_2^2}{n\|x^*\|_2^2}\right\}.$$

Conditioned on the event $\mathcal{E}(\delta)$, it can be verified that condition (2.5) implies condition (2.15). We also have

$$\begin{aligned}\Pr\{\mathcal{E}(\delta)\} &= 1 - \Pr\left\{\frac{\|Ax^*\|_2^2}{n\|x^*\|_2^2} > 1+\delta\right\}\\ &\overset{\text{(i)}}{\geq} 1 - c'e^{-cn\delta},\end{aligned}$$

where step (i) follows by using the sub-exponential tail bound (see Lemma A.2.2 in Appendix A.2.2), since $\frac{\|Ax^*\|_2^2}{\|x^*\|_2^2} \sim \chi_n^2$.

Putting together the pieces, we have that provided condition (2.5) holds,

$$\begin{aligned}\Pr\{\widehat{\Pi} \neq \Pi^*\} &\geq \Pr\{\widehat{\Pi} \neq \Pi^* | \mathcal{E}(\delta)\} \Pr\{\mathcal{E}(\delta)\}\\ &= (1 - 2\cdot 2^{-n\delta/2})(1 - c'e^{-cn\delta})\\ &\geq 1 - c'e^{-cn\delta}.\end{aligned}$$

□

We now move on to the proof of Proposition 2.3.1.

Proof of Proposition 2.3.1

Proposition 2.3.1 corresponds to a weak converse in the scenario where the estimator is also provided with the side information that Π^* lies within a Hamming ball of radius 2 around the identity. We carry out the proof for the more general case where $\mathsf{d_H}(\Pi^*, I) \leq \bar{h}$ for *any* $\bar{h} \geq 2$, later specializing to the case where $\bar{h} = 2$. We denote such a Hamming ball by $\mathbb{B}_{\mathsf{H}}(\bar{h}) := \{\Pi \in \mathfrak{S}_n \mid \mathsf{d_H}(\Pi, I) \leq \bar{h}\}$.

For the sake of the lower bound, assume that our observation model takes the form

$$y = \Pi^* A x^* + \Pi^* w. \tag{2.16}$$

Since $w \in \mathbb{R}^n$ is i.i.d. standard normal, model (2.16) has a distribution equivalent to that of model (2.1).

Notice that the ML estimation problem is essentially a multi-way hypothesis testing problem among permutations in $\mathbb{B}_{\mathsf{H}}(\bar{h})$, and so Fano's method is directly applicable. As before, we assume that the estimator knows x^*, and consider a uniformly random choice of $\Pi^* \in \mathbb{B}_{\mathsf{H}}(\bar{h})$.

Now note that the observation vector y is drawn from the mixture distribution

$$\mathbb{M}(\bar{h}) = \frac{1}{|\mathbb{B}_{\mathsf{H}}(\bar{h})|} \sum_{\Pi \in \mathfrak{S}_n} \mathbb{P}_\Pi, \tag{2.17}$$

where $\mathbb{P}_\Pi$ denotes the Gaussian distribution $\mathcal{N}(\Pi A x^*, \sigma^2 I_n)$. Lemma A.1.4 stated and proved in Section A.1.3 provides a crucial statistic for our bounds.

$$\det \mathbb{E}\left[yy^\top\right] \leq (\sigma^2 + \|x^*\|_2^2)^n \,(1+n) \left(\frac{\bar{h}}{n}\right)^{n-1}. \tag{2.18}$$

We now use Fano's inequality to bound the probability of error of any tester random ensemble of Π^*. In particular, for any estimator $\widehat{\Pi}$ that is a measurable function of the pair (y, A), we have

$$\Pr\{\widehat{\Pi} \neq \Pi^*\} \geq 1 - \frac{I(\Pi^*; y, A) + \log 2}{\log |\mathbb{B}_{\mathsf{H}}(\bar{h})|}. \tag{2.19}$$

Applying the chain rule for mutual information yields

$$\begin{aligned} I(\Pi^*; y, A) &= I(\Pi^*; y|A) + I(\Pi^*, A) \\ &\overset{\text{(i)}}{=} I(\Pi^*; y|A) \\ &= \mathbb{E}_A \left[I(\Pi^*; y|A = \alpha) \right], \end{aligned} \tag{2.20}$$

where step (i) follows since Π^* is chosen independently of A. We now evaluate the mutual information term $I(\Pi^*; y|A = \alpha)$, which we denote by $I_\alpha(\Pi^*; y)$. Letting $H_\alpha(y) := H(y|A = \alpha)$ denote the conditional entropy of y given a fixed realization of A, we have

$$\begin{aligned} I_\alpha(\Pi^*; y) &= H_\alpha(y) - H_\alpha(y|\Pi^*) \\ &\overset{\text{(ii)}}{\leq} \frac{1}{2} \log \det \operatorname{cov} yy^\top - \frac{n}{2} \log \sigma^2, \end{aligned}$$

where the covariance is evaluated with $A = \alpha$, and in step (ii), we have used two key facts:
(a) Gaussians maximize entropy for a fixed covariance, which bounds the first term, and
(b) For a fixed realization of Π^*, the vector y is composed of n uncorrelated Gaussians. This leads to an explicit evaluation of the second term.

Now taking expectations over A and noting that $\operatorname{cov} yy^\top \preceq \mathbb{E}_w \left[yy^\top \right]$, we have from the concavity of the log determinant function and Jensen's inequality that

$$\begin{aligned} I(\Pi^*; y|A) &= \mathbb{E}_A \left[I_\alpha(\Pi^*; y) \right] \\ &\leq \frac{1}{2} \log \det \mathbb{E} \left[yy^\top \right] - \frac{n}{2} \log \sigma^2, \end{aligned} \tag{2.21}$$

where the expectation in the last line is now taken over randomness in both A and w.

Applying Lemma A.1.4, we can then substitute inequality (2.18) into bound (2.21), which yields

$$\mathbb{E}_A \left[I_\alpha(\Pi^*; y) \right] \leq \frac{n}{2} \log \left(1 + \frac{\|x^*\|_2^2}{\sigma^2} \right) + \frac{n-1}{2} \log \frac{\bar{h}}{n} + \frac{1}{2} \log (1 + n).$$

Finally, substituting into the Fano bound (2.19) yields the bound

$$\begin{aligned} 1 - \Pr\{\widehat{\Pi} \neq \Pi^*\} &\leq \frac{\frac{n}{2} \log \left(1 + \frac{\|x^*\|_2^2}{\sigma^2} \right) + \frac{n-1}{2} \log \frac{\bar{h}}{n} + \frac{1}{2} \log (1 + n) + \log 2}{\log |\mathbb{B}_{\mathsf{H}}(\bar{h})|} \\ &\overset{\text{(ii)}}{\leq} \frac{\frac{n}{2} \log \left(1 + \frac{\|x^*\|_2^2}{\sigma^2} \right) + \frac{n-1}{2} \log \frac{\bar{h}}{n} + \frac{1}{2} \log (1 + n) + \log 2}{\bar{h} \log(n/e)}, \end{aligned}$$

where in step (ii), we have used the fact that

$$|\mathbb{B}_{\mathsf{H}}(\bar{h})| = \binom{n}{\bar{h}} \cdot \bar{h}! \geq \left(n/\bar{h}\right)^{\bar{h}} \left(\frac{\bar{h}}{e}\right)^{\bar{h}}.$$

In other words, whenever the SNR is upper bounded as

$$\log\left[4\left(1+\frac{\|x^*\|_2^2}{\sigma^2}\right)\right] \leq \frac{\bar{h}}{n}\log(n/e) + \frac{n-1}{n}\log(n/\bar{h}) - \frac{\log\left(1+n\right)}{n}, \tag{2.22}$$

then $\Pr\{\widehat{\Pi} \neq \Pi^*\} \geq 1/2$ for any estimator $\widehat{\Pi}$. Evaluating condition (2.22) for $\bar{h} = 2$ yields the required result. □

2.4.4 Proof of Theorem 2.3.3

We now prove Theorem 2.3.3 for approximate permutation recovery. For any estimator $\widehat{\Pi}$, we denote by the indicator random variable $E(\widehat{\Pi}, D)$ whether or not the $\widehat{\Pi}$ has acceptable distortion, i.e., $E(\widehat{\Pi}, D) = \mathbb{I}[\mathsf{d}_{\mathsf{H}}(\widehat{\Pi}, \Pi^*) \geq D]$, with $E = 1$ representing the error event. For Π^* picked uniformly at random in $\mathfrak{S}_n$, Lemma A.1.5 stated and proved in Section A.1.4 lower bounds the probability of error as:

$$\Pr\{E(\widehat{\Pi}, D) = 1\} \geq 1 - \frac{I(\Pi^*; y, A) + \log 2}{\log n! - \log \frac{n!}{(n-D+1)!}}.$$

The proof of the theorem follows by upper bounding the mutual information term. In particular, letting $I_\alpha(\Pi^*; y)$ denote $I(\Pi^*; y, A | A = \alpha)$ and using inequality (2.20), we have

$$\begin{aligned} I(\Pi^*; y, A) &\leq \mathbb{E}_A\left[I_\alpha(\Pi^*; y)\right] \\ &\leq \frac{1}{2}\log\det\mathbb{E}\left[yy^\top\right] - \frac{n}{2}\log\sigma^2 \\ &\overset{\text{(i)}}{\leq} \frac{n}{2}\log\left(1+\frac{\|x^*\|_2^2}{\sigma^2}\right), \end{aligned}$$

where the expectation on the RHS is taken over both Π^* and A. Also, step (i) follows from the AM-GM inequality for PSD matrices $\det X \leq \left(\frac{1}{n}\operatorname{trace} X\right)^n$, and by noting that the diagonal entries of the matrix $\mathbb{E}\left[yy^\top\right]$ are all equal to $\|x^*\|_2^2 + \sigma^2$.

Combining the pieces, we now have that $\Pr\{\widehat{\Pi} \neq \Pi^*\} \geq 1/2$ if

$$n\log\left(1+\frac{\|x^*\|_2^2}{\sigma^2}\right) \leq (n-D+1)\log\left(\frac{n-D+1}{2e}\right), \tag{2.23}$$

which completes the proof. □

2.5 Summary and open questions

We analyzed the problem of exact permutation recovery in the linear regression model, and provided necessary and sufficient conditions that are tight in most regimes of n and d. We also provided a converse for the problem of approximate permutation recovery to within some Hamming distortion. It is still an open problem to characterize the fundamental limits of exact and approximate permutation recovery for all regimes of n, d and the allowable distortion D. In the context of exact permutation recovery, we believe that the limit suggested by Theorem 2.3.1 is tight for all regimes of n and d, but showing this will likely require a different technique. In particular, as pointed out in Remark 2.3.1, all of our lower bounds assume that the estimator is provided with x^* as side information; it is an interesting question as to whether stronger lower bounds can be obtained without this side information.

On the computational front, many open questions remain. The primary question concerns the design of computationally efficient estimators that succeed in similar SNR regimes. We have already shown that the maximum likelihood estimator, while being statistically optimal for moderate d, is computationally hard to compute in the worst case. Showing a corresponding hardness result for random A with noise is also an open problem. Finally, while this chapter mainly addresses the problem of permutation recovery, the complementary problem of recovering x^* is also interesting.

In the broader context of this thesis, this chapter serves as an in-depth investigation into some statistical questions surrounding this permutation-based vector regression model. In the next chapter, we study a permutation-based matrix model, and consider both statistical and computational aspects.

Chapter 3

Fast and sample-efficient matrix completion

This chapter studies a permutation-based model for matrix-valued data, in which the latent permutations transparently model unknown orderings (or rankings) in the problem and facilitate the use of nonparametric assumptions.

3.1 Introduction

Structured matrices with unknown permutations acting on their rows and columns arise in multiple applications, including estimation from pairwise comparisons [40, 274] and crowd-labeling [80, 275]. Traditional parametric models (e.g., [40, 80, 204, 299]) assume that these matrices are obtained from rank-one or rank-two matrices via a known link function. Aided by tools such as maximum likelihood estimation and spectral methods, researchers have made significant progress in studying both statistical and computational aspects of these parametric models [74, 134, 165, 202, 229, 259, 273, 350] and their low-rank generalizations [164, 230, 260].

On the other hand, evidence from empirical studies suggests that real-world data is not always well-described by such parametric models [16, 220]. With the goal of increasing model flexibility, a recent line of work has studied the class of *permutation-based* models [59, 274, 275]. Rather than imposing parametric conditions on the matrix entries, these models impose only shape constraints on the matrix, such as monotonicity, before unknown permutations act on the rows and columns of the matrix. On one hand, this more flexible class reduces modeling bias compared to its parametric counterparts while, perhaps surprisingly, producing models that can be estimated at rates that differ only by logarithmic factors from the classical parametric models. On the other hand, these advantages of permutation-based models are accompanied by significant computational challenges. The unknown permutations make the parameter space highly non-convex, so that efficient maximum likelihood estimation is unlikely. Moreover, spectral methods are often sub-optimal in approximating shape-constrained sets of matrices [59, 274]. Consequently, results from many recent papers show a non-trivial statistical-computational gap in estimation rates for models with latent permutations [55, 104, 247, 274, 275]; the previous chapter also presents one such example but in a different metric.

Related work While the primary motivation of our work comes from nonparametric methods for aggregating pairwise comparisons, we begin by discussing a few other lines of related work. The current chapter lies at the intersection of shape-constrained estimation and latent permutation learning. Shape-constrained estimation has long been a major topic in nonparametric statistics, and of particular relevance to our work is the estimation of a bivariate isotonic matrix without latent permutations [56]. There, it was shown that the minimax rate of estimating an $n_1 \times n_2$ matrix from noisy observations of all its entries is $\widetilde{\Theta}((n_1 n_2)^{-1/2})$. The upper bound is achieved by the least squares estimator, which is efficiently computable due to the convexity of the parameter space.

Shape-constrained matrices with permuted rows or columns also arise in applications such as seriation [104, 105], feature matching [69], and graphon estimation [53]. In particular, the monotone subclass of the statistical seriation model [104] contains $n \times n$ matrices that have increasing columns, and an unknown row permutation. Flammarion et al. [104] established the minimax rate $\widetilde{\Theta}(n^{-2/3})$ for estimating matrices in this class and proposed a computationally efficient algorithm with rate $\widetilde{\mathcal{O}}(n^{-1/2})$. For the subclass of such matrices where in addition, the rows are also monotone, the results of this chapter improve the two rates to $\widetilde{\Theta}(n^{-1})$ and $\widetilde{\mathcal{O}}(n^{-3/4})$ respectively.

Graphon estimation has seen its own extensive literature, and we only list those papers that are most relevant to our setting. In essence, these problems involve nonparametric estimation of a bivariate function f from noisy observations of $f(\xi_i, \xi_j)$ with the design points $\{\xi_i\}_{i=1}^n$ drawn i.i.d. from some distribution supported on the interval $[0, 1]$. In contrast to nonparametric estimation, however, the design points in graphon estimation are unobserved, which gives rise to the underlying latent permutation. Modeling the function f as monotone recovers the model studied in this chapter, but other settings have been studied by various authors: notably those where the function f is Lipschitz [53], block-wise constant [32] (also known as the stochastic block model), or with f satisfying other smoothness assumptions [113]. There are many interesting statistical-computational gaps also known to exist in many of these problems.

Another related model in the pairwise comparison literature is that of noisy sorting [42], which involves a latent permutation but no shape-constraint. In this prototype of a permutation-based ranking model, we have an unknown, $n \times n$ matrix with constant upper and lower triangular portions whose rows and columns are acted upon by an unknown permutation. The hardness of recovering any such matrix in noise lies in estimating the unknown permutation. As it turns out, this class of matrices can be estimated efficiently at minimax optimal rate $\widetilde{\Theta}(n^{-1})$ by multiple procedures: the original work by Braverman and Mossel [42] proposed an algorithm with time complexity $\mathcal{O}(n^c)$ for some unknown and large constant c, and recently, an $\widetilde{\mathcal{O}}(n^2)$-time algorithm was proposed by Mao et al. [212]. These algorithms, however, do not generalize beyond the noisy sorting class, which constitutes a small subclass of an interesting class of matrices that we describe next.

The most relevant body of work to the current chapter is that on estimating matrices satisfying the *strong stochastic transitivity* condition, or SST for short. This class of matrices contains all $n \times n$ bivariate isotonic matrices with unknown permutations acting on their rows and columns, with an additional skew-symmetry constraint. The first theoretical study of these matrices was carried out by Chatterjee [59], who showed that a spectral algorithm achieved the rate $\widetilde{\mathcal{O}}(n^{-1/4})$ in the normalized, squared Frobenius norm. Shah et al. [274] then showed that the minimax rate of estimation is given by $\widetilde{\Theta}(n^{-1})$, and also improved the analysis of the spectral estimator of Chatterjee to obtain the computationally efficient rate $\widetilde{\mathcal{O}}(n^{-1/2})$. In follow-up work [277], they also showed a second CRL estimator based on the Borda count that achieved the same rate. In related work, Chatterjee and Mukherjee [55] analyzed a variant of the CRL estimator, showing that for sub-classes of SST matrices, it achieved rates that were faster than $\mathcal{O}(n^{-1/2})$. In a complementary direction, a superset of the current authors [244] analyzed the estimation problem under an observation model with structured missing data, and showed that for many observation patterns, a variant of the CRL estimator was minimax optimal.

Shah et al. [277] also showed that conditioned on the planted clique conjecture, it is impossible to improve upon a certain notion of adaptivity of the CRL estimator in polynomial time. Such results have prompted various authors [104, 277] to conjecture that a similar statistical-computational gap also exists when estimating SST matrices in the Frobenius norm.

Contributions In this chapter, we study the problem of estimating a bivariate isotonic matrix with unknown permutations acting on its rows and columns, given noisy, (possibly) partial observations of its entries; this matrix class strictly contains the SST model [59, 274] for ranking from pairwise comparisons. We also study a sub-class of such matrices motivated by applications in crowd-labeling, which consists of bivariate isotonic matrices with one unknown permutation acting on its rows.

We begin by characterizing, in the Frobenius norm metric, the fundamental limits of estimation of both classes of matrices; the former result significantly generalizes those obtained by Shah et al. [274]. In particular, our results hold for arbitrary matrix dimensions and sample sizes, and also extend results of Chatterjee, Guntuboyina and Sen [56] for estimating the sub-class of bivariate isotonic matrices without unknown permutations. We then present computationally efficient algorithms for estimating both classes of matrices; these algorithms are novel in the sense that they are neither spectral in nature, nor simple variations of the Borda count estimator that was previously employed. They are also tractable in practice and show significant improvements over state-of-the-art estimators; Figure 3.1 presents such a comparison for our algorithm specialized to SST matrices with (roughly) one observation per entry.

These algorithms are then analyzed in the the Frobenius error metric, and Theorem 3.4.2 constitutes our main contribution of this chapter. Two consequences of this result are noteworthy. First, our algorithm for bivariate isotonic matrices with unknown permutations attains the rate $\widetilde{\mathcal{O}}(n^{-3/4})$ in the squared Frobenius error, provided we have a full observation of an $n \times n$ matrix. Moreover, our algorithms for both classes of matrices are minimax-optimal when the number of observations grows to be sufficiently large; notably, this stands in stark contrast to existing computationally efficient algorithms, which are not minimax-optimal in *any* regime of the problem.

Chapter-specific notation: Recall the notational convention introduced in Section 1.4. We complement this notation with a few other definitions that are used solely in this chapter and the corresponding technical proof section in Appendix A.3. For a vector $v \in \mathbb{R}^n$, define its variation as $\mathrm{var}(v) = v_(1) - v_(n)$. We denote the i-th row of a matrix M by M_i, unless otherwise specified.

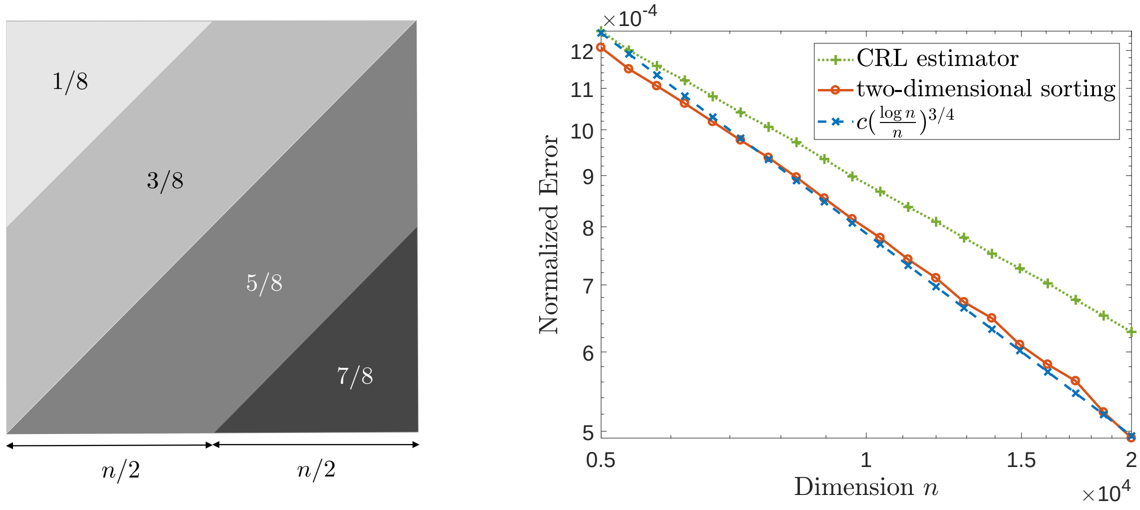


Figure 3.1: **Left:** A bivariate isotonic matrix; the ground truth $M^* \in [0,1]^{n \times n}$ is a row and column permuted version of such a matrix. **Right:** A log-log plot of the rescaled squared Frobenius error $\frac{1}{n^2}\|\widehat{M} - M^*\|_F^2$ versus the matrix dimension n. For each value of the dimension, the error is averaged over 10 experiments each using n^2 Bernoulli observations, and the estimator $\widehat{M}$ is either the two-dimensional sorting estimator that we introduce in Section 3.4.2, or the CRL estimator from past work [277].

3.2 Background and problem setup

In this section, we present the relevant technical background and notation on permutation-based models, and introduce the observation model and error metrics of interest. We also elaborate on how exactly these models arise in practice.

3.2.1 Matrix models

Our main focus is on designing efficient algorithms for estimating a bivariate isotonic matrix with unknown permutations acting on its rows and columns. Formally, we define $\mathbb{C}_{\mathsf{BISO}}$ to be the class of matrices in $[0,1]^{n_1 \times n_2}$ with non-decreasing rows and non-decreasing columns. For readability and without loss of generality, we assume frequently (in particular, everywhere except for Proposition 3.4.1 and Section 3.4.1) that $n_1 \geq n_2$; our results can be straightforwardly extended to the other case. Given a matrix $M \in \mathbb{R}^{n_1 \times n_2}$ and permutations $\pi \in \mathfrak{S}_{n_1}$ and $\sigma \in \mathfrak{S}_{n_2}$, we define the matrix $M(\pi, \sigma) \in \mathbb{R}^{n_1 \times n_2}$ by specifying its entries as

$$[M(\pi, \sigma)]_{i,j} = M_{\pi(i), \sigma(j)} \text{ for } i \in [n_1], j \in [n_2].$$

Also define the class $\mathbb{C}_{\mathsf{BISO}}(\pi, \sigma) := \{M(\pi, \sigma) : M \in \mathbb{C}_{\mathsf{BISO}}\}$ as the set of matrices that are bivariate isotonic when viewed along the row permutation π and column permutation σ, respectively.

The classes of matrices that we are interested in estimating are given by

$$\mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}} := \bigcup_{\substack{\pi \in \mathfrak{S}_{n_1} \\ \sigma \in \mathfrak{S}_{n_2}}} \mathbb{C}_{\mathsf{BISO}}(\pi, \sigma), \qquad \text{and its subclass} \qquad \mathbb{C}^{\mathsf{r}}_{\mathsf{Perm}} := \bigcup_{\pi \in \mathfrak{S}_{n_1}} \mathbb{C}_{\mathsf{BISO}}(\pi, \mathsf{id}).$$

The former class contains bivariate isotonic matrices with both rows and columns permuted, and the latter contains those with only rows permuted.

3.2.2 Observation model and error metric

In order to study estimation from noisy observations of a matrix M^* in either of the classes $\mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}}$ or $\mathbb{C}^{\mathsf{r}}_{\mathsf{Perm}}$, we suppose that N noisy entries are sampled independently and uniformly with replacement from all entries of M^*. This sampling model is popular in the matrix completion literature, and is a special case of the *trace regression model* [174, 231]. It has also been used in the context of permutation-based models by Mao et al. [212] to study the noisy sorting class.

More precisely, let $E^{(i,j)}$ denote the $n_1 \times n_2$ matrix with 1 in the (i, j)-th entry and 0 elsewhere, and suppose that X_ℓ is a random matrix sampled independently and uniformly from the set $\{E^{(i,j)} : i \in [n_1],\, j \in [n_2]\}$. We observe N independent pairs $\{(X_\ell, y_\ell)\}_{\ell=1}^N$ from the model

$$y_\ell = \operatorname{trace}(X_\ell^\top M^*) + z_\ell, \tag{3.1}$$

where the observations are contaminated by independent, zero-mean, sub-exponential noise z_ℓ with parameter ζ, that is,

$$\mathbb{E} \exp(s z_\ell) \le \exp(\zeta^2 s^2) \quad \text{for all } s \text{ such that } |s| \le 1/\zeta. \tag{3.2}$$

Note that if (3.2) holds for all $s \in \mathbb{R}$, then z_ℓ is called sub-Gaussian, which is a stronger condition. We assume for convenience that an upper bound on the parameter ζ is known to our estimators; this assumption is mild since for many standard noise distributions, such an upper bound is either immediate (in the case of any bounded distribution) or the standard deviation of the noise is a proxy, up to a universal constant factor, for the parameter ζ (in the case of the Gaussian or Poisson noise models) and can be estimated very accurately by a variety of methods[1].

It is important to note at this juncture that although the observation model (3.1) is motivated by the matrix completion literature, we make no assumptions of partial observability here. In particular, our results hold for all tuples (N, n_1, n_2), with the sample size N allowed to grow larger than the effective dimension $n_1 n_2$.

Besides the standard Gaussian observation model, in which

$$z_\ell \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1), \tag{3.3a}$$

[1] For instance, one could implement one of many consistent estimators for M^* to obtain a matrix $\widehat{M}$, and use the quantity $\{\frac{1}{N} \sum_{\ell=1}^N [y_\ell - \operatorname{trace}(X_\ell^\top \widehat{M})]^2\}^{1/2}$ as an estimate of the standard deviation.

another noise model of interest is one which arises in applications such as crowd-labeling and ranking from pairwise comparisons. Here, for every $x \in \{E^{(i,j)} : i \in [n_1],\ j \in [n_2]\}$ and conditioned on $X_\ell = x$, our observations take the form

$$y_\ell \sim \mathsf{Ber}\big(\operatorname{trace}(x^\top M^*)\big), \tag{3.3b}$$

and consequently, the sub-exponential parameter ζ is bounded. For a discussion of other regimes of noise in a related matrix model, see Gao [111].

For analytical convenience, we employ the standard trick of Poissonization, whereby we assume throughout this chapter that $N' = \mathsf{Poi}(N)$ random observations are drawn according to the trace regression model (3.1), with the Poisson random variable drawn independently of everything else. Upper and lower bounds derived under this model carry over with loss of constant factors to the model with exactly N observations; for a detailed discussion, see the full paper [211].

Now given $N' = \mathsf{Poi}(N)$ observations $\{(X_\ell, y_\ell)\}_{\ell=1}^{N'}$, let us define the matrix of observations $Y = Y\left(\{(X_\ell, y_\ell)\}_{\ell=1}^{N'}\right)$, with entry (i,j) given by

$$Y_{i,j} = \frac{n_1 n_2}{N} \sum_{\ell=1}^{N'} y_\ell \, \mathbf{1}\{X_\ell = E^{(i,j)}\}. \tag{3.4}$$

In other words, we simply average the observations at each entry by the expected number of observations, so that $\mathbb{E}[Y] = M^*$. Moreover, we may write the model in the linearized form $Y = M^* + W$, where W is a matrix of additive noise having independent, zero-mean entries thanks to Poissonization.[2] To be more precise, we can decompose the noise at each entry as

$$\begin{aligned} W_{i,j} &= Y_{i,j} - M^*_{i,j} \\ &= \frac{n_1 n_2}{N} \sum_{\ell=1}^{N'} z_\ell \cdot \mathbf{1}\{X_\ell = E^{(i,j)}\} + M^*_{i,j} \frac{n_1 n_2}{N} \Big(\sum_{\ell=1}^{N'} \mathbf{1}\{X_\ell = E^{(i,j)}\} - \frac{N}{n_1 n_2} \Big). \end{aligned}$$

By Poissonization, the quantities $\sum_{\ell=1}^{N'} \mathbf{1}\{X_\ell = E^{(i,j)}\}$ for $(i,j) \in [n_1] \times [n_2]$ are i.i.d. $\mathsf{Poi}(\frac{N}{n_1 n_2})$ random variables, so the second term above is simply the deviation of a Poisson variable from its mean. On the other hand, the first term is a normalized sum of independent sub-exponential noise. Therefore, this linearized and decomposed form of noise provides an amenable starting point for our analysis.

Let us now turn to the metric that we use to assess the error of our estimators. For a tuple of "proper" estimates $(\widehat{M}, \widehat{\pi}, \widehat{\sigma})$, in that $\widehat{M}(\widehat{\pi}, \widehat{\sigma}) \in \mathbb{C}_{\mathsf{BISO}}(\widehat{\pi}, \widehat{\sigma})$ (and $\widehat{\sigma} = \mathsf{id}$ if we are estimating over the class $\mathbb{C}^{\mathsf{r}}_{\mathsf{Perm}}$), the normalized squared Frobenius error is given by the random variable

$$\mathcal{F}\big(M^*, \widehat{M}(\widehat{\pi}, \widehat{\sigma})\big) = \frac{1}{n_1 n_2} \big\| \widehat{M}(\widehat{\pi}, \widehat{\sigma}) - M^* \big\|_F^2.$$

[2] See, e.g, Shah et al. [274] for a justification of such a decomposition in the fully observed setting.

In the paper that contains a superset of the results presented in this chapter, we also consider the max-row-norm approximation error of the permutation estimate $\widehat{\pi}$, which has applications to learning mixtures of rankings from pairwise comparisons [288]. Studying the problem in this metric also provides new results for testing problems on cones, while shedding light on why prior work on this problem was unable to surpass what was perceived as a fundamental gap in estimation in the Frobenius error.

3.2.3 Applications

The matrix models studied in this chapter arise in crowd-labeling and estimation from pairwise comparisons, and can be viewed as generalizations of low-rank matrices of a particular type.

Let us first describe their relevance to the crowd-labeling problem [275]. Here, there is a set of n_2 questions of a binary nature; the true answers to these questions can be represented by a vector $x^* \in \{0, 1\}^{n_2}$, and our goal is to estimate this vector by asking these questions to n_1 *workers* on a crowdsourcing platform. Since workers have varying levels of expertise, it is important to *calibrate* them, i.e., to obtain a good estimate of which workers are reliable and which are not. This is typically done by asking them a set of gold standard questions, which are expensive to generate, and sample efficiency is an extremely important consideration. Indeed, gold standard questions are carefully chosen to control for the level of difficulty and diversity [240]. Worker calibration is seen as a key step towards improving the quality of samples collected in crowdsourcing applications [241].

Mathematically, we may model worker abilities via the probabilities with which they answer questions correctly, and collect these probabilities within a matrix $M^* \in [0, 1]^{n_1 \times n_2}$. The entries of this matrix are latent, and must be learned from observing workers' answers to questions. In the calibration problem, we know the answers to the questions; from these, we can estimate worker abilities and question difficulties, or more generally, the entries of the matrix M^*. In many applications, we also have additional knowledge about gold standard questions; for instance, in addition to the true answers, we may also know the relative difficulties of the questions themselves.

Imposing sensible constraints on the matrix M^* in these applications goes back to classical work on the subject, with the majority of models of a *parametric* nature; for instance, the Dawid-Skene model [80] is widely used in crowd-labeling applications, and its variants have been analyzed by many authors (e.g., [74, 165, 202]). However, in a parallel line of work, generalizations of the parametric Dawid-Skene model have been empirically evaluated on a variety of crowd-labeling tasks [332], and shown to achieve performance superior to the Dawid-Skene model for many such tasks. The permutation-based model of Shah et al. [275] is one such generalization, and was proven to alleviate some important pitfalls of parametric models from both the statistical and algorithmic standpoints. Operationally, such a model assumes that workers have a total ordering π of their abilities, and that questions have a total ordering σ of their difficulties. The matrix M^* is thus bivariate isotonic when the rows are ordered in increasing order of worker ability, and columns are ordered in decreasing order of question difficulty. However, since worker abilities and question difficulties are unknown *a priori*, the matrix of probabilities obeys the inclusion $M^* \in \mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}}$. In the particular case where we also know the relative difficulties of the questions themselves, we may assume that the column permutation is known, so that our estimation problem is now over the class $\mathbb{C}^{\mathsf{r}}_{\mathsf{Perm}}$.

Let us now discuss the application to estimation from pairwise comparisons. An interesting subclass of $\mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}}$ are those matrices that are square ($n_1 = n_2 = n$), and also skew symmetric. More precisely, let us define $\mathbb{C}'_{\mathsf{BISO}}$ analogously to the class $\mathbb{C}_{\mathsf{BISO}}$, except with matrices having columns that are non-increasing instead of non-decreasing. Also define the class

$$\mathbb{C}_{\mathsf{skew}}(n) := \{M \in [0,1]^{n \times n} : M + M^\top = 11^\top\}, \tag{3.5a}$$

as well as the *strong stochastic transitivity* class

$$\mathbb{C}_{\mathsf{SST}}(n) := \left(\bigcup_{\pi \in \mathfrak{S}_n} \mathbb{C}'_{\mathsf{BISO}}(\pi, \pi) \right) \bigcap \mathbb{C}_{\mathsf{skew}}(n). \tag{3.5b}$$

The class $\mathbb{C}_{\mathsf{SST}}(n)$ is useful as a model for estimation from pairwise comparisons [59, 274], and was proposed as a strict generalization of parametric models for this problem [40, 229, 259]. In particular, given n items obeying some unknown underlying ranking π, entry (i, j) of a matrix $M^* \in \mathbb{C}_{\mathsf{SST}}(n)$ represent the probability $\Pr(i \succ j)$ with which item i beats item j in a comparison. The shape constraint encodes the transitivity condition that for all triples (i, j, k) obeying $\pi(i) < \pi(j) < \pi(k)$, we must have

$$\Pr(i \succ k) \geq \max\{\Pr(i \succ j), \Pr(j \succ k)\}.$$

For a more classical introduction to these models, see the papers [16, 103, 220] and the references therein. Our task is to estimate the underlying ranking from results of passively chosen pairwise comparisons[3] between the n items, or more generally, to estimate the underlying probabilities M^* that govern these comparisons[4]. All results in this chapter stated for the more general matrix model $\mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}}$ apply to the class $\mathbb{C}_{\mathsf{SST}}(n)$ with minimal modifications.

To conclude, we mention a final applications in which the flexibility afforded by nonparametric models with latent permutations has been noticed and exploited. In psychometric item-response theory, the Mokken double-monotonicity model is identical to . This model is known to be significantly more robust to misspecification than the parametric Rasch model; see [314] for an introduction and survey.

3.3 Fundamental limits of estimation

We begin by characterizing the fundamental limits of estimation under the trace regression observation model (3.1) with $N' = \mathsf{Poi}(N)$ observations. We define the least squares estimator over a

[3] Such a passive, simultaneous setting should be contrasted with the *active* case (e.g., [96]), where we may sequentially choose pairs of items to compare depending on the results of previous comparisons.

[4] Accurate, proper estimates of M^* in the Frobenius error metric translate to accurate estimates of the ranking π (see Shah et al. [274]).

closed set $\mathbb{C}$ of $n_1 \times n_2$ matrices as the projection

$$\widehat{M}_{\mathsf{LS}}(\mathbb{C}, Y) \in \arg\min_{M \in \mathbb{C}} \|Y - M\|_F^2.$$

The projection is a non-convex problem when the class $\mathbb{C}$ is given by either the class $\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}}$ or $\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r}}$, and is unlikely to be computable exactly in polynomial time. However, studying this estimator allows us to establish a baseline that characterizes the best achievable statistical rate. In the following theorem, we characterizes the Frobenius risk of the least squares estimator, and also provide a minimax lower bound. These results hold for any sample size N of the problem. Also recall the shorthand $Y = Y\left(\{X_\ell, y_\ell\}_{\ell=1}^{N'}\right)$, and let $\bar{\zeta} := \zeta \vee 1$ denote the proxy for the noise that accounts for missing data.

Theorem 3.3.1. *(a) Suppose that $n_2 \leq n_1$. There is an absolute constant $c_1 > 0$ such that for any matrix $M^* \in \mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}}$, we have*

$$\mathcal{F}\big(M^*, \widehat{M}_{\mathsf{LS}}(\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}}, Y)\big) \leq c_1 \Bigg\{ \bar{\zeta}^2 \frac{n_1}{N} \log n_1 + \bar{\zeta} \frac{n_2}{N} (\log n_1)^2 \\ + \left(\bar{\zeta} \frac{1}{\sqrt{N}} (\log n_1)^2\right) \wedge \left(\bar{\zeta}^2 \frac{n_2}{N} \log n_1\right)^{2/3} \wedge \left(\bar{\zeta}^2 \frac{n_1 n_2}{N} \log N\right) \Bigg\} \wedge 1 \tag{3.6}$$

with probability at least $1 - n_1^{-n_1}$.
(b) Suppose that $n_2 \leq n_1$, and that $N' \sim \mathsf{Poi}(N)$ independent samples are drawn under the standard Gaussian observation model (3.3a) *or the Bernoulli observation model* (3.3b). *Then there exists an absolute constant $c_2 > 0$ such that any estimator $\widehat{M}$ satisfies*

$$\sup_{M^* \in \mathbb{C}_{\mathsf{Perm}}^{\mathsf{r}}} \mathbb{E}\left[\mathcal{F}(M^*, \widehat{M})\right] \geq c_2 \left\{ \left[\frac{n_1}{N} + \left(\frac{1}{\sqrt{N}} \wedge \left(\frac{n_2}{N}\right)^{2/3} \wedge \frac{n_1 n_2}{N} \right) \right] \wedge 1 \right\}. \tag{3.7}$$

When interpreted in the context of square matrices under partial observations, our result should be viewed as paralleling that of Shah et al. [274]. In addition, however, the result also provides a generalization in several directions. First, the upper bound holds under the general sub-exponential noise model. Second, the lower bound holds for the class $\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r}}$, which is strictly smaller than the class $\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}}$. Third, and more importantly, we study the problem for arbitrary tuples (N, n_1, n_2), and this allows us to uncover interesting phase transitions in the rates that were previously unobserved[5]; these are discussed in more detail below.

Via the inclusion $\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r}} \subseteq \mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}}$, we observe that for both classes $\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r}}$ and $\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}}$, the upper and lower bounds match up to logarithmic factors in all regimes of (N, n_1, n_2) under the standard Gaussian or Bernoulli noise model. Such a poly-logarithmic gap between the upper and lower

[5] The regime $N \geq n_1 n_2$ is interesting for the problems of ranking and crowd-labeling that motivate our work, since it is pertinent to compare items or ask workers to answer questions multiple times in order to reduce the noisiness of the gathered data. From a theoretical standpoint, the large N regime isolates multiple non-asymptotic behaviors on the way to asymptotic consistency as $N \to \infty$, and is related in spirit to recent work on studying the high signal-to-noise ratio regime in ranking models [111], where phase transitions were also observed.

bounds is related to corresponding gaps that exist in upper and lower bounds on the metric entropy of bounded bivariate isotonic matrices [56]. Closing this gap is known to be an important problem in the study of these shape-constrained objects (see, e.g. [114]).

Let us now interpret the theorem in more detail. First, note that the risk of any proper estimator is bounded by 1 since the entries of M^* are bounded in the interval $[0,1]$. We thus focus on the remaining terms of the bound. Up to a poly-logarithmic factor, the upper bound can be simplified to $\bar\zeta^2 \frac{n_1}{N} + \big(\bar\zeta \frac{1}{\sqrt{N}}\big) \wedge \big(\bar\zeta^2 \frac{n_2}{N}\big)^{2/3} \wedge \big(\bar\zeta^2 \frac{n_1 n_2}{N}\big)$. The first term $\bar\zeta^2 \frac{n_1}{N}$ is due to the unknown permutation on the rows (which dominates the unknown column permutation since $n_1 \geq n_2$). The remaining terms, corresponding to the minimum of three rates, stem from estimating the underlying bivariate isotonic matrix, so we make a short digression to state a corollary specialized to this setting. Recall that $\mathbb{C}_{\mathsf{BISO}}$ was used to denote the class of $n_1 \times n_2$ bivariate isotonic matrices. Owing to the convexity of the set $\mathbb{C}_{\mathsf{BISO}}$, the least squares estimator $\widehat{M}_{\mathsf{LS}}(\mathbb{C}_{\mathsf{BISO}}, Y)$ is computable efficiently [182]. Moreover, we use the shorthand notation

$$\vartheta(N, n_1, n_2, \zeta) := \bar\zeta \frac{n_2}{N} (\log n_1)^2 + \\ \Big(\bar\zeta \frac{1}{\sqrt{N}} (\log n_1)^2\Big) \wedge \Big(\bar\zeta^2 \frac{n_2}{N} \log n_1\Big)^{2/3} \wedge \Big(\bar\zeta^2 \frac{n_1 n_2}{N} \log N\Big). \tag{3.8}$$

Corollary 3.3.1. *(a) Suppose that $n_2 \leq n_1$. Then there is a universal constant $c_1 > 0$ such that for any matrix $M^* \in \mathbb{C}_{\mathsf{BISO}}$, we have*

$$\mathcal{F}\big(M^*, \widehat{M}_{\mathsf{LS}}(\mathbb{C}_{\mathsf{BISO}}, Y)\big) \leq c_1 \vartheta(N, n_1, n_2, \zeta) \wedge 1$$

with probability at least $1 - n_1^{-3n_1}$.
(b) Suppose that $n_2 \leq n_1$, and that $N' \sim \mathsf{Poi}(N)$ independent samples are drawn under the standard Gaussian observation model (3.3a) *or the Bernoulli observation model* (3.3b). *Then there exists an absolute constant $c_2 > 0$ such that any estimator $\widehat{M}$ satisfies*

$$\sup_{M^* \in \mathbb{C}_{\mathsf{BISO}}} \mathbb{E}\Big[\mathcal{F}(M^*, \widehat{M})\Big] \geq c_2 \left\{ \frac{1}{\sqrt{N}} \wedge \Big(\frac{n_2}{N}\Big)^{2/3} \wedge \frac{n_1 n_2}{N} \wedge 1 \right\}.$$

Corollary 3.3.1 should be viewed as paralleling the results of Chatterjee et al. [56] (see, also, Han et al. [137]) under a slightly different noise model, while providing some notable extensions once again. Firstly, we handle sub-exponential noise; secondly, the bounds hold for all tuples (N, n_1, n_2) and are optimal up to a logarithmic factor provided the sample size N is sufficiently large. In more detail, the nonparametric rate $\bar\zeta \frac{1}{\sqrt{N}}$ was also observed in Theorems 2.1 and 2.2 of the paper [56] provided in the fully observed setting ($N = n_1 n_2$), with the lower bound additionally requiring that the matrix was not extremely skewed. In addition to this rate, we also isolate two other interesting regimes when $N \geq n_2^2$ and $1/\sqrt{N}$ is no longer the minimizer of the three terms above. The first of these regimes is the rate $(\bar\zeta^2 \frac{n_2}{N})^{2/3}$, which is also nonparametric; notably, it corresponds to the rate achieved by decoupling the structure across columns and treating the problem as n_2 separate isotonic regression problems [233, 347]. This suggests that if the matrix is extremely skewed or if N grows very large, monotonicity along the smaller dimension is no longer as helpful; the canonical example of this is when $n_2 = 1$, in which case we are left with the (univariate) isotonic regression problem[6]. The final rate $\bar{\zeta}^2 \frac{n_1 n_2}{N}$ is parametric and comparatively trivial, as it can be achieved by an estimator that simply averages observations at each entry. This suggests that when the number of samples grows extremely large, we can ignore all structure in the problem and still be optimal at least in a minimax sense.

Let us now return to a discussion of Theorem 3.3.1. To further clarify the rates and transitions between them, we simplify the discussion by focusing on two regimes of matrix dimensions.

Example 1: $n_2 \leq n_1 \leq n_2^2$ Here, by treating ζ as a constant, we may simplify the minimax rate (up to logarithmic factors) as

$$\inf_{\widehat{M}} \sup_{M^* \in \mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}}} \mathbb{E}\left[\mathcal{F}(M^*, \widehat{M})\right] \asymp \begin{cases} 1 & \text{for } N \leq n_1 \\ \frac{n_1}{N} & \text{for } n_1 \leq N \leq n_1^2 \\ \frac{1}{\sqrt{N}} & \text{for } n_1^2 \leq N \leq n_2^4 \\ (\frac{n_2}{N})^{2/3} & \text{for } n_2^4 \leq N \leq n_1^3 n_2 \\ \frac{n_1 n_2}{N} & \text{for } N \geq n_1^3 n_2 \end{cases} \tag{3.9}$$

which delineates five distinct regimes depending on the sample size N. The first regime is the trivial rate. The second regime $n_1 \leq N \leq n_1^2$ is when the error due to the latent permutation dominates, while the third regime $n_1^2 \leq N \leq n_2^4$ corresponds to when the hardness of the problem is dominated by the structure inherent to bivariate isotonic matrices. For N larger than n_2^4, the effect of *bivariate* isotonicity disappears, at least in a minimax sense. Namely, in the fourth regime $n_2^4 \leq N \leq n_1^3 n_2$, the rate $(n_2/N)^{2/3}$ is the same as if we treat the problem as n_2 separate n_1-dimensional isotonic regression problems with an unknown permutation [104]. For even larger sample size $N \geq n_1^3 n_2$, in the fifth regime, the minimax-optimal rate $n_1 n_2 / N$ is trivially achieved by ignoring all structure and outputting the matrix Y alone. ♣

Example 2: $n_2 \leq n_1 \leq C n_2$ In this near-square regime, we may once again simplify the bound and obtain (up to logarithmic factors) that

$$\inf_{\widehat{M}} \sup_{M^* \in \mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}}} \mathbb{E}\left[\mathcal{F}(M^*, \widehat{M})\right] \asymp \begin{cases} 1 & \text{for } N \leq n_1 \\ \frac{n_1}{N} & \text{for } n_1 \leq N \leq n_1^2 \\ \frac{1}{\sqrt{N}} & \text{for } n_1^2 \leq N \leq n_1^2 n_2^2 \\ \frac{n_1 n_2}{N} & \text{for } N \geq n_1^2 n_2^2 \end{cases} \tag{3.10}$$

[6] Indeed, in a similar regime, Chatterjee et al. [56] show in their Theorem 2.3 that the upper bound is achieved by an estimator that performs univariate regression along each row, followed by a projection onto the set of BISO matrices. On the other hand, our results establish near-optimal minimax rates in a unified manner through metric entropy estimates, so that the upper bounds are sharp in all regimes simultaneously, and hold for the least squares estimator under sub-exponential noise.

so that two of the cases from before now collapse into one. Ignoring the trivial constant rate, we thus observe a transition from a parametric rate to a nonparametric rate, and back to the trivial parametric rate. ♣

3.4 Efficient algorithms

Our algorithms belong to a broader family of algorithms that rely on two distinct steps: first, estimate the unknown permutation(s) defining the problem; then project onto the class of matrices that are bivariate isotonic when viewed along the estimated permutations. Formally, any such algorithm is described by the meta-algorithm below.

Algorithm 1 (meta-algorithm)

- Step 1: Use any algorithm to obtain permutation estimates $(\widehat{\pi}, \widehat{\sigma})$, setting $\widehat{\sigma} = \mathsf{id}$ if estimating over class $\mathbb{C}^{\mathsf{r}}_{\mathsf{Perm}}$.
- Step 2: Return the matrix estimate $\widehat{M}(\widehat{\pi}, \widehat{\sigma}) := \arg\min_{M \in \mathbb{C}_{\mathsf{BISO}}(\widehat{\pi}, \widehat{\sigma})} \|Y - M\|_F^2$.

Owing to the convexity of the set $\mathbb{C}_{\mathsf{BISO}}(\widehat{\pi}, \widehat{\sigma})$, the projection operation in Step 2 of the algorithm can be computed in polynomial (sub-quadratic) time [182]. The following result, a variant of Proposition 4.2 of Chatterjee and Mukherjee [55], allows us to characterize the error rate of any such meta-algorithm as a function of the permutation estimates $(\widehat{\pi}, \widehat{\sigma})$.

Recall the definition of the set $\mathbb{C}_{\mathsf{BISO}}(\pi, \sigma) := \{M(\pi, \sigma) : M \in \mathbb{C}_{\mathsf{BISO}}\}$ as the set of matrices that are bivariate isotonic when viewed along the row permutation π and column permutation σ, respectively. In particular, we have the inclusion $M^* \in \mathbb{C}_{\mathsf{BISO}}(\pi^*, \sigma^*)$ where π^* and σ^* are unknown permutations in $\mathfrak{S}_{n_1}$ and $\mathfrak{S}_{n_2}$, respectively. In the following proposition, we also do not make the assumption $n_2 \leq n_1$; recall our shorthand notation $\vartheta(N, n_1, n_2, \zeta)$ defined in equation (3.8).

Proposition 3.4.1. *There exists an absolute constant $C > 0$ such that for all $M^* \in \mathbb{C}_{\mathsf{BISO}}(\pi^*, \sigma^*)$, the estimator $\widehat{M}(\widehat{\pi}, \widehat{\sigma})$ obtained by running the meta-algorithm satisfies*

$$\begin{aligned}\mathcal{F}\big(M^*, \widehat{M}(\widehat{\pi}, \widehat{\sigma})\big) \leq C\Big\{&\vartheta(N, n_1 \vee n_2, n_1 \wedge n_2, \zeta) \\ &+ \frac{1}{n_1 n_2}\big\|M^*(\pi^*, \sigma^*) - M^*(\widehat{\pi}, \sigma^*)\big\|_F^2 + \frac{1}{n_1 n_2}\big\|M^*(\pi^*, \sigma^*) - M^*(\pi^*, \widehat{\sigma})\big\|_F^2\Big\}\end{aligned} \tag{3.11}$$

with probability exceeding $1 - n_1^{-n_1}$.

A few comments are in order. The term $\vartheta(N, n_1 \vee n_2, n_1 \wedge n_2, \zeta)$ on the upper line of the RHS of the bound (3.11) corresponds to an estimation error, if the true permutations π and σ were known a priori (see Corollary 3.3.1), and the latter terms on the lower line correspond to an approximation error that we incur as a result of having to estimate these permutations from data. Comparing the bound (3.11) to the minimax lower bound (3.7), we see that up to a poly-logarithmic factor, the estimation error terms of the bound (3.11) are unavoidable, and so we can restrict our attention to obtaining good permutation estimates $(\widehat{\pi}, \widehat{\sigma})$. We now present two permutation estimation procedures that can be plugged into Step 1 of the meta-algorithm.

3.4.1 Matrices with ordered columns

As a stepping stone to our main algorithm, which estimates over the class $\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}}$, we first consider the estimation problem when the permutation along one of the dimensions is known. This corresponds to estimation over the subclass $\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r}}$, and following the meta-algorithm above, it suffices to provide a permutation estimate $\widehat{\pi}$. The result of this section holds without the assumption $n_1 \geq n_2$.

We need more notation to facilitate the description of the algorithm. We say that $\mathsf{bl} = \{B_k\}_{k=1}^{|\mathsf{bl}|}$ is a *partition* of $[n_2]$, if $[n_2] = \bigcup_{k=1}^{|\mathsf{bl}|} B_k$ and $B_j \cap B_k = \varnothing$ for $j \neq k$. Moreover, we group the columns of a matrix $Y \in \mathbb{R}^{n_1 \times n_2}$ into $|\mathsf{bl}|$ blocks according to their indices in bl, and refer to bl as a partition or *blocking* of the columns of Y. In the algorithm, partial row sums of Y are computed on indices contained in each block.

Algorithm 2 (sorting partial sums)

- Step 1: Choose a partition $\mathsf{bl}_{\mathsf{ref}}$ of the set $[n_2]$ consisting of contiguous blocks, such that each block B in $\mathsf{bl}_{\mathsf{ref}}$ has size

$$\frac{1}{2} n_2 \sqrt{\frac{n_1}{N} \log(n_1 n_2)} \leq |B| \leq n_2 \sqrt{\frac{n_1}{N} \log(n_1 n_2)}.$$

- Step 2: Given the observation matrix Y, compute the row sums

$$S(i) = \sum_{j \in [n_2]} Y_{i,j} \quad \text{for each } i \in [n_1],$$

and the partial row sums within each block

$$S_B(i) = \sum_{j \in B} Y_{i,j} \quad \text{for each } i \in [n_1] \text{ and } B \in \mathsf{bl}_{\mathsf{ref}}.$$

Create a directed graph G with vertex set $[n_1]$, where an edge $u \to v$ is present if either

$$S(v) - S(u) > 16(\zeta + 1)\left(\sqrt{\frac{n_1 n_2^2}{N} \log(n_1 n_2)} + \frac{n_1 n_2}{N} \log(n_1 n_2)\right), \text{ or}$$

$$S_B(v) - S_B(u) > 16(\zeta + 1)\left(\sqrt{\frac{n_1 n_2}{N} |B| \log(n_1 n_2)} + \frac{n_1 n_2}{N} \log(n_1 n_2)\right)$$

$$\text{for some } B \in \mathsf{bl}_{\mathsf{ref}}.$$

- Step 3: Return any topological sort $\widehat{\pi}_{\mathsf{ref}}$ of the graph G; if none exists, return a uniformly random permutation $\widehat{\pi}_{\mathsf{ref}}$.

We now turn to a detailed discussion of the running time of Algorithm 2. A topological sort of a generic graph $G(V, E)$ can be found via Kahn's algorithm [159] in time $\mathcal{O}(|V| + |E|)$. In our context, the topological sort operation translates to a running time of $\mathcal{O}(n_1^2)$. In Step 2, constructing the graph G takes time $\mathcal{O}(n_1^2 n_2^{1/2})$, since there are at most $\mathcal{O}(n_2^{1/2})$ blocks. This leads to a total complexity of the order $\mathcal{O}(n_1^2 n_2^{1/2})$.

Let us now give an intuitive explanation for the algorithm. While algorithms in past work [55, 244, 277] sort the rows of the matrix according to the full Borda counts $S(i)$ defined in Step 2, they are limited by the high standard deviation in these estimates. Our key observation is that when the columns are perfectly ordered, judiciously chosen partial row sums (which are less noisy than full row sums) also contain information that can help estimate the underlying row permutation. The thresholds on the score differences in Step 2 are chosen to be comparable to the standard deviations of the respective estimates, with additional logarithmic factors that allow for high-probability statements via application of Bernstein's bounds.

Theorem 3.4.1. *There exists an absolute constant $c_1 > 0$ such that for any matrix $M^* \in \mathbb{C}^{\mathsf{r}}_{\mathsf{Perm}}$, we have*

$$\mathcal{F}(M^*, \widehat{M}(\widehat{\pi}_{\mathsf{ref}}, \mathsf{id})) \leq 1 \wedge c_1 \Big\{ \bar{\zeta}^2 \left(\frac{n_1 \log n_1}{N} \right)^{3/4} + \vartheta(N, n_1 \vee n_2, n_1 \wedge n_2, \zeta) \Big\} \tag{3.12}$$

with probability at least $1 - 3(n_1 n_2)^{-2}$.

In order to evaluate our Frobenius error guarantee, it is helpful to specialize to the regime $n_2 \leq n_1 \leq Cn_2$.

Example: $n_2 \leq n_1 \leq Cn_2$ In this case, the Frobenius error guarantee simplifies to

$$\frac{1}{(\log n_1)^2} \mathcal{F}(M^*, \widehat{M}(\widehat{\pi}_{\mathsf{ref}}, \mathsf{id})) \lesssim \begin{cases} 1 & \text{for } N \leq n_1 \\ \left(\frac{n_1}{N}\right)^{3/4} & \text{for } n_1 \leq N \leq n_1^3 \\ \frac{1}{\sqrt{N}} & \text{for } n_1^3 \leq N \leq n_1^4 \\ \frac{n_1 n_2}{N} & \text{for } n_1^4 \leq N \leq n_1^5 \\ \left(\frac{n_1}{N}\right)^{3/4} & \text{for } N \geq n_1^5. \end{cases} \tag{3.13}$$

We may compare the bounds (3.10) and (3.13); note that when $n_1^3 \leq N \leq n_1^5$, our estimator achieves the minimax lower bound given by $\frac{1}{\sqrt{N}} \wedge \frac{n_1 n_2}{N}$ (up to poly-logarithmic factors). As alluded to before, when $N \geq n_1^5$, we may switch to trivially outputting the matrix Y, and so the sub-optimality in the large N regime can be completely avoided. On the other hand, no such modification can be made in the small sample regime $n_1 \leq N \leq n_1^3$, and the estimator falls short of being optimal in the Frobenius error in this case. ♣

Closing the aforementioned gap in the Frobenius error in the small sample regime is an interesting open problem. Having established guarantees for our algorithm, we now turn to using the intuition gained from these guarantees to provide estimators for matrices in the larger class $\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}}$.

3.4.2 Two-dimensional sorting for class $\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}}$

We reinstate the assumption $n_2 \leq n_1$ in this section. The algorithm in the previous section cannot be immediately extended to the class $\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}}$, since it assumes that the matrix is perfectly sorted along one of the dimensions. However, it suggests a plug-in procedure that can be described informally as follows.

1. Sort the columns of the matrix Y according to its column sums.
2. Apply Algorithm 2 to the column-sorted matrix to obtain a row permutation estimate.
3. Repeat Steps 1 and 2 with Y transposed to obtain a column permutation estimate.

Although the columns of Y are only approximately sorted in the first step, the hope is that the finer row-wise control given by Algorithm 2 is able to improve the row permutation estimate. The actual algorithm, provided below, essentially implements this intuition, but with a careful data-dependent blocking procedure that we describe next. Given a data matrix $Y \in \mathbb{R}^{n_1 \times n_2}$, the following blocking subroutine returns a column partition $\mathsf{BL}(Y)$.

Subroutine 1 (blocking)

- Step 1: Compute the column sums $\{C(j)\}_{j=1}^{n_2}$ of the matrix Y as

$$C(j) = \sum_{i=1}^{n_1} Y_{i,j}.$$

  Let $\widehat{\sigma}_{\mathsf{pre}}$ be a permutation along which the sequence $\{C(\widehat{\sigma}_{\mathsf{pre}}(j))\}_{j=1}^{n_2}$ is non-decreasing.

- Step 2: Set $\tau = 16(\zeta + 1)\Big(\sqrt{\frac{n_1^2 n_2}{N} \log(n_1 n_2)} + \frac{n_1 n_2}{N} \log(n_1 n_2)\Big)$ and $K = \lceil n_2/\tau \rceil$. Partition the columns of Y into K blocks by defining

$$\begin{aligned} \mathsf{bl}_1 &= \{j \in [n_2] : C(j) \in (-\infty, \tau)\}, \\ \mathsf{bl}_k &= \big\{j \in [n_2] : C(j) \in \big[(k-1)\tau, k\tau\big)\big\} \text{ for } 1 < k < K, \text{ and} \\ \mathsf{bl}_K &= \{j \in [n_2] : C(j) \in [(K-1)\tau, \infty)\}. \end{aligned}$$

  Note that each block is contiguous when the columns are permuted by $\widehat{\sigma}_{\mathsf{pre}}$.

- Step 3 (aggregation): Set $\beta = n_2\sqrt{\frac{n_1}{N}\log(n_1 n_2)}$. Call a block bl_k "large" if $|\mathsf{bl}_k| \geq \beta$ and "small" otherwise. Aggregate small blocks in bl while leaving the large blocks as they are, to obtain the final partition BL.

  More precisely, consider the matrix $Y' = Y(\mathsf{id}, \widehat{\sigma}_{\mathsf{pre}})$ having non-decreasing column sums and contiguous blocks. Call two small blocks "adjacent" if there is no other small block between them. Take unions of adjacent small blocks to ensure that the size of each resulting block is in the range $[\frac{1}{2}\beta, 2\beta]$. If the union of all small blocks is smaller than $\frac{1}{2}\beta$, aggregate them all.

  Return the resulting partition $\mathsf{BL}(Y) = \mathsf{BL}$.

Ignoring Step 3 for the moment, we see that the blocking bl is analogous to the blocking $\mathsf{bl}_{\mathsf{ref}}$ of Algorithm 2, along which partial row sums may be computed. While the blocking $\mathsf{bl}_{\mathsf{ref}}$ was chosen in a data-independent manner due to the columns being sorted exactly, the blocking bl is chosen based on approximate estimation of the column permutation. However, some of these K blocks may be too small, resulting in noisy partial sums; in order to mitigate this issue, Step 3 aggregates small blocks into large enough ones. We are now in a position to describe the two-dimensional sorting algorithm.

Algorithm 3 (two-dimensional sorting)

- Step 0: Split the observations into two independent sub-samples of equal size, and form the corresponding matrices $Y^{(1)}$ and $Y^{(2)}$ according to equation (3.4).
- Step 1: Apply Subroutine 1 to the matrix $Y^{(1)}$ to obtain a partition $\mathsf{BL} = \mathsf{BL}(Y^{(1)})$ of the columns. Let K be the number of blocks in BL.
- Step 2: Using the second sample $Y^{(2)}$, compute the row sums

  $$S(i) = \sum_{j \in [n_2]} Y^{(2)}_{i,j} \text{ for each } i \in [n_1],$$

  and the partial row sums within each block

  $$S_{\mathsf{BL}_k}(i) = \sum_{j \in \mathsf{BL}_k} Y^{(2)}_{i,j} \text{ for each } i \in [n_1], k \in [K].$$

  Create a directed graph G with vertex set $[n_1]$, where an edge $u \to v$ is present if either

  $$S(v) - S(u) > 16(\zeta + 1)\Bigg(\sqrt{\frac{n_1 n_2^2}{N}\log(n_1 n_2)} + \frac{n_1 n_2}{N}\log(n_1 n_2)\Bigg), \text{ or} \tag{3.14a}$$

  $$S_{\mathsf{BL}_k}(v) - S_{\mathsf{BL}_k}(u) > 16(\zeta + 1)\Bigg(\sqrt{\frac{n_1 n_2}{N}|\mathsf{BL}_k|\log(n_1 n_2)} + \frac{n_1 n_2}{N}\log(n_1 n_2)\Bigg) \tag{3.14b}$$

  for some $k \in [K]$.

- Step 3: Compute a topological sort $\widehat{\pi}_{\mathsf{tds}}$ of the graph G; if none exists, set $\widehat{\pi}_{\mathsf{tds}} = \mathsf{id}$.
- Step 4: Repeat Steps 1–3 with $(Y^{(i)})^\top$ replacing $Y^{(i)}$ for $i = 1, 2$, the roles of n_1 and n_2 switched, and the roles of π and σ switched, to compute the permutation estimate $\widehat{\sigma}_{\mathsf{tds}}$.
- Step 5: Return the permutation estimates $(\widehat{\pi}_{\mathsf{tds}}, \widehat{\sigma}_{\mathsf{tds}})$.

The topological sorting step once again takes time $\mathcal{O}(n_1^2)$ and reading the matrix takes time $n_1 n_2$. Consequently, since $n_1 \geq n_2$, the construction of the graph G in Step 2 dominates the computational complexity, and takes time $\mathcal{O}(n_1^2 n_2/\beta) = \mathcal{O}(n_1^2 n_2^{1/2})$. Computing judiciously chosen partial row sums once again captures much more of the signal in the problem than entire row sums alone, and we obtain the following guarantee.

Theorem 3.4.2. *Suppose that $n_2 \leq n_1$. Then there exists an absolute constant $c_1 > 0$ such that for any matrix $M^* \in \mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}}$, we have*

$$\mathcal{F}(M^*, \widehat{M}(\widehat{\pi}_{\mathsf{tds}}, \widehat{\sigma}_{\mathsf{tds}})) \leq 1 \wedge c_1 \Big\{ \bar{\zeta}^2 \Big(\frac{n_1 \log n_1}{N} \Big)^{3/4} + \vartheta(N, n_1, n_2, \zeta) \Big\} \tag{3.15}$$

with probability exceeding $1 - 10(n_1 n_2)^{-3}$.

To interpret our rate in the Frobenius error, it is once again helpful to specialize to the case $n_2 \leq n_1 \leq C n_2$.

Example: $n_2 \leq n_1 \leq C n_2$ In this case, the Frobenius error guarantee simplifies exactly as before to

$$\frac{1}{(\log n_1)^2} \mathcal{F}(M^*, \widehat{M}(\widehat{\pi}_{\mathsf{tds}}, \widehat{\sigma}_{\mathsf{tds}})) \lesssim \begin{cases} 1 & \text{for } N \leq n_1 \\ \left(\frac{n_1}{N}\right)^{3/4} & \text{for } n_1 \leq N \leq n_1^3 \\ \frac{1}{\sqrt{N}} & \text{for } n_1^3 \leq N \leq n_1^4 \\ \frac{n_1 n_2}{N} & \text{for } n_1^4 \leq N \leq n_1^5 \\ \left(\frac{n_1}{N}\right)^{3/4} & \text{for } N \geq n_1^5. \end{cases} \tag{3.16}$$

Once again, comparing the bounds (3.10) and (3.16), we see that when $N \geq n_1^3$, our estimator (when combined with outputting the Y matrix when $N \geq n_1^5$) achieves the minimax lower bound up to poly-logarithmic factors.

This optimality is particularly notable because existing estimators [55, 59, 244, 274, 277] for the class $\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}}$ are only able to attain, in the regime $n_1 = n_2$, the rate

$$\frac{1}{(\log n_1)^2} \mathcal{F}(M^*, \widehat{M}_{\mathsf{prior}}) \lesssim \begin{cases} 1 & \text{for } N \leq n_1 \\ \left(\frac{n_1}{N}\right)^{1/2} & \text{for } N \geq n_1, \end{cases} \tag{3.17}$$

where we have used $\widehat{M}_{\mathsf{prior}}$ to indicator any such estimator from prior work. Thus, the nonparametric rates observed when N is large are completely washed out by the rate $\left(\frac{n_1}{N}\right)^{1/2}$, and so this prevents existing estimators from achieving minimax optimality in *any* regime of N.

Returning to our estimator, we see that in the regime $n_1 \leq N \leq n_1^2$, it falls short of being minimax-optimal, but breaks the conjectured statistical-computational barrier alluded to in the introduction. ♣

Note that Theorem 3.4.2 extends to estimation of matrices in the class $\mathbb{C}_{\mathsf{SST}}(n)$. In particular, we have $n_1 = n_2 = n$, and either of the two estimates $\widehat{\pi}_{\mathsf{tds}}$ or $\widehat{\sigma}_{\mathsf{tds}}$ may be returned as an estimate of the permutation π while preserving the same guarantees.

We conclude by noting that the sub-optimality of our estimator in the small-sample regime is not due to a weakness in the analysis. In particular, our analysis in this regime is also optimal up to poly-logarithmic factors; the rate $(\frac{n_1}{N})^{3/4}$ is indeed the rate attained by the two-dimensional sorting algorithm for the noisy sorting subclass of $\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}}$. In fact, a variant of this algorithm was used in a recursive fashion to successively improve the rate for noisy sorting matrices [212]; the first step of this algorithm generates an estimate with rate exactly $(\frac{n_1}{N})^{3/4}$.

3.5 Proofs of main results

In this appendix, we present all of our proofs. We begin by stating and proving technical lemmas that are used repeatedly in the proofs. We then prove our main results in the order in which they were stated.

Throughout the proofs, we assume without loss of generality that $M^* \in \mathbb{C}_{\mathsf{BISO}}(\mathsf{id}, \mathsf{id}) = \mathbb{C}_{\mathsf{BISO}}$. Because we are interested in rates of estimation up to universal constants, we assume that each independent sub-sample contains $N' = \mathsf{Poi}(N)$ observations (instead of $\mathsf{Poi}(N)/2$). We use the shorthand $Y = Y\big(\{(X_\ell, y_\ell)\}_{\ell=1}^{N'}\big)$, throughout. Our proof makes use of a few preliminary lemmas that are stated and proved in Appendix A.3.

3.5.1 Proof of Theorem 3.3.1

We split the proof into three distinct parts. We first prove the upper bound in part (a) of the theorem, and then prove the lower bound in part (b) in two separate sections. The proof in each section encompasses more than one regime of the theorem, and allows us to precisely pinpoint the source of the minimax lower bound.

Proof of part (a)

As argued, the bound is trivially true when $N \leq n_1$ by the boundedness of the set $\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}}$, so we assume that $N \geq n_1 \geq n_2$. Under our observation model (3.4), if we define $W = Y - M^*$, then Lemma A.3.1 implies that the assumption of Lemma A.3.2 is satisfied with $\alpha = c_1(\zeta + 1)\sqrt{\frac{n_1 n_2}{N}}$ and $\beta = c_1(\zeta + 1)\frac{n_1 n_2}{N}$ for a universal constant $c_1 > 0$. Therefore, Lemma A.3.2(a) yields that with probability at least $1 - n_1^{-3n_1}$, we have

$$\|\widehat{M}_{\mathsf{LS}}(\mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}}, Y) - M^*\|_F^2 \lesssim (\zeta+1)^2 \frac{n_1^2 n_2}{N} \log n_1 + (\zeta+1)\frac{n_1 n_2^2}{N}(\log n_1)^2$$
$$+ \left[(\zeta+1)\frac{n_1 n_2}{\sqrt{N}}(\log n_1)^2\right] \wedge \left[(\zeta+1)^2 \frac{n_1^2 n_2^2}{N}\log N\right] \wedge \left[(\zeta+1)^{4/3}\frac{n_1 n_2^{5/3}}{N^{2/3}}(\log n_1)^{2/3}\right].$$

Normalizing the bound by $1/(n_1 n_2)$ completes the proof. □

Proof of part (b): permutation error

We start with the term $\frac{n_1}{N} \wedge 1$ of the lower bound which stems from the unknown permutation on the rows.

Its proof is an application of Fano's lemma. The technique is standard, and we briefly review it here. Suppose we wish to estimate a parameter θ over an indexed class of distributions $\mathcal{P} = \{\mathbb{P}_\theta \mid \theta \in \Theta\}$ in the square of a (pseudo-)metric ρ. We refer to a subset of parameters $\{\theta^1, \theta^2, \ldots, \theta^K\}$ as a local (δ, ϵ)-packing set if

$$\min_{i,j\in[K],\, i\neq j} \rho(\theta^i, \theta^j) \geq \delta \qquad \text{and} \qquad \frac{1}{K(K-1)} \sum_{i,j\in[K],\, i\neq j} D(\mathbb{P}_{\theta^i} \| \mathbb{P}_{\theta^j}) \leq \epsilon.$$

Note that this set is a δ-packing in the metric ρ with the average Kullback-Leibler (KL) divergence bounded by ϵ. The following result is a straightforward consequence of Fano's inequality (see [302, Theorem 2.5]):

Lemma 3.5.1 (Local packing Fano lower bound)**.** *For any (δ, ϵ)-packing set of cardinality K, we have*

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \Theta} \mathbb{E}\left[\rho(\widehat{\theta}, \theta^*)^2\right] \geq \frac{\delta^2}{2}\left(1 - \frac{\epsilon + \log 2}{\log K}\right). \tag{3.18}$$

In addition, the Gilbert-Varshamov bound [121, 316] guarantees the existence of binary vectors $\{v^1, v^2, \ldots, v^K\} \subseteq \{0,1\}^{n_1}$ such that

$$K \geq 2^{c_1 n_1} \text{ and} \tag{3.19a}$$
$$\|v^i - v^j\|_2^2 \geq c_2 n_1 \text{ for each } i \neq j, \tag{3.19b}$$

for some fixed tuple of constants (c_1, c_2). We use this guarantee to design a packing of matrices in the class $\mathbb{C}^{\mathsf{r}}_{\mathsf{Perm}}$. For each $i \in [K]$, fix some $\delta \in [0, 1/4]$ to be precisely set later, and define the matrix M^i having identical columns, with entries given by

$$M^i_{j,k} = \begin{cases} 1/2, & \text{if } v^i_j = 0 \\ 1/2 + \delta, & \text{otherwise.} \end{cases} \tag{3.20}$$

Clearly, each of these matrices $\{M^i\}_{i=1}^K$ is a member of the class $\mathbb{C}^{\mathsf{r}}_{\mathsf{Perm}}$, and each distinct pair of matrices (M^i, M^j) satisfies the inequality $\|M^i - M^j\|_F^2 \geq c_2 n_1 n_2 \delta^2$.

Let $\mathbb{P}_M$ denote the probability distribution of the observations in the model (3.1) with underlying matrix $M \in \mathbb{C}^{\mathsf{r}}_{\mathsf{Perm}}$. Our observations are independent across entries of the matrix, and so the KL divergence tensorizes to yield

$$D(\mathbb{P}_{M^i} \| \mathbb{P}_{M^j}) = \sum_{\substack{k \in [n_1] \\ \ell \in [n_2]}} D(\mathbb{P}_{M^i_{k,\ell}} \| \mathbb{P}_{M^j_{k,\ell}}). \tag{3.21}$$

Let us now examine one term of this sum. Note that we observe $\kappa_{k,\ell} \sim \mathsf{Poi}(\frac{N}{n_1 n_2})$ samples of each entry (k, ℓ).

Under the Bernoulli observation model (3.3b), conditioned on the event $\kappa_{k,\ell} = \kappa$, we have the distributions

$$\mathbb{P}_{M^i_{k,\ell}} = \mathsf{Bin}(\kappa, M^i_{k,\ell}), \quad \text{and} \quad \mathbb{P}_{M^j_{k,\ell}} = \mathsf{Bin}(\kappa, M^j_{k,\ell}).$$

Consequently, the KL divergence conditioned on $\kappa_{k,\ell} = \kappa$ is given by

$$D(\mathbb{P}_{M^i_{k,\ell}} \| \mathbb{P}_{M^j_{k,\ell}}) = \kappa D(M^i_{k,\ell} \| M^j_{k,\ell}),$$

where we have used $D(p\|q) = p \log(\frac{p}{q}) + (1-p)\log(\frac{1-p}{1-q})$ to denote the KL divergence between the Bernoulli random variables Ber(p) and Ber(q).

Note that for $p, q \in [1/2, 3/4]$, we have

$$\begin{aligned} D(p\|q) &= p \log\left(\frac{p}{q}\right) + (1-p)\log\left(\frac{1-p}{1-q}\right) \\ &\overset{\text{(i)}}{\leq} p\left(\frac{p-q}{q}\right) + (1-p)\left(\frac{q-p}{1-q}\right) \\ &= \frac{(p-q)^2}{q(1-q)} \\ &\overset{\text{(ii)}}{\leq} \frac{16}{3}(p-q)^2. \end{aligned}$$

Here, step (i) follows from the inequality $\log x \leq x - 1$, and step (ii) from the assumption $q \in [\frac{1}{2}, \frac{3}{4}]$. Taking the expectation over κ, we have

$$D(\mathbb{P}_{M^i_{k,\ell}} \| \mathbb{P}_{M^j_{k,\ell}}) \leq \frac{16}{3}\frac{N}{n_1 n_2}(M^i_{k,\ell} - M^j_{k,\ell})^2 \leq \frac{16}{3}\frac{N}{n_1 n_2}\delta^2, \tag{3.22}$$

Summing over $k \in [n_1], \ell \in [n_2]$ yields $D(\mathbb{P}_{M^i} \| \mathbb{P}_{M^j}) \leq \frac{16}{3} N \delta^2$.

Under the standard Gaussian observation model (3.3a), a similar argument yields the bound $D(\mathbb{P}_{M^i} \| \mathbb{P}_{M^j}) \leq \frac{1}{2} N \delta^2$, since we have $D(\mathcal{N}(p, 1) \| \mathcal{N}(q, 1)) = (p-q)^2/2$.

Substituting into Fano's inequality (3.18), we have

$$\inf_{\widehat{M}} \sup_{M^* \in \mathbb{C}^{\mathsf{r}}_{\mathsf{Perm}}} \mathbb{E}\left[\|\widehat{M} - M^*\|_F^2\right] \geq \frac{c_2 n_1 n_2 \delta^2}{2}\left(1 - \frac{\frac{16}{3}N\delta^2 + \log 2}{c_3 n_1}\right).$$

Finally, if $N \geq c_4 n_1$ for a sufficiently large constant $c_4 > 0$, then we obtain the lower bound of order n_1/N by choosing $\delta^2 = c\frac{n_1}{N}$ for some constant $c > 0$ and normalizing by $1/(n_1 n_2)$. If $N \leq c_4 n_1$, then we simply choose δ to be a sufficiently small constant and normalize to obtain the lower bound of constant order. □

Proof of part (b): estimation error

We now turn to the term $\frac{1}{\sqrt{N}} \wedge \left(\frac{n_2}{N}\right)^{2/3} \wedge \frac{n_1 n_2}{N}$ of the lower bound which stems from estimation of the underlying bivariate isotonic matrix even if the permutations are given. This lower bound is partly known for the model of one observation per entry under Gaussian noise [56], and it suffices to slightly extend their proof to fit our model. The proof is based on Assouad's lemma.

Lemma 3.5.2 (Assouad's Lemma)**.** *Consider a parameter space Θ. Let $\mathbb{P}_\theta$ denote the distribution of the model given that the true parameter is $\theta \in \Theta$. Let $\mathbb{E}_\theta$ denote the corresponding expectation. Suppose that for each $\tau \in \{-1, 1\}^d$, there is an associated $\theta^\tau \in \Theta$. Then it holds that*

$$\inf_{\tilde{\theta}} \sup_{\theta^* \in \Theta} \mathbb{E}_{\theta^*} \ell^2(\tilde{\theta}, \theta^*) \geq \frac{d}{8} \min_{\tau \neq \tau'} \frac{\ell^2(\theta^\tau, \theta^{\tau'})}{d_H(\tau, \tau')} \min_{d_H(\tau, \tau')=1} (1 - \|\mathbb{P}_{\theta^\tau} - \mathbb{P}_{\theta^{\tau'}}\|_{TV}),$$

where ℓ denotes any distance function on Θ, d_H denotes the Hamming distance, $\|\cdot\|_{TV}$ denotes the total variation distance, and the infimum is taken over all estimators $\tilde{\theta}$ measurable with respect to the observation.

To apply the lemma, we construct a mapping from the hypercube to $\mathbb{C}_{\mathsf{BISO}} \subset \mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}}$. Consider integers $k_1 \in [n_1]$ and $k_2 \in [n_2]$, and let $m_1 = n_1/k_1$ and $m_2 = n_2/k_2$. Assume without loss of generality that m_1 and m_2 are integers. For each $\tau \in \{-1, 1\}^{k_1 \times k_2}$, define $M^\tau \in \mathbb{C}_{\mathsf{BISO}}$ in the following way. For $i \in [n_1]$ and $j \in [n_2]$, first identify the unique $u \in [k_1], v \in [k_2]$ for which $(u-1)m_1 < i \leq u m_1$ and $(v-1)m_2 < j \leq v m_2$, and then take

$$M^\tau_{i,j} = \lambda\Big(\frac{u + v - 1}{k_1 + k_2} + \frac{\tau_{u,v}}{2(k_1 + k_2)}\Big)$$

for $\lambda \in (0, 1]$ to be chosen. It is not hard to verify that $M^\tau \in \mathbb{C}_{\mathsf{BISO}}$.

We now proceed to show the lower bound using Assouad's lemma, which in our setup states that

$$\inf_{\tilde{M}} \sup_{M^* \in \mathbb{C}_{\mathsf{BISO}}} \mathbb{E}_{M^*}\|\tilde{M} - M^*\|_F^2 \geq \tag{3.23}$$
$$\frac{k_1 k_2}{8} \min_{\tau \neq \tau'} \frac{\|M^\tau - M^{\tau'}\|_F^2}{d_H(\tau, \tau')} \min_{d_H(\tau, \tau')=1} (1 - \|\mathbb{P}_{M^\tau} - \mathbb{P}_{M^{\tau'}}\|_{TV}).$$

For any $\tau, \tau' \in \{-1, 1\}^{k_1 \times k_2}$, it holds that

$$\begin{aligned}\|M^\tau - M^{\tau'}\|_F^2 &= \sum_{i,j} (M^\tau_{i,j} - M^{\tau'}_{i,j})^2 \\ &= \sum_{u \in [k_1], v \in [k_2]} \sum_{(u-1)m_1 < i \le u m_1} \sum_{(v-1)m_2 < j \le v m_2} (M^\tau_{i,j} - M^{\tau'}_{i,j})^2 \\ &= \sum_{u \in [k_1], v \in [k_2]} m_1 m_2 \frac{\lambda^2 (\tau_{u,v} - \tau'_{u,v})^2}{4(k_1 + k_2)^2} \\ &= \frac{\lambda^2 m_1 m_2}{(k_1 + k_2)^2} d_H(\tau, \tau'). \end{aligned} \tag{3.24}$$

As a result, we have

$$\min_{\tau \neq \tau'} \frac{\|M^\tau - M^{\tau'}\|_F^2}{d_H(\tau, \tau')} = \frac{\lambda^2 m_1 m_2}{(k_1 + k_2)^2}. \tag{3.25}$$

To bound $\|\mathbb{P}_{M^\tau} - \mathbb{P}_{M^{\tau'}}\|_{TV}$, we use Pinsker's inequality

$$\|\mathbb{P}_{M^\tau} - \mathbb{P}_{M^{\tau'}}\|_{TV}^2 \le \frac{1}{2} D(\mathbb{P}_{M^\tau} \| \mathbb{P}_{M^{\tau'}}).$$

Under either the Bernoulli or the standard Gaussian observation model, we have seen that by combining (3.21) and (3.22) (which hold even in this regime of N), the KL divergence can be bounded as

$$\begin{aligned} D(\mathbb{P}_{M^\tau} \| \mathbb{P}_{M^{\tau'}}) &\le \sum_{i \in [n_1],\, j \in [n_2]} \frac{16}{3} \frac{N}{n_1 n_2} (M^\tau_{i,j} - M^{\tau'}_{i,j})^2 \\ &= \frac{16}{3} \frac{N}{n_1 n_2} \|M^\tau - M^{\tau'}\|_F^2 \\ &= \frac{16}{3} \frac{N}{n_1 n_2} \frac{\lambda^2 m_1 m_2}{(k_1 + k_2)^2} = \frac{16}{3} \frac{\lambda^2 N}{k_1 k_2 (k_1 + k_2)^2}, \end{aligned}$$

where the second-to-last equality follows from (3.24) for any τ and τ' such that $d_H(\tau, \tau') = 1$. Therefore, it holds that

$$\min_{d_H(\tau, \tau') = 1} (1 - \|\mathbb{P}_{M^\tau} - \mathbb{P}_{M^{\tau'}}\|_{TV}) \ge \left(1 - \sqrt{\frac{8N}{3 k_1 k_2}} \frac{\lambda}{k_1 + k_2}\right). \tag{3.26}$$

Plugging (3.25) and (3.26) into Assouad's lemma (3.23), we obtain

$$\inf_{\tilde{M}} \sup_{M^* \in \mathbb{C}_{\mathsf{BISO}}} \mathbb{E}_{M^*} \|\tilde{M} - M^*\|_F^2 \ge \frac{\lambda^2 n_1 n_2}{8(k_1 + k_2)^2} \left(1 - \sqrt{\frac{8N}{3 k_1 k_2}} \frac{\lambda}{k_1 + k_2}\right). \tag{3.27}$$

Note that the bound (3.27) holds for all tuples (N, n_1, n_2). In order to obtain the various regimes, we must set particular values of k_1 and k_2. If $N \leq 1$, then setting $k_1 = k_2 = 1$ and λ to be a sufficiently small constant clearly gives the trivial bound. If $1 \leq N \leq n_2^4$, then we take $k_1 = k_2 = \lfloor N^{1/4} \rfloor \leq n_2$ and λ to be a sufficiently small positive constant so that

$$\inf_{\tilde{M}} \sup_{M^* \in \mathbb{C}_{\mathsf{BISO}}} \mathbb{E}_{M^*} \|\tilde{M} - M^*\|_F^2 \geq c_1 \frac{n_1 n_2}{\sqrt{N}},$$

for a constant $c_1 > 0$. If $n_2^4 \leq N \leq n_1^3 n_2$, we take $k_1 = \lfloor (N/n_2)^{1/3} \rfloor \leq n_1$, $k_2 = n_2$ and λ to be a sufficiently small positive constant so that

$$\inf_{\tilde{M}} \sup_{M^* \in \mathbb{C}_{\mathsf{BISO}}} \mathbb{E}_{M^*} \|\tilde{M} - M^*\|_F^2 \geq c_2 \Big(\frac{n_2}{N}\Big)^{2/3},$$

for a constant $c_2 > 0$. Finally, if $N \geq n_1^3 n_2$, we choose $k_1 = n_1$, $k_2 = n_2$ and $\lambda = \sqrt{\frac{3n_1 n_2}{32N}}(n_1 + n_2)$ to conclude that

$$\inf_{\tilde{M}} \sup_{M^* \in \mathbb{C}_{\mathsf{BISO}}} \mathbb{E}_{M^*} \|\tilde{M} - M^*\|_F^2 \geq \frac{3(n_1 n_2)^2}{256N}.$$

Normalizing the above bounds by $1/(n_1 n_2)$ yields the theorem. □

3.5.2 Proof of Corollary 3.3.1

The proof of this corollary follows from the steps used to establish Theorem 3.3.1. In particular, for the upper bound, applying Lemma A.3.2(b) with the same parameter choices as above yields that with probability at least $1 - n_1^{-3n_1}$, we have

$$\begin{aligned}&\|\widehat{M}_{\mathsf{LS}}(\mathbb{C}_{\mathsf{BISO}}, Y) - M^*\|_F^2 \lesssim (\zeta+1)\frac{n_1 n_2^2}{N}(\log n_1)^2 \\ &+ \Big[(\zeta+1)\frac{n_1 n_2}{\sqrt{N}}(\log n_1)^2\Big] \wedge \Big[(\zeta+1)^2 \frac{n_1^2 n_2^2}{N}\log N\Big] \wedge \Big[(\zeta+1)^{4/3}\frac{n_1 n_2^{5/3}}{N^{2/3}}(\log n_1)^{2/3}\Big].\end{aligned}$$

Normalizing the bound by $1/(n_1 n_2)$ proves the upper bound.

The lower bound established in Section 3.5.1 is valid for the class $\mathbb{C}_{\mathsf{BISO}}$, so the proof is complete. □

3.5.3 Proof of Proposition 3.4.1

Recall the definition of $\widehat{M}(\widehat{\pi}, \widehat{\sigma})$ in the meta-algorithm, and additionally, define the projection of any matrix $M \in \mathbb{R}^{n_1 \times n_2}$ onto $\mathbb{C}_{\mathsf{BISO}}(\pi, \sigma)$ as

$$\mathcal{P}_{\pi,\sigma}(M) = \arg \min_{\widetilde{M} \in \mathbb{C}_{\mathsf{BISO}}(\pi,\sigma)} \|M - \widetilde{M}\|_F^2.$$

Letting $W = Y^{(2)} - M^*$, we have

$$\begin{aligned}
\|\widehat{M}(\widehat{\pi}, \widehat{\sigma}) - M^*\|_F^2 &\overset{\text{(i)}}{\leq} 2\|\mathcal{P}_{\widehat{\pi},\widehat{\sigma}}(M^* + W) - \mathcal{P}_{\widehat{\pi},\widehat{\sigma}}(M^*(\widehat{\pi}, \widehat{\sigma}) + W)\|_F^2 \\
&\qquad + 2\|\mathcal{P}_{\widehat{\pi},\widehat{\sigma}}(M^*(\widehat{\pi}, \widehat{\sigma}) + W) - M^*\|_F^2 \\
&\overset{\text{(ii)}}{\leq} 2\|M^*(\widehat{\pi}, \widehat{\sigma}) - M^*\|_F^2 + 2\|\mathcal{P}_{\widehat{\pi},\widehat{\sigma}}(M^*(\widehat{\pi}, \widehat{\sigma}) + W) - M^*\|_F^2 \\
&\overset{\text{(iii)}}{\leq} 4\|\mathcal{P}_{\widehat{\pi},\widehat{\sigma}}(M^*(\widehat{\pi}, \widehat{\sigma}) + W) - M^*(\widehat{\pi}, \widehat{\sigma})\|_F^2 + 6\|M^*(\widehat{\pi}, \widehat{\sigma}) - M^*\|_F^2, \qquad (3.28)
\end{aligned}$$

where step (ii) follows from the non-expansiveness of a projection onto a convex set, and steps (i) and (iii) from the triangle inequality.

The first term in equation (3.28) is the estimation error of a bivariate isotonic matrix with known permutations. Therefore, applying Corollary 3.3.1 with a union bound over all permutations $\widehat{\pi} \in \mathfrak{S}_{n_1}$ and $\widehat{\sigma} \in \mathfrak{S}_{n_2}$ yields the bound

$$\|\mathcal{P}_{\widehat{\pi},\widehat{\sigma}}(M^*(\widehat{\pi}, \widehat{\sigma}) + W) - M^*(\widehat{\pi}, \widehat{\sigma})\|_F^2 \leq C\vartheta(N, n_1 \vee n_2, n_1 \wedge n_2, \zeta)$$

on the estimation error with probability at least[7] $1 - n_1^{-n_1}$.

The approximation error can be split into two components: one along the rows of the matrix, and the other along the columns. More explicitly, we have

$$\begin{aligned}
\|M^* - M^*(\widehat{\pi}, \widehat{\sigma})\|_F^2 &\leq 2\|M^* - M^*(\widehat{\pi}, \mathsf{id})\|_F^2 + 2\|M^*(\widehat{\pi}, \mathsf{id}) - M^*(\widehat{\pi}, \widehat{\sigma})\|_F^2 \\
&= 2\|M^* - M^*(\widehat{\pi}, \mathsf{id})\|_F^2 + 2\|M^* - M^*(\mathsf{id}, \widehat{\sigma})\|_F^2.
\end{aligned}$$

This completes the proof of the proposition. □

3.5.4 Proof of Theorem 3.4.1

In order to ease the notation, we adopt the shorthand

$$\eta := 16(\zeta + 1)\Big(\sqrt{\frac{n_1 n_2^2}{N}\log(n_1 n_2)} + \frac{n_1 n_2}{N}\log(n_1 n_2)\Big),$$

and for each block $B \in \mathsf{bl}_{\mathsf{ref}}$ in Algorithm 2, we use the shorthand

$$\eta_B := 16(\zeta + 1)\Big(\sqrt{\frac{|B| n_1 n_2}{N}\log(n_1 n_2)} + \frac{n_1 n_2}{N}\log(n_1 n_2)\Big) \qquad (3.29)$$

throughout the proof. Applying Lemma A.3.1 with $\mathcal{S} = \{i\} \times [n_2]$ and then with $\mathcal{S} = \{i\} \times B$ for each $i \in [n_1]$ and $B \in \mathsf{bl}_{\mathsf{ref}}$, we obtain

$$\Pr\left\{\Big|S(i) - \sum_{\ell \in [n_2]} M^*_{i,\ell}\Big| \geq \frac{\eta}{2}\right\} \leq 2(n_1 n_2)^{-4}, \qquad (3.30a)$$

[7] Choosing the constant C to be twice the constant in Corollary 3.3.1, we can boost the probability of the good event in Corollary 3.3.1 to $1 - n_1^{-2n_1}$, and applying a union bound over at most $2n_1!$ permutations yields the claimed result.

and

$$\Pr\left\{\left|S_B(i) - \sum_{\ell \in B} M^*_{i,\ell}\right| \geq \frac{\eta_B}{2}\right\} \leq 2(n_1 n_2)^{-4}. \tag{3.30b}$$

A union bound over all rows and blocks yields that $\Pr\{\mathcal{E}\} \geq 1 - 2(n_1 n_2)^{-3}$, where we define the event

$$\mathcal{E} := \left\{\left|S(i) - \sum_{\ell \in [n_2]} M^*_{i,\ell}\right| \leq \frac{\eta}{2} \text{ and } \left|S_B(i) - \sum_{\ell \in B} M^*_{i,\ell}\right| \leq \frac{\eta_B}{2} \text{ for all } i \in [n_1], B \in \mathsf{bl}_{\mathsf{ref}}\right\}.$$

We now condition on event $\mathcal{E}$. Applying the triangle inequality, if

$$S(v) - S(u) > \eta \quad \text{or} \quad S_B(v) - S_B(u) > \eta_B,$$

then we have

$$\sum_{\ell \in [n_2]} M^*_{v,\ell} - \sum_{\ell \in [n_2]} M^*_{u,\ell} > 0 \quad \text{or} \quad \sum_{\ell \in B} M^*_{v,\ell} - \sum_{\ell \in B} M^*_{u,\ell} > 0.$$

It follows that $u < v$ since M^* has non-decreasing columns. Thus, by the choice of thresholds η and η_B in the algorithm, we have guaranteed that every edge $u \to v$ in the graph G is consistent with the underlying permutation id, so a topological sort exists on event $\mathcal{E}$.

We need the following lemma, whose proof is deferred to Section A.4.

Lemma 3.5.3. *Let B be a subset of $[n_2]$ and let η_B be defined by* (3.29). *Suppose that $\widehat{\pi}$ is a topological sort of a graph G, where an edge $u \to v$ is present whenever*

$$\sum_{\ell \in B} M^*_{v,\ell} - \sum_{\ell \in B} M^*_{u,\ell} > 2\eta_B.$$

Then for all $i \in [n_1]$, we have

$$\sum_{j \in B} \left|M^*_{\widehat{\pi}(i),j} - M^*_{i,j}\right| \leq 96(\zeta + 1)\sqrt{\frac{n_1 n_2}{N}|B| \log(n_1 n_2)}.$$

If we have

$$\sum_{\ell \in [n_2]} M^*_{v,\ell} - \sum_{\ell \in [n_2]} M^*_{u,\ell} > 2\eta \quad \text{or} \quad \sum_{\ell \in B} M^*_{v,\ell} - \sum_{\ell \in B} M^*_{u,\ell} > 2\eta_B,$$

then the triangle inequality implies that

$$S(v) - S(u) > \eta \quad \text{or} \quad S_B(v) - S_B(u) > \eta_B.$$

Hence, the edge $u \to v$ is present in the graph G. As we defined $\widehat{\pi}_{\mathsf{ref}}$ as a topological sort of G, Lemma 3.5.3 implies that

$$\sum_{j \in [n_2]} \left| M^*_{\widehat{\pi}_{\mathsf{ref}}(i),j} - M^*_{i,j} \right| \leq 96(\zeta+1)\sqrt{\frac{n_1 n_2^2}{N}\log(n_1 n_2)} \quad \text{for all } i \in [n_1], \text{ and} \tag{3.31a}$$

$$\sum_{j \in B} \left| M^*_{\widehat{\pi}_{\mathsf{ref}}(i),j} - M^*_{i,j} \right| \leq 96(\zeta+1)\sqrt{\frac{n_1 n_2}{N}|B|\log(n_1 n_2)} \quad \text{for all } i \in [n_1], B \in \mathsf{bl}_{\mathsf{ref}}. \tag{3.31b}$$

Critical to the rest of our analysis is the following lemma:

Lemma 3.5.4. *For a vector $v \in \mathbb{R}^n$, define its variation as $\mathrm{var}(v) = \max_i v_i - \min_i v_i$. Then we have*

$$\|v\|_2^2 \leq \mathrm{var}(v)\|v\|_1 + \|v\|_1^2/n.$$

See Section A.4 for the proof of this lemma.

For each $i \in [n_1]$, define Δ^i to be the i-th row difference $M^*_{\widehat{\pi}_{\mathsf{ref}}(i)} - M^*_i$, and for each block $B \in \mathsf{bl}_{\mathsf{ref}}$, denote the restriction of Δ^i to B by Δ^i_B. Lemma 3.5.4 applied with $v = \Delta^i_B$ yields

$$\begin{aligned}
\|\Delta^i\|_2^2 &= \sum_{B \in \mathsf{bl}_{\mathsf{ref}}} \|\Delta^i_B\|_2^2 \\
&\leq \sum_{B \in \mathsf{bl}_{\mathsf{ref}}} \mathrm{var}(\Delta^i_B)\|\Delta^i_B\|_1 + \sum_{B \in \mathsf{bl}_{\mathsf{ref}}} \frac{\|\Delta^i_B\|_1^2}{|B|} \\
&\leq \left(\max_{B \in \mathsf{bl}_{\mathsf{ref}}} \|\Delta^i_B\|_1\right)\left(\sum_{B \in \mathsf{bl}_{\mathsf{ref}}} \mathrm{var}\left(\Delta^i_B\right)\right) + \frac{\max_{B \in \mathsf{bl}_{\mathsf{ref}}} \|\Delta^i_B\|_1}{\min_{B \in \mathsf{bl}_{\mathsf{ref}}} |B|} \sum_{B \in \mathsf{bl}_{\mathsf{ref}}} \|\Delta^i_B\|_1.
\end{aligned} \tag{3.32}$$

We now analyze the quantities in inequality (3.32). By the definition of the blocking BL, we have

$$\frac{1}{2} n_2 \sqrt{\frac{n_1}{N}\log(n_1 n_2)} \leq |B| \leq n_2 \sqrt{\frac{n_1}{N}\log(n_1 n_2)}.$$

Additionally, the bounds (3.31a) and (3.31b) imply that

$$\sum_{B \in \mathsf{bl}_{\mathsf{ref}}} \|\Delta^i_B\|_1 = \|\Delta^i\|_1 \leq 96(\zeta+1) n_2 \sqrt{\frac{n_1}{N}\log(n_1 n_2)}, \text{ and}$$

$$\|\Delta^i_B\|_1 \leq 96(\zeta+1) n_2 \Big(\frac{n_1}{N}\log(n_1 n_2)\Big)^{3/4} \text{ for all } B \in \mathsf{bl}_{\mathsf{ref}}.$$

Moreover, we have

$$\begin{aligned}
\sum_{B \in \mathsf{bl}_{\mathsf{ref}}} \mathrm{var}\left(\Delta^i_B\right) &\leq \sum_{B \in \mathsf{bl}_{\mathsf{ref}}} \left[\mathrm{var}(M^*_{i,B}) + \mathrm{var}(M^*_{\widehat{\pi}_{\mathsf{ref}}(i),B})\right] \\
&\leq \mathrm{var}(M^*_i) + \mathrm{var}(M^*_{\widehat{\pi}_{\mathsf{ref}}(i)}) \leq 2,
\end{aligned}$$

because M^* has monotone rows in $[0,1]^{n_2}$. Finally, plugging all the pieces into equation (3.32) yields

$$\|\Delta^i\|_2^2 \lesssim (\zeta \vee 1) n_2 \left(\frac{n_1 \log n_1}{N}\right)^{3/4}.$$

Normalizing this bound by $1/n_2$, summing over the rows, and applying Proposition 3.4.1, we obtain the bound (3.12) on the Frobenius error. □

3.5.5 Proof of Theorem 3.4.2

The beginning of the proof proceeds in the same way as the proof of Theorem 3.4.1, so that we provide only a sketch. We apply Lemma A.3.1 with $\mathcal{S} = \{i\} \times [n_2]$ and $\mathcal{S} = \{i\} \times \mathsf{BL}_k$ for each tuple $i \in [n_1], k \in [K]$, and use the fact that $K \le n_2/\beta \le n_2^{1/2}$, to obtain that with probability at least $1 - 2(n_1 n_2)^{-3}$, all the full row sums of $Y^{(2)}$ and all the partial row sums over the column blocks concentrate well around their means. By virtue of the conditions (3.14a) and (3.14b), we see that every edge $u \to v$ in the graph G is consistent with the underlying permutation so that a topological sort exists with probability at least $1 - 2(n_1 n_2)^{-3}$. Additionally, it follows from Lemma 3.5.3 and the same argument leading to equations (3.31a) and (3.31b) that for all $i \in [n_1]$, we have

$$\Big|\sum_{j \in [n_2]} (M^*_{\widehat{\pi}_{\mathsf{tds}}(i),j} - M^*_{i,j})\Big| \le 96(\zeta+1)\sqrt{\frac{n_1 n_2^2}{N}\log(n_1 n_2)}, \text{ and} \tag{3.33a}$$

$$\Big|\sum_{j \in \mathsf{BL}_k} (M^*_{\widehat{\pi}_{\mathsf{tds}}(i),j} - M^*_{i,j})\Big| \le 96(\zeta+1)\sqrt{\frac{n_1 n_2}{N}|\mathsf{BL}_k|\log(n_1 n_2)} \quad \text{for all } k \in [K], \tag{3.33b}$$

with probability at least $1 - 2(n_1 n_2)^{-3}$.

On the other hand, we apply Lemma A.3.1 with $\mathcal{S} = [n_1] \times \{j\}$ to obtain concentration for the column sums of $Y^{(1)}$:

$$\Big|C(j) - \sum_{i=1}^{n_1} M^*_{i,j}\Big| \le 8(\zeta+1)\left(\sqrt{\frac{n_1^2 n_2}{N}\log(n_1 n_2)} + \frac{n_1 n_2}{N}\log(n_1 n_2)\right) \tag{3.34}$$

for all $j \in [n_2]$ with probability at least $1 - 2(n_1 n_2)^{-3}$. We carry out the remainder of the proof conditioned on the event of probability at least $1 - 4(n_1 n_2)^{-3}$ that inequalities (3.33a), (3.33b) and (3.34) hold.

Having stated the necessary bounds, we now split the remainder of the proof into two parts for convenience. In order to do so, we first split the set BL into two disjoint sets of blocks, depending on whether a block comes from an originally large block (of size larger than $\beta = n_2\sqrt{\frac{n_1}{N}\log(n_1 n_2)}$ as in Step 3 of Subroutine 1) or from an aggregation of small blocks. More formally, define the sets

$$\mathsf{BL}^{\mathbb{L}} := \{B \in \mathsf{BL} : B \text{ was not obtained via aggregation}\}, \text{ and}$$
$$\mathsf{BL}^{\mathbb{S}} := \mathsf{BL} \setminus \mathsf{BL}^{\mathbb{L}}.$$

For a set of blocks B, define the shorthand $\cup\mathsf{B} = \bigcup_{B \in \mathsf{B}} B$ for convenience. We begin by focusing on the blocks $\mathsf{BL}^{\mathbb{L}}$.

Error on columns indexed by $\cup\mathsf{BL}^{\mathbb{L}}$

Recall that when the columns of the matrix are ordered according to $\widehat{\sigma}_{\mathsf{pre}}$, the blocks in $\mathsf{BL}^{\mathbb{L}}$ are contiguous and thus have an intrinsic ordering. We index the blocks according to this ordering as $B_1, B_2, \ldots, B_\ell$ where $\ell = |\mathsf{BL}^{\mathbb{L}}|$. Now define the disjoint sets

$$\mathsf{BL}^{(1)} := \{B_k \in \mathsf{BL}^{\mathbb{L}} : k = 0 \pmod 2\}, \text{ and}$$
$$\mathsf{BL}^{(2)} := \{B_k \in \mathsf{BL}^{\mathbb{L}} : k = 1 \pmod 2\}.$$

Let $\ell_t = |\mathsf{BL}^{(t)}|$ for each $t = 1, 2$.

Recall that each block B_k in $\mathsf{BL}^{\mathbb{L}}$ remains unchanged after aggregation, and that the threshold we used to block the columns is $\tau = 16(\zeta + 1)\big(\sqrt{\frac{n_1^2 n_2}{N} \log(n_1 n_2)} + 2\frac{n_1 n_2}{N} \log(n_1 n_2)\big)$. Hence, applying the concentration bound (3.34) together with the definition of blocks in Step 2 of Subroutine 1 yields

$$\Big| \sum_{i=1}^{n_1} M^*_{i,j_1} - \sum_{i=1}^{n_1} M^*_{i,j_2} \Big| \le 96(\zeta + 1)\sqrt{\frac{n_1^2 n_2}{N} \log(n_1 n_2)} \quad \text{for all } j_1, j_2 \in B_k, \tag{3.36}$$

where we again used the argument leading to the bounds (3.31a) and (3.31b) to combine the two terms. Moreover, since the threshold is twice the concentration bound, it holds that under the true ordering id, every index in B_k precedes every index in B_{k+2} for any $k \in [K-2]$. By definition, we have thus ensured that the blocks in $\mathsf{BL}^{(t)}$ do not “mix” with each other.

The rest of the argument hinges on the following lemma, which is proved in Section A.4.

Lemma 3.5.5. *For $m \in \mathbb{Z}_+$, let $J_1 \sqcup \cdots \sqcup J_\ell$ be a partition of $[m]$ such that each J_k is contiguous and J_k precedes J_{k+1}. Let $a_k = \min J_k$, $b_k = \max J_k$ and $m_k = |J_k|$. Let A be a matrix in $[0,1]^{n \times m}$ with non-decreasing rows and non-decreasing columns. Suppose that*

$$\sum_{i=1}^{n} (A_{i,b_k} - A_{i,a_k}) \le \chi \text{ for each } k \in [\ell] \text{ and some } \chi \ge 0.$$

Additionally, suppose that there are positive reals $\rho, \rho_1, \rho_2, \ldots, \rho_\ell$, and a permutation π such that for any $i \in [n]$, we have (i)$\sum_{j=1}^{m} |A_{\pi(i),j} - A_{i,j}| \le \rho$, and (ii)$\sum_{j \in J_k} |A_{\pi(i),j} - A_{i,j}| \le \rho_k$ for each $k \in [\ell]$. Then it holds that

$$\sum_{i=1}^{n} \sum_{j=1}^{m} (A_{\pi(i),j} - A_{i,j})^2 \le 2\chi \sum_{k=1}^{\ell} \rho_k + n\rho \max_{k \in [\ell]} \frac{\rho_k}{m_k}.$$

We apply the lemma as follows. For $t = 1, 2$, let the matrix $M^{(t)}$ be the sub-matrix of M^* restricted to the columns indexed by the indices in $\cup\mathsf{BL}^{(t)}$. The matrix $M^{(t)}$ has non-decreasing rows and columns by assumption. We have shown that the blocks in $\mathsf{BL}^{(t)}$ do not mix with each other, so they are contiguous and correctly ordered in $M^{(t)}$. Moreover, the inequality assumptions of the lemma correspond to the bounds (3.36), (3.33a) and (3.33b) respectively, by setting $J_1, \dots, J_\ell$ to be the blocks in $\mathsf{BL}^{(t)}$, and with the substitutions $A = M^{(t)}$, $n = n_1$, $m = | \cup \mathsf{BL}^{(t)}|$,

$$\begin{aligned}\chi &= 96(\zeta + 1)\sqrt{\frac{n_1^2 n_2}{N} \log(n_1 n_2)},\\ \rho &= 96(\zeta + 1)\sqrt{\frac{n_1 n_2^2}{N} \log(n_1 n_2)}, \text{ and}\\ \rho_k &= 96(\zeta + 1)\sqrt{\frac{n_1 n_2}{N} |J_k| \log(n_1 n_2)}.\end{aligned}$$

Moreover, we have

$$\begin{aligned}\chi \sum_{k=1}^{\ell} \rho_k &\lesssim (\zeta^2 \vee 1) \frac{n_1^{3/2} n_2}{N} \log(n_1 n_2) \sum_{B \in \mathsf{BL}^{(t)}} \sqrt{|B|}\\ &\overset{\text{(i)}}{\leq} (\zeta^2 \vee 1) \frac{n_1^{3/2} n_2}{N} \log(n_1 n_2) \sqrt{\sum_{B \in \mathsf{BL}^{(t)}} |B|} \sqrt{\ell_t}\\ &\overset{\text{(ii)}}{\leq} \frac{(\zeta^2 \vee 1)}{\sqrt{\beta}} \frac{n_1^{3/2} n_2^2}{N} \log(n_1 n_2),\end{aligned} \tag{3.37}$$

where step (i) follows from the Cauchy–Schwarz inequality, and step (ii) from the fact that $\sum_{B \in \mathsf{BL}^{(t)}} |B| \leq n_2$ and that by assumption of large blocks, we have $\min_{B \in \mathsf{BL}^{(t)}} |B| \geq \beta$ so that $\ell_t \leq n_2/\beta$.

We also have

$$\begin{aligned}n\rho \max_{k \in [\ell]} \frac{\rho_k}{m_k} &= (\zeta^2 \vee 1) \frac{n_1^2 n_2^{3/2}}{N} \log(n_1 n_2) \max_{B \in \mathsf{BL}^{(t)}} \frac{\sqrt{|B|}}{|B|}\\ &\leq \frac{(\zeta^2 \vee 1)}{\sqrt{\beta}} \frac{n_1^2 n_2^{3/2}}{N} \log(n_1 n_2),\end{aligned} \tag{3.38}$$

where we have again used the fact that $\min_{B \in \mathsf{BL}^{(t)}} |B| \geq \beta$. Putting together the bounds (3.37) and (3.38) and applying Lemma 3.5.5 yields

$$\sum_{i \in [n_1]} \sum_{j \in \cup \mathsf{BL}^{(t)}} (M^*_{\widehat{\pi}_{\mathsf{tds}}(i), j} - M^*_{i,j})^2 \lesssim \frac{(\zeta^2 \vee 1)}{\sqrt{\beta}} (n_1 n_2)^{3/2} (n_1 \vee n_2)^{1/2} \frac{\log(n_1 n_2)}{N}. \tag{3.39}$$

Substituting $\beta = n_2 \sqrt{\frac{n_1}{N} \log(n_1 n_2)}$ and normalizing by $n_1 n_2$ proves the result for the set of blocks $\mathsf{BL}^{(t)}$. Summing over $t = 1, 2$ then yields a bound of twice the size for columns of the matrix indexed by $\cup\mathsf{BL}^{\mathbb{L}}$.

Error on columns indexed by $\cup\mathsf{BL}^{\mathbb{S}}$

Next, we bound the approximation error of each row of the matrix with column indices restricted to the union of all small blocks. Roughly speaking, all small (sub-)blocks are aggregated into those that have size of order β; by definition of the blocking, this implies that the ambiguity in the column permutation for the aggregated block only exists within the small sub-blocks, and in that sense, this column permutation can be thought of as "essentially known". Thus, the proof resembles that of Theorem 3.4.1: it is sufficient (for the eventual bound we target) to bound the Frobenius error by the maximum error on the rows. Note that we must also make modifications that account for the fact that the column permutation is only approximately known. We split the proof into two cases.

Case 1 Let us first address the easy case where $\mathsf{BL}^{\mathbb{S}}$ contains a single block of size less than $\frac{\beta}{2} = \frac{1}{2}n_2\sqrt{\frac{n_1}{N}\log(n_1n_2)}$. Here, we have

$$\begin{aligned}
\sum_{i\in[n_1]}\sum_{j\in\cup\mathsf{BL}^{\mathbb{S}}}(M^*_{\widehat{\pi}_{\mathsf{tds}}(i),j} - M^*_{i,j})^2 &\overset{\text{(i)}}{\le} \sum_{i\in[n_1]}\sum_{j\in\cup\mathsf{BL}^{\mathbb{S}}}\left|M^*_{\widehat{\pi}_{\mathsf{tds}}(i),j} - M^*_{i,j}\right| \\
&\overset{\text{(ii)}}{=} \sum_{i\in[n_1]}\Big|\sum_{j\in\cup\mathsf{BL}^{\mathbb{S}}}(M^*_{\widehat{\pi}_{\mathsf{tds}}(i),j} - M^*_{i,j})\Big| \\
&\overset{\text{(iii)}}{\le} \sum_{i\in[n_1]} 96(\zeta+1)\sqrt{\frac{n_1n_2}{2N}\beta\log(n_1n_2)} \\
&= 48\sqrt{2}(\zeta+1)\frac{n_1^{3/2}n_2\,(n_1\vee n_2)^{1/4}}{N^{3/4}}\log^{3/4}(n_1n_2),
\end{aligned}$$

where step (i)follows from the Hölder's inequality and the fact that $M^*\in[0,1]^{n_1\times n_2}$, step (ii)from the monotonicity of the columns of M^*, and step (iii)from equation (3.33b). We have thus proved the theorem for this case.

Case 2 Let us now consider the case where $\mathsf{BL}^{\mathbb{S}}$ contains multiple blocks. For each $i\in[n_1]$, define Δ^i to be the restriction of the i-th row difference $M^*_{\widehat{\pi}_{\mathsf{tds}}(i)} - M^*_i$ to the union of blocks $\cup\mathsf{BL}^{\mathbb{S}}$. For each block $B\in\mathsf{BL}^{\mathbb{S}}$, denote the restriction of Δ^i to B by Δ^i_B. Lemma 3.5.4 applied with $v=\Delta^i$ yields

$$\begin{aligned}
\|\Delta^i\|_2^2 &= \sum_{B\in\mathsf{BL}^{\mathbb{S}}}\|\Delta^i_B\|_2^2 \\
&\le \sum_{B\in\mathsf{BL}^{\mathbb{S}}}\mathrm{var}(\Delta^i_B)\|\Delta^i_B\|_1 + \sum_{B\in\mathsf{BL}^{\mathbb{S}}}\frac{\|\Delta^i_B\|_1^2}{|B|} \\
&\le \left(\max_{B\in\mathsf{BL}^{\mathbb{S}}}\|\Delta^i_B\|_1\right)\left(\sum_{B\in\mathsf{BL}^{\mathbb{S}}}\mathrm{var}\left(\Delta^i_B\right)\right) + \frac{\max_{B\in\mathsf{BL}^{\mathbb{S}}}\|\Delta^i_B\|_1}{\min_{B\in\mathsf{BL}^{\mathbb{S}}}|B|}\sum_{B\in\mathsf{BL}^{\mathbb{S}}}\|\Delta^i_B\|_1. \qquad (3.40)
\end{aligned}$$

We now analyze the quantities in inequality (3.40). By the aggregation step of Subroutine 1, we have $\frac{1}{2}\beta \le |B| \le 2\beta$, where $\beta = n_2\sqrt{\frac{n_1}{N}\log(n_1 n_2)}$. Additionally, the bounds (3.33a) and (3.33b) imply that

$$\sum_{B \in \mathsf{BL}^{\mathbb{S}}} \|\Delta^i_B\|_1 = \|\Delta^i\|_1 \le 96(\zeta+1)\sqrt{\frac{n_1 n_2^2}{N}\log(n_1 n_2)} \lesssim (\zeta+1)\beta, \text{ and}$$

$$\begin{aligned} \|\Delta^i_B\|_1 &\le 96(\zeta+1)\sqrt{\frac{n_1 n_2}{N}|B|\log(n_1 n_2)} \\ &\le 96\sqrt{2}(\zeta+1)\sqrt{\frac{n_1 n_2}{N}\beta\log(n_1 n_2)} \text{ for all } B \in \mathsf{BL}^{\mathbb{S}}. \end{aligned}$$

Moreover, to bound the quantity $\sum_{B\in\mathsf{BL}^{\mathbb{S}}} \mathrm{var}\left(\Delta^i_B\right)$, we proceed as in the proof for the large blocks in $\mathsf{BL}^{\mathbb{L}}$. Recall that if we permute the columns by $\widehat{\sigma}_{\mathsf{pre}}$ according to the column sums, then the blocks in $\mathsf{BL}^{\mathbb{S}}$ have an intrinsic ordering, even after adjacent small blocks are aggregated. Let us index the blocks in $\mathsf{BL}^{\mathbb{S}}$ by $B_1, B_2, \ldots, B_m$ according to this ordering, where $m = |\mathsf{BL}^{\mathbb{S}}|$. As before, the odd-indexed (or even-indexed) blocks do not mix with each other under the true ordering id, because the threshold used to define the blocks is larger than twice the column sum perturbation. We thus have

$$\begin{aligned} \sum_{B\in\mathsf{BL}^{\mathbb{S}}} \mathrm{var}\left(\Delta^i_B\right) &= \sum_{\substack{k\in[m] \\ k \text{ odd}}} \mathrm{var}(\Delta^i_{B_k}) + \sum_{\substack{k\in[m] \\ k \text{ even}}} \mathrm{var}(\Delta^i_{B_k}) \\ &\le \sum_{\substack{k\in[m] \\ k \text{ odd}}} \left[\mathrm{var}(M^*_{i,B_k}) + \mathrm{var}(M^*_{\widehat{\pi}_{\mathsf{tds}}(i),B_k})\right] \\ &\qquad + \sum_{\substack{k\in[m] \\ k \text{ even}}} \left[\mathrm{var}(M^*_{i,B_k}) + \mathrm{var}(M^*_{\widehat{\pi}_{\mathsf{tds}}(i),B_k})\right] \\ &\overset{\text{(i)}}{\le} 2\,\mathrm{var}(M^*_i) + 2\,\mathrm{var}(M^*_{\widehat{\pi}_{\mathsf{tds}}(i)}) \overset{\text{(ii)}}{\le} 4, \end{aligned}$$

where inequality (i)holds because the odd (or even) blocks do not mix, and inequality (ii)holds because M^* has monotone rows in $[0,1]^{n_2}$.

Finally, putting together all the pieces, we can substitute for β, sum over the indices $i \in n_1$, and normalize by $n_1 n_2$ to obtain

$$\frac{1}{n_1 n_2}\sum_{i\in[n_1]} \|\Delta^i\|_2^2 \lesssim (\zeta^2 \vee 1)\left(\frac{n_1\log(n_1 n_2)}{N}\right)^{3/4}, \tag{3.41}$$

and so the error on columns indexed by the set $\cup\mathsf{BL}^{\mathbb{S}}$ is bounded as desired.

Combining the bounds (3.39) and (3.41), we conclude that

$$\frac{1}{n_1 n_2}\|M^*(\widehat{\pi}_{\mathsf{tds}}, \mathsf{id}) - M^*\|_F^2 \lesssim (\zeta^2 \vee 1) n_1^{1/4}(n_1 \vee n_2)^{1/2}\left(\frac{\log(n_1 n_2)}{N}\right)^{3/4}$$

with probability at least $1 - 4(n_1 n_2)^{-3}$. The same proof works with the roles of n_1 and n_2 switched and all the matrices transposed, so we have

$$\frac{1}{n_1 n_2} \| M^*(\mathsf{id}, \widehat{\sigma}_{\mathsf{tds}}) - M^* \|_F^2 \lesssim (\zeta^2 \vee 1) n_2^{1/4} \, (n_1 \vee n_2)^{1/2} \left(\frac{\log(n_1 n_2)}{N} \right)^{3/4}$$

with the same probability. Consequently,

$$\frac{1}{n_1 n_2} \left(\| M^*(\widehat{\pi}_{\mathsf{tds}}, \mathsf{id}) - M^* \|_F^2 + \| M^*(\mathsf{id}, \widehat{\sigma}_{\mathsf{tds}}) - M^* \|_F^2 \right) \lesssim (\zeta^2 \vee 1) \Big(\frac{n_1 \log n_1}{N} \Big)^{3/4}$$

with probability at least $1 - 8(n_1 n_2)^{-3}$, where we have used the relation $n_1 \geq n_2$. Applying Proposition 3.4.1 completes the proof. □

3.6 Summary and open questions

We have studied the class of permutation-based models in two distinct metrics. A notable consequence of our results is that our polynomial-time algorithms are able to achieve the minimax lower bound in the Frobenius error up to a poly-logarithmic factor provided the sample size grows to be large. Moreover, we have overcome a crucial bottleneck in previous analyses that underlay a statistical-computational gap; see the full paper [211] for a more detailed discussion. Several intriguing questions related to estimating such matrices remain:

What is the fastest Frobenius error rate achievable by computationally efficient estimators in the partially observed setting when N is small?

As a partial answer to the first question, it can be shown that when the informal algorithm described at the beginning of Section 3.4.2 is recursed in the natural way and applied to the noisy sorting subclass of the SST model, it yields another minimax-optimal estimator for noisy sorting, similar to the multistage algorithm of Mao et al. [212]. However, this same guarantee is preserved for neither the larger class of matrices $\mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}}$, nor for its sub-class $\mathbb{C}^{\mathsf{r}}_{\mathsf{Perm}}$. Improving the rate will likely require techniques that are beyond the reach of those introduced in this chapter. Indeed, in very recent work, Liu and Moitra [201] provide a more involved algorithm that uses "blocks" of partially sorted forms of the matrix to produce a minimax optimal estimator over the class $\mathbb{C}^{\mathsf{r}}_{\mathsf{Perm}}$ and another estimator that further narrows the statistical-computational gap in estimating matrices in the class $\mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}}$.

It is also worth noting that model (3.1) allowed us to perform sample-splitting in Algorithm 3 to preserve independence across observations, so that Step 2 is carried out on a sample that is independent of the blocking generated in Step 1. Thus, our proofs also hold for the observation model where we have *exactly* 2 independent samples per entry of the matrix.

It is natural to wonder if just one independent sample per entry suffices, and whether sample splitting is required at all. Reasoning heuristically, one way to handle the dependence between the two steps is to prove a union bound over exponentially many possible realizations of the blocking; unfortunately, this fails since the desired concentration of partial row sums fails with polynomially small probability. Thus, addressing the original sampling model [59, 274] (with one sample per entry) presents an interesting technical challenge that may also involve its own statistical-computational trade-offs [226].

In the broader context of this thesis, this chapter touches upon both the statistical and computational aspects of performing estimation in matrix-valued permutation-based models. While we did not dwell on the question of adaptation in this chapter, it can be shown (using the techniques of Chatterjee and Mukherjee [55]) that our two-dimensional sorting estimator is in fact optimally adaptive if the underlying matrix belongs a parametric family such as the BTL model. Other parametric structure is also interesting to study: for instance, what happens in the case where the underlying matrix is a permuted version of a block-wise constant matrix? This is a useful model for when pairwise comparisons are performed between clusters of similar objects [277]. In the next chapter, we address questions of this form for the even broader class of tensor-valued permutation-based models.

Chapter 4

Adaptive algorithms for tensor estimation

Consider the problem of estimating a d-dimensional, real-valued tensor $\theta^* \in \mathbb{R}^{n_1 \times \cdots \times n_d}$, whose entries are observed in noise. As in many problems in high-dimensional statistics, this tensor estimation problem requires prohibitively many samples to solve without the imposition of further structure, and consequently, many structural constraints have been placed in particular applications of tensor estimation. For instance, low "rank" structure is common in chemistry applications [4], blockwise constant structure is common in applications to clustering and classification of relational data [352], sparsity is commonly used in data mining applications [171], and variants and combinations of such assumptions have also appeared in other contexts [353]. In this chapter, we continue our study of permutation-based models from Chapter 3 to study a flexible, nonparametric structural assumption that generalizes parametric assumptions in applications of tensor models to discrete choice data.

4.1 Introduction

Suppose we are interested in modeling ordinal data, which arises in applications ranging from information retrieval [91] and assortment optimization [170] to crowdsourcing [60]; in a generic such problem, we have n_1 "items", subsets of which are evaluated using a multiway comparison. In particular, each datum takes the form of a tuple containing d of these items, and a single item that is chosen from the tuple as the "winner" of this comparison. A mathematical model for such data is a stochastic model of choice: For each tuple A and each item $i \in A$, suppose that i wins the comparison with probability $p(i, A)$. The winner of each comparison is then modeled as a random variable; equivalently, the overall statistical model is described by a d-dimensional mean tensor $\theta^* \in \mathbb{R}^{n_1 \times \cdots \times n_1}$, where $\theta^*(i_1, \ldots, i_d) = p(i_1, (i_1, \ldots, i_d))$, and our data consists of noisy observations of entries of this tensor. Imposing sensible constraints on the tensor θ^* in these applications goes back to classical, axiomatic work on the subject due to Luce [204] and Plackett [252]. A natural and flexible assumption is given by *simple scalability* [177, 219, 305]: suppose that each of the n_1 items can be associated with some scalar utility (item i with utility u_i), and that the comparison probability is given by

$$\theta^*(i_1, \ldots, i_d) = f(u_{i_1}, \ldots, u_{i_d}), \tag{4.1}$$

where f is a non-decreasing function of its first argument and a coordinate-wise non-increasing function of the remaining arguments. Operationally, an item should not have a lower chance of being chosen as a winner if—all else remaining equal—its utility were to be increased.

There are many models that satisfy the nonparametric simple scalability assumption; in particular, *parametric* assumptions in which a specific form of the function f is posited. The simplest parameterization is given by $f(u_1, \ldots, u_d) = \frac{u_1}{\sum_{j=1}^{d} u_d}$, which dates back to Luce [204]. A logarithmic transformation of Luce's parameterization leads to the multinomial logit (MNL) model, which has seen tremendous popularity in applications ranging from transportation [24] to marketing [54]. See, e.g., McFadden [218] for a classical but comprehensive introduction to this class of models. However the parametric assumptions of the MNL model have been called into question by a line of work, showing that more flexibility in modeling can lead to improved results in many applications (see, e.g., Farias et al. [97] and references therein). The simple scalability (SS) assumption in the special case $d = 2$ is equivalent to the strong stochastic transitivity, or SST, assumption studied in Chapter 3. In this special case, the MNL model is equivalent to the popular Bradley–Terry–Luce model [40, 204].

Let us state an equivalent formulation of the SS assumption in terms of structure on the tensor θ^*. Recall that for two vectors of equal dimension, we use $x \preceq y$ to denote that $x - y \leq 0$ entrywise. By ordering the items by their utilities, the monotonicity of the function f in the SS assumption (4.1) ensures that there is a permutation π that arranges them from "worst" to "best", such that

$$\theta^*\big(\pi(i_1), \pi^{-1}(i_2) \ldots, \pi^{-1}(i_d)\big) \leq \theta^*\big(\pi(i'_1), \pi^{-1}(i'_2), \ldots, \pi^{-1}(i'_d)\big) \quad \text{for all} \quad (i_1, \ldots, i_d) \preceq (i'_1, \ldots, i'_d). \tag{4.2}$$

Crucially, since the utilities themselves are latent, the permutation π is *unknown*—indeed, it represents the ranking that must be estimated from our data—and so θ^* is a coordinate-wise isotonic tensor with unknown permutations. In the multiway comparison problem, this tensor represents the stochastic model underlying our data, and accurate knowledge of these probabilities is useful in applications such as assortment optimization [170] and determine, for instance, pricing and revenue management decisions. While multiway comparisons form our primary motivation for studying this problem, the flexibility afforded by nonparametric models with latent permutations has also been noticed and exploited in other applications; see Section 3.2.3. Besides these, there are also several other examples of tensor estimation problems in which parametric structure is frequently assumed; for example, in random hypergraph models [120]. Similarly to before, nonparametric structure has the potential to generalize and lend flexibility to these parametric models.

It is worth noting that in many of the aforementioned applications, the underlying objects can be clustered into near identical sets. For example, there is evidence that indifference sets of items exist in crowdsourcing (see [277, Figure 1] for an illuminating example) and peer review [239] applications involving comparison data; clustering is often used in the application of psychometric evaluation methods [140], and many models for communities in hypergraphs posit the existence of such clusters of nodes [120]. For a precise mathematical definition of indifference sets and how they induce further structure in the tensor θ^*, see Section 4.2. Whenever such additional structure exists, it is conceivable that estimation can be performed in a more sample-efficient manner; we will precisely quantify such a phenomenon shortly.

Using these applications as motivation, our goal in this chapter is to study the tensor estimation problem under the nonparametric structural assumptions (4.2) of monotonicity constraints and unknown permutations.

4.1.1 Related work

We focus our discussion on the sub-class of such problems involving monotonic shape-constraints and (vector/matrix/tensor) estimation. When $d = 1$, the assumption (4.2) corresponds to the "uncoupled" or "shuffled" univariate isotonic regression problem [51]. Here, an estimator based on Wasserstein deconvolution is known to attain the minimax rate $\log \log n / \log n$ in (normalized) squared ℓ_2-error for estimation of the underlying (sorted) vector of length n [262]. A very recent paper [11] has also considered this problem, with a focus on isolating the effect of the noise distribution on the deconvolution procedure. A multivariate version of this problem (estimating multiple isotonic functions under a common unknown permutation of coordinates) has been studied under the moniker of "statistical seriation", and has been shown to have applications to archaeology and genome assembly [104, 206]. The case $d = 2$ was the focus of Chapter 3 of this thesis. To the best of our knowledge, analogs of these results have not been explored in the multivariate setting $d \geq 3$, although a significant body of literature has studied parametric models for choice data in this case (see, e.g., Negahban et al. [230] and references therein).

4.1.2 Overview of contributions

We begin by considering the minimax risk of estimating bounded tensors satisfying assumption (4.2), and show in Theorem 4.3.1 that it is dominated by the risk of estimating the underlying *ordered* coordinate-wise isotonic tensor. In other words, the latent permutations do not significantly influence the statistical difficulty of the problem. We also study the fundamental limits of estimating tensors having indifference set structure, and this allows us to assess the ability of an estimator to adapt to such structure via its *adaptivity index* (to be defined precisely in equation (4.3)). We establish two surprising phenomena in this context: First, we show in Proposition 4.3.1 that the fundamental limits of estimating these objects preclude a parametric rate, in sharp contrast to the case without unknown permutations. Second, we prove in Theorem 4.3.2 that the adaptivity index exhibits a statistical-computational gap under the assumption of a widely-believed conjecture in average-case complexity. In particular, we show that the adaptivity index of any polynomial-time computable estimator must grow at least polynomially in n, assuming the hypergraph planted clique conjecture [43]. Our results also have interesting consequences for the isotonic regression problem without unknown permutations (see Corollaries 4.3.1 and 4.3.2).

Having established these fundamental limits, we then turn to our main methodological contribution. We propose and analyze—in Theorem 4.4.1—an estimator based on Mirsky's partitioning algorithm [224] that estimates the underlying tensor (a) at the minimax rate for each $d \geq 3$ whenever this tensor has bounded entries, and (b) with the best possible adaptivity index for polynomial-time procedures for all $d \geq 2$. The first of these findings is particularly surprising because it shows that the case $d \geq 3$ of this problem is distinctly different from the bivariate case studied in Chapter 3, in that the minimax risk is achievable with a computationally efficient algorithm. This in in spite of the fact that there are more permutations to estimate as the dimension gets larger, which, at least in principle, ought to make the problem more difficult both statistically and computationally.

In addition to its favorable risk properties, the Mirsky partition estimator also has several other advantages: it is computable in time sub-quadratic in the size of the input, and its *computational complexity* also adapts to underlying indifference set structure. In particular, when there are a fixed number of indifference sets, the estimator has almost linear computational complexity with high probability. When specialized to $d = 2$, this estimator exhibits significantly better adaptation properties to indifference set structure than known estimators that were designed specifically for this purpose; see the full paper [245] for statements and discussions of these results.

To complement our upper bounds on the Mirsky partition estimator, we also show, somewhat surprisingly, that many other estimators proposed in the literature [55, 274, 277], and natural variants thereof, suffer from an extremely large adaptivity index. In particular, they are unable to attain the polynomial-time optimal adaptivity index (given by the fundamental limit established by Theorem 4.3.2) for any $d \geq 4$. This is in spite of the fact that some of these estimators are minimax optimal for estimation over the class of bounded tensors (see Proposition 4.5.1 and Corollary 4.3.1) for all $d \geq 3$. Thus, we see that simultaneously achieving good worst-case risk properties while remaining computationally efficient and adaptive to structure is a challenging requirement, providing further evidence of the value of the Mirsky partitioning estimator.

4.2 Background and problem formulation

Let $\mathfrak{S}_k$ denote the set of all permutations on the set $[k] := \{1, \ldots, k\}$. We interpret $\mathbb{R}^{n_1 \times \cdots \times n_d}$ as the set of all real-valued, tensors of dimension $n_1 \times \cdots \times n_d$. For a set of positive integers $i_j \in [n_j]$, we use $T(i_1, \ldots, i_d)$ to index entry $i_1, \ldots, i_d$ of a tensor $T \in \mathbb{R}^{n_1 \times \cdots \times n_d}$.

The set of all real-valued, coordinate-wise isotonic functions on the set $[0, 1]^d$ is denoted by

$$\mathcal{F}_d := \left\{ f : [0,1]^d \to \mathbb{R} : f(x_1, x_2, \ldots, x_d) \leq f^*(x_1', x_2', \ldots, x_d') \quad \text{when } x_j \leq x_j' \text{ for } j \in [d] \right\}.$$

Let n_j denote the number of observations along dimension j, with the total number of observations given by $n := \prod_{j=1}^d n_j$. For $n_1, \ldots, n_d \in \mathbb{N}$, let $\mathbb{L}_{d,n_1,\ldots,n_d} := \prod_{j=1}^d [n_j]$ denote the d-dimensional lattice. With this notation, we assume access to a tensor of observations $Y \in \mathbb{R}^{n_1 \times \cdots \times n_d}$, where

$$Y(i_1, \ldots, i_d) = f^* \left(\frac{\pi_1^*(i_1)}{n_1}, \frac{\pi_2^*(i_2)}{n_2}, \ldots, \frac{\pi_d^*(i_d)}{n_d} \right) + \epsilon(i_1, \ldots, i_d) \quad \text{for each } i_j \in [n_j],\ j \in [d].$$

Here, the function $f^* \in \mathcal{F}_d$ is unknown, and for each $j \in [d]$, we also have an unknown permutation $\pi_j^* \in \mathfrak{S}_{n_j}$. The tensor $\epsilon \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ represents noise in the observation process, and we assume that its entries are given by independent standard normal random variables[1]. Denote the noiseless observations on the lattice by

$$\theta^*(i_1, \ldots, i_d) := f^* \left(\frac{\pi_1^*(i_1)}{n_1}, \frac{\pi_2^*(i_2)}{n_2}, \ldots, \frac{\pi_d^*(i_d)}{n_d} \right) \qquad \text{for each } i_j \in [n_j],\ j \in [d];$$

this is precisely the nonparametric structure that was posited in equation (4.2).

It is also convenient to define the set of tensors that can be formed by permuting evaluations of a coordinate-wise monotone function on the lattice by the permutations $(\pi_1, \ldots, \pi_d)$. Denote this set by

$$\mathcal{M}(\mathbb{L}_{d,n_1,\ldots,n_d}; \pi_1, \ldots, \pi_d) := \Bigg\{ \theta \in \mathbb{R}^{n_1 \times \cdots \times n_d} : \ \exists f \in \mathcal{F}_d \text{ such that} \\ \theta(i_1, \ldots, i_d) := f\left(\frac{\pi_1(i_1)}{n_1}, \ldots, \frac{\pi_d(i_d)}{n_d} \right) \ \text{ for each } i_j \in [n_j],\ j \in [d] \Bigg\}.$$

We use the shorthand $\mathcal{M}(\mathbb{L}_{d,n_1,\ldots,n_d})$ to denote this set when the permutations are all the identity. Also define the set

$$\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n_1,\ldots,n_d}) := \bigcup_{\pi_1 \in \mathfrak{S}_{n_1}} \cdots \bigcup_{\pi_d \in \mathfrak{S}_{n_d}} \mathcal{M}(\mathbb{L}_{d,n_1,\ldots,n_d}; \pi_1, \ldots, \pi_d)$$

of tensors that can be formed by permuting evaluations of any coordinate-wise monotone function.

For a set of permutations $\{\pi_j \in \mathfrak{S}_{n_j}\}_{j=1}^d$ and a tensor $T \in \mathbb{R}^{n_1 \times \cdots \times n_d}$, we let $T\{\pi_1, \ldots, \pi_d\}$ denote the tensor T viewed along permutation π_j on dimension j. Specifically, we have

$$T\{\pi_1, \ldots, \pi_d\}(i_1, \ldots, i_d) = T(\pi_1(i_1), \ldots, \pi_d(i_d)) \quad \text{for each} \quad i_j \in [n_j],\ \ j \in [d].$$

With this notation, note the inclusion $\theta^*\{(\pi_1^*)^{-1}, \ldots, (\pi_d^*)^{-1}\} \in \mathcal{M}(\mathbb{L}_{d,n_1,\ldots,n_d})$. However, since we do not know the permutations $\pi_1^*, \ldots, \pi_d^*$ a-priori, we may only assume the inclusion $\theta^* \in \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n_1,\ldots,n_d})$, and our goal is to denoise our observations and produce an estimate of θ^*. Call any such tensor estimate $\widehat{\theta} \in \mathbb{R}^{n_1 \times \cdots \times n_d}$. We study its empirical L_2 risk, given by

$$\mathcal{R}_n(\widehat{\theta}, \theta^*) := \mathbb{E}\left[\ell_n^2(\widehat{\theta}, \theta^*) \right], \quad \text{ where } \quad \ell_n^2(\theta_1, \theta_2) := \frac{1}{n} \sum_{j=1}^{d} \sum_{i_j=1}^{n_j} \left(\theta_1(i_1, \ldots, i_d) - \theta_2(i_1, \ldots, i_d) \right)^2.$$

Note that the expectation is taken over both the noise ϵ and any randomness used to compute the estimate $\widehat{\theta}$. In the case where we have the inclusion $\widehat{\theta} \in \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n_1,\ldots,n_d})$, we also produce a function estimate $\widehat{f} \in \mathcal{F}_d$ and permutation estimates $\widehat{\pi}_j \in \mathfrak{S}_{n_j}$ for $j \in [d]$, with

$$\widehat{\theta}(i_1, \ldots, i_d) := \widehat{f}\left(\frac{\widehat{\pi}_1(i_1)}{n_1}, \frac{\widehat{\pi}_2(i_2)}{n_2}, \ldots, \frac{\widehat{\pi}_d(i_d)}{n_d} \right) \qquad \text{for each } i_j \in [n_j],\ j \in [d].$$

[1] We study the canonical Gaussian setting for convenience, but all of our our results extend straightforwardly to sub-Gaussian noise distributions.

Note that in general, the resulting estimates $\widehat{f}, \widehat{\pi}_1, \ldots, \widehat{\pi}_d$ need not be unique, but this identifiability issue will not concern us since we are only interested in the tensor $\widehat{\theta}$ as an estimate of the tensor θ^*.

As alluded to in the introduction, it is common in multiway comparisons for there to be *indifference sets* of items that all behave identically. These sets are easiest to describe in the space of functions. For each $j \in [d]$ and $s_j \in [n_j]$, let $I^j_1, \ldots, I^j_{s_j}$ denote a disjoint set of s_j intervals such that $[0, 1] = \cup_{\ell=1}^{s_j} I^j_\ell$. Suppose that for each ℓ, the length of the interval I^j_ℓ exceeds $1/n_j$, so that we are assured that the intersection of I^j_ℓ with the set $\frac{1}{n_j}\{1, \ldots, n_j\}$ is non-empty. With a slight abuse of terminology, we also refer to this intersection as an interval, and let the tuple $\mathbf{k}^j = (k^j_1, \ldots, k^j_{s_j})$ denote the cardinalities of these intervals, with $\sum_{\ell=1}^{s_j} k^j_\ell = n_j$. Let $\mathbf{K}_{s_j}$ denote the set of all such tuples, and define $k^j_{\max} := \max_{\ell \in [s_j]} \mathbf{k}^j_\ell$. Collect $\{\mathbf{k}^j\}_{j=1}^d$ in a tuple $\mathbb{k} = (\mathbf{k}^1, \ldots, \mathbf{k}^d)$, and the d values $\{s_j\}_{j=1}^d$ in a tuple $\mathrm{s} = (s_1, \ldots, s_d)$. Let $\mathbb{K}_{\mathrm{s}}$ denote the set of all such tuples $\mathbb{k}$, and note that the possible values of s range over the lattice $\mathbb{L}_{d,n_1,\ldots,n_d}$. Finally, let $k^* := \min_{j \in [d]} k^j_{\max}$.

If, for each $j \in [d]$, dimension j of the domain is partitioned into the intervals $I^j_1, \ldots, I^j_{s_j}$, then the set $[0, 1]^d$ is partitioned into $s := \prod_{j=1}^d s_j$ hyper-rectangles, where each hyper-rectangle takes the form $\prod_{j=1}^d I^j_{\ell_j}$ for some sequence of indices $\ell_j \in [s_j], j \in [d]$. We refer to the intersection of a hyper-rectangle with the lattice $\mathbb{L}_{d,n_1,\ldots,n_j}$ also as a hyper-rectangle, and note that $\mathbb{k}$ fully specifies such a hyper-rectangular partition. Denote by $\mathcal{M}^{\mathbb{k},\mathrm{s}}(\mathbb{L}_{d,n_1,\ldots,n_d})$ the set of all $\theta \in \mathcal{M}(\mathbb{L}_{d,n_1,\ldots,n_d})$ that are piecewise constant on a hyper-rectangular partition specified by $\mathbb{k}$—we have chosen to be explicit about the tuple s in our notation for clarity. Let $\mathcal{M}^{\mathbb{k},\mathsf{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n_1,\ldots,n_d})$ denote the set of all coordinate-wise permuted versions of $\theta \in \mathcal{M}^{\mathbb{k},\mathrm{s}}(\mathbb{L}_{d,n_1,\ldots,n_d})$.

For the rest of this chapter, we operate in the *uniform, or balanced* case $2 \leq n_1 = \cdots = n_d = n^{1/d}$, which is motivated by the comparison setting introduced in Section 4.1. We use the shorthand $\mathbb{L}_{d,n}$ to represent the uniform lattice and $\mathbb{R}_{d,n}$ to represent balanced tensors. We continue to use the notation n_j in some contexts since this simplifies our exposition, and also continue to accommmodate distinct permutations $\pi^*_1, \ldots, \pi^*_d$ and cardinalities of indifference sets $s_1, \ldots, s_d$ along the different dimensions for flexibility.

Let $\widehat{\Theta}$ denote the set of all estimators of θ^*, i.e. the set of all measurable functions (of the observation tensor Y) taking values in $\mathbb{R}_{d,n}$. Denote the minimax risk over the class of tensors in the set $\mathcal{M}^{\mathbb{k},\mathsf{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n})$ by

$$\mathfrak{M}_{d,n}(\mathbb{k}, \mathrm{s}) := \inf_{\widehat{\theta} \in \widehat{\Theta}} \sup_{\theta^* \in \mathcal{M}^{\mathbb{k},\mathsf{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n})} \mathcal{R}_n(\widehat{\theta}, \theta^*).$$

Note that $\mathfrak{M}_{d,n}(\mathbb{k}, \mathrm{s})$ measures the smallest possible risk achievable with a-priori knowledge of the inclusion $\theta^* \in \mathcal{M}^{\mathbb{k},\mathsf{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n})$. On the other hand, we are interested in estimators that *adapt* to hyper-rectangular structure without knowing of its existence in advance. One way to measure the extent of adaptation of an estimator $\widehat{\theta}$ is in terms of its *adaptivity index* to indifference set sizes $\mathbb{k}$, defined as

$$\mathfrak{A}^{\mathbb{k},\mathrm{s}}(\widehat{\theta}) := \frac{\sup_{\theta^* \in \mathcal{M}^{\mathbb{k},\mathsf{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n})} \mathcal{R}_n(\widehat{\theta}, \theta^*)}{\mathfrak{M}_{d,n}(\mathbb{k}, \mathrm{s})}.$$

A large value of this index indicates that the estimator $\widehat{\theta}$ is unable to adapt satisfactorily to the set $\mathcal{M}^{\mathbb{k},\mathsf{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n})$, since a much lower risk is achievable when the inclusion $\theta^* \in \mathcal{M}^{\mathbb{k},\mathsf{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n})$ is known in advance. The *global* adaptivity index of $\widehat{\theta}$ is then given by

$$\mathfrak{A}(\widehat{\theta}) := \max_{\mathsf{s}\in\mathbb{L}_{d,n}} \max_{\mathbb{k}\in\mathbb{K}_{\mathsf{s}}} \mathfrak{A}^{\mathbb{k},\mathsf{s}}(\widehat{\theta}). \tag{4.3}$$

We note that similar definitions of an adaptivity index or factor have appeared in the literature; our definition most closely resembles the index defined by Shah et al. [277], but similar concepts go back at least to Lepski and co-authors [193, 194].

While the above notions of an adaptivity index deal solely with the ratio of risks, one can additionally demand that adaptation occurs with high probability. In particular, for some confidence level $\delta \in (0, 1)$, denote the $(1 - \delta)$-quantile of the loss by

$$\mathcal{R}_n(\widehat{\theta}, \theta^*; \delta) := \inf \left\{ r \in (0, \infty) : \Pr\{\ell_n^2(\widehat{\theta}, \theta^*) > r\} \leq \delta \right\},$$

and define the high-confidence analogs of the adaptivity indices

$$\mathfrak{A}^{\mathbb{k},\mathsf{s}}(\widehat{\theta}; \delta) := \frac{\sup_{\theta^* \in \mathcal{M}^{\mathbb{k},\mathsf{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n})} \mathcal{R}_n(\widehat{\theta}, \theta^*; \delta)}{\mathfrak{M}_{d,n}(\mathbb{k}, \mathsf{s})} \quad \text{and} \quad \mathfrak{A}(\widehat{\theta}; \delta) := \max_{\mathsf{s}\in\mathbb{L}_{d,n}} \max_{\mathbb{k}\in\mathbb{K}_{\mathsf{s}}} \mathfrak{A}^{\mathbb{k},\mathsf{s}}(\widehat{\theta}; \delta). \tag{4.4}$$

In the discussions to follow, we will typically be concerned with cases where $\delta = \delta_n \asymp 1/n$, that is, we demand that the event on which adaptation does not occur has probability that decays polynomially in the sample size n.

Finally, for a tensor $X \in \mathbb{R}_{d,n}$ and closed set $\mathcal{C} \subseteq \mathbb{R}_{d,n}$, it is useful to define the projection of X onto $\mathcal{C}$ by

$$\widehat{\theta}_{\mathsf{LSE}}(\mathcal{C}, X) \in \operatorname*{argmin}_{\theta\in\mathcal{C}} \; \ell_n^2(X, \theta). \tag{4.5}$$

In our exposition to follow, the set $\mathcal{C}$ will either be compact or a finite union of closed convex sets, and so the projection is guaranteed to exist. When $\mathcal{C}$ is closed and convex, the projection is additionally unique.

Chapter-specific notation: Recall the notational convention introduced in Section 1.4. We complement this notation with a few other definitions that are used solely in this chapter and the corresponding technical proof section in Appendices A.5 and A.6. Let $\mathbb{B}_\infty(t)$ and $\mathbb{B}_2(t)$ denote the ℓ_∞ and ℓ_2 closed balls of radius t in $\mathbb{R}_{d,n}$, respectively, and denote by $\mathbb{1}_{d,n} \in \mathbb{R}_{d,n}$ the all-ones tensor. We use the symbols c_d, C_d to denote constants that depend on d alone; their values will typically change from line to line.

Let us now turn to statements and discussions of our main results. We begin by characterizing the fundamental limits of estimation and adaptation, and then turn to developing an estimator that achieves these limits. Finally, we analyze variants of existing estimators from this point of view.

4.3 Limits of estimation and adaptation

In this subsection, our focus is on characterizing the fundamental limits of estimation over various parameter spaces without imposing any computational constraints on our procedures. We begin by characterizing the minimax risk over the class of bounded, coordinate-wise isotonic tensors with unknown permutations.

Theorem 4.3.1. *There is a universal positive constant C such that for each $d \geq 2$, we have*

$$c_d \cdot n^{-1/d} \leq \inf_{\widehat{\theta} \in \widehat{\Theta}} \sup_{\theta^* \in \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)} \mathcal{R}_n(\widehat{\theta}, \theta^*) \leq C \cdot n^{-1/d} \log^2 n, \tag{4.6}$$

where $c_d > 0$ depends on d alone.

The lower bound on the minimax risk in equation (4.6) follows immediately from known results on estimating bounded monotone functions on the lattice without unknown permutations [83, 137]. The upper bound is our main contribution to the theorem, and is achieved by the bounded least squares estimator

$$\widehat{\theta}_{\mathsf{BLSE}} := \widehat{\theta}_{\mathsf{LSE}}\big(\mathcal{M}_{\mathsf{perm}}\big(\mathbb{L}_{d,n}\big) \cap \mathbb{B}_\infty(1), Y\big). \tag{4.7}$$

In fact, the risk of $\widehat{\theta}_{\mathsf{BLSE}}$ can be expressed as a sum of two terms:

$$\mathcal{R}_n(\widehat{\theta}_{\mathsf{BLSE}}, \theta^*) \leq C\big(n^{-1/d} \log^2 n + n^{-(1-1/d)} \log n\big). \tag{4.8}$$

The first term corresponds to the error of estimating the unknown isotonic function, and the second to the price paid for having unknown permutations. Such a characterization was known in the case $d = 2$ [211, 274], and our result shows that a similar decomposition holds for all d. Note that for all $d \geq 2$, the first term of equation (4.8) dominates the bound, and this is what leads to Theorem 4.3.1.

Although the bounded LSE (4.7) achieves the worst case risk (4.6), we may use its analysis as a vehicle to obtaining risk bounds for the vanilla least squares estimator without imposing any boundedness constraints. This results in the following corollary.

Corollary 4.3.1. *There is a universal positive constant C such that for each $d \geq 2$:*
(a) The least squares estimator over the set $\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n})$ has worst case risk bounded as

$$\sup_{\theta^* \in \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)} \mathcal{R}_n\big(\widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}), Y), \theta^*\big) \leq C n^{-1/d} \log^{5/2} n. \tag{4.9a}$$

(b) The isotonic least squares estimator over the set $\mathcal{M}(\mathbb{L}_{d,n})$ has worst case risk bounded as

$$\sup_{\theta^* \in \mathcal{M}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)} \mathcal{R}_n\big(\widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}(\mathbb{L}_{d,n}), Y), \theta^*\big) \leq C n^{-1/d} \log^{5/2} n. \tag{4.9b}$$

Part (a) of Corollary 4.3.1 deals with the LSE computed over the entire set $\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n})$, and appears to be new even when $d = 2$; to the best of our knowledge, prior work [211, 274] has only considered the bounded LSE $\widehat{\theta}_{\mathsf{BLSE}}$ (4.7).

Part (b) of Corollary 4.3.1, on the other hand, provides a risk for the vanilla isotonic least squares estimator when estimating functions in the set $\mathcal{M}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)$. This estimator has a long history in both the statistics and computer science communities [56, 92, 137, 182, 265, 291], and unlike the other estimators considered so far, the isotonic LSE is the solution to a convex optimization problem and can be computed in time polynomial in n. Bounds on the worst case risk of this estimator are also known: results for $d = 1$ are classical (see, e.g., the papers [47, 233, 347]); when $d = 2$, risk bounds were derived by Chatterjee et al. [56]; and the general case $d \geq 2$ was considered by Han et al. [137]. Corollary 4.3.1(b) improves the logarithmic factor in the latter two papers from $\log^4 n$ to $\log^{5/2} n$, and is obtained via a different proof technique involving a truncation argument.

Two other comments are worth making. First, it should be noted that there are other estimators for tensors in the set $\mathcal{M}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)$ besides the isotonic LSE. The block-isotonic estimator of Deng and Zhang [83], first proposed by Fokianos et al. [107], enjoys a risk bound of the order $C_d \cdot n^{-1/d}$ for all $d \geq 2$, where $C_d > 0$ is a d-dependent constant. This eliminates the logarithmic factor entirely, and matches the minimax lower bound up to a d-dependent constant. In addition, the block-isotonic estimator also enjoys significantly better adaptation properties than the isotonic LSE. Two issues, however, require further exploration. The best known algorithm to compute the block-isotonic estimator takes time $\mathcal{O}(n^3)$, while the isotonic LSE can be computed in time $\widetilde{\mathcal{O}}(n^{3/2})$ [182]. In addition, the behavior of the block-isotonic estimator under mis-specification is not yet well understood; the usual oracle inequality that can be shown for the isotonic LSE has not yet—to the best of our knowledge—been shown to hold for the block-isotonic estimator (see the discussion in [83, Section 3.7]).

Second, we note that when the design is random in the setting without unknown permutations [136, Theorem 3.9] improves, at the expense of a d-dependent constant, the logarithmic factors in the risk bounds of prior work [137]. His proof techniques are based on the concentration of empirical processes on upper and lower sets of $[0, 1]^d$, and do not apply to the lattice setting considered here. On the other hand, our proof works on the event on which the LSE is suitably bounded, and is not immediately applicable to the random design setting. Both of these techniques should be viewed as particular ways of establishing the optimality of global empirical risk minimization procedures even when the entropy integral for the corresponding function class diverges; this runs contrary to previous heuristic beliefs about the suboptimality of these procedures (see, e.g., the papers [35], [309, pp. 121–122], [261], and [136] for further discussion).

Let us now turn to establishing the fundamental limits of estimation over the class $\mathcal{M}^{\mathbb{k},\mathrm{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n})$. The following proposition characterizes the minimax risk $\mathfrak{M}_{d,n}(\mathbb{k}, \mathrm{s})$. Recall that we have $s = \prod_{j=1}^d s_j$ and $k^* = \min_{j \in [d]} \max_{\ell \in [s_j]} k^j_\ell$.

Proposition 4.3.1. *There is a pair of universal positive constants* (c, C) *such that for each* $d \geq 1$*,*

$\mathbb{s} \in \mathbb{L}_{d,n}$, *and* $\mathbb{k} \in \mathbb{K}_{\mathbb{s}}$, *the minimax risk* $\mathfrak{M}_{d,n}(\mathbb{k}, \mathbb{s})$ *is sandwiched as*

$$\frac{c}{n} \cdot \Big(s + (n_1 - k^*)\Big) \leq \; \mathfrak{M}_{d,n}(\mathbb{k}, \mathbb{s}) \; \leq \frac{C}{n} \cdot \Big(s + (n_1 - k^*) \log n\Big). \tag{4.10}$$

A few comments are in order. As before, the risk can be decomposed into two terms: the first term represents the *parametric* rate of estimating a tensor with s constant pieces, and the second term is the price paid for unknown permutations. When the underlying set is bounded, such a decomposition does not occur transparently even in the special case $d = 2$ [277]. Also note that when $s = \mathcal{O}(1)$ and $n_1 - k^* = \omega(1)$, the second term of the bound (4.10) dominates and the minimax risk is no longer of the parametric form s/n. This is in sharp contrast to isotonic regression without unknown permutations, where there are estimators that achieve the parametric risk up to poly-logarithmic factors [83]. Thus, the fundamental adaptation behavior that we expect changes significantly in the presence of unknown permutations.

Second, note that when $s_j = n_1$ for all $j \in [d]$, we have $\mathcal{M}^{\mathbb{k},\mathbb{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n}) = \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n})$, in which case the result above shows that consistent estimation is impossible over the set of all isotonic tensors with unknown permutations. This does *not* contradict Theorem 4.3.1, since Proposition 4.3.1 computes the minimax risk over isotonic tensors without imposing boundedness constraints.

Finally, we note that Proposition 4.3.1 yields the following corollary in the setting where we do not have unknown permutations. With a slight abuse of notation, we let

$$\mathcal{M}^{s}(\mathbb{L}_{d,n}) := \bigcup_{\mathbb{s}\,:\, \prod_{j=1}^{d} s_j = s} \; \bigcup_{\mathbb{k} \in \mathbb{K}_{\mathbb{s}}} \; \mathcal{M}^{\mathbb{k},\mathbb{s}}(\mathbb{L}_{d,n})$$

denote the set of all coordinate-wise monotone tensors that are piecewise constant on a d-dimensional partition having s pieces.

Corollary 4.3.2. *There is a pair of universal positive constants* (c, C) *such that for each* $d \geq 1$, *the following statements hold.*
(a) For each $\mathbb{s} \in \mathbb{L}_{d,n}$ *and* $\mathbb{k} \in \mathbb{K}_{\mathbb{s}}$, *we have*

$$c \cdot \frac{s}{n} \leq \inf_{\widehat{\theta} \in \widehat{\Theta}} \; \sup_{\theta^* \in \mathcal{M}^{\mathbb{k},\mathbb{s}}(\mathbb{L}_{d,n})} \; \mathcal{R}_n(\widehat{\theta}, \theta^*) \; \leq C \cdot \frac{s}{n}. \tag{4.11a}$$

(b) For each $s \in [n]$, *we have*

$$c \cdot \frac{s}{n} \leq \inf_{\widehat{\theta} \in \widehat{\Theta}} \; \sup_{\theta^* \in \mathcal{M}^{s}(\mathbb{L}_{d,n})} \; \mathcal{R}_n(\widehat{\theta}, \theta^*) \; \leq C \cdot \frac{s \log n}{n}. \tag{4.11b}$$

Let us interpret this corollary in the context of known results. When $d = 1$ and there are no permutations, Bellec and Tsybakov [23] established minimax lower bounds of order s/n and upper bounds of the order $s \log n/n$ for estimating s-piece monotone functions, and the bound (4.11b) recovers this result. The problem of estimating a univariate isotonic vector with s pieces was also considered by Gao et al. [112], who showed a rate-optimal characterization of the minimax risk that exhibits an iterated logarithmic factor in the sample size whenever $s \geq 3$. When $d \geq 2$, however, the results of Corollary 4.3.2 are new to the best of our knowledge. While it is possible that the bound (4.11b) can be improved to an iterated logarithmic factor in the multidimensional setting, our analysis is not refined enough to capture it. Understanding the exact scaling of the minimax risk in the multidimensional setting is an interesting open problem.

The fundamental limits of estimation over the class $\mathcal{M}_{\mathsf{perm}}^{\mathbb{k},\mathsf{s}}(\mathbb{L}_{d,n})$ will allow us to assess the adaptivity index of particular estimators. Before we do that, however, we establish a baseline for adaptation by proving a lower bound on the adaptivity index of polynomial time estimators.

4.3.1 Lower bounds on polynomial time adaptation

We now turn to our average-case reduction showing that any computationally efficient estimator cannot have a small adaptivity index. Our primitive is the hypergraph planted clique conjecture HPC_D, which is a hypergraph extension of the planted clique conjecture. Let us introduce the testing, or decision, version of this conjecture. Denote the set of D-uniform hypergraphs on N vertices (hypergraphs in which each hyper-edge is incident on D vertices) by $\mathbb{H}_{D,N}$. Define, via their generative models, the random hypergraphs

1. $\mathcal{G}_D(N, p)$: Generate each hyperedge independently with probability p, and

2. $\mathcal{G}_D(N, p; K)$: Choose $K \geq D$ vertices uniformly at random and form a clique, adding all $\binom{K}{D}$ possible hyperedges between them. Add each remaining hyperedge independently with probability p.

Given an instantiation of a random hypergraph $X \in \mathbb{H}_{D,N}$, the testing problem is to distinguish the hypotheses $H_0 : X \sim \mathcal{G}_D(N, p)$ and $H_1 : X \sim \mathcal{G}_D(N, p; K)$. The error of any test $\psi : \mathbb{H}_{D,N} \mapsto \{0, 1\}$ is given by

$$\mathcal{E}(\psi) := \frac{1}{2}\, \mathbb{E}_{H_0}\left[\psi(X)\right] + \frac{1}{2}\, \mathbb{E}_{H_1}\left[(1 - \psi(X))\right]. \tag{4.12}$$

Conjecture 4.3.1 (HPC_D conjecture)**.** *Suppose that $p = 1/2$. There is a universal positive constant c such that for each $K \leq c\sqrt{N}$, any test ψ that is computable in time polynomial in N^D must satisfy*

$$\mathcal{E}(\psi) > 2/3.$$

Note that when $D = 2$, Conjecture 4.3.1 is equivalent to the planted clique conjecture, which is a widely believed conjecture in average-case complexity [18, 98, 152]. The HPC_3 conjecture was used by Zhang and Xia [346] to show statistical-computational gaps for third order tensor completion; their evidence for the validity of this conjecture was based on the threshold at which the natural spectral method for the problem fails. Brennan and Bresler [43] recently showed that the planted clique conjecture with "secret leakage" can be reduced to HPC_D. They also provided evidence (see Section K of their paper) for the validity of the HPC_D conjecture, showing that the decision problem has close connections to the widely believed low-degree conjecture (for a discussion of this conjecture, see the papers [143, 179] and references therein). Luo and Zhang [205] also provide similar evidence in support of the conjecture. Recall our definition of the high-probability adaptivity index (4.4); the HPC_D conjecture implies the following computational lower bound.

Theorem 4.3.2. *Let $\delta_n := (10n)^{-1}$ and suppose that Conjecture 4.3.1 holds. Then there is a constant $c_d > 0$ depending on d alone such that any estimator $\widehat{\theta}$ that is computable in time polynomial in n must satisfy*

$$\mathfrak{A}(\widehat{\theta}; \delta_n) \geq c_d \cdot n^{\frac{1}{2}(1-\frac{1}{d})} \log^{-2} n. \tag{4.13}$$

Assuming Conjecture 4.3.1, Theorem 4.3.2 thus posits that the (high-probability) adaptivity index of any computationally efficient estimator must grow polynomially in n, thereby precluding the existence of efficient estimators with adaptivity index bounded poly-logarithmically in n. Contrast this with the case of isotonic regression without unknown permutations, where the block-isotonic estimator has adaptivity index[2] of the order $\mathcal{O}(\log^d n)$ [83]. This demonstrates yet another salient difference in adaptation behavior with and without unknown permutations.

Finally, while Theorem 4.3.2 is fully novel for all $d \geq 3$, we note that when $d = 2$, it resembles the computational lower bound established by Shah et al. [277] (which assumes the planted clique conjecture). Some connections between the two results are worth highlighting. While in our definition, indifference sets along the different dimensions may have different sizes, Shah et al. [277] restrict to the case where a single tuple of sizes parameterizes indifference sets along all dimensions simultaneously. Thus, when the noise distribution is Bernoulli and $d = 2$, the result of Shah et al. [277] implies ours, since our adaptivity index (4.3) is computed with the maximum ranging over a strictly larger set of configurations. Furthermore, the lower bound of Shah et al. [277] applies to the expected adaptivity index, and we note that a version of Theorem 4.3.2 also holds for the expected adaptivity index when $d = 2$. The question of whether a similar bound on the expected adaptivity index holds for $d \geq 3$ poses an interesting open problem. Having said this, it should be noted that Theorem 4.3.2 applies in the case of Gaussian noise, an extension that is accomplished via the machinery of Gaussian rejection kernels introduced by Brennan et al. [44]. This device shares many similarities with other reduction "gadgets" used in earlier arguments (e.g., [26, 207, 328]).

We have thus established both the fundamental limits of estimation without computational considerations (4.6), and a lower bound on the adaptivity index of polynomial time estimators (4.13). Next, we show that a simple, efficient estimator attains both lower bounds simultaneously for all $d \geq 3$.

4.4 Achieving the fundamental limits in polynomial time

We begin with notation that will be useful in defining our estimator. We say that a tuple $\mathsf{bl} = (S_1, \ldots, S_L)$ is a *one-dimensional ordered partition* of the set $[n_1]$ of size L if the sets $S_1, \ldots, S_L$

[2] Deng and Zhang [83] consider the more general case where the hyper-rectangular partition need not be consistent with the Cartesian product of one-dimensional partitions, but the adaptivity index claimed here can be obtained as a straightforward corollary of their results.

are pairwise disjoint, with $[n_1] = \bigcup_{\ell=1}^{L} S_\ell$. Equivalently, any such tuple may be viewed as the decomposition of a partial order on the set $[n_1]$ into disjoint antichains. Denote the set of all one-dimensional ordered partitions of size L by $\mathfrak{P}_L$, and let $\mathfrak{P} = \bigcup_{L=1}^{n_1} \mathfrak{P}_L$.

Note that any one-dimensional ordered partition induces a map $\sigma_{\mathsf{bl}} : [n_1] \to [n_1]$, where $\sigma_{\mathsf{bl}}(i)$ is the index ℓ of the set $S_\ell \ni i$. Now given d ordered partitions $\mathsf{bl}_1, \ldots, \mathsf{bl}_d \in \mathfrak{P}$, define the set

$$\mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) := \Bigg\{ \theta \in \mathbb{R}_{d,n} : \;\; \exists f \in \mathcal{F}_d \;\text{ such that} \\ \theta(i_1, \ldots, i_d) := f\left(\frac{\sigma_{\mathsf{bl}_1}(i_1)}{n_1}, \ldots, \frac{\sigma_{\mathsf{bl}_d}(i_d)}{n_d} \right) \;\text{ for each } i_j \in [n_j], \; j \in [d] \Bigg\}.$$

In other words, the set[3] $\mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)$ represents all tensors that are piecewise constant on the hyper-rectangles $\prod_{j=1}^{d} \mathsf{bl}_j$, while also being coordinate-wise isotonic on the partial orders specified by $\mathsf{bl}_1, \ldots, \mathsf{bl}_d$. We refer to any such hyper-rectangular partition of the lattice $\mathbb{L}_{d,n}$ that can be written in the form $\prod_{j=1}^{d} \mathsf{bl}_d$ as a *d-dimensional ordered partition*.

Our estimator computes various statistics of the observation tensor Y, and we require some more terminology to define these precisely. For each $j \in [d]$, define the vector $\widehat{\tau}_j \in \mathbb{R}^{n_j}$ of "scores", whose k-th entry is given by

$$\widehat{\tau}_j(k) := \sum_{j=1}^{d} \sum_{i_j=1}^{n_j} Y(i_1, \ldots, i_d) \cdot \mathbf{1}\{i_j = k\}. \tag{4.14a}$$

The score vector $\widehat{\tau}_j$ provides noisy information about the permutation π_j^*. In order to see this clearly, it is helpful to specialize to the noiseless case $Y = \theta^*$, in which case we obtain the population scores

$$\tau_j^*(k) := \sum_{j=1}^{d} \sum_{i_j=1}^{n_j} \theta^*(i_1, \ldots, i_d) \cdot \mathbf{1}\{i_j = k\}. \tag{4.14b}$$

One can verify that the entries of the vector τ_j^* are increasing when viewed along permutation π_j^*, i.e., that $\tau_j^*(\pi_j^*(1)) \leq \cdots \leq \tau_j^*(\pi_j^*(n_j))$.

For each pair $k, \ell \in [n_j]$, also define the pairwise statistics

$$\widehat{\Delta}_j^{\mathsf{sum}}(k, \ell) := \widehat{\tau}_j(\ell) - \widehat{\tau}_j(k) \quad \text{and} \tag{4.15a}$$

$$\widehat{\Delta}_j^{\max}(k, \ell) := \max_{q \in [d] \setminus \{j\}} \max_{i_q \in [n_1]} \left\{ Y(i_1, \ldots, i_d) \cdot \mathbf{1}\{i_j = \ell\} - Y(i_1, \ldots, i_d) \cdot \mathbf{1}\{i_j = k\} \right\}. \tag{4.15b}$$

Given that the scores provide noisy information about the unknown permutation, the statistic $\widehat{\Delta}_j^{\mathsf{sum}}(k, \ell)$ provides noisy information about the event $\{\pi_j^*(k) < \pi_j^*(\ell)\}$, i.e., a large positive value of $\widehat{\Delta}_j^{\mathsf{sum}}(k, \ell)$ provides evidence that $\pi_j^*(k) < \pi_j^*(\ell)$ and a large negative value indicates otherwise.

[3] Note that we have abused notation in defining the sets $\mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)$ and $\mathcal{M}(\mathbb{L}_{d,n}; \pi_1, \ldots, \pi_d)$ similarly to each other. The reader should be able to disambiguate the two from context, depending on whether the arguments are ordered partitions or permutations.

Now clearly, the scores are not the sole carriers of information about the unknown permutations; for instance, the statistic $\widehat{\Delta}_j^{\max}(k, \ell)$ measures the maximum difference between *individual* entries and a large, positive value of this statistic once again indicates that $\pi_j^*(k) < \pi_j^*(\ell)$. The statistics (4.15) thus allow us to distinguish pairs of indices, and our algorithm is based on precisely this observation.

We also need some graph theoretic terminology: Recall that an antichain of a graph is any set of nodes that is incomparable in the partial order, i.e., for any pair of nodes in the antichain, there is no directed path in the graph going from one node to the other. Having set up the necessary notation, we are now ready to describe the algorithm formally.

Algorithm: Mirsky partition estimator

I. (Partition estimation): For each $j \in [d]$, perform the following steps:

a. Create a directed graph G_j' with vertex set $[n_j]$ and add the edge $u \to v$ if either

$$\widehat{\Delta}_j^{\mathsf{sum}}(u, v) > 8\sqrt{\log n} \cdot n^{\frac{1}{2}(1-1/d)} \quad \text{or} \quad \widehat{\Delta}_j^{\max}(u, v) > 8\sqrt{\log n}. \tag{4.16a}$$

If G_j' has cycles, then prune the graph and only keep the edges corresponding to the first condition above, i.e.,

$$u \to v \qquad \text{iff} \qquad \widehat{\Delta}_j^{\mathsf{sum}}(u, v) > 8\sqrt{\log n} \cdot n^{\frac{1}{2}(1-1/d)}. \tag{4.16b}$$

Let G_j denote the pruned graph.

b. Compute a one-dimensional ordered partition $\widehat{\mathsf{bl}}_j$ as the decomposition of the vertices of G_j into disjoint antichains, via Mirsky's algorithm [224].

II. (Piecewise constant isotonic regression): Project the observations on the set of isotonic functions that are consistent with the blocking obtained in step I to obtain

$$\widehat{\theta}_{\mathsf{MP}} = \operatorname*{argmin}_{\theta \in \mathcal{M}(\mathbb{L}_{d,n}; \widehat{\mathsf{bl}}_1, \ldots, \widehat{\mathsf{bl}}_d)} \ell_n^2(Y, \theta).$$

Note that at the end of step Ia, the graph G_j is guaranteed to have no cycles, since the pruning step is based exclusively on the score vector $\widehat{\tau}_j$. Owing to its acyclic structure, the graph G_j can always be decomposed as the union of disjoint antichains.

Let us now describe the intuition behind the estimator as a whole. On each dimension j, we produce a partial order on the set $[n_j]$. We employ the statistics (4.15) in order to determine such a partial order, with two indices placed in the same block if they cannot be distinguished based on these statistics. This partitioning step serves a dual purpose: first, it discourages us from committing to orderings over indices when our observations on these indices look similar, and second, it serves to cluster indices that belong to the same indifference set, since the statistics (4.15) computed on pairs of indices lying in the same indifference set are likely to have small magnitudes. Once we have determined the partial order via Mirsky's algorithm, we project our observations onto isotonic tensors that are piecewise constant on the d-dimensional partition specified by the individual partial orders. The Mirsky partition estimator presented here derives inspiration from some existing estimators in prior work. For instance, the idea of associating a partial order with the indices has appeared before [211, 244], and variants of the pairwise statistics (4.15) have been used in prior work for permutation estimation [104, 211]; see, for instance, Chapter 3. However, to the best of our knowledge, no existing estimator computes a partition of the indices into antichains: a natural idea that significantly simplifies both the algorithm—speeding it up considerably when there are a small number of indifference sets (see the following paragraph for a discussion)—and its analysis.

We now turn to a discussion of the computational complexity of this estimator. Suppose that we compute the score vectors $\widehat{\tau}_j, j = 1, \ldots, d$ first, which takes $\mathcal{O}(dn)$ operations. Now for each $j \in [d]$, step I of the estimator can be computed in time $\mathcal{O}(n_j^2)$, since it takes $\mathcal{O}(n_j^2)$ operations to form the graph G_j, and Mirsky's algorithm [224] for the computation of a "dual Dilworth" decomposition into antichains runs in time $\mathcal{O}(n_j^2)$. Thus, the total computational complexity of step I is given by $\mathcal{O}(d \cdot n_1^2)$. Step II of the estimator involves an isotonic projection onto a partially ordered set. As we establish in Lemma A.5.1 in the appendix, such a projection can be computed by first averaging the entries of Y on the hyper-rectangular blocks formed by the d-dimensional ordered partition $\prod_{j=1}^d \widehat{\mathsf{bl}}_j$, and then performing multivariate isotonic regression on the result. The first operation takes linear time $\mathcal{O}(n)$, and the second operation is a weighted isotonic regression problem that can be computed in time $\widetilde{\mathcal{O}}(B^{3/2})$ if there are B blocks in the d-dimensional ordered partition [182]. Now clearly, $B \leq n$, so that step II of the Mirsky partition estimator has worst-case complexity $\widetilde{\mathcal{O}}(n^{3/2})$. Thus, the overall estimator (from start to finish) has worst-case complexity $\widetilde{\mathcal{O}}(n^{3/2})$. Furthermore, we show in Lemma 4.6.3 that if $\theta^* \in \mathcal{M}_{\mathsf{perm}}^{\mathbb{k},\mathsf{s}}(\mathbb{L}_{d,n})$, then $B \leq s$ with high probability, and on this event, step II only takes time $\mathcal{O}(n) + \widetilde{\mathcal{O}}(s^{3/2})$. When s is small, the overall complexity of the Mirsky partition procedure is therefore dominated by that of computing the scores, and given by $\mathcal{O}(dn)$ with high probability. Thus, the computational complexity also adapts to underlying structure.

Having discussed its algorithmic properties, let us now turn to the risk bounds enjoyed by the Mirsky partition estimator. Recall, once again, the notation $k^* = \min_{j \in [d]} k_{\max}^j$.

Theorem 4.4.1. *There is a universal positive constant C such that for all $d \geq 2$:*
(a) We have the worst-case risk bound

$$\sup_{\theta^* \in \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)} \mathcal{R}_n(\widehat{\theta}_{\mathsf{MP}}, \theta^*) \leq C \left\{ n^{-1/d} \log^{5/2} n + d^2 n^{-\frac{1}{2}(1-1/d)} \log n \right\}. \tag{4.17}$$

(b) Let $\delta_n := (10n)^{-1}$. We have the adaptive bounds

$$\sup_{\theta^* \in \mathcal{M}_{\mathsf{perm}}^{\mathbb{k},\mathsf{s}}(\mathbb{L}_{d,n})} \left\{ \mathcal{R}_n(\widehat{\theta}_{\mathsf{MP}}, \theta^*) \vee \mathcal{R}_n(\widehat{\theta}_{\mathsf{MP}}, \theta^*; \delta_n) \right\} \leq \frac{C}{n} \left\{ s + d^2 (n_1 - k^*) \cdot n^{\frac{1}{2}\left(1 - \frac{1}{d}\right)} \right\} \log n. \tag{4.18a}$$

Consequently, the estimator $\widehat{\theta}_{\mathsf{MP}}$ has adaptivity indices bounded as

$$\mathfrak{A}(\widehat{\theta}_{\mathsf{MP}}) \vee \mathfrak{A}(\widehat{\theta}_{\mathsf{MP}}; \delta_n) \leq Cd^2 \cdot n^{\frac{1}{2}\left(1-\frac{1}{d}\right)} \log n. \tag{4.18b}$$

When taken together, the two parts of Theorem 4.4.1 characterize both the risk and adaptation behaviors of the Mirsky partition estimator $\widehat{\theta}_{\mathsf{MP}}$. Let us discuss some particular consequences of these results, starting with part (a) of the theorem. When $d = 2$, we see that the second term of equation (4.17) dominates the bound, leading to a risk of order $n^{-1/4}$. Comparing with the minimax lower bound (4.6), we see that this is sub-optimal by a factor $n^{1/4}$. There are other estimators that attain strictly better rates [201, 211], but to the best of our knowledge, it is not yet known whether the minimax lower bound (4.6) can be attained by an estimator that is computable in polynomial time. On the other hand, for $d \geq 3$, the first term of equation (4.17) dominates, and we achieve the lower bound on the minimax risk (4.6) up to a poly-logarithmic factor. Thus, the case $d \geq 3$ of this problem is distinctly different from the bivariate case: The minimax risk is achievable with a computationally efficient algorithm in in spite of the fact that there are more permutations to estimate in higher dimensions. This surprising behavior can be reconciled with prevailing intuition by two high-level observations. First, as d grows, the isotonic function becomes much harder to estimate, so we are able to tolerate more sub-optimality in estimating the permutations. Second, in higher dimensional problems, a single permutation perturbs large blocks of the tensor, and this allows us to obtain more information about it than when $d = 2$. Both of these observations are made quantitative and precise in the proof.

As a side note, we believe that the logarithmic factor in the bound (4.17) can be improved; one way to do so is to use other isotonic regression estimators (like the bounded LSE) in step II of our algorithm. But since our notion of adaptation requires an estimator that performs well even when the signal is unbounded, we have used the vanilla isotonic LSE in step II.

Turning our attention now to part (b) of the theorem, notice that we achieve the the lower bound (4.13) on the adaptivity index of polynomial time procedures up to a poly-logarithmic factor. Such a result was not known, to the best of our knowledge, for any $d \geq 3$. Even when $d = 2$, the Count-Randomize-Least-Squares (CRL) estimator of Shah et al. [277] was shown to have adaptivity index bounded by $\widetilde{\mathcal{O}}(n^{1/4})$ over a sub-class of *bounded* bivariate isotonic matrices with unknown permutations that are also piecewise constant on two-dimensional ordered partitions $\mathcal{M}_{\mathsf{perm}}^{\mathbb{k},\mathsf{s}}(\mathbb{L}_{2,n}) \cap \mathbb{B}_\infty(1)$. As we show in the full paper [245], the Mirsky partition estimator is also adaptive in this case, and attains an adaptivity index that significantly improves upon that of the CRL estimator in terms of the logarithmic factor. An even starker difference between the adaptation properties of the CRL and Mirsky partition estimators is evident in higher dimensions. We show in Proposition 4.5.2 to follow that for higher dimensional problems with $d \geq 4$, the CRL estimator has strictly sub-optimal adaptivity index. Thus, in an overall sense, the Mirsky partition estimator is better equipped to adapt to indifference set structure than the CRL estimator.

Let us also briefly comment on the proof of part (b) of the theorem, which has several components that are novel to the best of our knowledge. We begin by employing a decomposition of the error of the estimator in terms of the sum of estimation and approximation errors; while there are also compelling aspects to our bound on the estimation error, let us showcase some interesting components involved in bounding the approximation error. The first key component is a certain structural result (collected as Lemma A.5.1 in Appendix A.5) that allows us to write step II of the algorithm as a composition of two simpler steps. Besides having algorithmic consequences (these were alluded to in our discussion of the running time of the Mirsky partition estimator), Lemma A.5.1 allows us to write the approximation error as a sum of two terms corresponding to the two simpler steps of this composition. In bounding these terms, we make repeated use of a second key component: Mirsky's algorithm groups the indices into clusters of disjoint antichains, and so our bound on the approximation error incurred on any single block of the partition critically leverages the condition (4.16a) used to accomplish this clustering. Our final key component, which is absent from proofs in the literature to the best of our knowledge, is to handle the approximation error on unbounded mean tensors θ^*, which is critical in establishing that the bound (4.18a) holds in expectation—this is, in turn, necessary to provide a bound on the adaptivity index. This component requires us to crucially leverage the pruning condition (4.16b) of the algorithm in conjunction with careful conditioning arguments.

Resurfacing from the specialized discussion above and considering both parts of Theorem 4.4.1 together, we have produced a computationally efficient estimator that is both worst-case optimal when $d \geq 3$ and optimally adaptive among the class of computationally efficient estimators. Let us now turn to other natural estimators for this problem, and assess their worst-case risk, computation, and adaptation properties.

4.5 Other natural estimators do not adapt

Arguably, the most natural estimator for this problem is the global least squares estimator, given by $\widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}), Y)$, which corresponds to the maximum likelihood estimator in our setting with Gaussian errors. The worst-case risk behavior of the LSE over the set $\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{n,d}) \cap \mathbb{B}_\infty(1)$ was already discussed in Corollary 4.3.1(a): It attains the minimax lower bound (4.6) up to a poly-logarithmic factor. However, computing such an estimator is NP-hard in the worst-case even when $d = 2$, since the notoriously difficult max-clique instance can be straightforwardly reduced to the corresponding quadratic assignment optimization problem (see, e.g., the book [251] for reductions of this type).

Another class of procedures consists of two-step estimators that first estimate the unknown permutations defining the model, and then the underlying isotonic function. Estimators of this form abound in prior work [55, 201, 211, 244, 277]. We unify such estimators under Definition 4.5.1 to follow, but first, let us consider a particular instance of such an estimator in which the permutation-estimation step is given by a multidimensional extension of the Borda or Copeland count. A close relative of such an estimator has been analyzed when $d = 2$ [55].

Algorithm: Borda count estimator

I. (Permutation estimation): Recall the score vectors $\widehat{\tau}_1, \ldots, \widehat{\tau}_d$ from (4.14a). Let $\widehat{\pi}_j^{\mathsf{BC}}$ be any permutation along which the entries of $\widehat{\tau}_j$ are non-decreasing; i.e.,

$$\widehat{\tau}_j\big(\widehat{\pi}_j^{\mathsf{BC}}(k)\big) \leq \widehat{\tau}_j\big(\widehat{\pi}_j^{\mathsf{BC}}(\ell)\big) \text{ for all } 1 \leq k \leq \ell \leq n_j.$$

II. (Isotonic regression): Project the observations onto the class of isotonic tensors that are consistent with the permutations obtained in step I to obtain

$$\widehat{\theta}_{\mathsf{BC}} = \mathop{\mathrm{argmin}}_{\theta \in \mathcal{M}(\mathbb{L}_{d,n}; \widehat{\pi}_1^{\mathsf{BC}}, \ldots, \widehat{\pi}_d^{\mathsf{BC}})} \ell_n^2(Y, \theta).$$

The rationale behind the estimator is simple: If we were given the true permutations $(\pi_1^*, \ldots, \pi_d^*)$, then performing isotonic regression on the permuted observations $Y\{(\pi_1^*)^{-1}, \ldots, (\pi_d^*)^{-1}\}$ would be the most natural thing to do. Thus, a natural idea is to *plug-in* permutation estimates $(\widehat{\pi}_1^{\mathsf{BC}}, \ldots, \widehat{\pi}_d^{\mathsf{BC}})$ of the true permutations. The computational complexity of this estimator is dominated by the isotonic regression step, and is thus given by $\widetilde{\mathcal{O}}(n^{3/2})$ [182]. The following proposition provides an upper bound on the worst-case risk of this estimator over bounded tensors in the set $\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n})$.

Proposition 4.5.1. *There is a universal positive constant C such that for each $d \geq 2$, we have*

$$\sup_{\theta^* \in \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)} \mathcal{R}_n(\widehat{\theta}_{\mathsf{BC}}, \theta^*) \leq C \cdot \left(n^{-1/d} \log^{5/2} n + d^2 n^{-\frac{1}{2}(1-1/d)}\right). \tag{4.19}$$

A few comments are in order. First, note that a variant of this estimator has been analyzed previously in the case $d = 2$, but with the bounded isotonic LSE in step II instead of the (unbounded) isotonic LSE [55]. When $d = 2$, the second term of equation (4.19) dominates the bound and Proposition 4.5.1 establishes the rate $n^{-1/4}$, without the logarithmic factor present in the paper [55]. It should be noted that this improvement was already shown for a variant of the Borda count estimator [244].

Second, note that when $d \geq 3$, the first term of equation (4.19) dominates the bound, and comparing this bound with the minimax lower bound (4.6), we see that the Borda count estimator is minimax optimal up to a poly-logarithmic factor for all $d \geq 3$. In this respect, it resembles both the full least squares estimator $\widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}), Y)$ and the Mirsky partition estimator $\widehat{\theta}_{\mathsf{MP}}$.

Unlike the Mirsky partition estimator, however, both the global LSE and the Borda count estimator are unable to adapt optimally to indifference sets. This is a consequence of a more general result that we state after the following definition.

Definition 4.5.1 (Permutation-projection based estimator)**.** *We say that an estimator $\widehat{\theta}$ is permutation-projection based if it can be written as either*

$$\widehat{\theta} = \mathop{\mathrm{argmin}}_{\theta \in \mathcal{M}(\mathbb{L}_{d,n}; \widehat{\pi}_1, \ldots, \widehat{\pi}_d)} \ell_n^2(Y, \theta) \qquad \textit{or} \qquad \widehat{\theta} = \mathop{\mathrm{argmin}}_{\theta \in \mathcal{M}(\mathbb{L}_{d,n}; \widehat{\pi}_1, \ldots, \widehat{\pi}_d) \cap \mathbb{B}_\infty(1)} \ell_n^2(Y, \theta)$$

for a tuple of permutations $(\widehat{\pi}_1, \ldots, \widehat{\pi}_d)$. These permutations could also be chosen in a data-dependent fashion.

Clearly, the bounded LSE (4.7), the global LSE, and the Borda count estimator are permutation-projection based, as is the CRL estimator [277]. The Mirsky partition estimator, on the other hand, is not. The following proposition proves a lower bound on the adaptivity index of any permutation-projection based estimator, both in expectation and high probability.

Proposition 4.5.2. *Let $\delta_n = (10n)^{-1}$. For each $d \geq 4$, there is a pair of constants (c_d, C_d) that depend only on the dimension d such that for each $n \geq C_d$ and any permutation-projection based estimator $\widehat{\theta}$, we have*

$$\{\mathfrak{A}(\widehat{\theta}) \vee \mathfrak{A}(\widehat{\theta}; \delta_n)\} \geq c_d \cdot n^{1-2/d}.$$

For each $d \geq 4$, we have $n^{1-2/d} \gg n^{\frac{1}{2}(1-1/d)}$, and so comparing Proposition 4.5.2 with Theorem 4.3.2, we see that no permutation-projection based estimator can attain the smallest adaptivity index possible for polynomial time algorithms. In fact, even the global LSE, which is not computable in polynomial time to the best of our knowledge, falls short of the polynomial-time benchmark of Theorem 4.3.2.

On the other hand, when $d = 2$, we note once again that the paper [277] leveraged the favorable adaptation properties of the bivariate isotonic LSE [56] to show that their CRL estimator has the optimal adaptivity index for polynomial time algorithms over the class $\mathcal{M}^{\mathbb{k},\mathsf{s}}_{\mathsf{perm}}(\mathbb{L}_{2,n}) \cap \mathbb{B}_\infty(1)$. They also showed that the bounded LSE (4.7) does not adapt optimally in this case. In higher dimensions, however, even the isotonic LSE—which must be employed within any permutation-projection based estimator—has poor adaptation properties [137], and this leads to our lower bound in Proposition 4.5.2.

The case $d = 3$ represents a transition between these two extremes, where the isotonic LSE adapts sub-optimally, but a good enough adaptivity index is still achievable owing to the lower bound of Theorem 4.3.2. Indeed, we show in the full paper [245] that a variant of the CRL estimator also attains the polynomial-time optimal adaptivity index for this case. Consequently, a result as strong as Proposition 4.5.2—valid for all permutation-projection based estimators—cannot hold when $d = 3$.

4.6 Proofs of main results

We now turn to proofs of our main results, beginning with some quick notes to the reader. Throughout, the values of universal constants $c, C, c_1, \ldots$ may change from line to line. Also recall our notation for inequalities holding up to a constant factor: For two sequences of non-negative reals $\{f_n\}_{n\geq 1}$ and $\{g_n\}_{n\geq 1}$, we use $f_n \lesssim g_n$ to indicate that there is a universal constant C such that $f_n \leq C g_n$ for all $n \geq 1$. We also require a bit of additional notation. For a tensor $T \in \mathbb{R}_{d,n}$, we write $\|T\|_2 = \sqrt{\sum_{x \in \mathbb{L}_{d,n}} T_x^2}$, so that $\ell_n^2(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|_2^2$. For a pair of binary vectors (v_1, v_2) of equal dimension, we let $\mathsf{d_H}(v_1, v_2)$ denote the Hamming distance between them. We use the abbreviation "wlog" for "without loss of generality". Finally, since we assume throughout that $n_1 \geq 2$ and $d \geq 2$, we will use the fact that $n \geq 4$ repeatedly and without explicit mention.

We repeatedly employ two elementary facts about ℓ_2 projections, which are stated below for convenience. Recall our notation for the least squares estimator (4.5) as the ℓ_2 projection onto a closed set $\mathcal{C} \subseteq \mathbb{R}_{d,n}$, and assume that the projection exists. For tensors $T_1 \in \mathcal{C}$ and $T_2 \in \mathbb{R}_{d,n}$, we have

$$\|\widehat{\theta}_{\mathsf{LSE}}(\mathcal{C}, T_1 + T_2) - T_1\|_2 \leq 2\|T_2\|_2. \tag{4.20a}$$

The proof of this statement is straightforward; by the triangle inequality, we have

$$\begin{aligned}\|\widehat{\theta}_{\mathsf{LSE}}(\mathcal{C}, T_1 + T_2) - T_1\|_2 &\leq \|\widehat{\theta}_{\mathsf{LSE}}(\mathcal{C}, T_1 + T_2) - (T_1 + T_2)\|_2 + \|(T_1 + T_2) - T_1\|_2 \\ &\leq 2\|(T_1 + T_2) - T_1\|_2 \\ &= 2\|T_2\|_2,\end{aligned}$$

where the second inequality follows since $\widehat{\theta}_{\mathsf{LSE}}(\mathcal{C}, T_1 + T_2)$ is the closest point in $\mathcal{C}$ to $T_1 + T_2$. If moreover, the set $\mathcal{C}$ is convex, then the projection is unique, and non-expansive:

$$\|\widehat{\theta}_{\mathsf{LSE}}(\mathcal{C}, T_1 + T_2) - T_1\|_2 \leq \|T_2\|_2. \tag{4.20b}$$

With this setup in hand, we are now ready to proceed to the proofs of the main results.

4.6.1 Proof of Theorem 4.3.1

It suffices to prove the upper bound, since the minimax lower bound was already shown by [137] for isotonic regression without unknown permutations. We also focus on the case $d \geq 3$ since the result is already available for $d = 2$ [211, 274]. Our proof proceeds in two parts. First, we show that the bounded least squares estimator over isotonic tensors (without unknown permutations) enjoys the claimed risk bound. We then use our proof of this result to prove the upper bound (4.8). The proof of the first result is also useful in establishing part (b) of Corollary 4.3.1.

Bounded LSE over isotonic tensors: For each $x_1, \ldots, x_{d-2} \in [n_1]$, let $\mathcal{M}(A_{x_1,\ldots,x_{d-2}})$ denote the set of bivariate isotonic tensors formed by fixing the first $d-2$ dimensions (variables) of a d-variate tensor to $x_1, \ldots, x_{d-2}$; we refer to this as the two-dimensional *slice* of the lattice $\mathbb{L}_{d,n}$ centered at $x_1, \ldots, x_{d-2}$. For convenience, let $\mathcal{M}(\mathbb{L}_{d,n} \mid r)$ and $\mathcal{M}(A_{x_1,\ldots,x_{d-2}} \mid r)$ denote the intersection of the respective sets with the ℓ_∞ ball of radius r. Letting $A - B$ denote the Minkowski difference between the sets A and B, define

$$\begin{aligned}\mathcal{M}^{\mathsf{diff}}(A_{x_1,\ldots,x_{d-2}} \mid r) &:= \mathcal{M}(A_{x_1,\ldots,x_{d-2}} \mid r) - \mathcal{M}(A_{x_1,\ldots,x_{d-2}} \mid r) \quad \text{and} \\ \mathcal{M}^{\mathsf{full}}(r) &:= \prod_{x_1,\ldots,x_{d-2}} \mathcal{M}(A_{x_1,\ldots,x_{d-2}} \mid r).\end{aligned}$$

In words, the set $\mathcal{M}^{\mathsf{diff}}(A_{x_1,\ldots,x_{d-2}} \mid r)$ denotes the set difference of two bounded, bivariate isotonic slices centered at $x_1, \ldots, x_{d-2}$, and $\mathcal{M}^{\mathsf{full}}(r)$ denotes the Cartesian product of all such two-dimensional slices. Note that by construction, we have ensured, for each $r \geq 0$, the inclusions

$$\mathcal{M}(\mathbb{L}_{d,n} \mid r) \subseteq \mathcal{M}^{\mathsf{full}}(r) \quad \text{and} \tag{4.21a}$$

$$\mathcal{M}(\mathbb{L}_{d,n} \mid r) - \mathcal{M}(\mathbb{L}_{d,n} \mid r) \subseteq \prod_{x_1,\ldots,x_{d-2}} \mathcal{M}^{\mathsf{diff}}(A_{x_1,\ldots,x_{d-2}} \mid r) = \mathcal{M}^{\mathsf{full}}(r) - \mathcal{M}^{\mathsf{full}}(r). \tag{4.21b}$$

With this notation at hand, let us now proceed to bound the risk of the bounded LSE. By definition, this estimator can be written as the projection of Y onto the set $\mathcal{M}(\mathbb{L}_{d,n} \mid 1)$, as so we have

$$\widehat{\theta}_{\mathsf{BLSE}} = \underset{\theta \in \mathcal{M}(\mathbb{L}_{d,n} \mid 1)}{\operatorname{argmin}} \|Y - \theta\|_2^2. \tag{4.22}$$

Letting $\widehat{\Delta} := \widehat{\theta}_{\mathsf{BLSE}} - \theta^*$, the optimality of $\widehat{\theta}_{\mathsf{BLSE}}$ and feasibility of θ^* in the objective (4.22) yield the basic inequality $\|\widehat{\Delta} - \epsilon\|_2^2 \leq \|\epsilon\|_2^2$, rearranging which we obtain

$$\frac{1}{2}\|\widehat{\Delta}\|_2^2 \leq \langle \epsilon, \widehat{\Delta} \rangle \leq \sup_{\substack{\theta \in \mathcal{M}(\mathbb{L}_{d,n} \mid 1) \\ \|\theta - \theta^*\|_2 \leq \|\widehat{\Delta}\|_2}} \langle \epsilon, \theta - \theta^* \rangle \leq \sup_{\substack{\Delta \in \mathcal{M}^{\mathsf{full}}(1) - \mathcal{M}^{\mathsf{full}}(1) \\ \|\Delta\|_2 \leq \|\widehat{\Delta}\|_2}} \langle \epsilon, \Delta \rangle.$$

For convenience, define for each $t \geq 0$ the random variable

$$\xi(t) = \sup_{\substack{\Delta \in \mathcal{M}^{\mathsf{full}}(1) - \mathcal{M}^{\mathsf{full}}(1) \\ \|\Delta\|_2 \leq t}} \langle \epsilon, \Delta \rangle.$$

Also note that the set $\mathcal{M}^{\mathsf{full}}(1) - \mathcal{M}^{\mathsf{full}}(1)$ is star-shaped and non-degenerate (see Definition A.6.1 in Appendix A.6.1). Now applying Lemma A.6.3 from the appendix—which is, in turn, based on [323, Theorem 13.5]—we see that

$$\mathbb{E}[\|\widehat{\Delta}\|_2^2] \leq C(t_n^2 + 1).$$

where t_n is the smallest (strictly) positive solution to the critical inequality

$$\mathbb{E}[\xi(t)] \leq \frac{t^2}{2}. \tag{4.23}$$

Thus, it suffices to produce a bound on $\mathbb{E}[\xi(t)]$, and in order to do so, we use Dudley's entropy integral along with a bound on the ℓ_2 metric entropy of the set $\big(\mathcal{M}^{\mathsf{full}}(1) - \mathcal{M}^{\mathsf{full}}(1)\big) \cap \mathbb{B}_2(t)$. Owing to the inclusions (4.21), we see that in order to cover the set $\mathcal{M}^{\mathsf{full}}(1) - \mathcal{M}^{\mathsf{full}}(1)$ in ℓ_2-norm at radius δ, it suffices to produce, for each $x_1, \ldots, x_{d-2}$, a cover of the set $\mathcal{M}(A_{x_1,\ldots,x_{d-2}} \mid 1)$ in ℓ_2-norm at radius $\delta' = n_1^{-\frac{d-2}{2}} \cdot \frac{\delta}{\sqrt{2}}$. This is because there are n_1^{d-2} unique slices of bivariate isotonic tensors and a δ-covering of the set $\mathcal{M}^{\mathsf{diff}}(A_{x_1,\ldots,x_{d-2}} \mid 1)$ can be accomplished using $\delta/\sqrt{2}$ coverings of the two copies of $\mathcal{M}(A_{x_1,\ldots,x_{d-2}} \mid 1)$ that are involved in the Minkowski difference. Thus, we have

$$N(\delta; \mathcal{M}^{\mathsf{full}}(1) - \mathcal{M}^{\mathsf{full}}(1), \|\cdot\|_2) \leq \prod_{x_1,\ldots,x_{d-2}} N(\delta'; \mathcal{M}(A_{x_1,\ldots,x_{d-2}} \mid 1), \|\cdot\|_2)^2. \tag{4.24}$$

Furthermore, by [114, Theorem 1.1] (see also [274, equation (29)]), we have

$$\log N(\tau; \mathcal{M}(A_{x_1,\ldots,x_{d-2}} \mid 1), \|\cdot\|_2) \lesssim \frac{n_1^2}{\tau^2} \log^2\left(\frac{n_1}{\tau}\right) \quad \text{for each } \tau > 0. \tag{4.25}$$

Putting together the pieces, we obtain

$$\begin{aligned}\log N(\delta; \mathcal{M}^{\mathsf{full}}(1) - \mathcal{M}^{\mathsf{full}}(1), \|\cdot\|_2) &\lesssim n_1^{d-2} \cdot \log N(\delta'; \mathcal{M}(A_{x_1,\ldots,x_{d-2}} \mid 1), \|\cdot\|_2) \\ &\overset{(\mathrm{i})}{\lesssim} n_1^{d-2} \cdot \frac{n}{\delta^2} \log^2\left(\frac{n}{\delta}\right),\end{aligned}$$

where in step (i), we have substituted the value of δ' and noted that $n_1^d = n$. Now the truncated form of Dudley's entropy integral (see, e.g., [323, Theorem 5.22]) yields, for each $t_0 \in [0, t]$, the bound

$$\begin{aligned}\mathbb{E}[\xi(t)] &\lesssim t_0 \cdot \sqrt{n} + \int_{t_0}^{t} \sqrt{\log N(\delta; \mathcal{M}^{\mathsf{full}}(1) - \mathcal{M}^{\mathsf{full}}(1) \cap \mathbb{B}_2(t), \|\cdot\|_2)} d\delta \\ &\leq t_0 \cdot \sqrt{n} + \int_{t_0}^{t} \sqrt{\log N(\delta; \mathcal{M}^{\mathsf{full}}(1) - \mathcal{M}^{\mathsf{full}}(1), \|\cdot\|_2)} d\delta\end{aligned}$$

Choose $t_0 = n^{-11/2}$, apply inequality (4.24), and note that $\log \frac{n}{\delta} \lesssim \log n$ for all $\delta \geq n^{-11/2}$ to obtain

$$\begin{aligned}\mathbb{E}[\xi(t)] &\lesssim n^{-5} + \int_{n^{-11/2}}^{t} \sqrt{n_1^{d-2}} \sqrt{n} \cdot (\log n) \cdot \delta^{-1} d\delta \\ &\lesssim n^{1-1/d} \cdot (\log n) \cdot (\log nt).\end{aligned}$$

Some algebraic manipulation then yields that the solution t_n to the critical inequality (4.23) must satisfy $t_n^2 \leq C n^{1-1/d} \cdot \log^2 n$. Putting together the pieces completes the proof of the claim

$$\mathbb{E}\left[\|\widehat{\Delta}\|_2^2\right] \leq C n^{1-1/d} \cdot \log^2 n. \tag{4.26}$$

Bounded least squares with unknown permutations: The proof for this case proceeds very similarly to before; the only additional effort is to bound the empirical process over a *union* of a large number of difference-of-monotone cones. Similarly to before, define $\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n} \mid r) := \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(r)$ and the sets $\mathcal{M}(\mathbb{L}_{d,n}; \pi_1, \ldots, \pi_d \mid r)$ analogously. Let $\widehat{\theta}_{\mathsf{BLSE}}$ now denote the bounded LSE with permutations (4.7). Proceeding similarly to before with $\widehat{\Delta} = \widehat{\theta}_{\mathsf{BLSE}} - \theta^*$ yields

$$\frac{1}{2}\|\widehat{\Delta}\|_2^2 \leq \sup_{\substack{\theta_1 \in \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}|1) \\ \theta_2 \in \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}|1) \\ \|\theta_1 - \theta_2\|_2 \leq \|\widehat{\Delta}\|_2}} \langle \epsilon, \theta_1 - \theta_2 \rangle = \max_{\pi_1,\ldots,\pi_d \in \mathfrak{S}_{n_1}} \max_{\pi_1',\ldots,\pi_d' \in \mathfrak{S}_{n_1}} \sup_{\substack{\theta_1 \in \mathcal{M}(\mathbb{L}_{d,n}; \pi_1,\ldots,\pi_d|1) \\ \theta_2 \in \mathcal{M}(\mathbb{L}_{d,n}; \pi_1',\ldots,\pi_d'|1) \\ \|\theta_1 - \theta_2\|_2 \leq \|\widehat{\Delta}\|_2}} \langle \epsilon, \theta_1 - \theta_2 \rangle \tag{4.27}$$

For convenience, denote the supremum of the empirical process localized at radius $t > 0$ by

$$\xi(t) := \sup_{\substack{\Delta \in \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}|1) - \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}|1) \\ \|\Delta\|_2 \le t}} \langle \epsilon, \Delta \rangle.$$

Since the set $\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n} \mid 1) - \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n} \mid 1)$ is star-shaped and non-degenerate (see Definition A.6.1), applying Lemma A.6.3 as before yields the risk bound $\mathbb{E}[\|\widehat{\Delta}\|_2^2] \le C(t_n^2 + 1)$, where t_n is the smallest positive solution to the critical inequality (4.23). Using the form of the empirical process in equation (4.27), notice that $\xi(t)$ is the supremum of a Gaussian process over the union of $K = (n_1!)^{2d}$ sets, each of which contains the origin and is contained in an ℓ_2 ball of radius t. We also have $\log K \le 2dn_1 \log n_1 = 2n_1 \log n$. Applying Lemma A.6.1 from the appendix, we thus obtain

$$\mathbb{E}[\xi(t)] \le \max_{\pi_1, \ldots, \pi_d \in \mathfrak{S}_{n_1}} \max_{\pi'_1, \ldots, \pi'_d \in \mathfrak{S}_{n_1}} \mathbb{E} \sup_{\substack{\theta_1 \in \mathcal{M}(\mathbb{L}_{d,n}; \pi_1, \ldots, \pi_d | 1) \\ \theta_2 \in \mathcal{M}(\mathbb{L}_{d,n}; \pi'_1, \ldots, \pi'_d | 1) \\ \|\theta_1 - \theta_2\|_2 \le t}} \langle \epsilon, \theta_1 - \theta_2 \rangle + Ct\sqrt{n_1 \log n}, \tag{4.28}$$

and so it suffices to bound the expectation of the supremum for a fixed pair of tuples $(\pi_1, \ldots, \pi_d)$ and $(\pi'_1, \ldots, \pi'_d)$. For convenience, let

$$\mathcal{D}(\pi_1, \ldots, \pi_d; \pi'_1, \ldots, \pi'_d) := \mathcal{M}(\mathbb{L}_{d,n}; \pi_1, \ldots, \pi_d \mid 1) - \mathcal{M}(\mathbb{L}_{d,n}; \pi'_1, \ldots, \pi'_d \mid 1),$$

and note the sequence of covering number bounds

$$\begin{aligned} N(\delta; \mathcal{D}(\pi_1, \ldots, \pi_d; \pi'_1, \ldots, \pi'_d) \cap \mathbb{B}_2(t), \|\cdot\|_2) &\le N(\delta; \mathcal{D}(\pi_1, \ldots, \pi_d; \pi'_1, \ldots, \pi'_d), \|\cdot\|_2) \\ &\overset{\text{(ii)}}{\le} \left[N(\delta/\sqrt{2}; \mathcal{M}(\mathbb{L}_{d,n} \mid 1), \|\cdot\|_2) \right]^2, \end{aligned}$$

where step (ii) follows since it suffices to cover $\mathcal{M}(\mathbb{L}_{d,n}; \pi_1, \ldots, \pi_d \mid 1)$ and $\mathcal{M}(\mathbb{L}_{d,n}; \pi'_1, \ldots, \pi'_d \mid 1)$ at radius $\delta/\sqrt{2}$, and each of these has covering number equal to that of $\mathcal{M}(\mathbb{L}_{d,n} \mid 1)$. Now proceeding exactly as in the previous calculation and performing the entropy integral, we have

$$\mathbb{E} \sup_{\substack{\theta_1 \in \mathcal{M}(\mathbb{L}_{d,n}; \pi_1, \ldots, \pi_d | 1) \\ \theta_2 \in \mathcal{M}(\mathbb{L}_{d,n}; \pi'_1, \ldots, \pi'_d | 1) \\ \|\theta_1 - \theta_2\|_2 \le t}} \langle \epsilon, \theta_1 - \theta_2 \rangle \lesssim n^{1-1/d} \cdot (\log n) \cdot (\log nt). \tag{4.29}$$

Putting together the pieces (4.28) and (4.29) along with the critical inequality (4.23) and some algebra completes the proof. □

4.6.2 Proof of Corollary 4.3.1

This proof utilizes Theorem 4.3.1 in conjunction with a truncation argument. We provide a full proof of part (a) of the corollary; the proof of part (b) is very similar and we sketch the differences. Recall throughout that by assumption, we have $\theta^* \in \mathbb{B}_\infty(1)$.

Recalling our notation (4.5) for least squares estimators, note that the global least squares estimator $\widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}), Y)$ belongs to the set

$$\left\{ \widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}(\mathbb{L}_{d,n}; \pi_1, \ldots, \pi_d), Y) \mid \pi_1, \ldots, \pi_d \in \mathfrak{S}_{n_1} \right\}.$$

Applying Lemma A.5.2 from the appendix, we see that the for each tuple of permutations $(\pi_1, \ldots, \pi_d)$, the projection onto the set $\mathcal{M}(\mathbb{L}_{d,n}; \pi_1, \ldots, \pi_d)$ is ℓ_∞-contractive, so that

$$\|\widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}(\mathbb{L}_{d,n}; \pi_1, \ldots, \pi_d), Y)\|_\infty \leq \|Y\|_\infty.$$

Consequently, we have $\|\widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}), Y)\|_\infty \leq \|Y\|_\infty \leq 1 + \|\epsilon\|_\infty$. By a union bound,

$$\Pr\{\|\epsilon\|_\infty \geq 4\sqrt{\log n}\} \leq n^{-7}.$$

Let $\psi_n := 4\sqrt{\log n} + 1$ for convenience. On the event $\mathcal{E} := \{\|\epsilon\|_\infty \leq 4\sqrt{\log n}\}$, we thus have $\|\widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}), Y)\|_\infty \leq \psi_n$. Therefore, on this event, we have an equivalence between the vanilla LSE and the bounded LSE:

$$\widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}), Y) = \widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(\psi_n), Y). \tag{4.30}$$

Now replicating the proof of Theorem 4.3.1 for the bounded LSE with ℓ_∞-radius[4] $r \in (0, n]$ yields the risk bound

$$\mathbb{E}\left[\|\widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(r), Y) - \theta^*\|_2^2\right] \leq c(rn^{1-1/d}\log^2 n + n^{1/d}\log n). \tag{4.31}$$

Finally, since the least squares estimator is a projection onto a union of convex sets, inequality (4.20a) yields the bound

$$\|\widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}), Y) - \theta^*\|_2^2 \leq 4\|\epsilon\|_2^2. \tag{4.32}$$

Using the bounds (4.30), (4.31) with $r = \psi_n$, and (4.32) in conjunction with Lemma A.6.6 from the appendix yields

$$\begin{aligned}\mathcal{R}_n(\widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}), Y), \theta^*) &\leq c \cdot \left\{\psi_n \cdot n^{1-1/d}\log^2 n + n^{1/d}\log n + n^{-7/2} \cdot \sqrt{\mathbb{E}[\|\epsilon\|_2^4]}\right\} \\ &\overset{\text{(i)}}{\leq} c \cdot \left\{\psi_n \cdot n^{1-1/d}\log^2 n + n^{1/d}\log n\right\},\end{aligned}$$

where step (i) follows since

$$\mathbb{E}[\|\epsilon\|_2^4] = n^2 + 2n \leq (n+1)^2. \tag{4.33}$$

[4] In more detail, note that by a rescaling argument, it suffices to replace τ in equation (4.25) with τ/r. Since $r \leq n$, note that $\log(rn) \lesssim \log n$.

In order to prove part (b) of the corollary, all steps of the previous argument can be reproduced verbatim with the set $\mathcal{M}(\mathbb{L}_{d,n})$ replacing $\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n})$. The risk bound for the estimator $\widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(r), Y)$ can be obtained from the first part of the proof of Theorem 4.3.1 (see equation (4.26)) and takes the form

$$\mathbb{E}\left[\|\widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(r), Y) - \theta^*\|_2^2\right] \leq crn^{1-1/d}\log^2 n \quad \text{for each } 0 < r \leq n. \tag{4.34}$$

Replacing equation (4.31) with (4.34), setting $r = \psi_n$, and putting together the pieces as before proves the claimed result. □

4.6.3 Proof of Proposition 4.3.1

We prove the upper and lower bounds separately. Recall our notation $\mathbf{K}_q$ for the set of all tuples of positive integers $\mathbf{k} = (k_1, \ldots, k_q)$ with $\sum_{\ell=1}^q k_\ell = n_1$. In this proof, we make use of notation that was defined in Section 4.4. Recall from this section our definition of a one-dimensional ordered partition, the set $\mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)$, and that $\mathfrak{P}_L$ denotes the set of all one-dimensional ordered partitions of size L. Also, let $\mathfrak{P}_k^{\max}$ denote all one-dimensional partitions of $[n_1]$ in which the largest block has size at least k.

Proof of upper bound

For each tuple $\mathbf{k} \in \mathbf{K}_q$, let $\beta(\mathbf{k}) \subseteq \mathfrak{P}_q$ denote the set of all one-dimensional ordered partitions that are consistent with the set sizes $\mathbf{k}$. Note the equivalence

$$\mathcal{M}_{\mathsf{perm}}^{\mathbb{k},\mathbf{s}}(\mathbb{L}_{d,n}) = \bigcup_{\mathsf{bl}_1 \in \beta(\mathbf{k}^1)} \cdots \bigcup_{\mathsf{bl}_d \in \beta(\mathbf{k}^d)} \mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d),$$

and also that $|\beta(\mathbf{k}^j)| \leq |\mathfrak{P}_{k^j_{\max}}^{\max}| \leq e^{3(n_1 - k^j_{\max})\log n_1}$; here, the final inequality follows from Lemma A.6.8(b) in the appendix. Recall that we have $Y = \theta^* + \epsilon$ for some tensor $\theta^* \in \mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1^*, \ldots, \mathsf{bl}_d^*)$, where $\mathsf{bl}_j^* \in \beta(\mathbf{k}^j)$ for each $j \in [d]$. The estimator that we analyze for the upper bound is the least squares estimator $\widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}_{\mathsf{perm}}^{\mathbb{k},\mathbf{s}}(\mathbb{L}_{d,n}), Y)$, which we denote for convenience by $\widehat{\theta}$ for this proof. Since we are analyzing a least squares estimator, our strategy for this proof will be to set up the appropriate empirical process and apply the variational inequality in Lemma A.3.3 in order to bound the error.

Specifically, for each $t \geq 0$, define the random variable

$$\xi(t) := \sup_{\substack{\theta \in \mathcal{M}_{\mathsf{perm}}^{\mathbb{k},\mathbf{s}}(\mathbb{L}_{d,n}) \\ \|\theta - \theta^*\|_2^2 \leq t}} \langle \epsilon, \theta - \theta^* \rangle = \max_{\substack{\mathsf{bl}_1, \ldots, \mathsf{bl}_d \\ \mathsf{bl}_j \in \beta(\mathbf{k}^j)}} \sup_{\substack{\theta \in \mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) \\ \|\theta - \theta^*\|_2 \leq t}} \langle \epsilon, \theta - \theta^* \rangle,$$

which is the pointwise maximum of $K = \prod_{j=1}^d |\beta(\mathbf{k}^j)|$ random variables. Note that we have

$$\log K = \sum_{j=1}^d \log|\beta(\mathbf{k}^j)| \leq \sum_{j=1}^d 3(n_1 - k^j_{\max})\log n_1 \leq 3(n_1 - k^*)\log n.$$

Applying Lemma A.6.1(a) from the appendix, we have that for each $u \geq 0$,

$$\Pr\left\{ \xi(t) \geq \max_{\substack{\mathsf{bl}_1, \ldots, \mathsf{bl}_d \\ \mathsf{bl}_j \in \beta(\mathbf{k}^j)}} \mathbb{E} \sup_{\substack{\theta \in \mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) \\ \|\theta - \theta^*\|_2 \leq t}} \langle \epsilon, \theta - \theta^* \rangle + Ct(\sqrt{(n_1 - k^*) \log n} + \sqrt{u}) \right\} \leq e^{-u}. \tag{4.35}$$

for some universal constant $C > 0$. Lemma 4.6.1, which is stated and proved at the end of this subsection, controls the expected supremum of the empirical process for a fixed choice of the partitions $\mathsf{bl}_1, \ldots, \mathsf{bl}_d$. Combining Lemma 4.6.1 with the high probability bound (4.35), we obtain, for each $u \geq 0$, the bound

$$\Pr\left\{ \xi(t) \geq Ct\left(\sqrt{s} + \sqrt{(n_1 - k^*) \log n} + \sqrt{u}\right) \right\} \leq e^{-u}. \tag{4.36}$$

Now define the function $f_{\theta^*}(t) := \xi(t) - \frac{t^2}{2}$; our goal—driven by Lemma A.3.3—is to compute a value t_* such that with high probability, $f_{\theta^*}(t) < 0$ for all $t \geq t_*$.

For a sufficiently large constant $C > 0$, define the scalar

$$t_u := C(\sqrt{s} + \sqrt{(n_1 - k^*) \log n} + \sqrt{u}) \quad \text{for each } u \geq 0.$$

We claim that on an event $\mathcal{E}$ occurring with probability at least $1 - Cn^{-10}$, the choice $u^* = C \log n$ ensures that

$$f_{\theta^*}(t) < 0 \qquad \text{simultaneously for all } t \geq t_{u^*}. \tag{4.37}$$

Taking this claim as given for the moment, the proof of the upper bound of the proposition follows straightforwardly: Applying Lemma A.3.3 and substituting the value $t^* = t_{u^*}$ yields the bound

$$\|\widehat{\theta} - \theta^*\|_2^2 \leq C\left(s + (n_1 - k^*) \log n\right)$$

with probability greater than $1 - Cn^{-10}$. In order to produce a bound that holds in expectation, note that since $\widehat{\theta}$ is obtained via a projection onto a union of convex sets, inequality (4.20a) yields the pointwise bound $\|\widehat{\theta} - \theta^*\|_2^2 \leq 4\|\epsilon\|_2^2$. Applying Lemma A.6.6 and combining the pieces yields

$$\begin{aligned} \mathbb{E}[\|\widehat{\theta} - \theta^*\|_2^2] &\leq C\left(s + (n_1 - k^*) \log n\right) + C'\sqrt{\mathbb{E}[\|\epsilon\|_2^4]} \cdot \sqrt{n^{-10}} \\ &\leq C\left(s + (n_1 - k^*) \log n\right), \end{aligned}$$

where the final inequality is a consequence of the bound (4.33).

It remains to establish claim (4.37). First, inequality (4.36) ensures that $\xi(t) < t^2/8$ for each *fixed* $t \geq t_u$ with probability at least $1 - e^{-u}$, thereby guaranteeing that $f_{\theta^*}(t) < 0$ for each fixed $t \geq t_u$. Moreover, the Cauchy–Schwarz inequality yields the pointwise bound $\xi(t) \leq t\|\epsilon\|_2$, so that applying the chi-square tail bound [185, Lemma 1] yields

$$\Pr\left\{ \xi(t) \leq t(\sqrt{n} + \sqrt{2u'}) \right\} \geq 1 - e^{-u'} \quad \text{for each } u' \geq 0.$$

Set $u' = u^*$, and note that on this event, we have $f_{\theta^*}(t) < 0$ *simultaneously* for all $t \geq t^{\#}_{u^*} := C(\sqrt{n} + \sqrt{\log n})$. It remains to handle the values of t between t_{u^*} and $t^{\#}_{u^*}$. We suppose that $t^{\#}_{u^*} \geq t_{u^*}$ without loss of generality—there is nothing to prove otherwise—and employ a discretization argument. Let $T = \{t^1, \ldots, t^L\}$ be a discretization of the interval $[t_{u^*}, t^{\#}_{u^*}]$ such that $t_{u^*} = t^1 < \cdots < t^L = t^{\#}_{u^*}$ and $2t^i \geq t^{i+1}$. Note that T can be chosen so that

$$L = |T| \leq \log_2 \frac{t^{\#}_{u^*}}{t_{u^*}} + 1 \leq c \log n.$$

Using the high probability bound $\xi(t) < t^2/8$ for each individual $t \geq t_{u^*}$ and a union bound over T, we obtain that with probability at least $1 - c \log n \cdot e^{-u^*}$, we have

$$\max_{t \in T} \left\{ \xi(t) - t^2/8 \right\} < 0.$$

On this event, we use the fact that $\xi(t)$ is (pointwise) non-decreasing and that $t^i \geq t^{i+1}/2$ to conclude that for each $i \in [L-1]$ and $t \in [t^i, t^{i+1}]$, we have

$$f_{\theta^*}(t) = \xi(t) - t^2/2 \leq \xi(t^{i+1}) - (t^i)^2/2 \leq \xi(t^{i+1}) - (t^{i+1})^2/8 \leq \max_{t \in T} \left\{ \xi(t) - t^2/8 \right\} < 0.$$

Putting together the pieces, we have shown that $f_{\theta^*}(t) < 0$ simultaneously for all $t \geq t_{u^*}$ with probability at least $1 - e^{-u^*} - c \log n \cdot e^{-u^*} \geq 1 - Cn^{-10}$. The final inequality is ensured by adjusting the constants appropriately. □

Lemma 4.6.1. *Let ϵ be the standard Gaussian tensor in $\mathbb{R}_{d,n}$. Suppose that* $\mathrm{s} = (s_1, \ldots, s_d)$ *satisfies* $\prod_{j=1}^d s_j = s$*, and that* $\mathbf{k}^j \in \mathbf{K}_{s_j}$ *for each* $j \in [d]$*. Then for any tensor* $\theta^* \in \mathbb{R}_{d,n}$ *and any sequence of ordered partitions* $\mathsf{bl}_1, \ldots, \mathsf{bl}_d$ *with* $\mathsf{bl}_j \in \beta(\mathbf{k}^j)$ *for all* $j \in [d]$*, we have*

$$\mathbb{E} \sup_{\substack{\theta \in \mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) \\ \|\theta - \theta^*\|_2 \leq t}} \langle \epsilon, \theta - \theta^* \rangle \leq t\sqrt{s} \quad \textit{for each } t \geq 0.$$

Proof. First, let $\overline{\theta} = \widehat{\theta}_{\mathsf{LSE}}(\mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d), \theta^*)$ denote the projection of θ^* onto the set $\mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)$. Let $\mathbb{B}_2(\theta, t)$ denote the ℓ_2 ball of radius t centered at θ. Since the ℓ_2 projection onto a convex set is non-expansive (4.20b), each $\theta \in \mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)$ satisfies $\|\theta - \overline{\theta}\|_2 \leq \|\theta - \theta^*\|_2$, and so we have the inclusion $\mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) \cap \mathbb{B}_2(\theta^*, t) \subseteq \mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) \cap \mathbb{B}_2(\overline{\theta}, t)$ for each $t \geq 0$. Consequently, we obtain

$$\begin{aligned} \sup_{\substack{\theta \in \mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) \\ \|\theta - \theta^*\|_2 \leq t}} \langle \epsilon, \theta - \theta^* \rangle &= \langle \epsilon, \overline{\theta} - \theta^* \rangle + \sup_{\substack{\theta \in \mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) \\ \|\theta - \theta^*\|_2 \leq t}} \langle \epsilon, \theta - \overline{\theta} \rangle \\ &\leq \langle \epsilon, \overline{\theta} - \theta^* \rangle + \sup_{\substack{\theta \in \mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) \\ \|\theta - \overline{\theta}\|_2 \leq t}} \langle \epsilon, \theta - \overline{\theta} \rangle. \end{aligned}$$

Since $\overline{\theta}$ is non-random, the term $\langle \epsilon, \overline{\theta} - \theta^* \rangle$ has expectation zero. Thus, taking expectations and applying Lemma A.6.2 from the appendix yields

$$\mathbb{E} \sup_{\substack{\theta \in \mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) \\ \|\theta - \theta^*\|_2 \leq t}} \langle \epsilon, \theta - \theta^* \rangle \leq \mathbb{E} \sup_{\substack{\theta \in \mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) \\ \|\theta - \overline{\theta}\|_2 \leq t}} \langle \epsilon, \theta - \overline{\theta} \rangle \leq t\sqrt{s},$$

thereby completing the proof. □

Proof of lower bound

Our proof proceeds in two parts: We separately establish the inequalities

$$\mathfrak{M}_{d,n}(\Bbbk, \mathrm{s}) \geq c_1 \cdot \frac{s}{n} \tag{4.38a}$$
$$\mathfrak{M}_{d,n}(\Bbbk, \mathrm{s}) \geq c_2 \cdot \frac{n_1 - k^*}{n} \tag{4.38b}$$

for a pair of universal positive constants (c_1, c_2) and each $\mathrm{s} \in \mathbb{L}_{d,n}$ and $\Bbbk \in \mathbb{K}_{\mathrm{s}}$. Combining the bounds (4.38) yields the claimed lower bound on the minimax risk.

Proof of claim (4.38a)**:** We show this lower bound over just the set $\mathcal{M}^{\Bbbk,\mathrm{s}}(\mathbb{L}_{d,n})$, without the unknown permutations. In order to simplify notation, we let $\phi_{\Bbbk} : \mathbb{R}_{d,n} \to \mathbb{R}_{d,s_1,\ldots,s_d}$ be a map that that collapses each hyper-rectangular block, defined by the tuple $\Bbbk$, of the input into a scalar that is equal to the average of the entries within that block. By construction, for an input tensor $\theta \in \mathcal{M}^{\Bbbk,\mathrm{s}}(\mathbb{L}_{d,n})$, we have the inclusion $\phi_{\Bbbk}(\theta) \in \mathcal{M}(\mathbb{L}_{d,s_1,\ldots,s_d})$. Let $\phi_{\Bbbk}^{-1} : \mathbb{R}_{d,s_1,\ldots,s_d} \to \mathbb{R}_{d,n}$ denote the inverse ("lifting") map obtained by populating each block of the output with identical entries. Furthermore, for each $x \in \mathbb{L}_{d,s_1,\ldots,s_d}$, let $b(x)$ denote the cardinality of block x specified by the tuple $\Bbbk$.

Case $s < 32$: The proof for this case follows from considering the case $s = 1$. In particular, let $\Bbbk_1$ denote the tuple $((n_1), \ldots, (n_d))$ corresponding to a single indifference set along all dimensions, with $\mathrm{s}^{(1)} := (1, \ldots, 1)$ denoting the corresponding tuple of indifference set cardinalities along the d dimensions. Note that $\mathcal{M}^{\Bbbk_1,\mathrm{s}^{(1)}}(\mathbb{L}_{d,n})$ consists of all constant tensors on the lattice. Clearly, we have the inclusion $\mathcal{M}^{\Bbbk_1,\mathrm{s}^{(1)}}(\mathbb{L}_{d,n}) \subseteq \mathcal{M}^{\Bbbk,s}(\mathbb{L}_{d,n})$, and the estimation problem over the class of tensors $\mathcal{M}^{\Bbbk_1,\mathrm{s}^{(1)}}(\mathbb{L}_{d,n})$ is equivalent to estimating a single scalar parameter from n i.i.d. observations with standard Gaussian noise. The minimax lower bound of order $1/n$ is classical, and adjusting the constant factor completes the proof for this case.

Case $s \geq 32$: In this case, we construct a packing of the set $\mathcal{M}(\mathbb{L}_{d,s_1,\ldots,s_d})$ and lift this packing into the space of interest. First, let $\alpha_0 \in \mathcal{M}(\mathbb{L}_{d,s_1,\ldots,s_d})$ denote a base tensor having entries

$$\alpha_0(i_1, \ldots, i_d) = \sum_{j=1}^{d} (i_j - 1) \qquad \text{for each } i_j \in [s_j],\ j \in d.$$

The Gilbert–Varshamov bound [121, 316] guarantees the existence a set of binary tensors on the lattice $\Omega \subseteq \{0,1\}^{\mathbb{L}_{d,s_1,\ldots,s_d}}$ such that the Hamming distance between each pair of distinct vectors $\omega, \omega' \in \Omega$ is lower bounded as $\mathsf{d}_{\mathsf{H}}(\omega, \omega') \geq s/4$ and

$$|\Omega| \geq \frac{2^s}{\sum_{i=0}^{s/4} \binom{s}{i}} \geq e^{s/8}.$$

In deriving the final inequality, we have used Hoeffding's inequality on the lower tail of the distribution $\mathsf{Bin}(s, 1/2)$ to deduce that $\sum_{i=0}^{(s-\alpha)/2} \binom{s}{i} \leq 2^s \exp(-\alpha^2/2s)$ for each $\alpha \geq 0$; see, also, [215, Lemma 4.7].

We now use the set Ω to construct a packing over $\mathcal{M}(\mathbb{L}_{d,s_1,\ldots,s_d})$: For a scalar $\delta \in (0,1]$ to be chosen shortly, define

$$\alpha^{\omega}(x) := \alpha_0(x) + \omega(x) \cdot \frac{\delta}{\sqrt{b(x)}} \qquad \text{for each} \quad x \in \mathbb{L}_{d,s_1,\ldots,s_d}.$$

By construction, the inclusion $\alpha^{\omega} \in \mathcal{M}(\mathbb{L}_{d,s_1,\ldots,s_d})$ holds for each $\omega \in \Omega$. Finally, define the tensors $\theta^{\omega} := \phi_{\Bbbk}^{-1}(\alpha^{\omega})$ for each $\omega \in \Omega$. Note that $\theta^{\omega} \in \mathcal{M}^{\Bbbk,\mathbf{s}}(\mathbb{L}_{d,n})$ for each $\omega \in \Omega$, and also that

$$\|\theta^{\omega} - \theta^{\omega'}\|_2^2 = \delta^2 \cdot \mathsf{d}_{\mathsf{H}}(\omega, \omega') \text{ for each distinct pair } \omega, \omega' \in \Omega.$$

Putting together the pieces, we see that we have constructed a local packing $\{\theta^{\omega}\}_{\omega \in \Omega}$ with $\log(|\Omega|) \geq s/8$ such that

$$\frac{s}{4}\delta^2 \leq \|\theta^{\omega} - \theta^{\omega'}\|_2^2 \leq s\delta^2 \text{ for each distinct pair } \omega, \omega' \in \Omega.$$

Employing Fano's method (see, e.g., [323, Proposition 15.12 and equation (15.34)]) then yields, for a universal positive constant c, the minimax risk lower bound

$$\inf_{\widehat{\theta} \in \widehat{\Theta}} \sup_{\theta^* \in \mathcal{M}^{\Bbbk,\mathbf{s}}(\mathbb{L}_{d,n})} \mathcal{R}_n(\widehat{\theta}, \theta^*) \geq c \cdot \delta^2 s \left(1 - \frac{\delta^2 s + \log 2}{\log(|\Omega|)}\right) \geq c \cdot \delta^2 s \left(3/4 - 8\delta^2\right),$$

where we have used the fact that $s \geq 32$ in order to write $\frac{\log 2}{s/8} \leq 1/4$. Choosing $\delta = 1/4$ completes the proof. □

Proof of claim (4.38b)**:** The proof of this claim uses the unknown permutations defining the model in order to construct a packing. A similar proof has appeared in the special case $d = 2$ [277]. Let us begin by defining some notation. For any tuple $\mathbf{k} \in \cup_{i=1}^{n_1} \mathbf{K}_i$, let

$$\begin{aligned} \Bbbk_j(\mathbf{k}) &= ((n_1), \ldots, (n_{j-1}), \mathbf{k}, (n_{j+1}), \ldots, (n_d)) \text{ and} \\ \mathbf{s}_j(\mathbf{k}) &= (\ 1, \quad \ldots \quad, 1\ , \quad |\mathbf{k}|\ ,\ 1, \quad \ldots \quad, 1\ ), \end{aligned}$$

respectively. In words, these denote the size tuple and cardinality tuple corresponding to a single indifference set along all dimensions except the j-th, along which we have $|\mathbf{k}|$ indifference sets with cardinalities given by the tuple $\mathbf{k}$.

Turning now to the problem at hand, consider $\mathbb{k} \in \mathbb{K}_{\mathbf{s}}$, and let $j^* \in \operatorname{argmin}_{j \in [d]} k^j_{\max}$ be any index that satisfies $k^{j^*}_{\max} = k^*$. Let $\widetilde{\mathbb{k}} = \mathbb{k}_{j^*}(\mathbf{k}^{j^*})$, and $\widetilde{\mathbf{s}} = \mathbf{s}_{j^*}(\mathbf{k}^{j^*})$. Finally, define the special tuple $\mathbf{s}^{(2)} \in \mathbb{N}^d$ by specifying, for each $j \in [d]$, its j-th entry as

$$[\mathbf{s}^{(2)}]_j := \begin{cases} 2 & \text{if } j = j^* \\ 1 & \text{otherwise.} \end{cases}$$

By definition, we have

$$\mathcal{M}^{\widetilde{\mathbb{k}}, \widetilde{\mathbf{s}}}_{\mathsf{perm}}(\mathbb{L}_{d,n}) \subseteq \mathcal{M}^{\mathbb{k}, \mathbf{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n}). \tag{4.39}$$

We require Lemma 4.6.2, which is stated and proved at the end of this subsection, and split the proof into two cases depending on a property of the size tuple $\mathbb{k}$.

Case $k^* > n_1/3$: In this case, set $\mathbf{k} = (k^*, n_1 - k^*)$ and note the inclusion

$$\mathcal{M}^{\mathbb{k}_{j^*}(\mathbf{k}), \mathbf{s}^{(2)}}_{\mathsf{perm}}(\mathbb{L}_{d,n}) \subseteq \mathcal{M}^{\widetilde{\mathbb{k}}, \widetilde{\mathbf{s}}}_{\mathsf{perm}}(\mathbb{L}_{d,n}).$$

Applying Lemma 4.6.2 in conjunction with the further inclusion (4.39) then yields the bound

$$\mathfrak{M}_{d,n}(\mathbb{k}, \mathbf{s}) \geq \frac{c}{n} \min\{k^*, n_1 - k^*\} \geq \frac{c}{2n}(n_1 - k^*),$$

where in the final inequality, we have used the bound $k^* > n_1/3$.

Case $k^* \leq n_1/3$: In this case, note that the largest indifference set defined by the size tuple $\mathbf{k}^{j^*}$ is at most $n_1/3$. Consequently, these indifference sets can be combined to form two indifference sets of sizes $(\widetilde{k}, n_1 - \widetilde{k})$ for some $n_1/3 \leq \widetilde{k} \leq 2n_1/3$. Now letting $\widetilde{\mathbf{k}} := (\widetilde{k}, n_1 - \widetilde{k})$, we have

$$\mathcal{M}^{\mathbb{k}_{j^*}(\widetilde{\mathbf{k}}), \mathbf{s}^{(2)}}_{\mathsf{perm}}(\mathbb{L}_{d,n}) \subseteq \mathcal{M}^{\widetilde{\mathbb{k}}, \widetilde{\mathbf{s}}}_{\mathsf{perm}}(\mathbb{L}_{d,n}),$$

and proceeding as before completes the proof for this case. □

Lemma 4.6.2. *Suppose* $\mathbf{k} = (k_1, k_2)$, *with* $\overline{k} := \max_{\ell=1,2} k_\ell$, *and let* $\mathcal{K}_{\mathsf{perm}} = \mathcal{M}^{\mathbb{k}_j(\mathbf{k}), \mathbf{s}_j(\mathbf{k})}_{\mathsf{perm}}(\mathbb{L}_{d,n})$ *for convenience. Then, for each* $j \in [d]$, *we have*

$$\inf_{\widehat{\theta} \in \widehat{\Theta}} \sup_{\theta^* \in \mathcal{K}_{\mathsf{perm}}} \mathcal{R}_n(\widehat{\theta}, \theta^*) \geq c \cdot \frac{n_1 - \overline{k}}{n},$$

where c *is a universal positive constant.*

Proof. Define the set $\mathcal{K} := \mathcal{M}^{\mathbb{k}_j(\mathbf{k}), \mathbf{s}_j(\mathbf{k})}(\mathbb{L}_{d,n})$ for convenience. Suppose wlog that $k_2 \leq k_1$, so that it suffices to prove the bound

$$\inf_{\widehat{\theta} \in \widehat{\Theta}} \sup_{\theta^* \in \mathcal{K}_{\mathsf{perm}}} \mathcal{R}_n(\widehat{\theta}, \theta^*) \geq c \cdot \frac{k_2}{n}.$$

Also note that by symmetry, it suffices to prove the bound for $j = 1$.
Case $k_2 < 32$: From the proof of claim (4.38a), recall the set $\mathcal{M}^{\mathbb{k}_1, \mathsf{s}^{(1)}}(\mathbb{L}_{d,n})$, noting the inclusion $\mathcal{M}^{\mathbb{k}_1, \mathsf{s}^{(1)}}(\mathbb{L}_{d,n}) \subseteq \mathcal{K}_{\mathsf{perm}}$. From the same proof, we thus have the bound $\inf_{\widehat{\theta} \in \widehat{\Theta}} \sup_{\theta^* \in \mathcal{K}_{\mathsf{perm}}} \mathcal{R}_n(\widehat{\theta}, \theta^*) \geq c \cdot \frac{1}{n}$, which suffices since the constant factors can be adjusted appropriately.
Case $k_2 \geq 32$: As before, we use the Gilbert–Varshamov bound [121, 316] to claim that there must exist a set of binary vectors $\Omega \subseteq \{0, 1\}^{k_2}$ such that $\log(|\Omega|) \geq k_2/8$ and $\mathsf{d}_{\mathsf{H}}(\omega, \omega') \geq k_2/4$ for each distinct $\omega, \omega' \in \Omega$. For a positive scalar δ to be specified shortly, construct a base tensor $\theta_0 \in \mathcal{K}$ by specifying its entries as

$$\theta_0(i_1, \ldots, i_d) := \begin{cases} \delta & \text{if } i_1 \leq k_2 \\ 0 & \text{otherwise.} \end{cases}$$

Now for each $\omega \in \Omega$, define the tensor $\theta^\omega \in \mathbb{R}_{d,n}$ via

$$\theta^\omega(i_1, \ldots, i_d) := \begin{cases} \delta \cdot \omega_{i_1} & \text{if } i_1 \leq k_2 \\ \delta \cdot (1 - \omega_{n_1 - i_1 + 1}) & \text{if } n_1 - k_2 + 1 \leq i_1 \leq n_1 \\ 0 & \text{otherwise.} \end{cases}$$

Since $k_2 \leq k_1$, we have $n_1 - k_2 \geq k_2$. Thus, an equivalent way to construct the tensor θ^ω is to specify a permutation using the vector ω (which flips particular entries depending of the value of ω on that entry), and then apply this permutation along the first dimension of θ_0. Consequently, we have $\theta^\omega \in \mathcal{K}_{\mathsf{perm}}$ for each $\omega \in \Omega$. Also, by construction, we have $\|\theta^\omega - \theta^{\omega'}\|_2^2 = \delta^2 \mathsf{d}_{\mathsf{H}}(\omega, \omega')$, so that the packing over the Hamming cube ensures that

$$\frac{k_2}{4} \cdot \delta^2 \leq \|\theta^\omega - \theta^{\omega'}\|_2^2 \leq k_2 \delta^2 \text{ for all distinct pairs } \omega, \omega' \in \Omega.$$

Thus, applying Fano's method as in the proof of the claim (4.38a) yields, for a small enough universal constant $c > 0$, the bound

$$\inf_{\widehat{\theta} \in \widehat{\Theta}} \sup_{\theta^* \in \mathcal{K}_{\mathsf{perm}}} \mathcal{R}_n(\widehat{\theta}, \theta^*) \geq c \cdot \delta^2 \frac{k_2}{n} \left(1 - \frac{\delta^2 k_2 + \log 2}{k_2/8}\right),$$

and choosing δ to be a small enough constant and noting that $k_2 \geq 32$ completes the proof. □

4.6.4 Proof of Corollary 4.3.2

We establish the two parts of the corollary separately.

Proof of part (a)

The lower bound follows immediately from our proof of claim (4.38a). Let us prove the upper bound. First, note that the set $\mathcal{M}^{\mathbb{k}, \mathsf{s}}(\mathbb{L}_{d,n}) - \mathcal{M}^{\mathbb{k}, \mathsf{s}}(\mathbb{L}_{d,n})$ is star-shaped and non-degenerate (see

Definition A.6.1). Thus, it suffices, as in the proof of Theorem 4.3.1, to bound the expectation of the random variable

$$\xi(t) = \sup_{\substack{\theta_1, \theta_2 \in \mathcal{M}^{\Bbbk, \mathbf{s}}(\mathbb{L}_{d,n}) \\ \|\theta_1 - \theta_2\|_2 \le t}} \langle \epsilon, \theta_1 - \theta_2 \rangle.$$

Applying Lemma A.6.2 from the appendix yields

$$\mathbb{E}[\xi(t)] \le t\sqrt{s},$$

and substituting into the critical inequality (4.23) completes the proof. □

Proof of part (b)

The lower bound follows directly from the corresponding lower bound in part (a) of the corollary. In order to establish the upper bound, we use an argument that is very similar to the proof of the corresponding upper bound in Proposition 4.3.1, and so we sketch the differences.

First, note that $\mathcal{M}^s(\mathbb{L}_{d,n})$ can be written as the union of convex sets. For convenience, let $\phi_s := \{\mathrm{s} \in \mathbb{L}_{d,n} : \prod_{j=1}^d s_j = s\}$. Proceeding as in the proof of Proposition 4.3.1, we see that it suffices to control the expectation of the random variable

$$\xi(t) = \max_{\mathbf{s} \in \phi_s} \max_{\Bbbk \in \mathbb{K}_{\mathbf{s}}} \sup_{\substack{\theta \in \mathcal{M}^{\Bbbk, \mathbf{s}}(\mathbb{L}_{d,n}) \\ \|\theta - \theta^*\|_2 \le t}} \langle \epsilon, \theta - \theta^* \rangle.$$

Now note that $|\mathbb{K}_{\mathbf{s}}|$ can be bounded by counting, for each $j \in [d]$, the number of s_j-tuples of positive integers whose sum is n_1. A stars-and-bars argument thus yields $|\mathbb{K}_{\mathbf{s}}| = \prod_{j=1}^d \binom{n_1 - 1}{s_j - 1} \le n_1^s$. Simultaneously, we also have $|\phi_s| \le s^d$. Putting together the pieces, we see that $\xi(t)$ is the maximum of at most $K := n_1^s \cdot s^d$ random variables. Note that $\log K = s \log n_1 + d \log s \lesssim s \log n$, where we have used the fact that $d \log s \lesssim ds \log n_1 = s \log n$.

Applying Lemma A.6.1(a) from the appendix in conjunction with Lemma 4.6.1 yields, for each $u \ge 0$, the tail bound

$$\Pr\left\{\xi(t) \ge t\sqrt{s} + Ct\left(\sqrt{s \log n} + \sqrt{u}\right)\right\} \le e^{-u}.$$

The rest of the proof is identical to the proof of Proposition 4.3.1, and putting together the pieces yields the claim. □

4.6.5 Proof of Theorem 4.3.2

Our proof of the theorem relies on a result of Brennan and Bresler [43] that reduces the hypergraph planted clique problem instance to an instance of testing between a random and planted tensor model in Gaussian noise. Let us introduce some notation to set up their result. Let $\mu_0 \in \mathbb{R}_{d,n}$ denote the all-zero tensor, and define the scalar $\rho := \frac{\log 2}{\sqrt{6 \log n + 2 \log 2}}$. For each $S \subseteq [n_1]$, define the tensor $\mu_S \in \mathbb{R}_{d,n}$ via

$$\mu_S(i_1, \ldots, i_d) := \begin{cases} \rho & \text{if } i_1, \ldots i_d \in S \\ 0 & \text{otherwise.} \end{cases}$$

With this notation at hand, we are now ready to formulate a conjecture that follows from the hypergraph planted clique conjecture; we define the appropriate hypotheses below. Let $\mathbf{S}_K$ denote a random subset chosen uniformly from all subsets of $[n_1]$ that have size K. Let $\emptyset$ denote the empty set. Given a random variable $Y \in \mathbb{R}_{d,n}$, we would like to distinguish the hypotheses $\widetilde{H}_0 : Y \sim \mathcal{N}(\mu_0, I)$ and $\widetilde{H}_1 : Y \sim \mathcal{N}(\mu_{\mathbf{S}_K}, I)$. The error of any test $\psi : \mathbb{R}_{d,n} \to \{0, 1\}$ is defined as before (4.12), but now with the hypotheses $\widetilde{H}_0$ and $\widetilde{H}_1$ representing the null and alternative, respectively.

Conjecture 4.6.1 (Gaussian planted tensor conjecture)**.** *Under the setup above, there is a pair of universal positive constants (c, C) such that if $n_1 \geq C$ and $K \leq c\sqrt{n_1}$, then any test ψ that is computable in time polynomial in n must satisfy*

$$\mathcal{E}(\psi) > (10n)^{-1}.$$

Conjecture 4.6.1 is a consequence of Conjecture 4.3.1; see the reduction in [43, Section J], and the rejection kernel framework [44]. We are now ready to prove Theorem 4.3.2 from Conjecture 4.6.1.

Let c denote the universal constant appearing in Conjecture 4.6.1, i.e., the conjecture holds provided $K \leq c\sqrt{n_1}$, and let $c_0 = c/2$. Our proof proceeds as follows. We first suppose that there exists a polynomial time algorithm having small adaptivity index. We then use the output of this algorithm, in conjunction with another polynomial-time decision rule, to distinguish the two hypotheses defining Conjecture 4.6.1 when $K = c_0\sqrt{n_1}$, thereby contradicting the conjecture. Let us now describe the details.

Begin by recalling the upper bound (4.10) on the minimax risk

$$\mathfrak{M}_{d,n}(\mathbb{k}, \mathbf{s}) \leq C_0 \left(\frac{s + (n_1 - k^*) \log n}{n} \right)$$

for a universal constant $C_0 > 0$. Recall that $\delta_n = (10n)^{-1}$, and let $\widehat{\theta}$ be a polynomial-time computable estimator with high-probability adaptivity index $\mathfrak{A}(\widehat{\theta}; \delta_n) \leq \frac{c_0^{d-1}}{64 C_0} \cdot n^{\frac{1}{2}(1-1/d)} (\log n)^{-2}$. By definition of the high-probability adaptivity index, this implies that for all $\theta^* \in \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n})$, we have

$$\|\widehat{\theta} - \theta^*\|_2^2 \leq \frac{c_0^{d-1}}{64 \log n} \cdot n^{\frac{1}{2}(1-1/d)} \cdot \{s(\theta^*) + (n_1 - k^*(\theta^*))\} \tag{4.40}$$

with probability greater than $1 - (10n)^{-1}$. Here, we have used $s(\theta)$ to denote the number of hyper-rectangular blocks partitioning $\theta \in \mathbb{R}_{d,n}$ and $k^*(\theta) := \min_{j \in [d]} k^j_{\max}(\theta)$. Here $k^j_{\max}(\theta)$ is the largest indifference set of θ along dimension j.

Given a value $K = c_0 \cdot \sqrt{n_1}$, let us now use this estimator $\widehat{\theta}$ in order to distinguish the two hypotheses appearing in Conjecture 4.6.1. Given a random tensor Y, recall that Conjecture 4.6.1 posits that any test that distinguishes the hypotheses $\widetilde{H}_0 : Y \sim \mathcal{N}(\mu_0, I)$ and $\widetilde{H}_1 : Y \sim \mathcal{N}(\mu_{\mathsf{S}_K}, I)$ and is computable in time polynomial in n must have error rate exceeding $(10n)^{-1}$. We now construct a decision rule from the estimator $\widehat{\theta}$, via

$$\psi(Y) := \mathbf{1}\left\{ \left\| \widehat{\theta} - \mu_0 \right\|_2^2 \geq \frac{1}{4}\rho^2 \cdot K^d \right\}.$$

Clearly, this decision rule is computable from $\widehat{\theta}$ in $\mathcal{O}(n)$ time. Let us analyze its error under the null and alternative hypotheses.

Null case: In this case, the mean of Y is given by the all-zero tensor μ_0, for which $s(\mu_0) = 1$ and $k^*(\mu_0) = n_1$. Equation (4.40) thus yields that with probability at least $1 - (10n)^{-1}$, we have

$$\|\widehat{\theta} - \mu_0\|_2^2 \leq \frac{c_0^{d-1}}{64 \log n} \cdot n^{\frac{1}{2}(1-1/d)} = \frac{1}{64 \log n} \cdot K^{d-1} < \frac{1}{4}\rho^2 \cdot K^d,$$

where the equality is a result of substituting our choice of K. Thus, on this event, we have $\left\| \widehat{\theta} - \mu_0 \right\|_2^2 < \frac{1}{4}\rho^2 \cdot K^d$, and so $\psi(Y) = 0$. Consequently, the test succeeds in the null case with probability at least $1 - \delta_n$, and so $\mathbb{E}_{H_0}[\psi(Y)] \leq \delta_n$.

Alternative case: In this case, we have $\mu_{\mathsf{S}_K} \in \mathcal{M}_{\mathsf{perm}}$ for each random choice of subset S_K with $s(\mu_{\mathsf{S}_K}) = 2^d$ and $k^*(\mu_{\mathsf{S}_K}) = n_1 - k$. Now suppose that $n \geq C_d$ for a large enough C_d, such that $2^d \leq K/3$. Thus, the bound (4.40) now yields that with probability greater than $1 - \delta_n$, we have

$$\|\widehat{\theta} - \mu_{\mathsf{S}_K}\|_2^2 \leq \frac{c_0^{d-1}}{64 \log n} \cdot (2^d + K) \cdot n^{\frac{1}{2}(1-1/d)} \leq \frac{1}{4}\rho^2 \cdot K^d.$$

Consequently, on this event and for $n \geq C_d$, we have

$$\begin{aligned} \left\| \widehat{\theta} - \mu_0 \right\|_2^2 &\geq \frac{1}{2} \|\mu_{\mathsf{S}_K} - \mu_0\|_2^2 - \|\widehat{\theta} - \mu_{\mathsf{S}_K}\|_2^2 \\ &\geq \frac{1}{2}\rho^2 \cdot K^d - \frac{1}{4}\rho^2 \cdot K^d = \frac{1}{4}\rho^2 \cdot K^d, \end{aligned}$$

in which case $\psi(Y) = 1$. Thus, the test succeeds in the alternative case with probability at least $1 - \delta_n$, and so $\mathbb{E}_{H_1}[\psi(Y)] \leq \delta_n$.

Putting together the two cases yields the bound $\mathcal{E}(\psi) \leq \delta_n$, which provides the required contradiction to Conjecture 4.6.1. □

4.6.6 Proof of Theorem 4.4.1

We first establish a certain error decomposition that results from our algorithm, and then proceed to proofs of the two parts of the theorem. Recall that the algorithm computes, from the score vectors, a set of estimated ordered partitions $\widehat{\mathsf{bl}}_1, \ldots, \widehat{\mathsf{bl}}_d$, and projects the observations onto the set $\mathcal{M}(\mathbb{L}_{d,n}; \widehat{\mathsf{bl}}_1, \ldots, \widehat{\mathsf{bl}}_d)$. Denote by $\mathfrak{B}(\theta; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)$ the tensor obtained by projecting $\theta \in \mathbb{R}_{n,d}$ onto the set $\mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)$; since the set is closed and convex, the projection is unique and given by

$$\mathfrak{B}(\theta; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) = \underset{\widetilde{\theta} \in \mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)}{\operatorname{argmin}} \|\theta - \widetilde{\theta}\|_2^2. \tag{4.41}$$

Additionally, for notational convenience, let $\widehat{\mathsf{B}} := \widehat{\mathsf{bl}}_1, \ldots, \widehat{\mathsf{bl}}_d$, so that $\widehat{\theta}_{\mathsf{MP}} = \mathfrak{B}(\theta^* + \epsilon; \widehat{\mathsf{B}})$. Applying the triangle inequality yields

$$\begin{aligned}
&\|\widehat{\theta}_{\mathsf{MP}} - \theta^*\|_2 \\
&\leq \|\widehat{\theta}_{\mathsf{MP}} - \mathfrak{B}(\mathfrak{B}(\theta^*; \widehat{\mathsf{B}}) + \epsilon; \widehat{\mathsf{B}})\|_2 + \|\mathfrak{B}(\mathfrak{B}(\theta^*; \widehat{\mathsf{B}}) + \epsilon; \widehat{\mathsf{B}}) - \mathfrak{B}(\theta^*; \widehat{\mathsf{B}})\|_2 + \|\mathfrak{B}(\theta^*; \widehat{\mathsf{B}}) - \theta^*\|_2 \\
&= \|\mathfrak{B}(\mathfrak{B}(\theta^*; \widehat{\mathsf{B}}) + \epsilon; \widehat{\mathsf{B}}) - \mathfrak{B}(\theta^* + \epsilon; \widehat{\mathsf{B}})\|_2 + \|\mathfrak{B}(\mathfrak{B}(\theta^*; \widehat{\mathsf{B}}) + \epsilon; \widehat{\mathsf{B}}) - \mathfrak{B}(\theta^*; \widehat{\mathsf{B}})\|_2 + \|\mathfrak{B}(\theta^*; \widehat{\mathsf{B}}) - \theta^*\|_2 && (4.42) \\
&\leq \underbrace{\|\mathfrak{B}(\mathfrak{B}(\theta^*; \widehat{\mathsf{B}}) + \epsilon; \widehat{\mathsf{B}}) - \mathfrak{B}(\theta^*; \widehat{\mathsf{B}})\|_2}_{\text{estimation error}} + \underbrace{2\|\mathfrak{B}(\theta^*; \widehat{\mathsf{B}}) - \theta^*\|_2}_{\text{approximation error}}, && (4.43)
\end{aligned}$$

where inequality (4.43) follows from the non-expansiveness of an ℓ_2-projection onto a convex set (4.20b), when applied to the first term in equation (4.42). We now state three lemmas that lead to the desired bounds in the various cases. For an ordered partition bl, denote by $\mathsf{card}(\mathsf{bl})$ the number of blocks in the partition, and let $\kappa^*(\mathsf{bl})$ denote the size of the largest block in the partition. Our first lemma captures some key structural properties of the estimated ordered partitions $\widehat{\mathsf{bl}}_1, \ldots, \widehat{\mathsf{bl}}_d$.

Lemma 4.6.3. *Suppose that $\theta^* \in \mathcal{M}^{\mathbb{k},\mathsf{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n})$. Then with probability at least $1 - 2n^{-7}$, the partition $\widehat{\mathsf{B}} = \widehat{\mathsf{bl}}_1, \ldots, \widehat{\mathsf{bl}}_d$ satisfies*

$$\mathsf{card}(\widehat{\mathsf{bl}}_j) \leq s_j \quad \textit{and} \quad \kappa^*(\widehat{\mathsf{bl}}_j) \geq k^j_{\max} \quad \textit{simultaneously for all } j \in [d]. \tag{4.44}$$

Our next lemma bounds the estimation error term in two different ways.

Lemma 4.6.4. *There is a universal positive constant C such that for all $u \geq 0$, each of the following statements holds with probability greater than $1 - e^{-u}$:*

(a) For any set of one-dimensional ordered partitions $\mathsf{bl}_1, \ldots, \mathsf{bl}_d$ satisfying $\mathsf{card}(\mathsf{bl}_j) \leq \widetilde{s}_j$ for all $j \in [d]$, and any tensor $\theta \in \mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)$, we have

$$\|\mathfrak{B}(\theta + \epsilon; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) - \theta\|_2^2 \leq C\left(\widetilde{s} + u\right),$$

where $\widetilde{s} = \prod_{j=1}^{d} \widetilde{s}_j$*.*

(b) For any set of one-dimensional ordered partitions $\mathsf{bl}_1, \ldots, \mathsf{bl}_d$ *and any tensor* $\theta \in \mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) \cap \mathbb{B}_\infty(1)$*, we have*

$$\|\mathfrak{B}(\theta + \epsilon; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) - \theta\|_2^2 \leq C(n^{1-1/d} \log^{5/2} n + u).$$

Our final lemma handles the approximation error term.

Lemma 4.6.5. *There is a universal positive constant* C *such that for each* $\theta^* \in \mathcal{M}^{\Bbbk,\mathbf{s}}_{\mathsf{perm}}(\mathbb{L}_{d,n})$*, we have*

$$\Pr\{\|\mathfrak{B}(\theta^*; \widehat{\mathsf{B}}) - \theta^*\|_2^2 \geq Cd^2(n_1 - k^*) \cdot n^{\frac{1}{2}(1-1/d)} \log n\} \leq 4n^{-7}, \text{ and} \tag{4.45a}$$

$$\mathbb{E}\left[\|\mathfrak{B}(\theta^*; \widehat{\mathsf{B}}) - \theta^*\|_2^2\right] \leq Cd^2(n_1 - k^*) \cdot n^{\frac{1}{2}(1-1/d)} \log n. \tag{4.45b}$$

We prove these lemmas in the subsections to follow. For now, let us use them to prove the two parts of Theorem 4.4.1.

Proof of Theorem 4.4.1, part (a)

Consider any tensor $\theta_0 \in \mathbb{B}_\infty(1)$. First, note that applying Lemma A.5.2 in the appendix, we obtain the inclusion $\mathfrak{B}(\theta_0; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) \in \mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) \cap \mathbb{B}_\infty(1)$, since the operator $\mathfrak{B}$ is ℓ_∞-contractive. Also, note from Lemma A.6.8(a) in the appendix that the total number of one-dimensional partitions satisfies $|\mathfrak{P}| = (n_1)^{n_1}$. Thus, we may apply Lemma 4.6.4(b) with the substitution $\theta = \mathfrak{B}(\theta_0; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)$ and $u = n_1 \log n + u'$. In conjunction with a union bound over at most $|\mathfrak{P}|^d$ possible choices of one-dimensional ordered partitions $\mathsf{bl}_1, \ldots, \mathsf{bl}_d \in \mathfrak{P}$, this yields the bound

$$\begin{aligned}\max_{\mathsf{bl}_1, \ldots, \mathsf{bl}_d \in \mathfrak{P}} &\|\mathfrak{B}(\mathfrak{B}(\theta_0; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) + \epsilon; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) - \mathfrak{B}(\theta_0; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)\|_2^2 \\ &\leq C \cdot (n^{1-1/d} \log^{5/2} n + n_1 \log n + u')\end{aligned}$$

with probability at least

$$1 - |\mathfrak{P}|^d \exp(-n_1 \log n - u') \geq 1 - (n_1)^{n_1 \cdot d} \cdot (n_1)^{-n_1 \cdot d} \cdot e^{-u'} = 1 - e^{-u'},$$

where we have used the fact that $\log (n_1)^{n_1 \cdot d} = n_1 \log n$ for each $n_1 \geq 2$. Integrating this tail bound and noting that $\widehat{\mathsf{bl}}_1, \ldots, \widehat{\mathsf{bl}}_d \in \mathfrak{P}$, we obtain

$$\mathbb{E}\left[\|\mathfrak{B}(\mathfrak{B}(\theta_0; \widehat{\mathsf{B}}) + \epsilon; \widehat{\mathsf{B}}) - \mathfrak{B}(\theta_0; \widehat{\mathsf{B}})\|_2^2\right] \leq Cn^{1-1/d} \log^{5/2} n$$

for any $\theta_0 \in \mathbb{B}_\infty(1)$. Choosing $\theta_0 = \theta^*$ and combining this with equation (4.43) and Lemma 4.6.5 completes the proof. □

Proof of Theorem 4.4.1, part (b)

We split the proof into two cases depending on the value of s.

Case 1: Let us first handle the case $s = 1$, in which case $n_1 - k^* = 0$. By Lemma 4.6.3, there is an event occurring with probability greater than $1 - 2n^{-7}$ such that on this event, our estimated blocks satisfy $\mathsf{card}(\widehat{\mathsf{bl}}_j) = 1$. On this event, the projection is a constant tensor, and each entry of the error tensor $\mathfrak{B}(\mathfrak{B}(\theta^*; \widehat{\mathsf{B}}) + \epsilon; \widehat{\mathsf{B}}) - \mathfrak{B}(\theta^*; \widehat{\mathsf{B}})$ is equal to $\overline{\epsilon} := n^{-1} \sum_{x \in \mathbb{L}_{d,n}} \epsilon_x$. But $\|\overline{\epsilon} \cdot \mathbb{1}_{d,n}\|_2^2 \sim \chi_1^2$, and so a tail bound for the standard Gaussian yields

$$\Pr\{\|\overline{\epsilon} \cdot \mathbb{1}_{d,n}\|_2 \geq t\} \leq e^{-t^2/2}.$$

Putting together the pieces with a union bound yields $\|\mathfrak{B}(\mathfrak{B}(\theta^*; \widehat{\mathsf{B}}) + \epsilon; \widehat{\mathsf{B}}) - \mathfrak{B}(\theta^*; \widehat{\mathsf{B}})\|_2^2 \leq 8 \log n$ with probability at least $1 - 2n^{-7} - n^{-4}$. In order to bound the error in expectation, first note that since the projection onto a convex set is non-expansive (4.20b), we have the pointwise bound

$$\|\mathfrak{B}(\mathfrak{B}(\theta^*; \widehat{\mathsf{B}}) + \epsilon; \widehat{\mathsf{B}}) - \mathfrak{B}(\theta^*; \widehat{\mathsf{B}})\|_2^2 \leq \|\epsilon\|_2^2.$$

Putting together the pieces and applying Lemma A.6.6 from the appendix then yields the bound

$$\mathbb{E}\left[\|\mathfrak{B}(\mathfrak{B}(\theta^*; \widehat{\mathsf{B}}) + \epsilon; \widehat{\mathsf{B}}) - \mathfrak{B}(\theta^*; \widehat{\mathsf{B}})\|_2^2\right] \leq 8 \log n + \sqrt{2n^{-7} + n^{-4}} \cdot \sqrt{\mathbb{E}[\|\epsilon\|_2^4]} \leq Cs \log n.$$

Combining with equation (4.43) and Lemma 4.6.5 completes the proof of the claim in expectation. The proof for this case is thus complete.

Case 2: In this case, $s \geq 2$, which ensures that $n_1 - k^* \geq 1$. Recall the set $\mathfrak{P}_{k^*}^{\max}$ defined in the proof of Proposition 4.3.1, and note that applying Lemma A.6.8(b) from the appendix yields the bound $|\mathfrak{P}_{k^*}^{\max}| \leq e^{3(n_1 - k^*) \log n_1}$. With this calculation in hand, the proof proceeds very similarly to that of Theorem 4.4.1(a).

We first apply Lemma 4.6.4(a) with the substitution $u = C(n_1 - k^*) \log n$ and take a union bound over at most $|\mathfrak{P}_{k^*}^{\max}|^d$ possible choices of ordered partitions $\mathsf{bl}_1, \ldots, \mathsf{bl}_d \in \mathfrak{P}_{k^*}^{\max}$ satisfying $\mathsf{card}(\mathsf{bl}_j) \leq s_j$ for all $j \in [d]$. This yields, for each $\theta^* \in \mathbb{R}_{d,n}$, the bound

$$\max_{\substack{\mathsf{bl}_1, \ldots, \mathsf{bl}_d \in \mathfrak{P}_{k^*}^{\max}: \\ \mathsf{card}(\mathsf{bl}_j) \leq s_j,\, j \in [d]}} \|\mathfrak{B}(\mathfrak{B}(\theta^*; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) + \epsilon; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) - \mathfrak{B}(\theta^*; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)\|_2^2$$
$$\leq C \cdot (s + (n_1 - k^*) \log n)$$

with probability exceeding

$$1 - |\mathfrak{P}_{k^*}^{\max}|^d \cdot \exp(-C(n_1 - k^*) \log n) \geq 1 - \exp(-(C - 3) \cdot (n_1 - k^*) \log n) \geq 1 - n^{-7}.$$

Here, the last inequality can be ensured by choosing a large enough constant C, since $n_1 - k^* \geq 1$.

Furthermore, Lemma 4.6.3 guarantees that with probability at least $1 - 2n^{-7}$, we have the inclusions $\widehat{\mathsf{bl}}_1, \ldots, \widehat{\mathsf{bl}}_d \in \mathfrak{P}_{k^*}^{\max}$ and that $\mathsf{card}(\widehat{\mathsf{bl}}_j) \leq s_j$ for all $j \in [d]$. Consequently, by applying a union bound, we obtain, for any $\theta^* \in \mathbb{R}_{d,n}$, the high probability bound

$$\Pr\left\{\|\mathfrak{B}(\mathfrak{B}(\theta^*; \widehat{\mathsf{B}}) + \epsilon; \widehat{\mathsf{B}}) - \mathfrak{B}(\theta^*; \widehat{\mathsf{B}})\|_2^2 \geq C \cdot (s + (n_1 - k^*) \log n)\right\} \leq 3n^{-7}. \tag{4.46}$$

Thus, we have succeeded in bounding the error with high probability. In order to bound the error in expectation, note once again that the projection onto a convex set is non-expansive (4.20b), and so we have the pointwise bound

$$\|\mathfrak{B}(\mathfrak{B}(\theta^*; \widehat{\mathsf{B}}) + \epsilon; \widehat{\mathsf{B}}) - \mathfrak{B}(\theta^*; \widehat{\mathsf{B}})\|_2^2 \leq \|\epsilon\|_2^2.$$

Putting together the pieces and applying Lemma A.6.6 from the appendix then yields

$$\begin{aligned}\mathbb{E}\left[\|\mathfrak{B}(\mathfrak{B}(\theta^*; \widehat{\mathsf{B}}) + \epsilon; \widehat{\mathsf{B}}) - \mathfrak{B}(\theta^*; \widehat{\mathsf{B}})\|_2^2\right] &\leq C \cdot \{s + (n_1 - k^*) \log n\} + \sqrt{3n^{-7}} \cdot \sqrt{\mathbb{E}[\|\epsilon\|_2^4]} \\ &\leq C \cdot \{s + (n_1 - k^*) \log n\}.\end{aligned}$$

Combining with equation (4.43) and Lemma 4.6.5 completes the proof in this case.

Combining the two cases completes the proof of the theorem. □

It remains to prove the three technical lemmas. Before we do so, we state and prove a claim that will be used in multiple proofs.

A preliminary result

Define two events

$$\mathcal{E}_1 := \left\{\|Y - \theta^*\|_\infty \leq 4\sqrt{\log n}\right\} \text{ and} \tag{4.47a}$$

$$\mathcal{E}_2 := \left\{\max_{1 \leq j \leq d} \|\widehat{\tau}_j - \tau_j^*\|_\infty \leq 4\sqrt{\log n} \cdot n^{\frac{1}{2}(1-1/d)}\right\}. \tag{4.47b}$$

and note that by a union bound, we have $\Pr\{\mathcal{E}_1 \cap \mathcal{E}_2\} \geq 1 - 2n^{-7}$. For the rest of this proof, we work on the event $\mathcal{E}_1 \cap \mathcal{E}_2$; recall the graphs G'_j and G_j obtained over the course of running the algorithm. We use the following fact guaranteed by step Ia of the algorithm.

Claim 4.6.1. *On the event $\mathcal{E}_1 \cap \mathcal{E}_2$, the following statements hold simultaneously for all $j \in [d]$:*

(a) The graph G'_j is a directed acyclic graph, and consequently $G_j = G'_j$.

(b) For each $\ell \in [s_j]$ and all pairs of indices $u, v \in I^j_\ell$, the edges $u \to v$ and $v \to u$ do not exist in graph G_j.

Proof. Recall our pairwise statistics $\widehat{\Delta}_j^{\mathsf{sum}}(u,v)$ and $\widehat{\Delta}_j^{\max}(u,v)$, and let $\Delta_j^{\mathsf{sum}}(u,v)$ and $\Delta_j^{\max}(u,v)$ denote their population versions, that is, with θ^* replacing Y in the definition (4.15). Note that by applying the triangle inequality, we obtain

$$|\widehat{\Delta}_j^{\mathsf{sum}}(u,v) - \Delta_j^{\mathsf{sum}}(u,v)| \leq |\widehat{\tau}_j(u) - \tau_j^*(u)| + |\widehat{\tau}_j(u) - \tau_j^*(u)| \text{ and} \tag{4.48a}$$

$$|\widehat{\Delta}_j^{\max}(u,v) - \Delta_j^{\max}(u,v)| \leq 2\|Y - \theta^*\|_\infty. \tag{4.48b}$$

We now prove each part of the claim separately.

Proof of part (a): Working on the event $\mathcal{E}_1 \cap \mathcal{E}_2$ and using equations (4.47) and (4.48), we see that if

$$\widehat{\Delta}_j^{\mathsf{sum}}(u,v) > 8\sqrt{\log n} \cdot n^{\frac{1}{2}(1-1/d)} \quad \text{or} \quad \widehat{\Delta}_j^{\max}(u,v) > 8\sqrt{\log n}$$

then

$$\Delta_j^{\mathsf{sum}}(u,v) = \tau_j^*(u) - \tau_j^*(v) > 0 \quad \text{or} \quad \Delta_j^{\max}(u,v) > 0.$$

In particular, the second relation implies that

$$\theta^*(i_1, \ldots, i_{j-1}, u, i_{j+1}, \ldots, i_d) - \theta^*(i_1, \ldots, i_{j-1}, v, i_{j+1}, \ldots, i_d) > 0 \text{ for some } i_\ell \in [n_1],\ \ell \in [d] \backslash j.$$

In either case, we have $\pi_j^*(u) < \pi_j^*(v)$ by the monotonicity property of θ^*. Thus, every edge $u \to v$ in the graph is consistent with the permutation π_j^*, and so the graph G_j' is acyclic.

Proof of part (b): Note that if $u, v \in I_\ell^j$ for some $\ell \in [s_j]$, then

$$\Delta_j^{\mathsf{sum}}(u,v) = \Delta_j^{\max}(u,v) = 0.$$

Therefore, on the event $\mathcal{E}_1 \cap \mathcal{E}_2$ and owing to the inequalities (4.47) and (4.48), we have

$$|\widehat{\Delta}_j^{\mathsf{sum}}(u,v)| \leq 8\sqrt{\log n} \cdot n^{\frac{1}{2}(1-1/d)} \quad \text{and} \quad |\widehat{\Delta}_j^{\max}(u,v)| \leq 8\sqrt{\log n}.$$

Consequently, neither of the edges $u \to v$ or $v \to u$ exists in the graph G_j. □

We are now ready to establish the individual lemmas.

Proof of Lemma 4.6.3

Suppose wlog that $\pi_1^* = \cdots = \pi_d^* = \mathsf{id}$, so that $\theta^* \in \mathcal{M}^{\mathbb{k},\mathsf{s}}(\mathbb{L}_{d,n})$. This implies that the true ordered partition bl_j^* consists of s_j intervals $(I_1^j, \ldots, I_{s_j}^j)$ of sizes $(k_1^j, \ldots, k_{s_j}^j)$, respectively. The size of the largest interval (call this $I_{\max}^j$) is given by $k_{\max}^j$.

By part (a) of Claim 4.6.1, the graph G_j' is a directed acyclic graph, and so $\mathsf{card}(\widehat{\mathsf{bl}}_j)$ is equal to the size of the minimal partition of the graph into disjoint antichains. But Claim 4.6.1(b) ensures that each of the sets $I_\ell^j, \ell \in [s_j]$ forms an antichain of graph G_j', and furthermore, these sets are disjoint and form a partition of $[n_j]$. Hence, $\mathsf{card}(\widehat{\mathsf{bl}}_j) \leq s_j$.

In order to show that $\kappa^*(\widehat{\mathsf{bl}}_j) \geq k_{\max}^j$, note that by Claim 4.6.1(b), the set $I_{\max}^j$ is an antichain of G_j' of size $k_{\max}^j$. □

Proof of Lemma 4.6.4

Recall that the estimator $\mathfrak{B}(\theta + \epsilon; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)$ is a projection onto a closed, convex set. Using this fact, let us prove the two parts of the lemma separately.

Proof of part (a): Recall that Corollary 4.3.2(a) already provides a bound on the error of interest in expectation. Combining this with Lemma A.6.4 from the appendix then yields the claimed high probability bound. □

Proof of part (b) Let us begin with a simple definition. Note that any one-dimensional ordered partition bl specifies a partial ordering over the set $[n_1]$. We say that a permutation π is *faithful* to the ordered partition bl if it is consistent with this partial ordering, and denote by $\mathcal{F}(\mathsf{bl})$ the set of all permutations that are faithful to bl. By definition, we have the inclusion $\mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) \subseteq \mathcal{M}(\mathbb{L}_{d,n}; \pi_1, \ldots, \pi_d)$ for any tuple $(\pi_1, \ldots, \pi_d)$ satisfying $\pi_j \in \mathcal{F}(\mathsf{bl}_j)$ for all $j \in [d]$.

We now turn to the proof of the lemma. Denote the error tensor by $\widehat{\Delta} = \mathfrak{B}(\theta+\epsilon; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) - \theta$; our key claim is that

$$\mathbb{E}[\|\widehat{\Delta}\|_2^2] \lesssim n^{1-1/d} \log^{5/2} n. \tag{4.49}$$

Indeed, with this claim in hand, the proof of the lemma follows by applying Lemma A.6.4 from the appendix.

We dedicate the rest of the proof to establishing claim (4.49). Our strategy is almost identical to the proof of Corollary 4.3.1; we first control the error of the bounded least squares estimator in this setting and then obtain claim (4.49) via a truncation argument.

Denote the bounded LSE for this setting by

$$\widehat{\theta}_{\mathsf{BLSE}}(r) = \operatorname*{argmin}_{\widetilde{\theta} \in \mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) \cap \mathbb{B}_\infty(r)} \|\theta + \epsilon - \widetilde{\theta}\|_2^2,$$

and let $\widehat{\Delta}_{\mathsf{blse}}(r) := \widehat{\theta}_{\mathsf{BLSE}}(r) - \theta$. Rearranging the basic inequality yields the bound

$$\frac{1}{2}\|\widehat{\Delta}_{\mathsf{blse}}(r)\|_2^2 \leq \sup_{\substack{\widetilde{\theta} \in \mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) \cap \mathbb{B}_\infty(r) \\ \|\widetilde{\theta} - \theta\|_2 \leq \|\widehat{\Delta}_{\mathsf{blse}}(r)\|_2}} \langle \epsilon, \widetilde{\theta} - \theta \rangle \leq \sup_{\substack{\widetilde{\theta} \in \mathcal{M}(\mathbb{L}_{d,n}; \pi_1, \ldots, \pi_d) \cap \mathbb{B}_\infty(r) \\ \|\widetilde{\theta} - \theta\|_2 \leq \|\widehat{\Delta}_{\mathsf{blse}}(r)\|_2}} \langle \epsilon, \widetilde{\theta} - \theta \rangle,$$

for permutations $(\pi_1, \ldots, \pi_d)$ satisfying $\pi_j \in \mathcal{F}(\mathsf{bl}_j)$ for all $j \in [d]$. Now since the inclusion $\theta \in \mathcal{M}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) \cap \mathbb{B}_\infty(1)$ holds, we also have $\theta \in \mathcal{M}(\mathbb{L}_{d,n}; \pi_1, \ldots, \pi_d) \cap \mathbb{B}_\infty(1)$. For convenience, let $\overline{\theta} = \theta\{\pi_1^{-1}, \ldots, \pi_d^{-1}\}$. Proceeding from the previous bound, we have

$$\begin{aligned}
\sup_{\substack{\widetilde{\theta} \in \mathcal{M}(\mathbb{L}_{d,n};\pi_1,\ldots,\pi_d) \cap \mathbb{B}_\infty(r) \\ \|\widetilde{\theta} - \theta\|_2 \leq \|\widehat{\Delta}_{\mathsf{blse}}(r)\|_2}} \langle \epsilon, \widetilde{\theta} - \theta \rangle &= \sup_{\substack{\widetilde{\theta} \in \mathcal{M}(\mathbb{L}_{d,n};\pi_1,\ldots,\pi_d) \cap \mathbb{B}_\infty(r) \\ \|\widetilde{\theta} - \theta\|_2 \leq \|\widehat{\Delta}_{\mathsf{blse}}(r)\|_2}} \langle \epsilon\{\pi_1, \ldots, \pi_d\}, (\widetilde{\theta} - \theta)\{\pi_1^{-1}, \ldots, \pi_d^{-1}\} \rangle \\
&= \sup_{\substack{\widetilde{\theta} \in \mathcal{M}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(r) \\ \|\widetilde{\theta} - \overline{\theta}\|_2 \leq \|\widehat{\Delta}_{\mathsf{blse}}(r)\|_2}} \langle \epsilon\{\pi_1, \ldots, \pi_d\}, \widetilde{\theta} - \theta \rangle \\
&\overset{d}{=} \sup_{\substack{\widetilde{\theta} \in \mathcal{M}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(r) \\ \|\widetilde{\theta} - \overline{\theta}\|_2 \leq \|\widehat{\Delta}_{\mathsf{blse}}(r)\|_2}} \langle \epsilon, \widetilde{\theta} - \theta \rangle,
\end{aligned}$$

where the equality in distribution follows from the exchangeability of the noise ϵ. Recall the notation $\mathcal{M}^{\mathsf{full}}(r)$ from the proof of Theorem 4.3.1. Since $\overline{\theta} \in \mathcal{M}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)$, we have $\widetilde{\theta} - \overline{\theta} \in \mathcal{M}^{\mathsf{full}}(r) - \mathcal{M}^{\mathsf{full}}(r)$. Let

$$\xi(t) = \sup_{\substack{\widetilde{\theta} \in \mathcal{M}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(r) \\ \|\widetilde{\theta} - \overline{\theta}\|_2 \leq t}} \langle \epsilon, \widetilde{\theta} - \theta \rangle,$$

and note from the proof of Theorem 4.3.1 that we have $\mathbb{E}[\xi(t)] \lesssim rn^{1-1/d} \log^2 n$ for each $r \in [0, n]$. Since the set $\mathcal{M}^{\mathsf{full}}(r) - \mathcal{M}^{\mathsf{full}}(r)$ is star-shaped and non-degenerate, applying Lemma A.6.3 then yields the bound

$$\mathbb{E}[\|\widehat{\Delta}_{\mathsf{blse}}(r)\|_2^2] \lesssim rn^{1-1/d} \log^2 n \text{ for each } r \in [0, n].$$

We now employ the truncation argument from the proof of Corollary 4.3.1 to bound $\mathbb{E}[\|\widehat{\Delta}\|_2^2]$. Lemma A.5.2 from the appendix guarantees the existence of an event $\mathcal{E}$ occurring with probability greater than $1 - n^{-7}$, on which $\|\mathfrak{B}(\theta + \epsilon; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)\|_\infty \leq \psi_n := 4\sqrt{\log n} + 1$. On this event, we therefore have

$$\mathfrak{B}(\theta + \epsilon; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) = \widehat{\theta}_{\mathsf{BLSE}}(\psi_n) \quad \text{and} \quad \|\widehat{\Delta}\|_2^2 = \|\widehat{\Delta}_{\mathsf{blse}}(\psi_n)\|_2^2.$$

Finally, since $\mathfrak{B}(\theta + \epsilon; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)$ is obtained via an ℓ_2-projection onto a convex set, we may apply inequality (4.20b) to obtain $\|\widehat{\Delta}\|_2^2 \leq \|\epsilon\|_2^2$ pointwise. Combining the pieces as in the proof of Corollary 4.3.1, we obtain claim (4.49). □

Proof of Lemma 4.6.5

Suppose wlog that $\pi_1^* = \cdots = \pi_d^* = \mathsf{id}$, so that $\theta^* \in \mathcal{M}^{\mathbb{k},\mathsf{s}}(\mathbb{L}_{d,n})$. Recall the set $\mathcal{F}(\mathsf{bl})$ containing permutations that are faithful to the ordered partition bl.

Applying Lemma A.5.1 in the appendix, we see that the projection onto $\mathcal{M}(\mathbb{L}_{d,n}; \widehat{\mathsf{bl}}_1, \ldots, \widehat{\mathsf{bl}}_d)$ can be written as successive projections $\mathfrak{B}(\,\cdot\,; \widehat{\mathsf{bl}}_1, \ldots, \widehat{\mathsf{bl}}_d) = \mathcal{P}(\,\mathcal{A}(\,\cdot\,; \widehat{\mathsf{bl}}_1, \ldots, \widehat{\mathsf{bl}}_d)\,; \widehat{\pi}_1, \ldots, \widehat{\pi}_d)$, where $\widehat{\pi}_1, \ldots, \widehat{\pi}_d$ is any set of permutations such that $\widehat{\pi}_j \in \mathcal{F}(\widehat{\mathsf{bl}}_j)$. Thus, from successive applications of the triangle inequality, we obtain

$$\begin{aligned}
&\|\mathfrak{B}(\theta^*; \widehat{\mathsf{bl}}_1, \ldots, \widehat{\mathsf{bl}}_d) - \theta^*\|_2 \\
&\quad \leq \|\mathfrak{B}(\theta^*; \widehat{\mathsf{bl}}_1, \ldots, \widehat{\mathsf{bl}}_d) - \mathcal{P}(\theta^*; \widehat{\pi}_1, \ldots, \widehat{\pi}_d)\|_2 + \|\mathcal{P}(\theta^*; \widehat{\pi}_1, \ldots, \widehat{\pi}_d) - \theta^*\{\widehat{\pi}_1, \ldots, \widehat{\pi}_d\}\|_2 \\
&\qquad\qquad + \|\theta^*\{\widehat{\pi}_1, \ldots, \widehat{\pi}_d\} - \theta^*\|_2 \\
&\quad \overset{\text{(i)}}{\leq} \|\mathcal{A}(\theta^*; \widehat{\mathsf{bl}}_1, \ldots, \widehat{\mathsf{bl}}_d) - \theta^*\|_2 + 2\|\theta^*\{\widehat{\pi}_1, \ldots, \widehat{\pi}_d\} - \theta^*\|_2,
\end{aligned}$$

where in step (i), we have twice used the fact that the projection onto a convex set is non-expansive (4.20b). We now bound these two terms separately, starting with the second term, but first, for each $j \in [d]$, define the random variables

$$T_j := 2\|\widehat{\tau}_j - \tau_j^*\|_\infty + 8\sqrt{\log n} \cdot n^{\frac{1}{2}(1-1/d)} \quad \text{and} \quad U := 2\|Y - \theta^*\|_\infty + 8\sqrt{\log n}. \tag{4.50}$$

Also recall the events $\mathcal{E}_1$ and $\mathcal{E}_2$ defined just above equation (4.47).

Bound on permutation error: Our proof proceeds by bounding this term in two different ways; let us now sketch it. We will establish the claims

$$\|\theta^*\{\widehat{\pi}_1, \ldots, \widehat{\pi}_d\} - \theta^*\|_2^2 \leq 2d \cdot (n_1 - k^*) \cdot \begin{cases} U \cdot \sum_{j=1}^d T_j & \text{conditioned on event } \mathcal{E}_1 \cap \mathcal{E}_2, \\ \sum_{j=1}^d T_j^2 & \text{pointwise.} \end{cases} \tag{4.51}$$

Let us take equation (4.51) as given for the moment and establish a bound on the permutation error. First, note that conditioned on $\mathcal{E}_1 \cap \mathcal{E}_2$, we have $T_j \lesssim \sqrt{\log n} \cdot n^{\frac{1}{2}(1-1/d)}$ and $U \lesssim \sqrt{\log n}$. Also note that $\Pr\{(\mathcal{E}_1 \cap \mathcal{E}_2)^c\} \leq 2n^{-7}$; this established the claimed high probability bound. On the other hand, the (unconditional) expectation can be bounded as

$$\mathbb{E}[T_j^2] \leq 4\mathbb{E}\left[\|\widehat{\tau}_j - \tau_j^*\|_\infty^2\right] + 128(\log n) \cdot n^{1-1/d} \lesssim \log n \cdot n^{1-1/d},$$

where the second inequality follows since $\widehat{\tau}_j \sim \mathcal{N}(\tau_j^*, n^{1-1/d} \cdot I)$ and so $\|\widehat{\tau}_j - \tau_j^*\|_\infty$ is maximum absolute deviation of n_1 i.i.d. Gaussian random variables with mean zero and variance $n^{1-1/d}$.

Putting together the pieces and applying Lemma A.6.6 from the appendix, we obtain

$$\begin{aligned}
\frac{\mathbb{E}\left[\|\theta^*\{\widehat{\pi}_1, \ldots, \widehat{\pi}_d\} - \theta^*\|_2^2\right]}{d^2(n_1 - k^*)} &\lesssim \left(\sqrt{\log n}\right) \cdot \left(\sqrt{\log n} \cdot n^{\frac{1}{2}(1-1/d)}\right) + (\log n) \cdot n^{1-1/d} \cdot n^{-7/2} \\
&\lesssim (\log n) \cdot \cdot n^{\frac{1}{2}(1-1/d)},
\end{aligned}$$

which is of the same order as the bound claimed by Lemma 4.6.5. It remains to establish claim (4.51).

In order to do so, we employ an inductive argument by peeling the approximation error along one dimension at a time. As a first step, we have

$$\begin{aligned}
\|\theta^*\{\widehat{\pi}_1, \ldots, \widehat{\pi}_d\} - \theta^*\|_2 &= \|\theta^*\{\widehat{\pi}_1, \ldots, \widehat{\pi}_d\} - \theta^*\{\pi_1^*, \widehat{\pi}_2, \ldots, \widehat{\pi}_d\}\|_2 + \|\theta^*\{\pi_1^*, \widehat{\pi}_2, \ldots, \widehat{\pi}_d\} - \theta^*\|_2 \\
&\overset{\text{(ii)}}{=} \|\theta^*\{\widehat{\pi}_1, \mathsf{id}, \ldots, \mathsf{id}\} - \theta^*\{\pi_1^*, \mathsf{id}, \ldots, \mathsf{id}\}\|_2 + \|\theta^*\{\pi_1^*, \widehat{\pi}_2, \ldots, \widehat{\pi}_d\} - \theta^*\|_2,
\end{aligned} \tag{4.52}$$

where step (ii) follows by the unitary invariance of the ℓ_2-norm. If we write P_j for the squared error peeled along the j-th dimension with $P_1 = \|\theta^*\{\widehat{\pi}_1, \mathsf{id}, \ldots, \mathsf{id}\} - \theta^*\{\pi_1^*, \mathsf{id}, \ldots, \mathsf{id}\}\|_2^2$, then peeling the error along the remaining dimensions using an inductive argument yields the bound

$$\|\theta^*\{\widehat{\pi}_1, \ldots, \widehat{\pi}_d\} - \theta^*\|_2^2 \leq \left(\sum_{j=1}^{d} \sqrt{P_j} \right)^2 \leq 2d \cdot \sum_{j=1}^{d} P_j.$$

Thus, our strategy to establish claim (4.51) will be to establish the sufficient claim

$$P_j \leq (n_1 - k^*) \cdot \begin{cases} U \cdot T_j & \text{conditioned on } \mathcal{E}_1 \cap \mathcal{E}_2 \\ T_j^2 & \text{pointwise.} \end{cases} \tag{4.53}$$

We establish this claim for $j = 1$; the general proof is identical. Letting $\overline{n} := n_1^{d-1}$ and recalling our assumption $\pi_j^* = \mathsf{id}$ for all $j \in [d]$, we have

$$\begin{aligned} P_1 &= \|\theta^*\{\widehat{\pi}_1, \mathsf{id}, \ldots, \mathsf{id}\} - \theta^*\{\mathsf{id}, \mathsf{id}, \ldots, \mathsf{id}\}\|_2^2 \\ &= \sum_{i_1=1}^{n_1} \sum_{(i_2,\ldots,i_d) \in \mathbb{L}_{\overline{n},d-1}} \left(\theta^*(\widehat{\pi}_1(i_1), i_2, \ldots, i_d) - \theta^*(i_1, i_2, \ldots, i_d)\right)^2. \end{aligned}$$

We now split the proof into the two cases of equation (4.53).

Case 1: In this case, we condition on the event $\mathcal{E}_1 \cap \mathcal{E}_2$. By Claim 4.6.1, we know that conditioned on this event, we have $G_j' = G_j$. Consequently, for any $\widehat{\pi}_1 \in \mathcal{F}(\widehat{\mathsf{bl}}_1)$, we have $\widehat{\pi}_1(k) < \widehat{\pi}_1(\ell)$ if

$$\widehat{\tau}_1(\ell) - \widehat{\tau}_1(k) > 8\sqrt{\log n} \cdot n^{\frac{1}{2}(1-1/d)} \quad \text{or} \quad \max_{i_2,\ldots,i_d} Y(\ell, i_2, \ldots, i_d) - Y(k, i_2, \ldots, i_d) > 8\sqrt{\log n}. \tag{4.54}$$

As a consequence of the second condition, for any fixed tuple $(i_2, \ldots, i_d)$, we have $\widehat{\pi}_1(k) < \widehat{\pi}_1(\ell)$ whenever $Y(\ell, i_2, \ldots, i_d) - Y(k, i_2, \ldots, i_d) > 8\sqrt{\log n}$. Applying Lemma A.6.5 from the appendix with the substitution $a = \theta^*(\cdot, i_2, \ldots, i_d)$, $b = Y(\cdot, i_2, \ldots, i_d)$, and $\tau = 8\sqrt{\log n}$ yields the bound

$$|\theta^*(\widehat{\pi}_1(i_1), i_2, \ldots, i_d) - \theta^*(i_1, i_2, \ldots, i_d)| \leq 2\|Y - \theta^*\|_\infty + 8\sqrt{\log n} = U.$$

We may now apply Hölder's inequality to obtain

$$\begin{aligned} &\sum_{(i_2,\ldots,i_d) \in \mathbb{L}_{\overline{n},d-1}} \left(\theta^*(\widehat{\pi}_1(i_1), i_2, \ldots, i_d) - \theta^*(i_1, i_2, \ldots, i_d)\right)^2 \\ &\qquad \leq U \cdot \sum_{(i_2,\ldots,i_d) \in \mathbb{L}_{\overline{n},d-1}} |\theta^*(\widehat{\pi}_1(i_1), i_2, \ldots, i_d) - \theta^*(i_1, i_2, \ldots, i_d)| \\ &\qquad \overset{\text{(iii)}}{=} U \cdot \left| \sum_{(i_2,\ldots,i_d) \in \mathbb{L}_{\overline{n},d-1}} \theta^*(\widehat{\pi}_1(i_1), i_2, \ldots, i_d) - \theta^*(i_1, i_2, \ldots, i_d) \right| \\ &\qquad = U \cdot |\tau_1^*(\widehat{\pi}_1(i_1)) - \tau_1^*(i_1)|, \end{aligned} \tag{4.55}$$

where step (iii) follows since the set of scalars

$$\Big\{\theta^*(\widehat{\pi}_1(i_1), i_2, \ldots, i_d) - \theta^*(i_1, i_2, \ldots, i_d)\Big\}_{(i_2,\ldots,i_d)\in\mathbb{L}_{\overline{n},d-1}}$$

all have the same sign by our monotonicity assumption $\theta^* \in \mathcal{M}(\mathbb{L}_{d,n})$.

Let $\mathcal{I}$ denote the set containing all indices i_1 for which $|\tau_1^*(\widehat{\pi}_1(i_1)) - \tau_1^*(i_1)|$ is non-zero. Since there is an indifference set of size $k^1_{\max}$ along the first dimension, a non-zero value can only occur if either i_1 or $\widehat{\pi}_1(i_1)$ belong to the $n_1 - k^1_{\max}$ indices that are not in the largest indifference set. Consequently, we obtain $|\mathcal{I}| \leq 2(n_1 - k^1_{\max}) \leq 2(n_1 - k^*)$. Moreover, by our conditions (4.54), we have $\widehat{\pi}_1(k) < \widehat{\pi}_1(\ell)$ whenever $\widehat{\tau}_1(\ell) - \widehat{\tau}_1(k) > 8\sqrt{\log n} \cdot n^{\frac{1}{2}(1-1/d)}$. Therefore,

$$\begin{aligned}\sum_{i_1=1}^{n_1} |\tau_1^*(\widehat{\pi}_1(i_1)) - \tau_1^*(i_1)| &= \sum_{i_1\in\mathcal{I}} |\tau_1^*(\widehat{\pi}_1(i_1)) - \tau_1^*(i_1)| \\ &\overset{\text{(iv)}}{\leq} (n_1 - k^*) \cdot \Big(2\|\widehat{\tau}_1 - \tau_1^*\|_\infty + 8\sqrt{\log n} \cdot n^{\frac{1}{2}(1-1/d)}\Big) \\ &= (n_1 - k^*) \cdot T_1, \end{aligned} \tag{4.56}$$

where step (iv) follows by applying Lemma A.6.5 once again, but now to the scores. This completes the proof of the first case in equation (4.53).

<u>Case 2:</u> In this case, our goal is to prove a pointwise bound that holds unconditionally. In order to do so, we appeal to the properties of our algorithm: by construction, the edges of the graph G_j are always consistent with the conditions imposed by the pairwise statistics $\widehat{\Delta}_1^{\mathsf{sum}}$, so that for any $\widehat{\pi}_1 \in \mathcal{F}(\widehat{\mathsf{bl}}_1)$, we have $\widehat{\pi}_1(k) < \widehat{\pi}_1(\ell)$ if $\widehat{\tau}_\ell - \widehat{\tau}_k > 8\sqrt{\log n} \cdot n^{\frac{1}{2}(1-1/d)}$. We now repeat the reasoning from before to obtain the (crude) sequence of bounds

$$\begin{aligned}&\sum_{(i_2,\ldots,i_d)\in\mathbb{L}_{\overline{n},d-1}} \left(\theta^*(\widehat{\pi}_1(i_1), i_2, \ldots, i_d) - \theta^*(i_1, i_2, \ldots, i_d)\right)^2 \qquad (4.57)\\ &\leq \left(\sum_{(i_2,\ldots,i_d)\in\mathbb{L}_{\overline{n},d-1}} |\theta^*(\widehat{\pi}_1(i_1), i_2, \ldots, i_d) - \theta^*(i_1, i_2, \ldots, i_d)|\right)^2 \\ &= \left|\sum_{(i_2,\ldots,i_d)\in\mathbb{L}_{\overline{n},d-1}} \theta^*(\widehat{\pi}_1(i_1), i_2, \ldots, i_d) - \theta^*(i_1, i_2, \ldots, i_d)\right|^2 \\ &= |\tau_1^*(\widehat{\pi}_1(i_1)) - \tau_1^*(i_1)|^2.\end{aligned}$$

Proceeding exactly as before then yields the bound

$$\sum_{i_1=1}^{n_1} |\tau_1^*(\widehat{\pi}_1(i_1)) - \tau_1^*(i_1)|^2 \leq (n_1 - k^*) \cdot T_1^2, \tag{4.58}$$

which establishes the second case of equation (4.53). □

Bound on averaging error: In order to prove a bound on the averaging error, we first set up some notation and terminology to write the averaging error in a manner that is very similar to the permutation error bounded above. Then the proof follows from the arguments above.

First, note that the averaging operator can be equivalently implemented by sequentially averaging the entries along one dimension at a time. Let us make this precise with some ancillary definitions. Let $\widehat{\mathcal{A}}_j(\theta)$ denote the average of $\theta \in \mathbb{R}_{d,n}$ along dimension j according to the partition specified by $\widehat{\mathsf{bl}}_j$, i.e.,

$$\widehat{\mathcal{A}}_j(\theta)(i_1, \ldots, i_d) = \frac{1}{|\widehat{\mathsf{bl}}_j(i_j)|} \sum_{\ell \in \widehat{\mathsf{bl}}_j(i_j)} \theta(i_1, \ldots, i_{j-1}, \ell, i_{j+1}, \ldots, i_d) \qquad \text{for each } i_1, \ldots, i_d \in [n_1].$$

As a straightforward consequence of the linearity of the averaging operation, we have

$$\mathcal{A}(\theta; \widehat{\mathsf{bl}}_1, \ldots, \widehat{\mathsf{bl}}_d) = \widehat{\mathcal{A}}_1 \circ \cdots \circ \widehat{\mathcal{A}}_d(\theta) \qquad \text{for each } \; \theta \in \mathbb{R}_{d,n}.$$

Consequently, we may peel off the first dimension from the error of interest to write

$$\begin{aligned} \|\mathcal{A}(\theta^*; \widehat{\mathsf{bl}}_1, \ldots, \widehat{\mathsf{bl}}_d) - \theta^*\|_2 &\leq \|\widehat{\mathcal{A}}_1 \circ \cdots \circ \widehat{\mathcal{A}}_d(\theta^*) - \widehat{\mathcal{A}}_2 \circ \cdots \circ \widehat{\mathcal{A}}_d(\theta^*)\|_2 + \|\widehat{\mathcal{A}}_2 \circ \cdots \circ \widehat{\mathcal{A}}_d(\theta^*) - \theta^*\|_2 \\ &= \|\widehat{\mathcal{A}}_1(\widehat{\theta}_{2:d}) - \widehat{\theta}_{2:d}\|_2 + \|\widehat{\theta}_{2:d} - \theta^*\|_2, \end{aligned} \tag{4.59}$$

where we have let $\widehat{\theta}_{2:d} := \mathcal{A}_2 \circ \cdots \circ \mathcal{A}_d(\theta^*)$. Note the similarity between equations (4.59) and (4.52). Indeed, if we now write P'_j for the squared error peeled along the j-th dimension with $P'_1 = \|\widehat{\mathcal{A}}_1(\widehat{\theta}_{2:d}) - \widehat{\theta}_{2:d}\|_2^2$, then peeling the error along the remaining dimensions using an inductive argument, we obtain (exactly as before)

$$\|\mathcal{A}(\theta^*; \widehat{\mathsf{bl}}_1, \ldots, \widehat{\mathsf{bl}}_d) - \theta^*\|_2 \leq \left(\sum_{j=1}^{d} \sqrt{P'_j} \right)^2 \leq 2d \cdot \sum_{j=1}^{d} P'_j.$$

We now claim that with the random variables U and T_j defined exactly as before, we have the (identical) bound

$$P'_j \leq (n_1 - k^*) \cdot \begin{cases} U \cdot T_j & \text{conditioned on } \mathcal{E}_1 \cap \mathcal{E}_2 \\ T_j^2 & \text{pointwise,} \end{cases} \tag{4.60}$$

from which the proof of Lemma 4.6.5 for the averaging term follows identically.

Let us now establish the bound (4.60) for $j = 1$, for which we require some ancillary definitions. For an ordered partition $\mathsf{bl} = (S_1, \ldots, S_L)$ and index $i \in [n_1]$, recall the notation $\sigma_{\mathsf{bl}}(i)$ as the index ℓ of the set $S_\ell \ni i$. Let $\mathsf{bl}(i) = S_{\sigma_{\mathsf{bl}}(i)}$ denote the block containing index i. Let $\mathfrak{S}_V$ denote the set of all permutations on a set $V \subseteq [n_1]$, and let $\mathcal{J}(\mathsf{bl})$ denote the set of all permutations $\pi \in \mathfrak{S}_{n_1}$ such that $\pi(i) \in \mathsf{bl}(i)$ for all $\ell \in [n_1]$. Note that any permutation in the set $\mathcal{J}(\mathsf{bl})$ is given by compositions of individual permutations in $\mathfrak{S}_V$ for $V \in \mathsf{bl}$.

With this notation, we have for each $\theta \in \mathbb{R}_{d,n}$, the bound

$$\begin{aligned}
\|\widehat{\mathcal{A}}_1(\theta) - \theta\|_2^2 &= \sum_{i_2,\ldots,i_d} \sum_{i_1=1}^{n} \left(\frac{1}{|\widehat{\mathsf{bl}}_1(i_1)|} \sum_{\ell \in \widehat{\mathsf{bl}}_1(i_1)} \theta(\ell, i_2, \ldots, i_d) - \theta(i_1, \ldots, i_d) \right)^2 \\
&= \sum_{i_2,\ldots,i_d} \sum_{V \in \widehat{\mathsf{bl}}_1} \sum_{i_1 \in V} \left(\frac{1}{|V|} \sum_{\ell \in V} \theta(\ell, i_2, \ldots, i_d) - \theta(i_1, \ldots, i_d) \right)^2 \\
&\overset{\text{(i)}}{\leq} \sum_{i_2,\ldots,i_d} \sum_{V \in \widehat{\mathsf{bl}}_1} \max_{\pi' \in \mathfrak{S}_V} \sum_{i_1 \in V} \left(\theta(\pi'(i_1), i_2, \ldots, i_d) - \theta(i_1, \ldots, i_d) \right)^2 \\
&= \max_{\pi \in \mathcal{J}(\widehat{\mathsf{bl}}_1)} \sum_{i_1,\ldots,i_d} (\theta(\pi(i_1), i_2, \ldots, i_d) - \theta(i_1, i_2, \ldots, i_d))^2. \qquad (4.61)
\end{aligned}$$

Here, step (i) follows from Lemma A.6.7 in the appendix.

It is also useful to note that $\widehat{\theta}_{2:d}$ enjoys some additional structure. In particular, the estimate $\widehat{\theta}_{2:d}$ satisfies some properties that are straightforward to verify:

1. For any $\theta^* \in \mathbb{R}_{d,n}$, the slices along the first dimension have the same sum as θ^*, i.e., for each index $\ell \in [n_1]$, we have

$$\sum_{j=1}^{d} \sum_{i_j=1}^{n_j} \widehat{\theta}_{2:d}(i_1, \ldots, i_d) \cdot \mathbf{1}\{i_1 = \ell\} = \sum_{j=1}^{d} \sum_{i_j=1}^{n_j} \theta^*(i_1, \ldots, i_d) \cdot \mathbf{1}\{i_1 = \ell\} = \tau_1^*(\ell), \qquad (4.62a)$$

   where the final equality holds by definition (4.14b).

2. If $\theta^* \in \mathcal{M}(\mathbb{L}_{d,n})$, then its monotonicity property is preserved along the first dimension, i.e., for each pair of indices $1 \leq k \leq \ell \leq n_1$, we have

$$\widehat{\theta}_{2:d}(k, i_2, \ldots, i_d) \leq \widehat{\theta}_{2:d}(\ell, i_2, \ldots, i_d) \quad \text{for all } i_2, \ldots, i_d \in [n_1]. \qquad (4.62b)$$

With these properties in hand, we are now ready to establish the proof of claim (4.60). First, use equation (4.61) and let $\overline{n} := n_1^{d-1}$ to obtain the pointwise bound

$$P_1' \leq \max_{\pi_1 \in \mathcal{J}(\widehat{\mathsf{bl}}_1)} \sum_{i_1=1}^{n_1} \sum_{(i_2,\ldots,i_d) \in \mathbb{L}_{\overline{n},d-1}} \left(\widehat{\theta}_{2:d}(\pi_1(i_1), i_2, \ldots, i_d) - \widehat{\theta}_{2:d}(i_1, i_2, \ldots, i_d) \right)^2. \qquad (4.63)$$

We now establish the two cases of equation (4.60) separately.

<u>Case 1:</u> In this case, we condition on the event $\mathcal{E}_1 \cap \mathcal{E}_2$, in which case the estimated blocks obey the conditions (4.54); in particular, two indices k, ℓ are placed in the same block of $\widehat{\mathsf{bl}}_1$ iff

$$|\widehat{\tau}_1(k) - \widehat{\tau}_1(\ell)| \leq 8\sqrt{\log n} \cdot n^{\frac{1}{2}(1-1/d)} \text{ and } \max_{i_2,\ldots,i_d} |Y(k, i_2, \ldots, i_d) - Y(\ell, i_2, \ldots, i_d)| \leq 8\sqrt{\log n}. \qquad (4.64)$$

Since the averaging operation is ℓ_∞ contractive, we have, for each $\pi \in \mathcal{J}(\widehat{\mathsf{bl}}_1)$, the sequence of bounds

$$\begin{aligned}
&|\widehat{\theta}_{2:d}(\pi_1(i_1), i_2, \ldots, i_d) - \widehat{\theta}_{2:d}(i_1, i_2, \ldots, i_d)| \\
&\quad \leq |\theta^*(\pi_1(i_1), i_2, \ldots, i_d) - \theta^*(i_1, i_2, \ldots, i_d)| \\
&\quad \leq |\theta^*(\pi_1(i_1), i_2, \ldots, i_d) - Y(\pi_1(i_1), i_2, \ldots, i_d)| + |Y(i_1, i_2, \ldots, i_d) - \theta^*(i_1, i_2, \ldots, i_d)| \\
&\qquad\qquad + |Y(\pi_1(i_1), i_2, \ldots, i_d) - Y(i_1, i_2, \ldots, i_d)| \\
&\quad \overset{\text{(ii)}}{\leq} 2\|Y - \theta^*\|_\infty + 8\sqrt{\log n} = U,
\end{aligned}$$

where step (ii) follows from the second condition (4.64).

Thus, we have

$$\begin{aligned}
\frac{P_1'}{U} &\leq \max_{\pi_1 \in \mathcal{J}(\widehat{\mathsf{bl}}_1)} \sum_{i_1=1}^{n_1} \sum_{(i_2,\ldots,i_d) \in \mathbb{L}_{\overline{n},d-1}} \left| \widehat{\theta}_{2:d}(\pi_1(i_1), i_2, \ldots, i_d) - \widehat{\theta}_{2:d}(i_1, i_2, \ldots, i_d) \right| \\
&\overset{\text{(iii)}}{=} \max_{\pi_1 \in \mathcal{J}(\widehat{\mathsf{bl}}_1)} \sum_{i_1=1}^{n_1} \left| \sum_{(i_2,\ldots,i_d) \in \mathbb{L}_{\overline{n},d-1}} \widehat{\theta}_{2:d}(\pi_1(i_1), i_2, \ldots, i_d) - \widehat{\theta}_{2:d}(i_1, i_2, \ldots, i_d) \right| \\
&= \max_{\pi_1 \in \mathcal{J}(\widehat{\mathsf{bl}}_1)} \sum_{i_1=1}^{n_1} |\tau_1^*(\pi_1(i_1)) - \tau_1^*(i_1)|, \qquad (4.65)
\end{aligned}$$

where step (iii) follows by the monotonicity property (4.62b) and step (iv) from property (4.62a).

Now for each $\pi_1 \in \mathcal{J}(\widehat{\mathsf{bl}}_1)$, we have

$$\begin{aligned}
|\tau_1^*(\pi_1(i_1)) - \tau_1^*(i_1)| &\leq |\tau_1^*(\pi_1(i_1)) - \widehat{\tau}_1(\pi(i_1))| + |\widehat{\tau}_1(i_1) - \tau_1^*(i_1)| + |\widehat{\tau}_1(\pi_1(i_1)) - \widehat{\tau}_1(i_1)| \\
&\overset{\text{(v)}}{\leq} 2\|\widehat{\tau}_1 - \tau_1^*\|_\infty + 8\sqrt{\log n} \cdot n^{\frac{1}{2}(1-1/d)} = T_1,
\end{aligned}$$

where step (v) is guaranteed by the first condition (4.64). Since there are at most $2(n_1 - k^*)$ indices in the sum (4.65) that are non-zero, putting together the pieces yields the bound

$$P_1' \leq U \cdot (n_1 - k^*) \cdot T_1,$$

and this completes the proof of the first case of equation (4.60).

Case 2: In this case, our goal is to establish a pointwise bound unconditionally. Once again, by construction, the estimated ordered partitions are always consistent with the pairwise statistics $\widehat{\Delta}_j^{\mathsf{sum}}$, so that two indices k, ℓ are placed within the same block of $\widehat{\mathsf{bl}}_1$ iff

$$|\widehat{\tau}_1(k) - \widehat{\tau}_1(\ell)| \leq 8\sqrt{\log n} \cdot n^{\frac{1}{2}(1-1/d)}. \qquad (4.66)$$

Consequently, proceeding from equation (4.63) and using the same properties as before, we have

$$\begin{aligned}
P_1' &\leq \max_{\pi_1 \in \mathcal{J}(\widehat{\mathsf{bl}}_1)} \sum_{i_1=1}^{n_1} \left(\sum_{(i_2,\ldots,i_d)\in\mathbb{L}_{\overline{n},d-1}} |\widehat{\theta}_{2:d}(\pi_1(i_1), i_2, \ldots, i_d) - \widehat{\theta}_{2:d}(i_1, i_2, \ldots, i_d)| \right)^2 \\
&= \max_{\pi_1 \in \mathcal{J}(\widehat{\mathsf{bl}}_1)} \sum_{i_1=1}^{n_1} \left| \sum_{(i_2,\ldots,i_d)\in\mathbb{L}_{\overline{n},d-1}} \widehat{\theta}_{2:d}(\pi_1(i_1), i_2, \ldots, i_d) - \widehat{\theta}_{2:d}(i_1, i_2, \ldots, i_d) \right|^2 \\
&= \max_{\pi_1 \in \mathcal{J}(\widehat{\mathsf{bl}}_1)} \sum_{i_1=1}^{n_1} |\tau_1^*(\pi_1(i_1)) - \tau_1^*(i_1)|^2.
\end{aligned}$$

Identically to before, for each $\pi_1 \in \mathcal{J}(\widehat{\mathsf{bl}}_1)$, we have

$$\begin{aligned}
|\tau_1^*(\pi_1(i_1)) - \tau_1^*(i_1)| &\leq |\tau_1^*(\pi_1(i_1)) - \widehat{\tau}_1(\pi(i_1))| + |\widehat{\tau}_1(i_1) - \tau_1^*(i_1)| + |\widehat{\tau}_1(\pi_1(i_1)) - \widehat{\tau}_1(i_1)| \\
&\leq 2\|\widehat{\tau}_1 - \tau_1^*\|_\infty + 8\sqrt{\log n} \cdot n^{\frac{1}{2}(1-1/d)} = T_1,
\end{aligned}$$

where the second inequality is guaranteed by condition (4.66). Since there are at most $2(n_1 - k^*)$ indices in the sum (4.65) that are non-zero, putting together the pieces yields the bound

$$P_1' \leq (n_1 - k^*) \cdot T_1^2,$$

and this completes the proof of the second case of equation (4.60). □

4.6.7 Proof of Proposition 4.5.1

At the heart of the proposition lies the following lemma, which bounds the ℓ_2 error as a sum of approximation and estimation errors.

Lemma 4.6.6. *There is a universal positive constant C such that for all $\theta^* \in \mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n}) \cap \mathbb{B}_\infty(1)$, we have*

$$\mathcal{R}_n(\widehat{\theta}_{\mathsf{BC}}, \theta^*) \leq C \left(n^{-1/d} \log^{5/2} n + \frac{d}{n} \sum_{j=1}^{d} \mathbb{E}\left[\|\widehat{\tau}_j - \tau_j^*\|_1\right] \right).$$

Taking this lemma as given for the moment, the proof of the proposition is straightforward. The random variable $\widehat{\tau}_j(k) - \tau_j^*(k)$ is the sum of $n^{1-1/d}$ independent standard Gaussians, so that

$$\mathbb{E}[|\widehat{\tau}_j(k) - \tau_j^*(k)|] = \sqrt{\frac{2}{\pi}} \cdot n^{\frac{1}{2}(1-1/d)} \text{ for each } k \in [n_1],\ j \in [d].$$

Summing over both $k \in [n_1]$ and $j \in [d]$ and normalizing, we have

$$\frac{d}{n} \sum_{j=1}^{d} \mathbb{E}\left[\|\widehat{\tau}_j - \tau_j^*\|_1\right] \leq C d^2 n^{\frac{1}{2}(1/d-1)}.$$

□

It remains to prove Lemma 4.6.6.

Proof of Lemma 4.6.6

In order to lighten notation in this section, we use the convenient shorthand $\widehat{\theta} \equiv \widehat{\theta}_{\mathsf{BC}}$ and $\widehat{\pi}_j \equiv \widehat{\pi}_j^{\mathsf{BC}}$ for each $j \in [d]$. Assume without loss of generality that $\pi_1^* = \cdots = \pi_d^* = \mathsf{id}$, so that $\theta^* \in \mathcal{M}(\mathbb{L}_{d,n})$. Let $\widetilde{\theta}$ denote the projection of the tensor $\theta^*\{\widehat{\pi}_1, \ldots, \widehat{\pi}_d\}$ onto the set $\mathcal{M}(\mathbb{L}_{d,n}; \widehat{\pi}_1, \ldots, \widehat{\pi}_d)$. With this setup, we have

$$\begin{aligned} \|\widehat{\theta} - \theta^*\|_2 &\leq \|\widehat{\theta} - \widetilde{\theta}\|_2 + \|\widetilde{\theta} - \theta^*\{\widehat{\pi}_1, \ldots, \widehat{\pi}_d\}\|_2 + \|\theta^*\{\widehat{\pi}_1, \ldots, \widehat{\pi}_d\} - \theta^*\|_2 \\ &\overset{\text{(i)}}{\leq} \|\theta^* + \epsilon - (\theta^*\{\widehat{\pi}_1, \ldots, \widehat{\pi}_d\} + \epsilon)\|_2 + \|\widetilde{\theta} - \theta^*\{\widehat{\pi}_1, \ldots, \widehat{\pi}_d\}\|_2 + \|\theta^*\{\widehat{\pi}_1, \ldots, \widehat{\pi}_d\} - \theta^*\|_2 \\ &= \underbrace{\|\widetilde{\theta} - \theta^*\{\widehat{\pi}_1, \ldots, \widehat{\pi}_d\}\|_2}_{\text{estimation error}} + \underbrace{2 \cdot \|\theta^*\{\widehat{\pi}_1, \ldots, \widehat{\pi}_d\} - \theta^*\|_2}_{\text{approximation error}}. \qquad (4.67) \end{aligned}$$

Here, step (i) follows since an ℓ_2-projection onto a convex set is always non-expansive (4.20b). We now bound the estimation and approximation error terms separately.

Bounding the estimation error: The key difficulty here is that the estimated permutations $\widehat{\pi}_1, \ldots, \widehat{\pi}_d$ *depend* on the noise tensor ϵ. Similarly to before, we handle this dependence by establishing a uniform result that holds simultaneously over all choices of permutations. In particular, letting $\widehat{\theta}_{\pi_1,\ldots,\pi_d}$ denote the ℓ_2 projection of the tensor $\theta^*\{\pi_1, \ldots, \pi_d\} + \epsilon$ onto the set $\mathcal{M}(\mathbb{L}_{d,n}; \widehat{\pi}_1, \ldots, \widehat{\pi}_d)$, we claim that for each $\theta^* \in \mathcal{M}(\mathbb{L}_{n,d}) \cap \mathbb{B}_\infty(1)$, we have

$$\mathbb{E}\left[\max_{1 \leq j \leq d} \max_{\pi_j \in \mathfrak{S}_{n_j}} \|\widehat{\theta}_{\pi_1,\ldots,\pi_d} - \theta^*\{\pi_1, \ldots, \pi_d\}\|_2^2\right] \leq C n^{1-1/d} \log^{5/2} n. \qquad (4.68)$$

Since $\widetilde{\theta} = \widehat{\theta}_{\widehat{\pi}_1,\ldots,\widehat{\pi}_d}$, equation (4.68) provides a bound on the estimation error that is of the claimed order.

Let us now prove claim (4.68). For each fixed tuple of permutations $(\pi_1, \ldots, \pi_d)$, combining Corollary 4.3.1(b) with Lemma A.6.4 yields the tail bound

$$\Pr\left\{\|\widehat{\theta}_{\pi_1,\ldots,\pi_d} - \theta^*\{\pi_1, \ldots, \pi_d\}\|_2^2 \geq C \cdot n^{1-1/d} \log^{5/2} n + 2u\right\} \leq \exp\{-u\} \text{ for each } u \geq 0.$$

Taking a union bound over all $\prod_{j=1}^d n_j! \leq \exp(dn_1 \log n_1) = \exp(n_1 \log n)$ permutations and setting $u = Cn_1 \log n + u'$ for a sufficiently large constant C, we obtain

$$\Pr\left\{\max_{1 \leq j \leq d} \max_{\pi_j \in \mathfrak{S}_{n_j}} \|\widehat{\theta}_{\pi_1,\ldots,\pi_d} - \theta^*\{\pi_1, \ldots, \pi_d\}\|_2^2 \geq C(n^{1-1/d} \log^{5/2} n + n_1 \log n) + u'\right\} \leq e^{-cu}. \qquad (4.69)$$

Finally, note that for each $d \geq 2$, we have $n_1 \leq n^{1-1/d}$ and integrate the tail bound (4.69) to complete the proof of claim (4.68).

Bounding the approximation error: Our bound on the approximation error proceeds very similarly to before, so we sketch the key differences. First, we have the decomposition

$$\|\theta^*\{\widehat{\pi}_1, \ldots, \widehat{\pi}_d\} - \theta^*\|_2 \le \|\theta^*\{\widehat{\pi}_1, \mathsf{id}, \ldots, \mathsf{id}\} - \theta^*\{\pi_1^*, \mathsf{id}, \ldots, \mathsf{id}\}\|_2 + \|\theta^*\{\pi_1^*, \widehat{\pi}_2, \ldots, \widehat{\pi}_d\} - \theta^*\|_2.$$

But since $\theta^* \in \mathbb{B}_\infty(1)$, now each scalar $\theta^*(\widehat{\pi}_1(i_1), i_2, \ldots, i_d) - \theta^*(\pi_1^*(i_1), i_2, \ldots, i_d)$ is bounded in the range $[-2, 2]$. Letting $\overline{n} := n_1^{d-1}$, and proceeding exactly as in equations (4.52)–(4.55), we have

$$\sum_{(i_2, \ldots, i_d) \in \mathbb{L}_{\overline{n}, d-1}} (\theta^*(\widehat{\pi}_1(i_1), i_2, \ldots, i_d) - \theta^*(\pi_1^*(i_1), i_2, \ldots, i_d))^2 \le 2|\tau_1^*(\widehat{\pi}_1(i_1)) - \tau_1^*(i_1)|.$$

Combining this inequality with the outer sum, we have

$$\begin{aligned}\sum_{i_1=1}^{n_1} |\tau_1^*(\widehat{\pi}_1(i_1)) - \tau_1^*(i_1)| &\le \sum_{i_1=1}^{n_1} |\widehat{\tau}_1(\widehat{\pi}_1(i_1)) - \tau_1^*(i_1)| + |\widehat{\tau}_1(\widehat{\pi}_1(i_1)) - \tau_1^*(\widehat{\pi}_1(i_1))| \\ &\overset{\text{(ii)}}{\le} \sum_{i_1=1}^{n_1} |\widehat{\tau}_1(i_1) - \tau_1^*(i_1)| + |\widehat{\tau}_1(\widehat{\pi}_1(i_1)) - \tau_1^*(\widehat{\pi}_1(i_1))| \\ &= 2\|\widehat{\tau}_1 - \tau_1^*\|_1\end{aligned}$$

where step (ii) follows from the rearrangement inequality for the ℓ_1 norm [320], since $\widehat{\tau}_1$ and τ_1^* are sorted in increasing order along the permutations $\widehat{\pi}_1$ and $\pi_1^* = \mathsf{id}$, respectively. Putting together the pieces, we have shown that

$$\|\theta^*\{\widehat{\pi}_1, \ldots, \widehat{\pi}_d\} - \theta^*\|_2 \le \sqrt{4\|\widehat{\tau}_1 - \tau_1^*\|_1} + \|\theta^*\{\pi_1^*, \widehat{\pi}_2, \ldots, \widehat{\pi}_d\} - \theta^*\|_2.$$

Proceeding inductively, we have

$$\|\theta^*\{\widehat{\pi}_1, \ldots, \widehat{\pi}_d\} - \theta^*\|_2^2 \le \left(\sum_{j=1}^{d} \sqrt{4\|\widehat{\tau}_j - \tau_j^*\|_1} \right)^2 \le 4d \sum_{j=1}^{d} \|\widehat{\tau}_j - \tau_j^*\|_1,$$

and this provides a bound on the approximation error that is of the claimed order. □

4.6.8 Proof of Proposition 4.5.2

We handle the case where the projection in Definition 4.5.1 is onto unbounded tensors; the bounded case follows identically. For each tuple of permutations $\pi_1, \ldots, \pi_d$, define the estimator

$$\widehat{\theta}_{\pi_1, \ldots, \pi_d} = \operatorname*{argmin}_{\theta \in \mathcal{M}(\mathbb{L}_{d,n}; \pi_1, \ldots, \pi_d)} \|Y - \theta\|_2.$$

By definition, any permutation-projection based estimator must equal $\widehat{\theta}_{\pi_1, \ldots, \pi_d}$ for some choice of permutations $\pi_1, \ldots, \pi_d \in \mathfrak{S}_{n_1}$. Our strategy will thus be to lower bound the risk $\min_{\pi_1, \ldots, \pi_d} \|\widehat{\theta}_{\pi_1, \ldots, \pi_d} -$

$\theta^*\|_2^2$ for a particular choice of θ^*. To that end, let us analyze the risk of the individual estimators around the point $\theta^* = 0$. Define the positive scalar t_0 via

$$\begin{aligned} t_0 &:= \operatorname*{argmax}_{t \geq 0} \left\{ \mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d,n}; \pi_1, \ldots, \pi_d) \cap \mathbb{B}_2(t)} \langle \epsilon, \theta \rangle - t^2/2 \right\} \\ &= \operatorname*{argmax}_{t \geq 0} \left\{ \mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d,n}) \cap \mathbb{B}_2(t)} \langle \epsilon, \theta \rangle - t^2/2 \right\} \\ &= \mathbb{E} \sup_{\theta \in \mathcal{M}(\mathbb{L}_{d,n}) \cap \mathbb{B}_2(1)} \langle \epsilon, \theta \rangle, \end{aligned}$$

where the final equality follows by a rescaling argument.

Applying [58, Theorem 1.1] yields that for each tuple $\pi_1, \ldots, \pi_d$, we have

$$\Pr\left\{ \left| \|\widehat{\theta}_{\pi_1, \ldots, \pi_d} - \theta^*\|_2 - t_0 \right| \geq u\sqrt{t_0} \right\} \leq 3 \exp\left(-\frac{u^4}{32(1 + u/\sqrt{t_0})^2} \right) \quad \text{for each } u \geq 0. \tag{4.70}$$

Furthermore, applying [137, Proposition 5] yields the lower bound

$$t_0 \geq c_d \cdot n^{1/2 - 1/d} \qquad \text{for each } d \geq 3. \tag{4.71}$$

Substituting the value $u = \sqrt{t_0}/2$ into the bound (4.70) and using the lower bound (4.71) on t_0 yields, for each fixed tuple of permutations $\pi_1, \ldots, \pi_d$, the high probability bound

$$\Pr\left\{ \|\widehat{\theta}_{\pi_1, \ldots, \pi_d} - \theta^*\|_2 \leq c_d \cdot n^{1/2 - 1/d} \right\} \leq 3 \exp\left\{ -c_d' \cdot n^{1 - 2/d} \right\},$$

where the pair (c_d, c_d') are different constants that depends on d alone. Applying a union bound over all choices of permutations now yields

$$\Pr\left\{ \min_{\pi_1, \ldots, \pi_d} \|\widehat{\theta}_{\pi_1, \ldots, \pi_d} - \theta^*\|_2^2 \leq c_d \cdot n^{1 - 2/d} \right\} \leq 3 \exp\left\{ -c_d' \cdot n^{1 - 2/d} + dn_1 \log n_1 \right\}.$$

Now for each $d \geq 4$, there is a large enough constant $C_d > 0$ depending on d alone such that if $n \geq C_d$, then $c_d' \cdot n^{1 - 2/d} \geq 2dn_1 \log n_1$. Consequently, if $n \geq C_d$, then

$$\Pr\left\{ \min_{\pi_1, \ldots, \pi_d} \|\widehat{\theta}_{\pi_1, \ldots, \pi_d} - \theta^*\|_2^2 \geq c_d \cdot n^{1 - 2/d} \right\} \geq 1/2 \quad \text{and} \quad \mathbb{E}\left[\min_{\pi_1, \ldots, \pi_d} \|\widehat{\theta}_{\pi_1, \ldots, \pi_d} - \theta^*\|_2^2 \right] \geq c_d \cdot n^{1 - 2/d}$$

for a sufficiently small constant $c_d > 0$ depending only on d. Thus, any permutation-projection based estimator $\widehat{\theta}$ must satisfy

$$\Pr\left\{ \|\widehat{\theta} - \theta^*\|_2^2 \geq c_d \cdot n^{1 - 2/d} \right\} \geq 1/2 \quad \text{and} \quad \mathbb{E}\left[\|\widehat{\theta} - \theta^*\|_2^2 \right] \geq c_d \cdot n^{1 - 2/d}.$$

On the other hand, we have $\theta^* \in \mathcal{M}_{\textsf{perm}}^{\mathbb{k}_0, \mathrm{s}_0}(\mathbb{L}_{d,n})$ with $\mathrm{s}_0 = (1, \ldots, 1)$ and $\mathbb{k}_0 = ((n_1), \ldots, (n_1))$. Thus, $s(\theta^*) = 1$ and $k^*(\theta^*) = n_1$, and Proposition 4.3.1 yields the upper bound $\mathfrak{M}_{d,n}(\mathbb{k}_0, \mathrm{s}_0) \lesssim 1/n$. Combining the pieces and noting that $\delta_n = (10n)^{-1} \leq 1/2$, the adaptivity index of any permutation-projection based estimator $\widehat{\theta}$ must satisfy

$$\mathfrak{A}(\widehat{\theta}; \delta_n) \geq \mathfrak{A}^{\mathbb{k}_0, \mathrm{s}_0}(\widehat{\theta}; \delta_n) \geq c_d \cdot n^{1 - 2/d} \quad \text{and} \quad \mathfrak{A}(\widehat{\theta}) \geq \mathfrak{A}^{\mathbb{k}_0, \mathrm{s}_0}(\widehat{\theta}) \geq c_d \cdot n^{1 - 2/d}. \tag{4.72}$$

□

4.7 Summary and open questions

We considered the problem of estimating a multivariate isotonic regression function on the lattice from noisy observations that were also permuted along each coordinate, and established several results surrounding statistics, computation, and adaptation for this class of models. First, we showed that unlike in the bivariate case considered in Chapter 3, computationally efficient estimators are able to achieve the minimax lower bound for estimation of bounded tensors in this class. Second, when the tensor is also structured, in that it is piecewise constant on a d-dimensional partition with a small number of blocks, we showed that the fundamental limits of adaptation are still nonparametric. Third, by appealing to the hypergraph planted clique conjecture, we also showed that the adaptivity index of polynomial time estimators is significantly poorer than that of their inefficient counterparts. The second and third phenomena are both significantly different from the case without unknown permutations. Fourth, we introduced a novel procedure that was simultaneously optimal both in worst-case risk and adaptation, while also being computable in sub-quadratic time. Our results for this algorithm are particularly surprising given that a large class of natural estimators does not exhibit fast adaptation in the multivariate case. Finally, we also established some risk bounds and structural properties (see Appendix A.5) for natural isotonic regression estimators without unknown permutations.

Our work raises many interesting questions from both the modeling and theoretical standpoints. From a modeling perspective, the isotonic regression model with unknown permutations should be viewed as just a particular nonparametric model for tensor data. There are many ways one may extend these models. For instance, taking a linear combination of $k > 1$ tensors in the set $\mathcal{M}_{\mathsf{perm}}(\mathbb{L}_{d,n})$ directly generalizes the class of nonnegative tensors of (canonical polyadic) rank k. Studying such models would parallel a similar investigation that was conducted in the case $d = 2$ for matrix estimation [276]. It would also be interesting to incorporate latent permutations within other multidimensional nonparametric function estimation tasks that are not shape constrained; as noted in Chapter 3, a similar study has been carried out in the case $d = 2$ for graphon estimation [113].

Methodological and theoretical questions also abound. First, note that in typical applications, the tensor dimension n_1 will be very large, and we will only observe a subset of entries chosen at random. Indeed, when $d = 2$, Chapter 3 shows that the fundamental limits of the problem exhibit an intricate dependence on the probability of observing each entry and the dimensions of the tensor. What are the analogs of these results when $d \geq 3$? The second question concerns adaptation. Our focus on indifference sets to define structure in the tensor was motivated by the application to multiway comparisons, but other structures are also interesting to study. For instance, what does a characterization of adaptation look like when there is simply a partition into hyper-rectangles—not necessarily Cartesian products of one-dimensional partitions—on which the tensor is piecewise constant? Such structure has been extensively studied in the isotonic regression literature [56, 83, 137]. What about cases where θ^* is a nonnegative tensor of rank 1? It would be worth studying spectral methods for tensor estimation for this problem, especially in the latter case.

Having studied the statistics, computation, and adaptation of permutation-based models, we now turn to the class of index models in the next part of the thesis. Our questions will once again be inspired by this broad perspective.

Part II

Index models

The surprising effectiveness of alternating projections

Chapter 5

Tractable algorithms for max-affine regression

As we alluded to in Chapter 1, this portion of the thesis will deal with index models. This particular chapter is focused on dimensionality reduction for problems of the convex regression type. These problems and further applications are introduced below.

5.1 Introduction

Max-affine regression refers to the regression model

$$Y = \max_{1 \leq j \leq k} \left(\langle X,\, \theta_j^* \rangle + b_j^* \right) + \epsilon \tag{5.1}$$

where Y is a univariate response, X is a d-dimensional vector of covariates and ϵ models zero-mean noise that is independent of X. We assume that $k \geq 1$ is a known integer and study the problem of estimating the unknown parameters $\theta_1^*, \ldots, \theta_k^* \in \mathbb{R}^d$ and $b_1^*, \ldots, b_k^* \in \mathbb{R}$ from independent observations $(x_1, y_1), \ldots, (x_n, y_n)$ drawn according to the model (5.1). Furthermore, we assume for concreteness[1] in this chapter that the covariate distribution is standard Gaussian, with $x_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_d)$.

Let us provide some motivation for studying the model (5.1). When $k = 1$, equation (5.1) corresponds to the classical linear regression model. When $k = 2$, the intercepts $b_2^* = b_1^* = 0$, and $\theta_2^* = -\theta_1^* = \theta^*$, the model (5.1) reduces to

$$Y = |\langle X,\, \theta^* \rangle| + \epsilon. \tag{5.2}$$

The problem of recovering θ^* from observations drawn according to the above model is known as (real) phase retrieval—variants of which arise in a diverse array of science and engineering

[1] Our companion paper [118] weakens distributional assumptions on the covariates, but this requires significantly more technical effort.

applications [101, 106, 141]—and has associated with it an extensive statistical and algorithmic literature.

To motivate the model (5.1) for general k, note that the function $x \mapsto \max_{1 \le j \le k}(\langle x,\, \theta_j^* \rangle + b_j^*)$ is always a convex function and, thus, estimation under the model (5.1) can be used to fit convex functions to the observed data. Indeed, the model (5.1) serves as a parametric approximation to the nonparametric convex regression model

$$Y = \phi^*(X) + \epsilon, \tag{5.3}$$

where $\phi^* : \mathbb{R}^d \to \mathbb{R}$ is an unknown convex function. It is well-known that convex regression suffers from the curse of dimensionality unless d is small, which is basically a consequence of the fact that the metric entropy of natural totally bounded sub-classes of convex functions grows exponentially in d (see, e.g., [46, 115, 131]). To overcome this curse of dimensionality, one would need to work with more structured sub-classes of convex functions. Since convex functions can be approximated to arbitrary accuracy by maxima of affine functions, it is reasonable to regularize the problem by considering only those convex functions that can be written as a maximum of a fixed number of affine functions. Constraining the number of affine pieces in the function therefore presents a simple method to enforce structure, and such function classes have been introduced and studied in the convex regression literature (see e.g., [138]). This assumption directly leads to our model (5.1), and it has been argued by [15, 139, 208] that the parametric model (5.1) is a tractable alternative to the full nonparametric convex regression model (5.3) in common applications of convex regression to data arising in economics, finance and operations research where d is often moderate to large.

Another motivation for the model (5.1) comes from the problem of estimating convex sets from support function measurements. The support function of a compact convex set $K \subseteq \mathbb{R}^d$ is defined by $h_K(x) := \sup_{u \in K} \langle x,\, u \rangle$ for d-dimensional unit vectors x. The problem of estimating an unknown compact, convex set K^* from noisy measurements of $h_{K^*}(\cdot)$ arises in certain engineering applications such as robotic tactile sensing and projection magnetic resonance imaging (see, e.g., [116, 125, 257]). Specifically, the model considered here is

$$Y = h_{K^*}(X) + \epsilon,$$

and the goal is to estimate the set $K^* \subseteq \mathbb{R}^d$. As in convex regression, this problem suffers from a curse of dimensionality unless d is small, as is evident from known minimax lower bounds [129]. To alleviate this curse, it is natural to restrict K^* to the class of all polytopes with at most k extreme points for a fixed k; such a restriction has been studied as a special case of enforcing structure in these problems by Soh and Chandrasekharan [287]. Under this restriction, one is led to the model (5.1) with $b_1^* = \cdots = b_k^* = 0$, since if K^* is the polytope given by the convex hull of $\theta_1^*, \ldots, \theta_k^* \in \mathbb{R}^d$, then its support function is equal to $x \mapsto \max_{1 \le j \le k} \langle x,\, \theta_j^* \rangle$.

Equipped with these motivating examples, our goal is to study a computationally efficient estimation methodology for the unknown parameters of the model (5.1) from i.i.d observations $(x_i, y_i)_{i=1}^n$. Before presenting our contributions, let us first rewrite the observation model (5.1) by using more convenient notation, and use it to describe existing estimation procedures for this model. Denote the unknown parameters by $\beta_j^* := (\theta_j^*, b_j^*) \in \mathbb{R}^{d+1}$ for $j = 1, \ldots, k$ and the observations by

(ξ_i, y_i) for $i = 1, \ldots, n$, where $\xi_i := (x_i, 1) \in \mathbb{R}^{d+1}$. In this notation, the observation model takes the form

$$y_i = \max_{1 \leq j \leq k} \langle \xi_i,\, \beta_j^* \rangle + \epsilon_i, \qquad \text{for } i = 1, 2, \ldots, n. \tag{5.4}$$

Throughout the paper, we assume that in addition to the covariates being i.i.d. standard Gaussian, the noise variables $\epsilon_1, \ldots, \epsilon_n$ are independent random variables drawn from a (univariate) distribution that is zero-mean and sub-Gaussian, with unknown sub-Gaussian parameter σ.

Let us now describe existing estimation procedures for max-affine regression. The most obvious approach is the global least squares estimator, defined as any minimizer of the least squares criterion

$$(\widehat{\beta}_1^{(\mathsf{ls})}, \ldots, \widehat{\beta}_k^{(\mathsf{ls})}) \in \operatorname*{argmin}_{\beta_1, \ldots, \beta_k \in \mathbb{R}^{d+1}} \sum_{i=1}^{n} \left(y_i - \max_{1 \leq j \leq k} \langle \xi_i,\, \beta_j \rangle \right)^2. \tag{5.5}$$

It is easy to see (see the full paper [119]) that a global minimizer of the least squares criterion above always exists but it will not—at least in general—be unique, since any relabeling of the indices of a minimizer will also be a minimizer. While the least squares estimator has appealing statistical properties (see, e.g. [129, 287, 310]), the optimization problem (5.5) is non-convex and, in general, NP-hard [100]. It is interesting to compare (5.5) to the optimization problem used to compute the least squares estimator in the more general convex regression model (5.3), given by

$$\widehat{\phi}^{(\mathsf{ls})} \in \operatorname*{argmin}_{\phi} \sum_{i=1}^{n} (y_i - \phi(x_i))^2, \tag{5.6}$$

where the minimization is over all convex functions ϕ. In sharp contrast to the problem (5.5), the optimization problem (5.6) is convex [198, 272] and can be solved efficiently for fairly large values of the pair (d, n) [216]. Unfortunately however, the utility of $\widehat{\phi}^{(\mathsf{ls})}$ in estimating the parameters of the max-affine model is debatable, as it is unclear how one may obtain estimates of the true parameters $\beta_1^*, \ldots, \beta_k^*$ from $\widehat{\phi}^{(\mathsf{ls})}$, which typically will *not* be a maximum of only k affine functions[2]. It is also worth mentioning that while the convex LSE is known to adapt to piece-wise linear structure when $d = 1$ [132], it was recently shown by Kur et al. [180] that parametric adaptation *cannot* occur for any $d \geq 5$. This provides further justification for studying the more explicit max-affine model (5.1) in even moderate-dimensional problems.

Three heuristic techniques for solving the non-convex optimization problem (5.5) were empirically evaluated by Balázs [15, Chapters 6 and 7], who compared running times and performance of these techniques on a wide variety of real and synthetic datasets for convex regression. The first technique is the alternating minimization algorithm of Magnani and Boyd [208], the second technique is the convex adaptive partitioning (or CAP) algorithm of Hannah and Dunson [139], and the third is the adaptive max-affine partitioning algorithm proposed by Balázs himself [15]. The simplest and most intuitive of these three methods is the first alternating minimization (AM) algorithm, which is an iterative algorithm for estimating the parameters $\beta_1^*, \ldots, \beta_k^*$ and forms the focus of our study.

[2] Notably, the convex LSE $\widehat{\phi}^{(\mathsf{ls})}$ is also the maximum of (at most n) affine functions.

In the t-th iteration of the algorithm, the current estimates $\beta_1^{(t)}, \ldots, \beta_k^{(t)}$ are used to partition the observation indices $1, \ldots, n$ into k sets $S_1^{(t)}, \ldots, S_k^{(t)}$ such that $j \in \operatorname{argmax}_{u \in [k]} \langle \xi_i, \beta_u^{(t)} \rangle$ for every $i \in S_j^{(t)}$. For each $1 \leq j \leq k$, the next estimate $\beta_j^{(t+1)}$ is then obtained by performing a least squares fit (or equivalently, linear regression) to the data $(\xi_i, y_i), i \in S_j^{(t)}$. More intuition and a formal description of the algorithm are provided in Section 5.2. Balázs found that when this algorithm was run on a variety of datasets with multiple random initializations, it compared favorably with the state of the art in terms of its final predictive performance—see, for example, Figures 7.4 and 7.5 in the thesis [15], which show encouraging results when the algorithm is used to fit convex functions to datasets of average wages and aircraft profile drag data, respectively. In the context of fitting convex sets to support function measurements, Soh and Chandrasekaran [287] recently proposed and empirically evaluated a similar algorithm in the case of isotropic covariates. However, to the best of our knowledge, no theoretical results exist to support the performance of such a technique.

In this chapter, we present a theoretical analysis of the AM algorithm for recovering the parameters of the max-affine regression model when the covariate distribution is Gaussian[3]. This assumption forms a natural starting point for the study of many iterative algorithms in related problems [14, 236, 327, 349], and is also quite standard in theoretical investigations of multidimensional regression problems. Note that the AM algorithm described above can be seen as a generalization of classical AM algorithms for (real) phase retrieval [102, 117], which have recently been theoretically analyzed in a series of papers [236, 327, 349] for Gaussian designs. The AM—and the closely related expectation maximization[4], or EM—methodology is widely used for parameter estimation in missing data problems [22, 142] and mixture models [336], including those with covariates such as mixtures-of-experts [158] and mixtures-of-regressions [52] models. Theoretical guarantees for such algorithms have been established in multiple statistical contexts [14, 68, 300, 333]; in the case when the likelihood is not unimodal, these are typically of the local convergence type. In particular, algorithms of the EM type return, for many such latent variable models, minimax-optimal parameter estimates when initialized in a neighborhood of the optimal solution (e.g., [52, 350, 351]); conversely, these algorithms can get stuck at spurious fixed points when initialized at random [155]. In some specific applications of EM to mixtures of two Gaussians [79, 335] and mixtures of two regressions [181], however, it has been shown that randomly initializing the EM algorithm suffices in order to obtain consistent parameter estimates. Here, we establish guarantees on the AM algorithm for max-affine regression that are of the former type: we prove local geometric convergence of the AM iterates when initialized in a neighborhood of the optimal solution. We analyze the practical variant of the algorithm in which the steps are performed without sample-splitting. As in the case of mixture models [52, 148], we use spectral methods to obtain such an initialization. In order to keep the narrative of this thesis coherent, our guarantees for this initialization step have been omitted, and can be found in the archival version of the paper [119].

[3] In our companion paper [118], we weaken this assumption on the covariate distribution.

[4] Indeed, for many problems, the EM algorithm reduces to AM in the noiseless limit, and AM should thus be viewed as a variant of EM that uses hard-thresholding to determine values of the latent variables.

Contributions Let us now describe our results in more detail. To simplify the exposition, we state simplified corollaries of our results; precise statements are presented shortly. We prove in Theorem 5.3.1 that for each $\epsilon > 0$, the parameter estimates $\beta_1^{(t)}, \dots, \beta_k^{(t)}$ returned by the AM algorithm at iteration t satisfy, with high probability, the inequality

$$\sum_{j=1}^{k} \|\beta_j^{(t)} - \beta_j^*\|^2 \le \epsilon + C(\beta_1^*, \dots, \beta_k^*) \frac{\sigma^2 kd}{n} \log(kd) \log\left(\frac{n}{kd}\right) \tag{5.7}$$

for every $t \ge \log_{4/3}\left(\frac{\sum_{j=1}^{k} \|\beta_j^{(0)} - \beta_j^*\|^2}{\epsilon}\right)$, provided that the sample size n is sufficiently large and that the initial estimates satisfy the condition

$$\min_{c>0} \max_{1 \le j \le k} \|c\beta_j^{(0)} - \beta_j^*\|^2 \le \frac{1}{k} c(\beta_1^*, \dots, \beta_k^*). \tag{5.8}$$

Here $C(\beta_1^*, \dots, \beta_k^*)$ and $c(\beta_1^*, \dots, \beta_k^*)$ are constants depending only on the true parameters $\beta_1^*, \dots, \beta_k^*$, and their explicit values are given in Theorem 5.3.1. The constant c in equation (5.8) endows the initialization with a scale-invariance property: indeed, scaling all parameters $\beta_1^{(0)}, \dots, \beta_k^{(0)}$ by the same positive constant c produces the same initial partition of subsets $S_1^{(0)}, \dots, S_k^{(0)}$, from which the algorithm proceeds identically.

Treating k as a fixed constant, inequality (5.7) implies, under the initialization condition (5.8), that the parameter estimates returned by AM converge geometrically to within a small ball of the true parameters, and that this error term is nearly the parametric risk $\frac{\sigma^2 d}{n}$ up to a logarithmic factor. The initialization condition (5.8) requires the distance between the initial estimates and the true parameters to be at most a specific (k-dependent) constant. It has been empirically observed that there exist bad initializations under which the AM algorithm behaves poorly (see, e.g., [15, 208]) and assumption (5.8) is one way to rule these out.

A natural question based on our Theorem 5.3.1 is whether it is possible to produce preliminary estimates $\beta_1^{(0)}, \dots, \beta_k^{(0)}$ satisfying the initialization condition (5.8). Indeed, one such method is to repeatedly initialize parameters (uniformly) at random within the unit ball $\mathbb{B}^{d+1}$; Balázs empirically observed in a close relative of such a scheme (see Figure 6.6 in his thesis [15]) that increasing the number of random initializations is often sufficient to get the AM algorithm to succeed. However, reasoning heuristically, the number of repetitions required to ensure that one such random initialization generates parameters that satisfy condition (5.8) increases exponentially in the ambient dimension d, and so it is reasonable to ask if, in large dimensions, there is some natural form of dimensionality reduction that allows us to perform this step in a lower-dimensional space.

When[5] $k < d$, we show that a natural spectral method (described formally in Algorithm 2) is able to reduce the dimensionality of our problem from d to k. In particular, this method returns an orthonormal basis of vectors $\widehat{U}_1, \dots, \widehat{U}_k$ such that the k-dimensional linear subspace spanned

[5] If $k \ge d$, then this dimensionality reduction step can be done away with and one can implement the random search routine directly.

by these vectors accurately estimates the subspace spanned by the vectors $\theta_1^*, \ldots, \theta_k^*$. We form the matrix $\widehat{U} := [\widehat{U}_1 : \cdots : \widehat{U}_k]$ by collecting these vectors as its columns, and in order to account for the intercepts, further append such a matrix to form the matrix $\widehat{V} := \begin{bmatrix} \widehat{U} & 0 \\ 0 & 1 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(k+1)}$. We then choose M random initializations in $(k+1)$ dimensions—the ℓ-th such initialization is given by a set of vectors $\nu_1^\ell, \ldots, \nu_k^\ell \in \mathbb{R}^{k+1}$ each chosen uniformly at random from the $(k+1)$-dimensional unit ball—so that the collection of k vectors $\{\widehat{V}\nu_j^\ell\}_{j=1}^k$ serves as our ℓ-th guess of the true parameters. In order to decide which of these random points to choose, we evaluate (on an independent set of samples) the goodness-of-fit statistic $\min_{c\geq 0} \sum_i (y_i - c \max_{1\leq j\leq k} \langle \xi_i, \widehat{V}\nu_j^\ell \rangle)^2$ for each $1 \leq \ell \leq M$, where the minimization over the constant c accounts for the scale-invariance property alluded to above. Letting ℓ^* denote the index with the smallest loss, we then return the initialization $\beta_j^{(0)} = \widehat{V}\nu_j^{\ell^*}$ for $j = 1, \ldots, k$.

Our algorithm can thus be viewed as a variant of the repeated random initialization evaluated by Balázs [15], but incurs significantly smaller computational cost, since we only run the full-blown iterative AM algorithm once. Note that our algorithm treats the number of initializations M as a tuning parameter to be chosen by the statistician, similar to Balázs [15], but we show a concrete upper bound on M that is sufficient to guarantee convergence. In particular, we show that in order to produce an initialization satisfying condition (5.8) with high probability, it suffices to choose M as a function *only* of the number of affine pieces k and other geometric parameters of the problem[6].

From a technical standpoint, our results for the AM algorithm are significantly more challenging to establish than related results in the literature [14, 279, 327, 334]. First, it is technically very challenging to compute the *population operator* [14]—corresponding to running the AM update in the infinite sample limit—in this setting, since the $\max$ function introduces intricate geometry in the problem that is difficult to reason about in closed form. Second, we are interested in analyzing the AM update without sample-splitting, and so cannot assume that the iterates are independent of the covariates; the latter assumption has been used fruitfully in the literature to simplify analyses of such algorithms [236, 334, 349]. Third, and unlike algorithms for phase retrieval [279, 327], our algorithm performs least squares using sub-matrices of the covariate matrix that are chosen *depending* on our random iterates. Accordingly, a key technical difficulty of the proof, which may be of independent interest, is to control the spectrum of these random matrices, rows of which are drawn from (randomly) truncated variants of the Gaussian distribution.

Chapter-specific notation: Recall the notational convention introduced in Section 1.4. We complement this notation with a few other definitions that are used solely in this chapter and the corresponding technical proof section in Appendix B.1. For a pair of vectors (u, v), we let $u \otimes v := uv^\top$ denote their outer product. Let $\lambda_i(\Gamma)$ denote the i-th largest eigenvalue of a symmetric matrix Γ.

[6] While we omit our theoretical results for the initialization step in this thesis, these can be found in the paper [119].

5.2 Background and problem formulation

In this section, we formally introduce the geometric parameters underlying the max-affine regression model, as well as the methodology we use to perform parameter estimation.

5.2.1 Model and Geometric Parameters

We work throughout with the observation model defined in equation (5.4); recall that our covariates are drawn i.i.d. from a standard Gaussian distribution, and that our noise is σ-sub-Gaussian. We let $X \in \mathbb{R}^{n \times d}$ denote the covariate matrix with row i given by the vector x_i, and collect the responses in a vector $y \in \mathbb{R}^n$.

Recall that $\xi_i = (x_i,\ 1) \in \mathbb{R}^{d+1}$ for each $i \in [n]$; the matrix of appended covariates $\Xi \in \mathbb{R}^{n \times (d+1)}$ is defined by appending a vector of ones to the right of the matrix X. Our primary goal is to use the data (X, y)—or equivalently, the pair (Ξ, y)—to estimate the underlying parameters $\{\beta^*_j\}_{j=1}^k$.

An important consideration in achieving such a goal is the "effective" sample size with which we observe the parameter β^*_j. Toward that end, for $X \sim \mathcal{N}(0, I_d)$, let

$$\pi_j(\beta^*_1, \ldots, \beta^*_k) := \Pr\left\{ \langle X,\, \theta^*_j \rangle + b^*_j = \max_{j' \in [k]} \ \left(\langle X,\, \theta^*_{j'} \rangle + b^*_{j'} \right) \right\} \tag{5.9}$$

denote the probability with which the j-th parameter $\beta^*_j = (\theta^*_j \ \ b^*_j)$ attains the maximum. Note that the event on which more than one of the parameters attains the maximum has measure zero, except in the case where $\beta^*_i = \beta^*_j$ for some $i \neq j$. We explicitly disallow this case and assume that the parameters $\beta^*_1, \ldots, \beta^*_k$ are distinct. Let

$$\pi_{\min}(\beta^*_1, \ldots, \beta^*_k) := \min_{j \in [k]} \pi_j(\beta^*_1, \ldots, \beta^*_k), \tag{5.10}$$

and assume that we have $\pi_{\min}(\beta^*_1, \ldots, \beta^*_k) > 0$; in other words, we ignore vacuous cases in which some parameter is never observed. Roughly speaking, the sample size of the parameter that is observed most rarely is given by $\min_{j \in [k]} \pi_j n \sim n \cdot \pi_{\min}(\beta^*_1, \ldots, \beta^*_k)$, and so the error in estimating this parameter should naturally depend on $\pi_{\min}(\beta^*_1, \ldots, \beta^*_k)$. By definition, we always have $\pi_{\min}(\beta^*_1, \ldots, \beta^*_k) \leq 1/k$.

Since we are interested in performing parameter estimation under the max-affine regression model, a few geometric quantities also appear in our bounds, and serve as natural notions of "signal strength" and "condition number" of the estimation problem. The signal strength is given by the minimum separation

$$\Delta(\beta^*_1, \ldots, \beta^*_k) = \min_{j, j' : j \neq j'} \left\| \theta^*_j - \theta^*_{j'} \right\|^2;$$

we also assume that Δ is strictly positive, since otherwise, a particular parameter is never observed. To denote a natural form of conditioning, define the quantities

$$\kappa_j(\beta^*_1, \ldots, \beta^*_k) = \frac{\max_{j' \neq j} \left\| \theta^*_j - \theta^*_{j'} \right\|^2}{\min_{j' \neq j} \left\| \theta^*_j - \theta^*_{j'} \right\|^2}, \qquad \text{with} \qquad \kappa(\beta^*_1, \ldots, \beta^*_k) = \max_{j \in [k]} \kappa_j(\beta^*_1, \ldots, \beta^*_k).$$

Finally, let $\mathsf{B}_{\mathsf{max}}(\beta_1^*, \ldots, \beta_k^*) := \max_{j \in [k]} \|\beta_j^*\|$ denote the maximum norm of any unknown parameter. We often use the shorthand

$$\begin{aligned} \pi_{\min} &= \pi_{\min}(\beta_1^*, \ldots, \beta_k^*), & & & \Delta &= \Delta(\beta_1^*, \ldots, \beta_k^*), \\ \kappa &= \kappa(\beta_1^*, \ldots, \beta_k^*), & &\text{and} & \mathsf{B}_{\mathsf{max}} &= \mathsf{B}_{\mathsf{max}}(\beta_1^*, \ldots, \beta_k^*) \end{aligned}$$

when the true parameters $\beta_1^*, \ldots, \beta_k^*$ are clear from context.

5.2.2 Methodology

As discussed in the introduction, the most natural estimation procedure from i.i.d. samples $(\xi_i, y_i)_{i=1}^n$ of the model (5.4) is the least squares estimator (5.5). The appendix of our full paper [119] establishes that this estimator always exists. Note, however, that it will not be unique in general since any relabeling of a minimizer is also a minimizer.

In spite of the fact that the least squares estimator always exists, the problem (5.5) is non-convex and NP-hard in general. The AM algorithm presents a tractable approach towards solving it in the statistical setting that we consider.

Alternating Minimization

We now formally describe the AM algorithm proposed by Magnani and Boyd [208]. For each $\beta_1, \ldots, \beta_k$, define the sets

$$S_j(\beta_1, \ldots, \beta_k) := \left\{ i \in [n] : j = \min \operatorname*{argmax}_{1 \le u \le k} \left(\langle \xi_i, \, \beta_u \rangle \right) \right\} \tag{5.11}$$

for $j = 1, \ldots, k$. In words, the set $S_j(\beta_1, \ldots, \beta_k)$ contains the indices of samples on which parameter β_j attains the maximum; in the case of a tie, samples having multiple parameters attaining the maximum are assigned to the set with the smallest corresponding index (i.e., ties are broken in the lexicographic order[7]). Thus, the sets $\{S_j(\beta_1, \ldots, \beta_k)\}_{j=1}^k$ define a partition of $[n]$. The AM algorithm employs an iterative scheme where one first constructs the partition $S_j\left(\beta_1^{(t)}, \ldots, \beta_k^{(t)}\right)$ based on the current iterates $\beta_1^{(t)}, \ldots, \beta_k^{(t)}$ and then calculates the next parameter estimate $\beta_j^{(t+1)}$ by a least squares fit to the dataset $\{(\xi_i, y_i), i \in S_j(\beta_1^{(t)}, \ldots, \beta_k^{(t)})\}$. The algorithm (also described below as Algorithm 1) is, clearly, quite intuitive and presents a natural approach to solving (5.5).

As a sanity check, we show in the full paper [119] that the global least squares estimator (5.5) is a fixed-point of this iterative scheme under a mild technical assumption.

We also note that the AM algorithm was proposed by Soh [286] in the context of estimating structured convex sets from support function measurements. It should be viewed as a generalization of a classical algorithm for (real) phase retrieval due to Fienup [102], which has been more recently

[7] In principle, it is sufficient to define the sets $S_j(\beta_1, \ldots, \beta_k), j \in [k]$ as any partition of $[n]$ having the property that $\langle \xi_i, \, \beta_j \rangle = \max_{u \in [k]} \langle \xi_i, \, \beta_u \rangle$ for every $j \in [k]$ and $i \in S_j(\beta_1, \ldots, \beta_k)$; here “any” means that ties can be broken according to an arbitrary rule, and we have chosen this rule to be the lexicographic order in equation (5.11).

Algorithm 1: Alternating minimization for estimating maximum of k affine functions

```
Input: Data $\{\xi_i, y_i\}_{i=1}^n$; initial parameter estimates $\beta_1^{(0)}, \ldots, \beta_k^{(0)}$; number of iterations $T$.
Output: Final estimator of parameters $\widehat{\beta}_1, \ldots, \widehat{\beta}_k$.
Initialize $t \leftarrow 0$.
repeat
    Compute maximizing index sets
        $\mathsf{S}_j^{(t)} = S_j(\beta_1^{(t)}, \ldots, \beta_k^{(t)}),$    (5.12a)
      for each $j \in [k]$, according to equation (5.11).
    Update
        $\beta_j^{(t+1)} \in \operatorname*{argmin}_{\beta \in \mathbb{R}^{d+1}} \sum_{i \in \mathsf{S}_j^{(t)}} \left(y_i - \langle \xi_i, \beta \rangle\right)^2,$    (5.12b)
      for each $j \in [k]$.
until $t = T$;
Return $\widehat{\beta}_j = \beta_j^{(T)}$ for each $j \in [k]$.
```

analyzed in a series of papers [236, 327] for Gaussian designs. While some analyses of AM algorithms assume sample-splitting across iterations (e.g. [236, 334, 349]), we consider the more practical variant of AM run without sample-splitting, since the update (5.12a)-(5.12b) is run on the full data (Ξ, y) in every iteration.

Initialization

The alternating minimization algorithm described above requires an initialization. While the algorithm was proposed to be run from a random initialization with restarts [208, 287], we propose to initialize the algorithm from parameter estimates that are sufficiently close to the optimal parameters. This is similar to multiple procedures to solve non-convex optimization problems in statistical settings (e.g., [14]), that are based on iterative algorithms that exhibit *local* convergence to the unknown parameters. Such algorithms are typically initialized by using a moment method, which (under various covariate assumptions) returns useful parameter estimates.

Our approach to the initialization problem is similar, in that we combine a moment method with random search in a lower-dimensional space. For convenience of analysis, we split the n samples into two equal parts—assume that n is even without loss of generality—and perform each of the above steps on different samples so as to maintain independence between the two steps. The formal algorithm is presented in two parts as Algorithms 2 and 3.

In related problems [52, 271, 334, 350], a combination of a second order and third order method (involving tensor decomposition) is employed to obtain parameter estimates in one shot. Take

```
Algorithm 2: PCA for $k$-dimensional subspace initialization
  Input: Data $\{\xi_i, y_i\}_{i=1}^n$.
  Output: Matrix $\widehat{U} \in \mathbb{R}^{d \times k}$ having orthonormal columns that (approximately) span the $k$
          dimensional subspace spanned by the vectors $\theta_1^*, \ldots, \theta_k^*$.
2 Compute the quantities
      $$\widehat{M_1} = \frac{2}{n} \sum_{i=1}^{n/2} y_i x_i \qquad \text{and} \qquad \widehat{M_2} = \frac{2}{n} \sum_{i=1}^{n/2} y_i \left( x_i x_i^\top - I_d \right), \qquad (5.13)$$
    and let $\widehat{M} = \widehat{M_1} \otimes \widehat{M_1} + \widehat{M_2}$; here, $I_d$ denotes the $d \times d$ identity matrix and $\otimes$ denotes the
    outer product.
4 Perform the eigendecomposition $\widehat{M} = \widehat{P}\widehat{\Lambda}\widehat{P}^\top$, and use the first $k$ columns of $\widehat{P}$
    (corresponding to the $k$ largest eigenvalues) to form the matrix $\widehat{U} \in \mathbb{R}^{d \times k}$. Return $\widehat{U}$.
```

the problem of learning generalized linear models [271] as an example; here, the analysis of the moment method relies on the link function being (at least) three times differentiable so that the population moment quantities can be explicitly computed. After showing that these expectations are closed form functions of the unknown parameters, matrix/tensor perturbation tools are then applied to show that the empirical moments concentrate about their population counterparts. However, in our setting, the $\max$ function is not differentiable, and so it is not clear that higher order moments return reasonable estimates even in expectation since Stein's lemma (on which many of these results rely) is not applicable[8] in this setting. Nevertheless, we show that the second order moment returns a k-dimensional subspace that is close to the true span of the parameters $\{\theta_j^*\}_{j=1}^k$; the degree of closeness depends only on the geometric properties of these parameters.

Let us also briefly discuss Algorithm 3, which corresponds to performing random search in $(k+1) \cdot k$ dimensional space to obtain the final initialization. In addition to the random initialization employed in step 1 of this algorithm, we also use the mean squared error on a holdout set (corresponding to samples $n/2+1$ through n) to select the final parameter estimates. In particular, we evaluate the error in a scale-invariant fashion; the computation of the optimal constant c in step 2 of the algorithm can be performed in closed form for each fixed index ℓ, since for a pair of vectors (u, v) having equal dimension, we have

$$\operatorname*{argmin}_{c \geq 0} \|u - cv\|^2 = \max \left\{ \frac{\langle u, \, v \rangle}{\|v\|^2}, 0 \right\}.$$

A key parameter that governs the performance of our search procedure is the number of initializations M; we show in the sequel that it suffices to take M to be a quantity that depends only on the number of affine pieces k, and on other geometric parameters in the problem.

[8] A natural workaround is to use Stein's lemma on the infinitely differentiable "softmax" surrogate function, but this approach also does not work for various technical reasons.

```
Algorithm 3: Low-dimensional random search
    Input: Data $\{\xi_i, y_i\}_{i=1}^n$, subspace estimate $\widehat{U} \in \mathbb{R}^{d \times k}$ having orthonormal columns that
           (approximately) span the $k$ dimensional subspace spanned by the vectors $\theta_1^*, \ldots, \theta_k^*$,
           and number of random initializations $M \in \mathbb{N}$.
    Output: Initial estimator of parameters $\beta_1^{(0)}, \ldots, \beta_k^{(0)}$.
  2 Choose $M \cdot k$ random points $\nu_j^\ell$ i.i.d. for $\ell \in M$ and $j \in [k]$, each uniformly from the
    $(k+1)$-dimensional unit ball $\mathbb{B}^{k+1}$. Let
        $$\widehat{V} = \begin{bmatrix} \widehat{U} & 0 \\ 0 & 1 \end{bmatrix}$$
    be a matrix in $\mathbb{R}^{(d+1) \times (k+1)}$ having orthonormal columns.
  4 Compute the index
        $$\ell^* \in \operatorname*{argmin}_{\ell \in [M]} \frac{2}{n} \left\{ \min_{c \geq 0} \sum_{i=n/2+1}^{n} \left( y_i - c \max_{j \in [k]} \langle \xi_i, \widehat{V} \nu_j^\ell \rangle \right)^2 \right\}.$$
  6 Return the $(d+1)$-dimensional parameters
        $$\beta_j^{(0)} = \widehat{V} \nu_j^{\ell^*} \qquad \text{for each } j \in [k].$$
```

Our overall algorithm should be viewed as a slight variation of the AM algorithm with random restarts. It inherits similar empirical performance (see the full paper [119]), while significantly reducing the computational cost, since operations are now performed in ambient dimension $k+1$, and the iterative AM algorithm is run only once overall. It also produces *provable* parameter estimates, and as we show in the sequel, the number of random initializations M can be set independently of the pair (n, d). Having stated the necessary background and described our methodology, we now proceed to statements and discussions of our main results. We focus on the AM algorithm in this thesis; results for the initialization step are also quite involved, and can be found in the paper [119].

5.3 Local geometric convergence of alternating minimization

We now establish local convergence results for the AM algorithm. Recall the definition of the parameters $(\pi_{\min}, \Delta, \kappa)$ introduced in Section 5.2, and the assumption that the covariates $\{x_i\}_{i=1}^n$ are drawn i.i.d. from the standard Gaussian distribution $\mathcal{N}(0, I_d)$. Throughout the paper, we assume that the true parameters $\beta_1^*, \ldots, \beta_k^*$ are fixed.

Theorem 5.3.1. *There exists a tuple of universal constants* (c_1, c_2) *such that if the sample size satisfies the bound*

$$n \geq c_1 \max\left\{d, 10 \log n\right\} \max\left\{\frac{k\kappa}{\pi_{\min}^3}, \; \frac{\log^2(1/\pi_{\min})}{\pi_{\min}^3}, \log(n/d), \; \sigma^2 \frac{k^3}{\Delta \pi_{\min}^9} \log(k/\pi_{\min}^3) \log(n/d)\right\},$$

then for all initializations $\beta_1^{(0)}, \ldots, \beta_k^{(0)}$ *satisfying the bound*

$$\min_{c>0} \max_{1 \leq j \neq j' \leq k} \frac{\left\| c\left(\beta_j^{(0)} - \beta_{j'}^{(0)}\right) - \left(\beta_j^* - \beta_{j'}^*\right)\right\|}{\|\theta_j^* - \theta_{j'}^*\|} \leq c_2 \frac{\pi_{\min}^3}{k\kappa} \log^{-3/2}\left(\frac{k\kappa}{\pi_{\min}^3}\right), \tag{5.14a}$$

the estimation error at all iterations $t \geq 1$ *is simultaneously bounded as*

$$\sum_{j=1}^k \|\beta_j^{(t)} - \beta_j^*\|^2 \leq \left(\frac{3}{4}\right)^t \left(\sum_{j=1}^k \|c^* \beta_j^{(0)} - \beta_j^*\|^2\right) + c_1 \sigma^2 \frac{kd}{\pi_{\min}^3 n} \log(kd) \log(n/kd) \tag{5.14b}$$

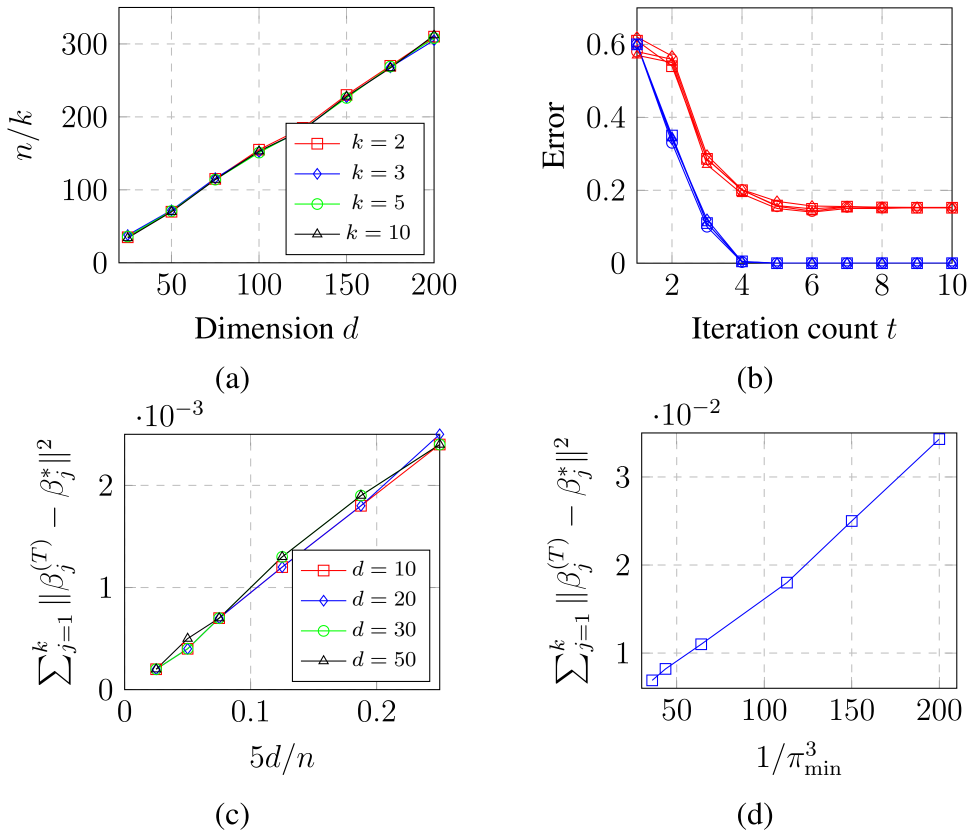
with probability exceeding $1 - c_2\left(k \exp\left(-c_1 n \frac{\pi_{\min}^4}{\log^2(1/\pi_{\min})}\right) + \frac{k^2}{n^7}\right)$. *Here, the positive scalar* c^* *minimizes the LHS of inequality* (5.14a).

Let us interpret the various facets of Theorem 5.3.1. As mentioned before, it is a local convergence result, which requires the initialization $\beta_1^{(0)}, \ldots, \beta_k^{(0)}$ to satisfy condition (5.14a). In the well-balanced case (with $\pi_{\min} \sim 1/k$) and treating k as a fixed constant, the initialization condition (5.14a) posits that the parameters are a constant "distance" from the true parameters. Notably, closeness is measured in a relative sense, and between pairwise differences of the parameter estimates as opposed to the parameters themselves; the intuition for this is that the initialization $\beta_1^{(0)}, \ldots, \beta_k^{(0)}$ induces the initial partition of samples $S_1(\beta_1^{(0)}, \ldots, \beta_k^{(0)}), \ldots, S_1(\beta_1^{(0)}, \ldots, \beta_k^{(0)})$, whose closeness to the true partition depends only on the relative pairwise differences between parameters, and is also invariant to a global scaling of the parameters. It is also worth noting that local geometric convergence of the AM algorithm is guaranteed uniformly from *all* initializations satisfying condition (5.14a). In particular, the initialization parameters are not additionally required to be independent of the covariates or noise, and this allows us to use the same n samples for initialization of the parameters.

Let us now turn our attention to the bound (5.14b), which consists of two terms. In the limit $t \to \infty$, the final parameters provide an estimate of the true parameters that is accurate to within the second term of the bound (5.14b). Up to a constant, this is the statistical error term

$$\delta_{n,\sigma}(d, k, \pi_{\min}) = \sigma^2 \frac{kd}{\pi_{\min}^3 n} \log(kd) \log(n/kd) \tag{5.15}$$

that converges to 0 as $n \to \infty$, thereby providing a consistent estimate in the large sample limit. Notice that the dependence of $\delta_{n,\sigma}(d, k, \pi_{\min})$ on the tuple (σ, d, n) is minimax-optimal up to the logarithmic factor $\log(n/d)$, since a matching lower bound can be proved for the linear regression



(a)

(b)

(c)

(d)

Figure 5.1: Convergence of the AM with Gaussian covariates— in panel (a), we plot the noiseless sample complexity of AM; we fix $\|\beta_i^*\| = 1$ for all $i \in [k]$, $\sigma = 0$ and $\pi_{\min} = 1/k$. We say β_i^* is recovered if $\left\|\beta_i^{(t)} - \beta_i^*\right\| \leq 0.01$. For a fixed dimension d, we run a linear search on the number of samples n, such that the empirical probability of success over 100 trials is more than 0.95, and output the least such n. In panel (b), we plot the optimization error (in blue) $\sum_{j=1}^{k} \|\beta_j^{(t)} - \beta_j^{(T)}\|^2$ and the deviation from the true parameters (in red) $\sum_{j=1}^{k} \left\|\beta_j^{(t)} - \beta_j^*\right\|^2 / \sigma^2$ over iterations t for different σ $(0.15, 0.25, 0.4, 0.5)$, with $k = 5$, $d = 100$, $T = 50$ and $n = 5d$, and averaged over 50 trials. Panel (c) shows that the estimation error at $T = 50$ scales at the parametric rate d/n, where we have chosen a fixed $k = 5$ and $\sigma = 0.25$. Panel (d) shows the variation of this error as a function of $\pi_{\min}$ where we fix $k = 3, d = 2, n = 10^3, \sigma = 0.4$.

problem when $k = 1$. In the paper [119], we also show a parametric lower bound on the minimax estimation error for general k, of the order $\sigma^2 kd/n$. Panel (c) of Figure 5.1 verifies in a simulation that the statistical error depends linearly on d/n. The dependence of the statistical error on the pair $(k, \pi_{\min})$ is more involved, and we do not yet know if these are optimal. As discussed before, a linear dependence of $\pi_{\min}$ is immediate from a sample-size argument; the cubic dependence arises because the sub-matrices of Ξ chosen over the course of the algorithm are not always well-conditioned, and their condition number scales (at most) as $\pi_{\min}^2$. The full version also contains a low-dimensional example (with $d = 2$ and $k = 3$) in which the least squares estimator incurs a parameter estimation error of the order $\frac{1}{\pi_{\min}^3 n}$ even when provided with the *true* partition of covariates $\{S_j(\beta_1^*, \dots, \beta_k^*)\}_{j=1}^3$. While this does not constitute an information theoretic lower bound, it provides strong evidence to suggest that our dependence on $\pi_{\min}$ is optimal at least when viewed in isolation. We verify this intuition via simulation: in panel (d) of Figure 5.1, we observe that on this example, the error of the final AM iterate varies linearly with the quantity $1/\pi_{\min}^3$.

The first term of the bound (5.14b) is an optimization error that is best interpreted in the noiseless case $\sigma = 0$, wherein the parameters $\beta_1^{(t)}, \dots, \beta_k^{(t)}$ converge at a geometric rate to the true parameters $\beta_1^*, \dots, \beta_k^*$, as verified in panel (a) of Figure 5.1. In particular, in the noiseless case, we obtain exact recovery of the parameters provided $n \geq C \frac{kd}{\pi_{\min}^3} \log(n/d)$. Thus, the "sample complexity" of parameter recovery is linear in the dimension d, which is optimal (panel (a) verifies this fact). In the well-balanced case, the dependence on k is quartic, but lower bounds based on parameter counting suggest that the true dependence ought to be linear. Again, we are not aware of whether the dependence on $\pi_{\min}$ in the noiseless case is optimal; our simulations shown in panel (a) suggests that the sample complexity depends inversely on $\pi_{\min}$, and so closing this gap is an interesting open problem. When $\sigma > 0$, we have an overall sample size requirement

$$n \geq c \max\left\{ \frac{kd\kappa}{\pi_{\min}^3}, \; \frac{d \log^2(1/\pi_{\min})}{\pi_{\min}^3}, d\log(n/d), \; \sigma^2 \frac{k^3 d}{\Delta \pi_{\min}^9} \log(k/\pi_{\min}^3) \log(n/d) \right\} := n_{\mathsf{AM}}(c). \tag{5.16}$$

As a final remark, note that Theorem 5.3.1 holds under Gaussian covariates and when the true parameters $\beta_1^*, \dots, \beta_k^*$ are fixed independently of the covariates. In the companion paper([..]), it is shown that both of these features of the result can be relaxed, i.e., AM converges geometrically even under a milder covariate assumption, and this convergence occurs *for all* true parameters that are geometrically similar.

5.3.1 Proof ideas and technical challenges

Let us first sketch, at a high level, the ideas required to establish guarantees on the AM algorithm. We need to control the iterates of the AM algorithm without sample-splitting across iterations, and so the iterates themselves are random and depend on the sequence of random variables $(\xi_i, \epsilon_i)_{i=1}^n$. A popular and recent approach to handling this issue in related iterative algorithms (e.g., [14]) goes through two steps: first, the population update, corresponding to running (5.12a)-(5.12b) in the case $n \to \infty$, is analyzed, after which the random iterates in the finite-sample case are shown to be close to their (non-random) population counterparts by using concentration bounds for the associated empirical process. The main challenge in our setting is that the population update is quite non-trivial to write down since it involves a delicate understanding of the geometry of the covariate distribution induced by the maxima of affine functions. We thus resort to handling the random iterates directly, thereby sidestepping the calculation of the population operator entirely.

Broadly speaking, we analyze the update (5.12a)-(5.12b) by relating the error of the parameters generated by this update to the error of the parameters from which the update is run. This involves three distinct technical steps. The first step (handled by Lemma 5.4.1) is to control the behavior of the noise in the problem. In order to do so, we apply standard concentration bounds for quadratic forms of sub-Gaussian random variables, in conjunction with bounds on the *growth functions* of multi-class classifiers [75]. Crucially, this affords a uniform bound on the noise irrespective of which iterate the alternating minimization update is run from. The second step corresponds (roughly) to controlling the prediction error in the noiseless problem, for which we show a quantitative result (in Lemma 5.4.2) that strictly generalize a result of Waldspurger [327]. Finally, in order to translate a prediction error guarantee into a guarantee on the estimation error, we invert specifically chosen sub-matrices of the covariate matrix Ξ over the course of the algorithm, and our bounds naturally depend on how these sub-matrices are conditioned. A key technical difficulty of the proof is therefore to control the spectrum of these random matrices, rows of which are drawn from (randomly) truncated variants of the Gaussian distribution. The expectation of such a random matrix can be characterized by appealing to tail bounds on the non-central χ^2 distribution, and the Gaussian covariate assumption additionally allows us to show that an analogous result holds for the random matrix with high probability (see Lemma 5.4.3). Here, our initialization condition is crucial: the aforementioned singular value control suffices for the sub-matrices formed by the *true* parameters, and we translate these bounds to the sub-matrices generated by random parameters by appealing to the fact that the initialization is sufficiently close to the truth.

5.4 Proof of main theorem

Let us begin by introducing some shorthand notation, and providing a formal statement of the probability bound guaranteed by the theorem. For a scalar w^*, vectors $u^* \in \mathbb{R}^d$ and $v^* = (u^*,\ w^*) \in \mathbb{R}^{d+1}$, and a positive scalar r, let $\mathcal{B}_{v^*}(r) = \left\{ v \in \mathbb{R}^{d+1} : \frac{\|v - v^*\|}{\|u^*\|} \leq r \right\}$, and let

$$\mathcal{I}\left(r; \{\beta_j^*\}_{j=1}^k\right) = \left\{ \beta_1, \ldots \beta_k \in \mathbb{R}^{d+1} : \exists c > 0 : c(\beta_i - \beta_j) \in \mathcal{B}_{\beta_i^* - \beta_j^*}(r) \text{ for all } 1 \leq i \neq j \leq k \right\}.$$

Also, use the shorthand

$$\vartheta_t\left(r; \{\beta_j^*\}_{j=1}^k\right) := \sup_{\beta_1^{(0)}, \ldots, \beta_k^{(0)} \in \mathcal{I}(r)} \sum_{j=1}^k \|\beta_j^{(t)} - \beta_j^*\|^2 - \left(\frac{3}{4}\right)^t \left(\sum_{j=1}^k \|c^* \beta_j^{(0)} - \beta_j^*\|^2 \right), \text{ and}$$

$$\delta_{n,\sigma}^{\mathcal{N}}(d, k, \pi_{\min}) := \sigma^2 \frac{kd}{\pi_{\min}^3 n} \log(kd) \log(n/kd)$$

to denote the error tracked over iterations (with c^* denoting the smallest $c > 0$ such that $c(\beta_i - \beta_j) \in \mathcal{B}_{\beta_i^* - \beta_j^*}(r)$ for all $1 \leq i \neq j \leq k$), and a proxy for the final statistical rate, respectively.

Theorem 5.3.1 states that there are universal constants c_1 and c_2 such that if the sample size obeys the condition $n \geq n_{\mathsf{AM}}(c_1)$, then we have

$$\Pr\left\{ \max_{t \geq 1} \, \vartheta_t \left(c_2 \frac{\pi_{\min}^3}{k\kappa}; \left\{\beta_j^*\right\}_{j=1}^k \right) \geq c_1 \delta_{n,\sigma}^{\mathcal{N}}(d, k, \pi_{\min}) \right\} \leq c_2 \left(k \exp\left(-c_1 n \frac{\pi_{\min}^4}{\log^2(1/\pi_{\min})} \right) + \frac{k^2}{n^7} \right). \tag{5.17}$$

Let us now proceed to a proof of the theorem, assuming without loss of generality that the scalar c^* above is equal to 1. It is convenient to state and prove another result that guarantees a one-step contraction, from which Theorem 5.3.1 follows as a corollary. In order to state this result, we assume that one step of the alternating minimization update (5.12a)-(5.12b) is run starting from the parameters $\{\beta_j\}_{j=1}^k$ to produce the next iterate $\left\{\beta_j^+\right\}_{j=1}^k$. In the statement of the proposition, we use the shorthand

$$\begin{aligned} v_{i,j}^* &= \beta_i^* - \beta_j^*, \\ v_{i,j} &= \beta_i - \beta_j, \text{ and} \\ v_{i,j}^+ &= \beta_i^+ - \beta_j^+. \end{aligned}$$

Also recall the definitions of the geometric quantities (Δ, κ). The following proposition guarantees the one step contraction bound.

Proposition 5.4.1. *There exist universal constants c_1 and c_2 such that*

(a) If the sample size satisfies the bound $n \geq c_1 \max\{d, 10 \log n\} \max\left\{ \frac{k}{\pi_{\min}^3}, \frac{\log^2(1/\pi_{\min})}{\pi_{\min}^3}, \log(n/d) \right\}$, then for all parameters $\{\beta_j\}_{j=1}^k$ satisfying

$$\max_{1 \leq j \neq j' \leq k} \frac{\left\| v_{j,j'} - v_{j,j'}^* \right\|}{\|\theta_j^* - \theta_{j'}^*\|} \log^{3/2}\left(\frac{\|\theta_j^* - \theta_{j'}^*\|}{\left\| v_{j,j'} - v_{j,j'}^* \right\|} \right) \leq c_2 \frac{\pi_{\min}^3}{k\kappa}, \tag{5.18a}$$

we have, simultaneously for all pairs $1 \leq j \neq \ell \leq k$, the bound

$$\frac{\left\| v_{j,\ell}^+ - v_{j,\ell}^* \right\|^2}{\|\theta_j^* - \theta_\ell^*\|^2} \leq \max\left\{ \frac{d\kappa}{\pi_{\min}^3 n}, \frac{1}{4k} \right\} \left(\sum_{j'=1}^k \frac{\left\| v_{j,j'} - v_{j,j'}^* \right\|^2}{\|\theta_j^* - \theta_{j'}^*\|^2} + \frac{\left\| v_{\ell,j'} - v_{\ell,j'}^* \right\|^2}{\|\theta_\ell^* - \theta_{j'}^*\|^2} \right) + c_1 \frac{\sigma^2}{\Delta} \frac{kd}{\pi_{\min}^3 n} \log(n/d) \tag{5.18b}$$

with probability exceeding $1 - c_1 \left(k \exp\left(-c_2 n \frac{\pi_{\min}^4}{\log^2(1/\pi_{\min})} \right) + \frac{k^2}{n^7} \right)$.

(b) If the sample size satisfies the bound $n \geq c_1 \max\left\{ \max\{d, 10 \log n\} \max\left\{ \frac{k}{\pi_{\min}^3}, \frac{\log^2(1/\pi_{\min})}{\pi_{\min}^3}, \log(n/d) \right\}, \frac{kd}{\pi_{\min}^3} \right\}$, then for all parameters $\{\beta_j\}_{j=1}^k$ satisfying

$$\max_{1 \leq j \neq j' \leq k} \frac{\left\| v_{j,j'} - v_{j,j'}^* \right\|}{\|\theta_j^* - \theta_{j'}^*\|} \log^{3/2}\left(\frac{\|\theta_j^* - \theta_{j'}^*\|}{\left\| v_{j,j'} - v_{j,j'}^* \right\|} \right) \leq c_2 \frac{\pi_{\min}^3}{k}, \tag{5.19a}$$

we have the overall estimation error bound

$$\sum_{i=1}^{k} \|\beta_j^+ - \beta_j^*\|^2 \leq \frac{3}{4} \cdot \left(\sum_{i=1}^{k} \|\beta_j - \beta_j^*\|^2 \right) + c_1 \sigma^2 \frac{kd}{\pi_{\min}^3 n} \log(k) \log(n/dk) \tag{5.19b}$$

with probability exceeding $1 - c_1 \left(k \exp\left(-c_2 n \frac{\pi_{\min}^4}{\log^2(1/\pi_{\min})} \right) + \frac{k^2}{n^7} \right)$.

Let us briefly comment on why Proposition 5.4.1 implies Theorem 5.3.1 as a corollary. Clearly, equations (5.19a) and (5.19b) in conjunction show that the estimation error decays geometrically after running one step of the algorithm. The only remaining detail to be verified is that the next iterates $\left\{\beta_j^+\right\}_{j=1}^k$ also satisfy condition (5.18a) provided the sample size is large enough; in that case, the one step estimation bound (5.19b) can be applied recursively to obtain the final bound (5.14b).

For the constant c_2 in the proposition, let r_b be the largest value in the interval $[0, e^{-3/2}]$ such that $r_b \log^{3/2}(1/r_b) \leq c_2 \frac{\pi_{\min}^3}{k}$. Similarly, let r_a be the largest value in the interval $[0, e^{3/2}]$ such that $r_a \log^{3/2}(1/r_a) \leq c_2 \frac{\pi_{\min}^3}{k\kappa}$.

Assume that the current parameters satisfy the bound (5.18a). Choosing $n \geq 4\kappa d/\pi_{\min}^3$ and applying inequality (5.18b), we have, for each pair $1 \leq j \neq \ell \leq k$, the bound

$$\begin{aligned} \frac{\left\|v_{j,\ell}^+ - v_{j,\ell}^*\right\|^2}{\|\theta_j^* - \theta_\ell^*\|^2} &\leq \frac{1}{4k} \left(\sum_{j'=1}^{k} \frac{\left\|v_{j,j'} - v_{j,j'}^*\right\|^2}{\|\theta_j^* - \theta_{j'}^*\|^2} + \frac{\left\|v_{\ell,j'} - v_{\ell,j'}^*\right\|^2}{\|\theta_\ell^* - \theta_{j'}^*\|^2} \right) + c_1 \frac{1}{\|\theta_j^* - \theta_\ell^*\|^2} \sigma^2 \frac{kd}{\pi_{\min}^3 n} \log(n/d) \\ &\leq \frac{1}{2} r_a^2 + c_1 \frac{\sigma^2}{\Delta} \frac{kd}{\pi_{\min}^3 n} \log(n/d). \end{aligned}$$

Further, if $n \geq C\sigma^2 \frac{k^3 \kappa^2 d}{\pi_{\min}^9 \Delta r_0^2} \log(k\kappa/\pi_{\min}^3) \log(n/d)$ for a sufficiently large constant C, we have

$$\frac{\left\|v_{j,\ell}^+ - v_{j,\ell}^*\right\|^2}{\|\theta_j^* - \theta_\ell^*\|^2} \leq r_a^2.$$

Thus, the parameters $\left\{\beta_j^+\right\}_{j=1}^k$ satisfy inequality (5.18a) for the sample size choice required by Theorem 5.3.1. Finally, noting, for a pair of small enough scalars (a, b), the implication

$$a \leq \frac{b}{2} \log^{-3/2}(1/b) \implies a \log^{3/2}(1/a) \leq b,$$

and adjusting the constants appropriately to simplify the probability statement completes the proof of the theorem.

5.4.1 Proof of Proposition 5.4.1

We use the shorthand notation $S_j := S_j(\beta_1, \ldots, \beta_k)$, and let P_{S_j} denote the projection matrix onto the range of the matrix Ξ_{S_j}. Recall our notation for the difference vectors.

Let y^* denote the vector with entry i given by $\max_{\ell \in [k]} \langle \xi_i,\, \beta^*_\ell \rangle$. We have

$$\begin{aligned}
\|\Xi_{S_j}(\beta^+_j - \beta^*_j)\|^2 &= \|P_{S_j} y_{S_j} - \Xi_{S_j}\beta^*_j\|^2 \\
&= \|P_{S_j} y^*_{S_j} + P_{S_j}\epsilon_{S_j} - \Xi_{S_j}\beta^*_j\|^2 \\
&\leq 2\|P_{S_j}(y^*_{S_j} - \Xi_{S_j}\beta^*_j)\|^2 + 2\|P_{S_j}\epsilon_{S_j}\|^2 \\
&\leq 2\|y^*_{S_j} - \Xi_{S_j}\beta^*_j\|^2 + 2\|P_{S_j}\epsilon_{S_j}\|^2, \qquad (5.20)
\end{aligned}$$

where we have used the fact that the projection operator is non-expansive on a convex set.

Let

$$\{\langle \xi_i,\, \beta_\ell \rangle = \max\} := \left\{ \langle \xi_i,\, \beta_\ell \rangle = \max_{u \in [k]} \langle \xi_i,\, \beta_u \rangle \right\}, \text{ for each } i \in [n], \ell \in [k]$$

denote a convenient shorthand for these events. The first term on the RHS of inequality (5.20) can be written as

$$\sum_{i \in S_j} (y^*_i - \langle \xi_i,\, \beta^*_j \rangle)^2 \leq \sum_{i=1}^n \sum_{j' : j' \neq j} \mathbf{1}\left\{ \langle \xi_i,\, \beta_j \rangle = \max \text{ and } \langle \xi_i,\, \beta^*_{j'} \rangle = \max \right\} \langle \xi_i,\, \beta^*_{j'} - \beta^*_j \rangle^2,$$

where the inequality accounts for ties. Each indicator random variable is bounded, in turn, as

$$\begin{aligned}
\mathbf{1}\left\{ \langle \xi_i,\, \beta_j \rangle = \max \text{ and } \langle \xi_i,\, \beta^*_{j'} \rangle = \max \right\} &\leq \mathbf{1}\left\{ \langle \xi_i,\, \beta_j \rangle \geq \langle \xi_i,\, \beta_{j'} \rangle \text{ and } \langle \xi_i,\, \beta^*_{j'} \rangle \geq \langle \xi_i,\, \beta^*_j \rangle \right\} \\
&= \mathbf{1}\left\{ \langle \xi_i,\, v_{j,j'} \rangle \cdot \langle \xi_i,\, v^*_{j,j'} \rangle \leq 0 \right\}.
\end{aligned}$$

Switching the order of summation yields the bound

$$\sum_{i \in S_j} (y^*_i - \langle \xi_i,\, \beta^*_j \rangle)^2 \leq \sum_{j' : j' \neq j} \sum_{i=1}^n \mathbf{1}\left\{ \langle \xi_i,\, v_{j,j'} \rangle \cdot \langle \xi_i,\, v^*_{j,j'} \rangle \leq 0 \right\} \langle \xi_i,\, v^*_{j,j'} \rangle^2.$$

Recalling our notation for the minimum eigenvalue of a symmetric matrix, the LHS of inequality (5.20) can be bounded as

$$\|\Xi_{S_j}(\beta^+_j - \beta^*_j)\|^2 \geq \lambda_{\min}\left(\Xi^\top_{S_j} \Xi_{S_j} \right) \cdot \|\beta^+_j - \beta^*_j\|^2.$$

Putting together the pieces yields, for each $j \in [k]$, the pointwise bound

$$\frac{1}{2}\lambda_{\min}\left(\Xi^\top_{S_j} \Xi_{S_j} \right) \cdot \|\beta^+_j - \beta^*_j\|^2 \leq \sum_{j' : j' \neq j} \sum_{i=1}^n \mathbf{1}\left\{ \langle \xi_i,\, v_{j,j'} \rangle \cdot \langle \xi_i,\, v^*_{j,j'} \rangle \leq 0 \right\} \langle \xi_i,\, v^*_{j,j'} \rangle^2 + \|P_{S_j}\epsilon_{S_j}\|^2. \qquad (5.21)$$

Up to this point, note that all steps of the proof were deterministic. In order to complete the proof, it suffices to show high probability bounds on the various quantities appearing in the bound (5.21).

Since the set S_j is in itself random and could depend on the pair (Ξ, ϵ), bounding individual terms is especially challenging. Our approach is to show bounds that hold uniformly over all parameters $\{\beta_j\}_{j=1}^k$ that are close to the true parameters.

Recall the notation

$$\mathcal{B}_{v^*}(r) = \left\{ v \in \mathbb{R}^{d+1} : \frac{\|v - v^*\|}{\|u^*\|} \leq r \right\}$$

introduced before, and the definitions of the pair of scalars (r_a, r_b).

To be agnostic to the scale invariance of the problem, we set $c^* = 1$ and define the set of parameters

$$\mathcal{I}(r) = \left\{ \beta_1, \ldots, \beta_k : v_{i,j} \in \mathcal{B}_{v^*_{i,j}}(r) \text{ for all } 1 \leq i \neq j \leq k \right\},$$

and use the shorthand $\mathcal{I}_a := \mathcal{I}(r_a)$ and $\mathcal{I}_b := \mathcal{I}(r_b)$, to denote the set of parameters satisfying conditions (5.18a) and (5.19a), respectively,

Recall that we denote by

$$S_j(\beta_1, \ldots, \beta_k) := \left\{ 1 \leq i \leq n : \langle \xi_i,\, \beta_j \rangle = \max_{1 \leq u \leq k} \left(\langle \xi_i,\, \beta_u \rangle \right) \right\},$$

the indices of the rows for which β_j attains the maximum, and we additionally keep this sets disjoint by breaking ties lexicographically. To lighten notation, we use the shorthand

$$\Xi^j(\beta_1, \ldots, \beta_k) := \Xi_{S_j(\beta_1, \ldots, \beta_k)}.$$

Having defined this notation, we are now ready to return to the proof of Proposition 5.4.1. We make the following claims to handle the three terms in the bound (5.21). First, we claim that the noise terms are uniformly bounded as

$$\Pr\left\{ \sup_{\beta_1, \ldots, \beta_k \in \mathbb{R}^{d+1}} \sum_{j=1}^{k} \|P_{\Xi^j(\beta_1, \ldots, \beta_k)} \epsilon_{S_j(\beta_1, \ldots, \beta_k)}\|^2 \geq 2\sigma^2 k(d+1) \log(kd) \log(n/kd) \right\} \leq \binom{n}{kd}^{-1}, \text{ and} \tag{5.22a.I}$$

$$\Pr\left\{ \sup_{\beta_1, \ldots, \beta_k \in \mathbb{R}^{d+1}} \|P_{\Xi^j(\beta_1, \ldots, \beta_k)} \epsilon_{S_j(\beta_1, \ldots, \beta_k)}\|^2 \geq 2\sigma^2 k(d+1) \log(n/d) \right\} \leq \binom{n}{d}^{-1} \text{ for each } j \in [k]. \tag{5.22a.II}$$

Second, we show that the indicator quantities are simultaneously bounded for all j, j' pairs. In particular, we claim that there exists a tuple of universal constants (C, c_1, c_2, c') such that for each positive scalar $r \leq 1/24$, we have

$$\begin{aligned} \Pr\Big\{ \exists 1 \leq j \neq j' \leq k,\ v_{j,j'} \in \mathfrak{B}_{v^*_{j,j'}}(r) : \sum_{j' : j' \neq j} \sum_{i=1}^{n} \mathbf{1}\left\{ \langle \xi_i,\, v_{j,j'} \rangle \cdot \langle \xi_i,\, v^*_{j,j'} \rangle \leq 0 \right\} \langle \xi_i,\, v^*_{j,j'} \rangle^2 \\ \geq C \max\{d, nr \log^{3/2}(1/r)\} \sum_{j' : j' \neq j} \|v_{j,j'} - v^*_{j,j'}\|^2 \Big\} \leq c_1 \binom{k}{2} \left\{ n e^{-c_2 n} + e^{-c' \max\{d, 10 \log n\}} \right\}. \end{aligned} \tag{5.22b}$$

Finally, we show a bound on the LHS of the bound (5.21) by handling the singular values of (random) sub-matrices of Ξ with a uniform bound. In particular, we claim that there are universal constants (C, c, c') such that if $n \geq Cd \max\left\{\frac{k}{\pi_{\min}^3}, \frac{\log^2(1/\pi_{\min})}{\pi_{\min}^3}, \log(n/d)\right\}$, then for each $j \in [k]$, we have

$$\Pr\left\{\inf_{\beta_1,\ldots,\beta_k \in \mathcal{I}_b} \quad \lambda_{\min}\left(\Xi^j(\beta_1,\ldots,\beta_k)^\top \cdot \Xi^j(\beta_1,\ldots,\beta_k)\right) \leq C\pi_{\min}^3 n\right\}$$
$$\leq c\exp\left(-cn\frac{\pi_{\min}^4}{\log^2(1/\pi_{\min})}\right) + c'\exp(-c'n\cdot\pi_{\min}). \qquad (5.22c)$$

Notice that claim (5.22a.I) implicitly defines a high probability event $\mathcal{E}^{(a.I)}$, claim (5.22a.II) defines high probability events $\mathcal{E}_j^{(a.II)}$, claim (5.22b) defines a high probability event $\mathcal{E}^{(b)}(r)$, and claim (5.22c) defines high probability events $\mathcal{E}_j^{(c)}$. Define the intersection of these events as

$$\mathcal{E}(r) := \mathcal{E}^{(a.I)} \bigcap \left(\bigcap_{j\in[k]} \mathcal{E}_j^{(a.II)}\right) \bigcap \mathcal{E}^{(b)}(r) \bigcap \left(\bigcap_{j\in[k]} \mathcal{E}_j^{(c)}\right),$$

and note that the claims in conjunction with the union bound guarantee that if the condition on the sample size $n \geq c_1 d \max\left\{\frac{k}{\pi_{\min}^3}, \frac{\log^2(1/\pi_{\min})}{\pi_{\min}^3}, \log(n/d)\right\}$ holds, then for all $r \leq r_b$, we have

$$\Pr\left\{\mathcal{E}(r)\right\} \geq 1 - c_1\left(k\exp\left(-c_2 n\frac{\pi_{\min}^4}{\log^2(1/\pi_{\min})}\right) + \frac{k^2}{n^7}\right),$$

where we have adjusted constants appropriately in stating the bound. We are now ready to prove the two parts of the proposition.

Proof of part (a): Work on the event $\mathcal{E}(r_a)$. Normalizing inequality (5.21) by n and using claims (5.22a.II). (5.22b), and (5.22c) with $r = r_a$ then yields, simultaneously for all $j \in [k]$, the bound

$$\|\beta_j^+ - \beta_j^*\|^2 \leq C\max\left\{\frac{d}{\pi_{\min}^3 n}, \frac{r_a}{\pi_{\min}^3}\log^{3/2}(1/r_a)\right\}\sum_{j':j'\neq j}\|v_{j,j'} - v_{j,j'}^*\|^2 + C'\sigma^2\frac{kd}{\pi_{\min}^3 n}\log(n/d)$$
$$\overset{\text{(i)}}{\leq} \max\left\{\frac{Cd}{\pi_{\min}^3 n}, \frac{1}{4k\kappa}\right\}\sum_{j':j'\neq j}\|v_{j,j'} - v_{j,j'}^*\|^2 + C'\sigma^2\frac{kd}{\pi_{\min}^3 n}\log(n/d),$$

where in step (i), we have used the definition of the quantity r_a. Using this bound for the indices j, ℓ in conjunction with the definition of the quantity κ proves inequality (5.18b). □

Proof of part (b): We now work on the event $\mathcal{E}(r_b)$ and proceed again from the bound

$$\|\beta_j^+ - \beta_j^*\|^2 \leq C \max\left\{ \frac{d}{\pi_{\min}^3 n}, \frac{r_b}{\pi_{\min}^3} \log^{3/2}(1/r_b) \right\} \sum_{j': j' \neq j} \|v_{j,j'} - v_{j,j'}^*\|^2 + \frac{C}{\pi_{\min}^3 n} \|P_{S_j} \epsilon_{S_j}\|^2.$$

Summing over $j \in [k]$ and using the Cauchy–Schwarz inequality, we obtain

$$\begin{aligned} \sum_{j=1}^k \|\beta_j^+ - \beta_j^*\|^2 &\leq C \max\left\{ \frac{kd}{\pi_{\min}^3 n}, \frac{k r_b}{\pi_{\min}^3} \log^{3/2}(1/r_b) \right\} \left(\sum_{j=1}^k \|\beta_j - \beta_j^*\|^2 \right) + \frac{C}{\pi_{\min}^3 n} \sum_{j \in [k]} \|P_{S_j} \epsilon_{S_j}\|^2 \\ &\overset{\text{(ii)}}{\leq} \frac{3}{4} \left(\sum_{j=1}^k \|\beta_j - \beta_j^*\|^2 \right) + C' \sigma^2 \frac{kd}{\pi_{\min}^3 n} \log(k) \log(n/kd), \end{aligned}$$

where in step (ii), we have used the definition of the quantity r_b, the bound $n \geq Ckd/\pi_{\min}^3$, and claim (5.22a.I). This completes the proof. □

We now prove each of the claims in turn. This constitutes the technical meat of our proof, and involves multiple technical lemmas whose proofs are postponed to the end of the section.

Proof of claims (5.22a.I) **and** (5.22a.II)**:** We begin by stating a general lemma about concentration properties of the noise.

Lemma 5.4.1. *Consider a random variable $z \in \mathbb{R}^n$ with i.i.d. σ-sub-Gaussian entries, and a fixed matrix $\Xi \in \mathbb{R}^{n \times (d+1)}$. Then, we have*

$$\sup_{\beta_1, \ldots, \beta_k \in \mathbb{R}^{d+1}} \sum_{j=1}^k \|P_{\Xi^j(\beta_1, \ldots, \beta_k)} z\|^2 \leq 2\sigma^2 k(d+1) \log(kd) \log(n/kd) \tag{5.32a}$$

with probability greater than $1 - \binom{n}{kd}^{-1}$ and

$$\sup_{\beta_1, \ldots, \beta_k \in \mathbb{R}^{d+1}} \max_{j \in [k]} \|P_{\Xi^j(\beta_1, \ldots, \beta_k)} z_{S_j(\beta_1, \ldots, \beta_k)}\|^2 \leq 2\sigma^2 k(d+1) \log(n/d) \tag{5.32b}$$

with probability greater than $1 - \binom{n}{d}^{-1}$.

The proof of the claims follows directly from Lemma 5.4.1, since the noise vector ϵ is independent of the matrix Ξ, and $\mathcal{I}_b \subseteq \left(\mathbb{R}^{d+1}\right)^{\otimes k}$. □

Proof of claim (5.22b)**:** We now state a lemma that directly handles indicator functions as they appear in the claim.

Lemma 5.4.2. *Let $u^* \in \mathbb{R}^d$ and $w^* \in \mathbb{R}$, and consider a fixed parameter $v^* = (u^*,\ w^*) \in \mathbb{R}^{d+1}$. Then there are universal constants (c_1, c_2, c_3, c_4) such that for all positive scalars $r \leq 1/24$, we have*

$$\sup_{v \in \mathfrak{B}_{v^*}(r)} \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left\{ \langle \xi_i,\ v \rangle \cdot \langle \xi_i,\ v^* \rangle \leq 0 \right\} \langle \xi_i,\ v^* \rangle^2 \right) / \|v - v^*\|^2 \leq c_1 \cdot \max\left\{ \frac{d}{n}, r \log^{3/2}\left(\frac{1}{r}\right) \right\}$$

with probability exceeding $1 - c_1 e^{-c_2 \max\{d, 10 \log n\}} - c_3 n e^{-c_4 n}$. Here, we adopt the convention that $0/0 = 0$.

Applying Lemma 5.4.2 with $v = v_{j,j'}$ and $v^* = v^*_{j,j'}$ for all pairs (j, j') and using a union bound directly yields the claim. □

Proof of claim (5.22c)**:** For this claim, we state three technical lemmas pertaining to the singular values of random matrices whose rows are formed by truncated Gaussian random vectors. We let $\mathrm{vol}(K)$ denote the volume of a set $K \subseteq \mathbb{R}^d$ with respect to d-dimensional standard Gaussian measure, i.e., with $\mathrm{vol}(K) = \Pr\{Z \in K\}$ for $Z \sim \mathcal{N}(0, I_d)$.

Lemma 5.4.3. *Suppose n vectors $\{x_i\}_{i=1}^n$ are drawn i.i.d. from $\mathcal{N}(0, I_d)$, and $K \subseteq \mathbb{R}^d$ is a fixed convex set. Then there exists a tuple of universal constants (c_1, c_2) such that if $\mathrm{vol}^3(K) n \geq c_1 d \log^2(1/\mathrm{vol}(K))$, then*

$$\lambda_{\min}\left(\sum_{i : x_i \in K} \xi_i \xi_i^\top \right) \geq c_2 \,\mathrm{vol}^3(K) \cdot n$$

with probability greater than $1 - c_1 \exp\left(-c_2 n \frac{\mathrm{vol}^4(K)}{\log^2(1/\mathrm{vol}(K))} \right) - c_1 \exp(-c_2 n \cdot \mathrm{vol}(K))$.

For a pair of scalars (w, w') and d-dimensional vectors (u, u'), define the *wedge* formed by the $d + 1$-dimensional vectors $v = (u,\ w)$ and $v' = (u',\ w')$ as the region

$$W(v, v') = \{ x \in \mathbb{R}^d : (\langle x,\ u \rangle + w) \cdot (\langle x,\ u' \rangle + w') \leq 0 \},$$

and let $\mathcal{W}_\delta = \{ W = W(v, v') : \mathrm{vol}(W) \leq \delta \}$ denote the set of all wedges with Gaussian volume less than δ. The next lemma bounds the maximum singular value of a sub-matrix formed by any such wedge.

Lemma 5.4.4. *There is a tuple of universal constants (c_1, c_2) such that if $n \geq c_1 d \log(n/d)$, then*

$$\sup_{W \in \mathcal{W}_\delta} \lambda_{\max}\left(\sum_{i : x_i \in W} \xi_i \xi_i^\top \right) \leq c_1 \left(\delta n + d + n\delta \log(1/\delta) \right)$$

with probability greater than $1 - 2\exp(-c_2 \delta n) - \binom{n}{c_2 \delta n}^{-1}$.

We are now ready to proceed to a proof of claim (5.22c). For convenience, introduce the shorthand notation

$$S_j^* := S_j\left(\beta_1^*, \ldots, \beta_k^*\right)$$

to denote the set of indices corresponding to observations generated by the true parameter β_j^*. Letting $A \Delta B := (A \setminus B) \bigcup (B \setminus A)$ denote the symmetric difference between two sets A and B, we have

$$\lambda_{\min}\left(\Xi_{S_j}^\top \Xi_{S_j}\right) \geq \lambda_{\min}\left(\Xi_{S_j^*}^\top \Xi_{S_j^*}\right) - \lambda_{\max}\left(\Xi_{S_j^* \Delta S_j}^\top \Xi_{S_j^* \Delta S_j}\right).$$

Recall that by definition, we have

$$\begin{aligned} S_j^* \Delta S_j &= \{i : \langle \xi_i,\, \beta_j^* \rangle\} = \max \text{ and } \langle \xi_i,\, \beta_j \rangle \neq \max\} \bigcup \{i : \langle \xi_i,\, \beta_j^* \rangle \neq \max \text{ and } \langle \xi_i,\, \beta_j \rangle = \max\} \\ &\subseteq \bigcup_{j' \in [k] \setminus j} \{i : \langle \xi_i,\, v_{j,j'}^* \rangle \cdot \langle \xi_i,\, v_{j,j'} \rangle < 0\} \\ &= \bigcup_{j' \in [k] \setminus j} \{i : x_i \in W\left(v_{j,j'}^*, v_{j,j'}\right)\}. \end{aligned} \tag{5.33}$$

Putting together the pieces, we have

$$\lambda_{\min}\left(\Xi_{S_j}^\top \Xi_{S_j}\right) \geq \lambda_{\min}\left(\Xi_{S_j^*}^\top \Xi_{S_j^*}\right) - \sum_{j' \neq j} \lambda_{\max}\left(\sum_{i : x_i \in W\left(v_{j,j'}^*, v_{j,j'}\right)} \xi_i \xi_i^\top\right). \tag{5.34}$$

Conditioned on the event guaranteed by Lemma 5.4.4 with $\delta = \text{vol}\left(W\left(v_{j,j'}^*, v_{j,j'}\right)\right)$ and for a universal constant C_1, we have the bound

$$\begin{aligned} &\sup_{v_{j,j'} \in \mathfrak{B}_{v_{j,j'}^*}(r_0)} \lambda_{\max}\left(\sum_{i : x_i \in W\left(v_{j,j'}^*, v_{j,j'}\right)} \xi_i \xi_i^\top\right) \\ &\qquad \leq \sup_{v_{j,j'} \in \mathfrak{B}_{v_{j,j'}^*}(r_0)} C_1(n \, \text{vol}(W\left(v_{j,j'}^*, v_{j,j'}\right)) \log(1/\text{vol}(W\left(v_{j,j'}^*, v_{j,j'}\right))) + d) \\ &\qquad \overset{\text{(i)}}{\leq} \sup_{v_{j,j'} \in \mathfrak{B}_{v_{j,j'}^*}(r_0)} C_1\left(n \frac{\left\|v_{j,j'} - v_{j,j'}^*\right\|}{\left\|u_{j,j'}^*\right\|} \log^{3/2} \frac{\left\|u_{j,j'}^*\right\|}{\left\|v_{j,j'} - v_{j,j'}^*\right\|} + d\right) \\ &\qquad \overset{\text{(ii)}}{\leq} n r_0 \log^{3/2}(1/r_0) + d \\ &\qquad \overset{\text{(iii)}}{\leq} nC \frac{\pi_{\min}^3}{k}, \end{aligned}$$

where in step (i), we have used Lemma B.1.1, and in step (ii), we have used the definition of the set $\mathfrak{B}$. Step (iii) uses the assumption $n \geq c_1 kd/\pi_{\min}^3$.

Moreover, Lemma 5.4.3 guarantees the bound $\lambda_{\min}\left(\Xi_{S_j^*}^\top \Xi_{S_j^*}\right) \geq c_2 n \cdot \pi_{\min}^3$, so that putting together the pieces, we have

$$\begin{aligned} \inf_{\beta_1,\ldots,\beta_k \in \mathcal{I}_b} \lambda_{\min}\left(\Xi_{S_j}^\top \Xi_{S_j}\right) &\geq c_2 n \pi_{\min}^3 - Cnk\frac{\pi_{\min}^3}{k} \\ &\geq C\pi_{\min}^3 n, \end{aligned} \tag{5.35}$$

with probability greater than $1 - c\exp\left(-cn\frac{\pi_{\min}^4}{\log^2(1/\pi_{\min})}\right) - c'\exp(-c'n \cdot \pi_{\min})$. These assertions hold provided $n \geq Cd\max\left\{\frac{k}{\pi_{\min}^3}, \frac{\log^2(1/\pi_{\min})}{\pi_{\min}^3}, \log(n/d)\right\}$, and this completes the proof. □

Having proved the claims, we turn to proofs of our technical lemmas.

Proof of Lemma 5.4.1

In this proof, we assume that $\sigma = 1$; our bounds can finally be scaled by σ^2.

It is natural to prove the bound (5.32b) first followed by bound (5.32a). First, consider a fixed set of parameters $\{\beta_1, \ldots, \beta_k\}$. Then, we have

$$\left\|P_{\Xi^j(\beta_1,\ldots,\beta_k)} z_{S^j}\right\|^2 = \left\|UU^\top z_{S^j}\right\|^2,$$

where $U \in \mathbb{R}^{|\Xi^j| \times (d+1)}$ denotes a matrix with orthonormal columns that span the range of $\Xi^j(\beta_1, \ldots, \beta_k)$.

Applying the Hanson-Wright inequality for independent sub-Gaussians (see [269, Theorem 2.1]) and noting that $\|UU^\top\|_{\text{fro}} \leq \sqrt{d+1}$ we obtain

$$\Pr\left\{\left\|UU^\top z_{S^j}\right\|^2 \geq (d+1) + t\right\} \leq e^{-ct},$$

for each $t \geq 0$. In particular, this implies that the random variable $\left\|UU^\top z_{S^j}\right\|^2$ is sub-exponential.

This tail bound holds for a fixed partition of the rows of Ξ; we now take a union bound over all possible partitions. Toward that end, define the sets

$$\mathcal{S}^j = \left\{S_j(\beta_1, \ldots, \beta_k) : \beta_1, \ldots, \beta_k \in \mathbb{R}^{d+1}\right\}, \text{ for each } j \in [k].$$

From Lemma B.1.3, we have the bound $|\mathcal{S}^j| \leq 2^{ckd\log(en/d)}$. Thus, applying the union bound, we obtain

$$\Pr\left\{\sup_{\beta_1,\ldots,\beta_k \in \mathbb{R}^{d+1}} \left\|P_{\Xi^j(\beta_1,\ldots,\beta_k)} z_{S^j}\right\|^2 \geq (d+1) + t\right\} \leq |\mathcal{S}^j| e^{-ct},$$

and substituting $t = ck(d+1)\log(n/d)$ and performing some algebra establishes bound (5.32b).

In order to establish bound (5.32a), we once again consider the term $\sum_{j=1}^k \left\|P_{\Xi^j(\beta_1,\ldots,\beta_k)} z_{S^j}\right\|^2$ for a fixed set of parameters $\{\beta_1, \ldots, \beta_k\}$. Note that this is the sum of k independent sub-exponential random variables and can be thought of as a quadratic form of the entire vector z. So once again from the Hanson-Wright inequality, we have

$$\Pr\left\{ \sup_{\beta_1,\ldots,\beta_k \in \mathbb{R}^{d+1}} \sum_{j=1}^{k} \left\| P_{\Xi^j(\beta_1,\ldots,\beta_k)} z_{S^j} \right\|^2 \geq k(d+1) + t \right\} \leq e^{-ct/k}$$

for all $t \geq 0$.

Also define the set of all possible partitions of the n points via the max-affine function; we have the set

$$\mathcal{S} = \left\{ S_1(\beta_1, \ldots, \beta_k), \ldots, S_k(\beta_1, \ldots, \beta_k) : \beta_1, \ldots, \beta_k \in \mathbb{R}^{d+1} \right\}.$$

Lemma B.1.4 yields the bound $|\mathcal{S}| \leq 2^{ckd \log(kd) \log(n/kd)}$, and combining a union bound with the high probability bound above establishes bound (5.32a) after some algebraic manipulation. □

Proof of Lemma 5.4.2

Let $\gamma_v = v - v^*$; we have

$$\begin{aligned} \mathbf{1}\left\{ \langle \xi_i,\, v \rangle \cdot \langle \xi_i,\, v^* \rangle \leq 0 \right\} \langle \xi_i,\, v^* \rangle^2 &\leq \mathbf{1}\left\{ \langle \xi_i,\, v \rangle \cdot \langle \xi_i,\, v^* \rangle \leq 0 \right\} \langle \xi_i,\, \gamma_v \rangle^2 \\ &\leq \mathbf{1}\left\{ \langle \xi_i,\, \gamma_v \rangle^2 \geq \langle \xi_i,\, v^* \rangle^2 \right\} \langle \xi_i,\, \gamma_v \rangle^2. \end{aligned}$$

Define the (random) set $K_v = \{ i : \langle \xi_i,\, \gamma_v \rangle^2 > \langle \xi_i,\, v^* \rangle^2 \}$; we have the bound

$$\frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\left\{ \langle \xi_i,\, v \rangle \cdot \langle \xi_i,\, v^* \rangle \leq 0 \right\} \langle \xi_i,\, v^* \rangle^2 \leq \frac{1}{n} \| \Xi_{K_v} \gamma_v \|^2.$$

We now show that the quantity $\| \Xi_{K_v} \gamma_v \|^2$ is bounded uniformly for all $v \in \mathfrak{B}_{v^*}(r)$ for small enough r. Recall that u^* is the "linear" portion of v^*, and let $m = \max\{ d, 10 \log n, n \cdot (16r \cdot \sqrt{\log(1/r)} \}$ (note that m depends implicitly on r). We claim that for all $r \in (0, 1/24]$, we have

$$\Pr\left\{ \sup_{v \in \mathfrak{B}_{v^*}(r)} |K_v| > m \right\} \leq 4 e^{-c \max\{d, 10 \log n\}} + cne^{-c'n}, \text{ and} \tag{5.36a}$$

$$\Pr\left\{ \bigcup_{\substack{T \subseteq [n]: \\ |T| \leq m}} \sup_{\substack{\omega \in \mathbb{R}^{d+1} \\ \omega \neq 0}} \frac{\| \Xi_T \omega \|^2}{\| \omega \|^2} \geq (d + 16m \log(n/m)) \right\} \leq e^{-c \max\{d, 10 \log n\}}. \tag{5.36b}$$

Taking these claims as given, the proof of the lemma is immediate, since $\frac{n}{m} \leq \frac{1}{16r \log(1/r)}$, so that $\log(n/m) \leq C \log(1/r)$.

Proof of claim (5.36a)**:** By definition of the set K_v, we have

$$\begin{aligned}
\Pr\{\sup_{v \in \mathfrak{B}_{v^*}(r)} |K_v| > m\} &\leq \sum_{\substack{T \subseteq [n]: \\ |T| > m}} \Pr\left\{\exists v \in \mathfrak{B}_{v^*}(r) : \|\Xi_T \gamma_v\|^2 \geq \|\Xi_T v^*\|^2\right\} \\
&= \sum_{\substack{T \subseteq [n]: \\ |T| > m}} \Pr\left\{\exists v \in \mathfrak{B}_{v^*}(r) : \frac{\|\gamma_v\|^2}{\|u^*\|^2} \frac{\|\Xi_T \gamma_v\|^2}{\|\gamma_v\|^2} \geq \frac{\|\Xi_T v^*\|^2}{\|u^*\|^2}\right\} \\
&\leq \sum_{\substack{T \subseteq [n]: \\ |T| > m}} \Pr\left\{\exists v \in \mathfrak{B}_{v^*}(r) : r^2 \frac{\|\Xi_T \gamma_v\|^2}{\|\gamma_v\|^2} \geq \frac{\|\Xi_T v^*\|^2}{\|u^*\|^2}\right\} \\
&\leq \sum_{\substack{T \subseteq [n]: \\ |T| > m}} \Bigg(\Pr\left\{\exists v \in \mathfrak{B}_{v^*}(r) : \frac{\|\Xi_T \gamma_v\|^2}{\|\gamma_v\|^2} \geq (\sqrt{d} + \sqrt{|T|} + t_T)^2\right\} \\
&\qquad + \Pr\left\{\frac{\|\Xi_T v^*\|^2}{\|u^*\|^2} \leq r^2 (\sqrt{d} + \sqrt{|T|} + t_T)^2\right\}\Bigg),
\end{aligned}$$

where the final step follows by the union bound and holds for all positive scalars $\{t_T\}_{T \subseteq [n]}$. For some *fixed* subset T of size ℓ, we have the tail bounds

$$\Pr\left\{\sup_{\substack{\omega \in \mathbb{R}^{d+1} \\ \omega \neq 0}} \frac{\|\Xi_T \omega\|^2}{\|\omega\|^2} (\sqrt{d} + \sqrt{\ell} + t)^2\right\} \overset{\text{(i)}}{\leq} 2e^{-t^2/2}, \text{ for all } t \geq 0, \text{ and} \tag{5.37a}$$

$$\Pr\left\{\frac{\|\Xi_T v^*\|^2}{\|u^*\|^2} \leq \delta\ell\right\} \overset{\text{(ii)}}{\leq} (e\delta)^{\ell/2} \text{ for all } \delta \geq 0, \tag{5.37b}$$

where step (i) follows from the sub-Gaussianity of the covariate matrix (see Lemma B.1.5), and step (ii) from a tail bound for the non-central χ^2 distribution (see Lemma B.1.6).

Substituting these bounds yields

$$\begin{aligned}
\Pr\{\sup_{v \in \mathfrak{B}_{v^*}(r)} |K_v| > m\} &\leq \sum_{\ell = m+1}^{n} \binom{n}{\ell} \left[2e^{-t_\ell^2/2} + \left(er^2 \cdot \frac{(\sqrt{d} + \sqrt{\ell} + t_\ell)^2}{\ell}\right)^{\ell/2}\right] \\
&\leq \sum_{\ell = m+1}^{n} \binom{n}{\ell} \left[2e^{-t_\ell^2/2} + \left(2r \cdot \frac{\sqrt{d} + \sqrt{\ell} + t_\ell}{\sqrt{\ell}}\right)^{\ell}\right].
\end{aligned}$$

Recall that t_ℓ was a free (non-negative) variable to be chosen. We now split the proof into two cases and choose this parameter differently for the two cases.

Case 1, $m \leq \ell < n/e$**:** Substituting the choice $t_\ell = 4\sqrt{\ell \log(n/\ell)}$, we obtain

$$\begin{aligned}
\binom{n}{\ell}\left[2e^{-t_\ell^2/2} + \left(2r \cdot \frac{\sqrt{d} + \sqrt{\ell} + t_\ell}{\sqrt{\ell}}\right)^\ell\right] &\leq \left(\frac{n}{\ell}\right)^{-c\ell} + \binom{n}{\ell} \cdot \left(2r \cdot \frac{\sqrt{d} + 5\sqrt{\ell \log(n/\ell)}}{\sqrt{\ell}}\right)^\ell \\
&\overset{\text{(i)}}{\leq} \left(\frac{n}{\ell}\right)^{-c\ell} + \binom{n}{\ell} \cdot \left(2r \cdot (1 + 5\sqrt{\log(n/\ell)})\right)^\ell \\
&\overset{\text{(ii)}}{\leq} \left(\frac{n}{\ell}\right)^{-c\ell} + \binom{n}{\ell} \cdot \left(12r \cdot \sqrt{\log(n/\ell)}\right)^\ell \\
&\leq \left(\frac{n}{\ell}\right)^{-c\ell} + \left(12\left(\frac{en}{\ell}\right) r \cdot \sqrt{\log(n/\ell)}\right)^\ell,
\end{aligned}$$

where step (i) follows from the bound $m \geq d$, and step (ii) from the bound $\ell \leq n/e$.

Now note that the second term is only problematic for small ℓ. For all $\ell \geq m = n \cdot (16r \cdot \sqrt{\log(1/r)})$, we have

$$\left(12\left(\frac{en}{\ell}\right) r \cdot \sqrt{\log(n/\ell)}\right)^\ell \leq (3/4)^\ell.$$

The first term, on the other hand, satisfies the bound $\left(\frac{n}{\ell}\right)^{-c\ell} \leq (3/4)^\ell$ for sufficiently large n.

Case 2, $\ell \geq n/e$**:** In this case, setting $t_\ell = 2\sqrt{n}$ for each ℓ yields the bound

$$\begin{aligned}
\binom{n}{\ell}\left[2e^{-t_\ell^2/2} + \left(2r \cdot \frac{\sqrt{d} + \sqrt{\ell} + t_\ell}{\sqrt{\ell}}\right)^\ell\right] &\leq 2\binom{n}{n/2} e^{-2n} + (12r)^\ell \\
&\leq ce^{-c'n},
\end{aligned}$$

where we have used the fact that $d \leq n/2$ and $r \leq 1/24$.

Putting together the pieces from both cases, we have shown that for all $r \in (0, 1/24]$, we have

$$\begin{aligned}
\Pr\{\sup_{v \in \mathfrak{B}_{v^*}(r)} |K_v| > m\} &\leq cne^{-c'n} + \sum_{\ell=m+1}^{n/e} (3/4)^\ell \\
&\leq cne^{-c'n} + 4(3/4)^{\max\{d, 10\log n\}},
\end{aligned}$$

thus completing the proof of the claim.

Proof of claim (5.36b)**:** The proof of this claim follows immediately from the steps used to establish the previous claim. In particular, writing

$$\begin{aligned}
\Pr\left\{ \bigcup_{\substack{T \subseteq [n]:\\ |T| \leq m}} \bigcup_{\omega : \|\omega\| = 1} \|\Xi_T \omega\|^2 \geq d + 16 m \log(n/m) \right\} \\
&\leq \Pr\left\{ \bigcup_{\substack{T \subseteq [n]:\\ |T| \leq m}} \bigcup_{\omega : \|\omega\| = 1} \|\Xi_T \omega\|^2 \geq \left(\sqrt{d} + \sqrt{m} + \sqrt{4 m \log(n/m)} \right)^2 \right\} \\
&\leq \sum_{\ell=1}^{m} \Pr\left\{ \bigcup_{\substack{T \subseteq [n]:\\ |T| = \ell}} \bigcup_{\omega : \|\omega\| = 1} \|\Xi_T \omega\|^2 \geq \left(\sqrt{d} + \sqrt{m} + \sqrt{4 m \log(n/m)} \right)^2 \right\} \\
&\overset{\text{(iv)}}{\leq} 2 \sum_{\ell=1}^{m} \binom{n}{\ell} \exp\{-2m \log(n/m)\} \\
&\leq 2 \left(\frac{n}{m} \right)^{-cm} \leq 2 e^{-c \max\{d, 10 \log n\}},
\end{aligned}$$

where step (iv) follows from the tail bound (5.37a). □

Proof of Lemma 5.4.3

The lemma follows from some structural results on the truncated Gaussian distribution. Using the shorthand $\mathrm{vol} := \mathrm{vol}(K)$ and letting ψ denote the d-dimensional Gaussian density, consider a random vector τ drawn from the distribution having density $h(y) = \frac{1}{\mathrm{vol}} \psi(y) \mathbf{1}\{y \in K\}$, and denote its mean and second moment matrix by μ_τ and Σ_τ, respectively. Also denote the recentered random variable by $\widetilde{\tau} = \tau - \mu_\tau$. We claim that

$$\|\mu_\tau\|^2 \leq C \log(1/\mathrm{vol}), \tag{5.38a}$$

$$C \,\mathrm{vol}^2 \cdot I \preceq \Sigma_\tau \preceq (1 + C \log(1/\mathrm{vol}))\, I, \text{ and} \tag{5.38b}$$

$$\widetilde{\tau} \text{ is } c\text{-sub-Gaussian for a universal constant } c. \tag{5.38c}$$

Taking these claims as given for the moment, let us prove the lemma.

The claims (5.38a) and (5.38c) taken together imply that the random variable τ is sub-Gaussian with ψ_2 parameter $\zeta^2 \leq 2c^2 + 2C \log(1/\mathrm{vol})$. Now consider m i.i.d. draws of τ given by $\{\tau_i\}_{i=1}^m$; standard results (see, e.g., Vershynin [319, Remark 5.40], or Wainwright [323, Theorem 6.2]) yield the bound

$$\Pr\left\{ \left\|\left\| \frac{1}{m} \sum_{i=1}^{m} \tau_i \tau_i^\top - \Sigma_\tau \right\|\right\|_{\mathrm{op}} \geq \zeta^2 \left(\frac{d}{m} + \sqrt{\frac{d}{m}} + \delta \right) \right\} \leq 2 \exp\left(-cn \min\{\delta, \delta^2\} \right).$$

Using this bound along with claim (5.38b) and Weyl's inequality yields

$$\lambda_{\min}\left(\frac{1}{m}\sum_{i=1}^{m}\tau_i\tau_i^\top\right) \geq C\,\mathrm{vol}^2 - \zeta^2\left(\frac{d}{m}+\sqrt{\frac{d}{m}}+\delta\right) \tag{5.39}$$

with probability greater than $1-2\exp\left(-cn\min\{\delta,\delta^2\}\right)$.

Furthermore, when n samples are drawn from a standard Gaussian distribution, the number m of them that fall in the set K satisfies $m \geq \frac{1}{2}n\cdot\mathrm{vol}$ with high probability. In particular, this follows from a straightforward binomial tail bound, which yields

$$\Pr\left\{m \leq \frac{n\cdot\mathrm{vol}}{2}\right\} \leq \exp(-cn\cdot\mathrm{vol}). \tag{5.40}$$

Recall our choice $n \geq Cd\frac{\log^2(1/\mathrm{vol})}{\mathrm{vol}^3}$, which in conjunction with the bound (5.40) ensures that $C\,\mathrm{vol}^2 \geq \frac{1}{8}\sigma^2\sqrt{\frac{d}{m}}$ with high probability. Setting $\delta = C\,\mathrm{vol}^2/\sigma^2$ in inequality (5.39), we have

$$\lambda_{\min}\left(\frac{1}{m}\sum_{i=1}^{m}\tau_i\tau_i^\top\right) \geq \frac{C}{2}\,\mathrm{vol}^2$$

with probability greater than $1-2\exp\left(-cn\,\mathrm{vol}^4/\sigma^4\right)$. Putting together the pieces thus proves the lemma. It remains to show the various claims. □

Proof of claim (5.38a) Let $\tau_{\mathcal{A}}$ denote a random variable formed as a result of truncating the Gaussian distribution to a (general) set $\mathcal{A}$ with volume vol. Letting $\mu_{\mathcal{A}}$ denote its mean, the dual norm definition of the ℓ_2 norm yields

$$\begin{aligned}\|\mu_{\mathcal{A}}\| &= \sup_{v\in\mathbb{S}^{d-1}}\langle v,\,\mu_{\mathcal{A}}\rangle\\ &\leq \sup_{v\in\mathbb{S}^{d-1}}\mathbb{E}|\langle v,\,\tau_{\mathcal{A}}\rangle|.\end{aligned}$$

Let us now evaluate an upper bound on the quantity $\mathbb{E}|\langle v,\,\tau_{\mathcal{A}}\rangle|$. In the calculation, for any d-dimensional vector y, we use the shorthand $y_v := v^\top y$ and $y_{\backslash v} := U_{\backslash v}^\top y$ for a matrix $U_{\backslash v} \in \mathbb{R}^{d\times(d-1)}$ having orthonormal columns that span the subspace orthogonal to v. Letting $\mathcal{A}_v \subseteq \mathbb{R}$ denote the projection of $\mathcal{A}$ onto the direction v, define the set $\mathcal{A}_{\backslash v}(w) \subseteq \mathbb{R}^{d-1}$ via

$$\mathcal{A}_{\backslash v}(w) = \{y_{\backslash v} \in \mathbb{R}^{d-1} : y \in \mathcal{A} \text{ and } y_v = w\}.$$

Letting ψ_d denote the d-dimensional standard Gaussian pdf, we have

$$\begin{aligned}\mathbb{E}|\langle v,\, \tau_{\mathcal{A}}\rangle| &= \frac{1}{\mathrm{vol}}\int_{y\in\mathcal{A}} |y^\top v|\psi_d(y)dy \\ &= \frac{1}{\mathrm{vol}}\int_{y\in\mathcal{A}} |y_v|\psi(y_v)\psi_{d-1}(y_{\setminus v})dy \\ &= \frac{1}{\mathrm{vol}}\int_{y_v\in\mathcal{A}_v} |y_v|\psi(y_v)\underbrace{\left(\int_{y_{\setminus v}\in\mathcal{A}_{\setminus v}(y_v)} \psi_{d-1}(y_{\setminus v}\in\mathcal{A}_{\setminus v}(y_v))dy_{\setminus v}\right)}_{f(y_v)} dy_v \\ &\overset{\text{(i)}}{\leq} \frac{1}{\mathrm{vol}}\int_{y_v\in\mathcal{A}_v} |y_v|\psi(y_v)dy_v, \end{aligned} \tag{5.41}$$

where step (i) follows since $f(y_v) \leq 1$ point-wise. On the other hand, we have

$$\mathrm{vol} = \int_{y_v\in\mathcal{A}_v} \psi(y_v)\left(\int_{y_{\setminus v}\in\mathcal{A}_{\setminus v}(y_v)} \psi_{d-1}dy_{\setminus v}\right)dy_v \leq \int_{y_v\in\mathcal{A}_v}\psi(y_v)dy_v. \tag{5.42}$$

Combining inequalities (5.41) and (5.42) and letting $w = y_v$, an upper bound on $\|\mu_\tau\|$ can be obtained by solving the one-dimensional problem given by

$$\begin{aligned}\|\mu_\tau\| \leq &\sup_{\mathcal{S}\subseteq\mathbb{R}} \frac{1}{\mathrm{vol}}\int_{w\in\mathcal{S}} |w|\psi(w)dw \\ &\text{s.t.} \int_{w\in\mathcal{S}}\psi(w)dw \geq \mathrm{vol}\,. \end{aligned}$$

It can be verified that the optimal solution to the problem above is given by choosing the truncation set $\mathcal{S} = (\infty, -\beta) \cup [\beta, \infty)$ for some threshold $\beta > 0$. With this choice, the constraint can be written as

$$\mathrm{vol} \leq \int_{|w|\geq\beta}\psi(w)dw \leq 2\sqrt{\frac{2}{\pi}}\frac{1}{\beta}e^{-\beta^2/2},$$

where we have used a standard Gaussian tail bound. Simplifying yields the bound

$$\beta \leq 2\sqrt{\log(C/\,\mathrm{vol})}.$$

Furthermore, we have

$$\begin{aligned}\frac{1}{\mathrm{vol}}\int_{|w|\geq\beta}|w|\psi(w)dw &= \frac{C}{\mathrm{vol}}e^{-\beta^2/2} \\ &\overset{\text{(ii)}}{\lesssim} \frac{\beta^3}{\beta^2 - 1} \\ &\leq c\sqrt{\log(1/\,\mathrm{vol})}, \end{aligned}$$

where step (ii) follows from the bound $\Pr\{Z \geq z\} \geq \psi(z)\left(\frac{1}{z} - \frac{1}{z^3}\right)$ valid for a standard Gaussian variate Z. Putting together the pieces, we have

$$\|\mu_\tau\|^2 \leq c \log(1/\operatorname{vol}).$$

□

Proof of claim (5.38b) Let us first show the upper bound. Writing $\operatorname{cov}(\tau)$ for the covariance matrix, we have

$$\begin{aligned}
|\!|\!|\Sigma_\tau|\!|\!|_{\text{op}} &\leq |\!|\!|\operatorname{cov}(\tau)|\!|\!|_{\text{op}} + \|\mu_\tau\|^2 \\
&\overset{\text{(iii)}}{\leq} |\!|\!|I|\!|\!|_{\text{op}} + C \log(1/\operatorname{vol}),
\end{aligned}$$

where step (iii) follows from the fact that $\operatorname{cov}(\tau) \preceq \operatorname{cov}(Z)$, since truncating a Gaussian to a convex set reduces its variance along all directions [163, 317].

We now proceed to the lower bound. Let $\mathbb{P}_K$ denote the Gaussian distribution truncated to the set K. Recall that we denoted the probability that a Gaussian random variable falls in the set K by $\operatorname{vol}(K)$; use the shorthand $\operatorname{vol} = \operatorname{vol}(K)$. Define the polynomial

$$p_u(x) = \langle x - \mathbb{E}_{X \sim \mathbb{P}_K}[X],\, u\rangle^2;$$

note that we are interested in a lower bound on $\inf_{u \in \mathbb{S}^{d-1}} \mathbb{E}_{X \sim \mathbb{P}_K}[p_u(X)]$.

For $\delta > 0$, define the set

$$S_\delta := \{x \in \mathbb{R}^d : p_u(x) \leq \delta\} \subseteq \mathbb{R}^d.$$

Letting Z denote a d-dimensional standard Gaussian random vector and using the shorthand $\alpha := \mathbb{E}_{X \sim \mathbb{P}_K}[X]$, we have

$$\begin{aligned}
\Pr\{Z \in S_\delta\} &= \Pr\left\{\langle Z - \alpha,\, u\rangle^2 \leq \delta\right\} &\quad (5.43)\\
&= \Pr\left\{\langle \alpha,\, u\rangle - \sqrt{\delta} \leq \langle Z,\, u\rangle \leq \langle \alpha,\, u\rangle + \sqrt{\delta}\right\} &\quad (5.44)\\
&= \int_{\langle \alpha,\, u\rangle - \sqrt{\delta}}^{\langle \alpha,\, u\rangle + \sqrt{\delta}} \psi(x) dx \leq \sqrt{\frac{2}{\pi}\delta}, &\quad (5.45)
\end{aligned}$$

where in the final step, we have used the fact that $\psi(x) \leq 1/\sqrt{2\pi}$ for all $x \in \mathbb{R}$.

Consequently, we have

$$\begin{aligned}\mathbb{E}_{X \sim \mathbb{P}_K}[p_u(X)] &= \frac{1}{\mathrm{vol}} \mathbb{E}_Z \left[p_u(Z) \mathbf{1} \left\{ Z \in K \right\} \right] \\ &\geq \frac{1}{\mathrm{vol}} \mathbb{E}_Z \left[p_u(Z) \mathbf{1} \left\{ Z \in K \cap S_\delta^c \right\} \right] \\ &\overset{\text{(iv)}}{\geq} \frac{1}{\mathrm{vol}} \mathbb{E}_Z \left[\delta \mathbf{1} \left\{ Z \in K \cap S_\delta^c \right\} \right] \\ &= \frac{\delta}{\mathrm{vol}} \Pr\{ Z \in K \cap S_\delta^c \} \\ &\overset{\text{(v)}}{\geq} \delta \frac{\mathrm{vol} - \sqrt{\frac{2}{\pi}} \delta}{\mathrm{vol}}.\end{aligned}$$

Here, step (iv) follows from the definition of the set S_δ, which ensures that $p_u(x) \geq \delta$ for all $x \in S_\delta^c$. Step (v) follows as a consequence of equation (5.45), since

$$\Pr\{ Z \in K \cap S_\delta^c \} = \Pr\{ Z \in K \} - \Pr\{ Z \in S_\delta \} \geq \mathrm{vol} - \sqrt{\frac{2}{\pi}} \delta.$$

Finally, choosing $\delta = c\,\mathrm{vol}^2$ for a suitably small constant c, we have $\mathbb{E}_{X \sim \mathbb{P}_K}[p_u(X)] \geq C\,\mathrm{vol}^2$ for a fixed $u \in \mathbb{S}^{d-1}$. Since u was chosen arbitrarily, this proves the claim. □

Proof of claim (5.38c) Since the random variable ξ is obtained by truncating a Gaussian random variable to a convex set, it is 1-strongly log-concave. Thus, standard results [190, Theorem 2.15] show that the random variable $\widetilde{\xi}$ is c-sub-Gaussian. □

Proof of Lemma 5.4.4

For a pair of $d + 1$-dimensional vectors (v, v'), denote by

$$n_{W(v,v')} = \#\{ i : x_i \in W(v, v') \} \tag{5.46}$$

the random variable that counts the number of points that fall within the wedge $W(v, v')$; recall our notation W_δ for the set of all wedges with Gaussian volume less than δ. Since each wedge is formed by the intersection of two hyperplanes, applying Lemmas B.1.2 and B.1.3 in conjunction yields that there are universal constants (c, c', C) such that

$$\sup_{W \in \mathcal{W}_\delta} n_W \leq c \delta n \tag{5.47}$$

with probability exceeding $1 - \exp(-c' n \delta^2)$, provided $n \geq \frac{C}{\delta^2} d \log(n/d)$. In words, the maximum number of points that fall in *any* wedge of volume δ is linear in δn with high probability.

It thus suffices to bound, simultaneously, the maximum singular value of every sub-matrix of Ξ having (at most) $c\delta n$ rows. For a *fixed* subset S of size $c\delta n$, standard bounds for Gaussian random matrices (see, e.g., [319])) yield the bound

$$\lambda_{\max}\left(\sum_{i\in S}\xi_i\xi_i^T\right) \leq c_1(\delta n + d + t)$$

with probability exceeding $1 - 2\exp(-C_1 t)$.

Furthermore, there are at most $c\delta n \cdot \binom{n}{c\delta n}$ subsets of size at most $c\delta n$; taking a union bound over all such subsets yields the bound

$$\Pr\left\{\max_{S:|S|\leq c\delta n}\lambda_{\max}\left(\sum_{i\in S}\xi_i\xi_i^\top\right) \geq c_1(\delta n + d + t)\right\} \leq 2c\delta n \cdot \binom{n}{c\delta n}\exp(-C_1 t).$$

Making the choice $t = 2c\delta n\log(1/c\delta)$ and putting together the pieces proves the lemma. □

5.5 Summary and open questions

We conclude this portion of the paper with short discussions of related models and future directions.

5.5.1 Related models

Models closely related to (5.1) also appear in second price auctions, where an item having d features is bid on and sold to the highest bidder at the second highest bid [221, 227]. Assuming that each of k user groups bids on an item and that each bid is a linear function of the features, one can use a variant of the model (5.1) with the $\max$ function replaced by the second order statistic to estimate the individual bids of the user groups based on historical data. Another related problem is that of multi-class classification [75], in which one of k labels is assigned to each sample based on the argmax function, i.e., for a class of functions $\mathcal{F}$, we have the model $Y = \operatorname{argmax}_{1\leq j\leq k} f_j(X)$ for j distinct functions $f_1, \ldots, f_k \in \mathcal{F}$. When $\mathcal{F}$ is the class of linear functions based on d features, this can be viewed as the "classification" variant of our regression problem.

While the connection of the max-affine model to multi-index models was discussed extensively in Chapter 1, the model (5.1) can also be seen as a special case of mixture-of-experts models [151]. In the mixture-of-experts model, the covariate space is partitioned into k regions via certain *gating functions*, and the observation model is given by k distinct regression functions: one on each region. The model (5.1) is clearly a member of this class, since the $\max(\cdot)$ function implicitly defines a partition of $\mathbb{R}^d$ depending on which of the k linear functions of X attains the maximum, and on each of these partitions, the regression function is linear in X.

5.5.2 Future directions

In this chapter, we analyzed a natural alternating minimization algorithm for estimating the maximum of unknown affine functions, and established that it enjoys local linear convergence to a ball around the optimal parameters. We also proposed an initialization based on PCA followed by random search in a lower-dimensional space. An interesting open question is if there are other efficient methods besides random search that work just as well post dimensionality reduction. Another interesting question has to do with the necessity of dimensionality reduction: in simulations (see the full paper [119]), we have observed that if the AM algorithm is repeatedly initialized in $(d+1)$-dimensional space without dimensionality reduction, then the number of repetitions required to obtain an initialization from which it succeeds (with high probability) is similar to the number of repetitions required after dimensionality reduction. This suggests that our (sufficient) initialization condition (5.14a) may be too stringent, and that the necessary conditions on the initialization to ensure convergence of the AM algorithm are actually much weaker. We leave such a characterization for future work, but note that some such conditions must exist: the AM algorithm when run from a single random initialization, for instance, fails with constant probability when $k \geq 3$. Understanding the behavior of the randomly initialized AM algorithm is also an open problem in the context of phase retrieval [327, 348].

In the broader context of max-affine estimation, it is also interesting to analyze other non-convex procedures (e.g. gradient descent) to obtain conditions under which they obtain accurate parameter estimates. The CAP estimator of Hannah and Dunson [139] and the adaptive max-affine partitioning algorithm of Balázs [15] are also interesting procedures for estimation under these models, and it would be interesting to analyze their performance when the number of affine pieces k is fixed and known. For applications in which the dimension d is very large, it is also interesting to study the model with additional restrictions of sparsity on the unknown parameters.

In the context of this dissertation, this chapter demonstrates a computationally efficient method that enjoys statistical estimation guarantees while solving a non-convex optimization problem. Our investigation is also motivated in part by the fact the convex LSE (5.6) does not adapt to piecewise affine structure for any $d \geq 5$ [180], and in that sense, the methodology introduced and analyzed in this chapter should be viewed as performing tractable statistical estimation over simpler sub-models when the natural procedure for the overall model—convex regression MLE in this case—does not come with any adaptation guarantees. In the next chapter, we will show that a close relative of the alternating minimization heuristic analyzed here has favorable properties even in semiparametric index models; in particular, that it adapts to the noise level in a subclass of single-index models.

Chapter 6

Adapting to noise level in semiparametric estimation

In classical nonparametric regression, we are interested in modeling the relationship between a d-dimensional covariate x and a scalar response y through a function $f : \mathbb{R}^d \mapsto \mathbb{R}$ that satisfies some regularity conditions. However, as alluded to briefly in the previous chapter, standard nonparametric function classes in high dimensions are extremely expressive and require prohibitively many samples—exponential in the dimension—to learn (see e.g., Tsybakov [302]). A popular dimensionality reduction technique is to assume that the function f is given by the composition of a lower-dimensional function $h : \mathbb{R}^k \mapsto \mathbb{R}$ with a linear model. Formally, we have

$$f(x) = h(\langle \theta_1, x \rangle, \langle \theta_2, x \rangle, \ldots, \langle \theta_k, x \rangle),$$

where the d-dimensional regression coefficients $\theta_1, \ldots, \theta_k$ span a k-dimensional subspace for some $k \ll d$. Such models are called *multi-index models*, since the functional relationship can be captured by a few *indices* that represent particular directions of the covariate space. Indeed, we saw an instance of such a model in the previous chapter where the nonlinear function h was known, and given by the $\max$ function. In this chapter, our focus will be on the *semiparametric* setting in which the nonlinearity is only known to belong to a class of functions.

6.1 Introduction

In this chapter, we focus on the special case of the multi-index model where $k = 1$, which results in the *single-index model*

$$y = g^*(\langle \theta^*, x \rangle) + \epsilon, \tag{6.1}$$

or SIM for short. Here g^* is a *univariate*, nonparametric link function, $\theta^* \in \mathbb{R}^d$ is the salient linear predictor, and ϵ is a random variable independent of everything else that captures the noise in the modeling process. The model (6.1) between the covariate and response should be seen as one of the most basic forms of non-linear dimensionality reduction, and as a step towards the broader goal of representation learning or feature engineering. In order to facilitate a concrete theoretical study, we also assume in this chapter that the covariates x are drawn from a normal distribution, and that the noise ϵ is sub-Gaussian with parameter σ; these are standard assumptions in many parts of the literature [45, 89, 195].

As stated, the single-index model (6.1) is classical, and there is an extensive body of literature spanning the statistics, econometrics, and geometric functional analysis communities that is dedicated to studying many aspects of this model. We provide an extensive survey of this literature in Section 6.1.2 to follow. For now, let us focus on the recent paper by Plan and Versyhnin [253], which studies the problem under further geometric constraints on the parameter θ^* and represents, to an extent, the state-of-the-art progress on this problem. Upon analyzing a moment-based method—whose roots go back to the classical work of Brillinger [45]—for recovering the "signal" θ^*, they point out that it is not necessary to explicitly model the non-linear link function g^*. To quote portions of their text:

> *"This leads to the intriguing conclusion that in the* **high noise** *regime, an unknown non-linearity in the observations does not significantly reduce one's ability to determine the signal... even when the non-linearity is not explicitly modeled."*

This surprising claim is somewhat counter-intuitive: after all, obtaining a "good" model for the function g^* should help in the estimation task, and this intuition has largely guided the extensive sub-field of *generalized linear modeling* [217], in which we assume the function g^* is known exactly. More generally, there ought to exist a trade-off between the approximation and estimation errors for this class of problems: on the one hand, we incur a certain approximation error (or bias) by treating the unknown function as linear, and our estimation error (or variance) behaves as though the true model is linear. The results of Plan and Vershynin—and of many other preceding papers in this general area—are intriguing because they show that for a large enough noise level, and provided that the function g^* is not "orthogonal" to the class of linear functions, a biased estimator for the parameter achieves error that is optimal up to constant factor, since the bias is of a smaller order than the variance.

On the other hand, one could instead ask what happens in the low noise, or *high signal* regime[1], when the errors made due to modeling the non-linear function as linear are no longer of the same order as the noise. Indeed, such a question is motivated by applications in which we often have significant side-information that allows us posit some function class $\mathcal{G}$ to which g^* belongs. By building better models for the non-linearity, it would stand to reason that the bias can be reduced and finally eliminated when $\mathcal{G} \supseteq g^*$. The major motivation for this chapter is to understand this phenomenon in quantitative terms. We study this issue in the context of parameter recovery, i.e., recovering θ^* from n i.i.d. samples drawn from the model. From a statistical perspective, we would like to derive precise bounds on the recovery error as a function of both the dimension d

[1] A natural measure of signal-to-noise ratio in the problem is given by $\|\theta^*\|/\sigma$. We set $\|\theta^*\| = 1$ for identifiability in the single-index model, and so the low noise regime in which $\sigma \to 0$ corresponds to high signal-to-noise ratio. Thus, we use the terms 'low noise' and 'high signal' interchangeably in this chapter.

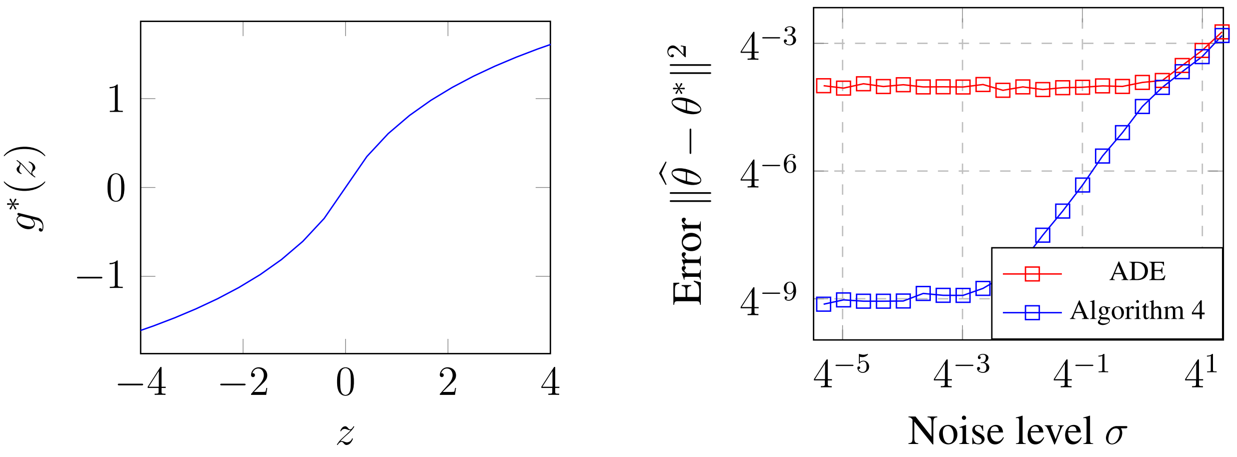


Figure 6.1: **Left:** The unknown, monotone function $g^*(z) = \mathsf{sgn}(z) \cdot \log(1 + |z|)$ used in the simulation. We collected i.i.d. samples from the single-index model defined by this function, corrupted by Gaussian noise of variance σ^2. **Right:** Simulations of the error of parameter estimation plotted against the noise level σ for (a) in red, the standard Average Derivative Estimator (ADE) [45, 254] and (b) in blue, our refined estimator from Algorithm 4 that employs the least squares estimator over monotone functions as the non-parametric estimate of the 'inverse' function. In this experiment, we set $p = 20$ and $n = 5000$, and the errors are averaged over 50 independent runs of the respective algorithms. The generalized Lasso estimator of Plan and Vershynin [253] also has very similar performance to the ADE method and so its error is not plotted here.

and the sample size n. In addition, we would like to be able to accomplish the estimation task in a computationally efficient manner, and by using fine-grained properties about the function class $\mathcal{G}$. For a comparison of our approach and motivation with those of related work, see Section 6.1.2. Overall, our approach formalizes a complementary notion to that articulated by Plan and Vershynin above. In particular, we show that when $g^* \in \mathcal{G}$, then leveraging certain structural properties of the class $\mathcal{G}$ through a natural, iterative algorithm can lead to uniformly faster rates of estimation for all noise levels. In particular, significant gains are obtainable in the high signal regime by methodology that *automatically adapts* to the noise level of the problem.

To foreshadow our results, let us illustrate in a simulation the quantitative benefit of using our iterative framework for a sample link function. Figure 6.1 plots the performance of our procedure along with a classical semiparametric estimate as a function of the noise parameter σ. In particular, while standard algorithms see a large error floor even as $\sigma \to 0$, our estimator achieves asymptotically better error in the high signal (or low noise) regime, while remaining competitive with the classical approach even for larger values of σ. It is also worth noting that even so, the error achieved by our estimator plateaus for small values of σ, leading to a non-zero *error floor* of the problem. This motivates our study of the special case $\sigma = 0$, which we show serves as a proxy for all values of σ that are "sufficiently small".

In the literature on statistical learning theory, the low-noise regime has received considerable attention for empirical risk minimization applied to regression problems (e.g., [189, 222]). In particular, Mendelson [222], noted that classical analyses of ERM are often very conservative in the low noise (or what is referred to in learning theory as the nearly-realizable) setting. He proposed a new "small-ball" method of analysis to derive rates for the problem that are usually *much* faster when the model is nearly-realizable. Our motivation should be viewed as analogous but as applied to semiparametric regression[2]. The low-noise regime has also been extensively studied in the literature on statistical signal processing, but as applied to specific models such as phase retrieval (in which g^* is the absolute value or square function) and its relatives. For the related (noisy) matrix completion problem, the recent paper [62] provides an analysis of a popular convex relaxation method in the low-noise regime via a delicate analysis of a non-convex optimization algorithm.

We now discuss our contributions in a little bit more detail in Section 6.1.1, before providing a survey of related work and applications in Section 6.1.2.

6.1.1 Contributions and Organization

Our approach to performing estimation under the single-index model is based on leveraging fine-grained structure in the function g^*, and we formalize the notion of structure that we require by assuming access to a certain "labeling" oracle that provides information about the "inverse" model $y \mapsto \mathbb{E}[\langle x,\, \theta^* \rangle | y]$. Loosely speaking, the labeling oracle helps us narrow our investigation to regions of the domain of the function g^* on which this conditional expectation is easy to reason about. This provides, in broad terms, a program for estimation under the semiparametric model (6.1) via a reduction to nonparametric regression over the inverse function class. While we omit it here for brevity, this intuition is illustrated in the full paper [243] via a warm-up exercise on the phase retrieval problem, where we implement a labeling oracle and reduce the problem to linear regression. This leads to a simple algorithm for phase retrieval that achieves optimal parameter estimation rates.

We present in Section 6.2 the precise labeling oracle that we require for general single-index models, and provide a flexible procedure for parameter estimation. This procedure assumes access to the labeling oracle and solves a non-convex problem via an iterative algorithm. It requires, as input, any nonparametric function estimation oracle over the inverse function class. Under standard assumptions on the observation model, our general result (Theorem 6.2.1) provides guarantees on the error attained by such an iterative algorithm for general SIMs as a function of the error rate of the nonparametric estimator provided as input to the procedure. We then leverage the vast literature on empirical risk minimization (ERM) for nonparametric function estimation in order to establish Theorem 6.4.1, which shows upper bounds on parameter estimation in terms of natural measures of complexity of the class of inverse functions.

In order to illustrate a concrete application of our framework, we consider a sub-class of monotone SIMs for which a labeling oracle can be implemented efficiently, and with no additional computational effort. This leads to a procedure with end-to-end guarantees for this class, which we present as Corollary 6.3.1. This result provides a sharper parameter estimate than classical procedures, and the gains are particularly significant when $\sigma \to 0$.

[2] Indeed, Mendelson's general techniques are also applicable to this problem via the reduction that we establish; see the full paper [243] for a discussion.

6.1.2 Related work

Single-index models have seen a concrete theoretical treatment across multiple, related communities. The classical viewpoint emerges from the statistics community, in which these have been studied under the broader umbrella of semiparametric estimation; the latter is broadly applied in microeconomics, finance, and the social sciences. Index models, in particular, have been used as a general-purpose, non-linear dimensionality reduction tool. We refer the the interested reader to the books by Bickel et al. [33] and Li and Racine [197] for a broad overview of classical methods for semiparametric estimation, their applications, and associated guarantees. In the context of the single-index model, a well known estimator for the index vector is the semiparametric maximum likelihood estimator (SMLE) [145], which solves the full-blown M-estimation problem, finding the function, index pair that maximizes the likelihood of the observed samples. The SMLE is known to have excellent statistical properties in the asymptotic regime where the ambient dimension is fixed and the number of samples goes to infinity—in particular, a parameter estimate obtained as a result of running these procedures is often "$\sqrt{n}$-consistent"—and succeeds with minimal assumptions on the covariate distribution [176, 266]. In addition to the SMLE, other influential approaches include gradient-based estimators [73, 146], moment-based estimators [45, 195], and slicing estimators [196], which have driven a lot of progress in the deployment of semiparametric models in practice. There has also been recent interest in studying some of these procedures under weak covariate assumptions [8]. Indeed, our general approach can be viewed as a more refined version of slicing; this is discussed in detail in Remark 6.2.2. We also note the recent work of Dudeja and Hsu, which falls under this broad umbrella and analyzes the single-index model with Gaussian covariates by expressing the unknown function in the Hermite polynomial basis. Their estimators may be viewed as higher-order moment methods, and they propose efficient, gradient-based algorithms to compute them.

There has also been a lot of recent interest in applying the double (or de-biased) machine learning approach to semiparametric models,especially in the high-dimensional regime [66, 67]. These papers are motivated by the fact that semiparametric estimation is a natural lens through which to view estimation problems with *nuisance* components, when the statistician is only interested in some *target* component; examples of such problems span the diverse fields of treatment effect estimation, policy learning, and domain adaptation. The classical notion of *Neyman orthogonality* [238] has re-emerged as a natural and flexible condition under which to study these problems. We do not survey this literature in detail, but refer the reader to the recent paper by Foster and Syrgkanis [108], which provides a general treatment of problems in this space. Focusing on proving excess risk bounds for problems with a nuisance component, these results show that a natural one-step meta-algorithm that splits samples between estimating the nuisance component and the target component (or parameter) is able to achieve oracle excess risk bounds in some settings. In particular, they show that if a Neyman orthogonality condition is satisfied and the class of nuisance components is not too large when compared to the target class, then *oracle risk* bounds[3] are achievable. The generality of these results is striking: they apply to a general class of problems, general loss functions, and general

[3] That is, the excess risk of estimating the target is of the same order as the risk attainable if the nuisance component were known exactly

data distributions, thereby providing a broad framework for the study of such models. Notably, the results are also reduction-based, in that they allow the statistician to use *any* procedure for estimation of the target and nuisance components, and derive bounds that depend on the rates at which these components can be estimated. In this last respect, our treatment is similar; however, our focus should be viewed as being complementary to this general theory. Some salient differences are worth highlighting: First, and foremost, we are interested primarily in understanding the rates of estimation as a function of the noise level in the problem, which was not the focus of these recent results. In particular, any one-step meta-algorithm will no longer be optimal (even in the special case of SIMs) over all noise levels. Second, we are interested in the rates of parameter estimation as in the semiparametric literature, and this requires us to impose stronger covariate assumptions. Finally, by specializing our model class to single-index models, we are able to simultaneously address issues of computational efficiency, statistical optimality, and adaptivity to the noise level.

A second perspective on single-index models emerges from the statistical signal processing literature[4]—or more broadly, the literature on geometric functional analysis and linear inverse problems—in which we are interested in imposing additional *structure* on the underlying parameter θ^*. While the application of geometric functional analysis to linear inverse problems is a relatively recent endeavor, the literature in this general space is already quite formidable; examples of results here can be found in the papers [110, 237, 253, 254, 296, 297, 339, 340, 342]. The focus in this area is on recovering the underlying "signal" θ^* at a rate that depends optimally on the properties of the set to which the signal belongs. This literature often places stronger assumptions on the measurements or covariates—often Gaussian, although some extensions to sub-Gaussian settings are available (e.g. [223]). Many of the algorithms in this space are based on convex relaxations, but in the case where there is no structure on θ^*, they reduce to more classical moment-based estimators. As mentioned in Section 6.1, a representative result in this space is that of Plan and Vershynin, which shows that provided the unknown link function has a non-zero "projection" onto the class of linear functions, a constrained variant of the standard (linear) least squares estimator recovers the true parameter at the optimal rate for large noise levels; in particular, this error rate depends precisely on the geometric properties of the set to which θ^* belongs. Extensions of this result are also available for cases when g^* is an even function [297], and are based on constrained versions of the Principal Hessian Directions (PHD) algorithm [195]. Besides convex relaxation approaches, there are also non-convex approaches to problems in this space; for example, Yang et al. [341] study a two-step non-convex optimization procedure for SIMs, and show that this algorithm is able to obtain a parameter estimate at the optimal $\frac{s \log d}{n}$ rate for s-sparse vectors θ^* under moment conditions on the link function.

Given that we specialize some of our results in the sequel to the class of monotone SIMs, let us now discuss some prior work in this space. The design of efficient algorithms for monotone single-index models was the focus of much work in the machine learning community [161, 162], where these models were introduced in order to account for mis-specification in generalized linear models with known link functions. The algorithms here—Isotron [162] and variants [161]—are inspired by

[4] Our division of related work under these two broad headings is somewhat arbitrary; the motivations of some of the papers listed in the geometric functional analysis literature were statistical, and vice versa.

the Perceptron algorithm and run variants of the stochastic gradient method. They obtain bounds on the excess risk incurred by the algorithm, showing bounds that are typically nonparametric. These models have also seen a more recent appearance in the literature on shape-constrained estimation, in which index-models and their relatives have emerged as natural means to alleviate the curse of dimensionality [12, 61, 178]. Broadly speaking, these papers analyze the consistency of the global SMLE for their respective problems, and propose heuristic algorithms—without provable guarantees—that solve this non-convex problem by alternating projection procedures. It should be noted that in the absence of smoothness assumptions, there are a multitude of technical obstacles that must be overcome in order to show that the SMLE is even consistent. The monotone single-index model, in particular, has been analyzed in recent papers by Balabdaoui et al. [12] and Groeneboom and Hendrickx [126]. In addition to providing fine-grained guarantees for the SMLE (e.g., the limiting distribution of the regression estimate at a point [127], or the prediction error of the "bundled" function $g^*(\langle\theta^*, \cdot\rangle)$), these papers also provide guarantees for the ADE approach and their guarantees hold under minimal assumptions on the underlying link function.

Having discussed the lay of the land, let us now put our contributions in context. In spite of the vast literature on single-index models, some important and fundamental questions remain unaddressed. In particular, our focus is on simultaneously tackling the following issues:

- **Leveraging structure in the class of link functions:** Moment and slicing based estimators, which form the foundation for the investigation of SIMs in the literature on linear inverse problems, completely ignore any fine-grained structure in the true function g^*. As alluded to earlier, they simply require g^* to obey certain moment conditions, and do not attempt to model it in any way. This leads to a "bias" in these estimators that becomes significant in the high signal regime, and indicates that better models for g^* can be leveraged to reduce this bias.

- **Adapting to the noise level:** As alluded to in the introduction, none of the computationally efficient estimators of θ^* obtain a provably optimal error bound as a function of the noise variance σ^2. In particular, the performance of estimators in the low noise setting is near-identical to their performance in the constant-noise setting. Take, for example, the recent results of Babichev and Bach [8] and Dudeja and Hsu [89], which show a bound of the form

$$\|\widehat{\theta} - \theta^*\|^2 \lesssim (\sigma^2 + c)\frac{d}{n} \tag{6.2}$$

  for their respective estimators, provided the function g^* satisfies certain conditions. The $\lesssim$ notation in these bounds hides logarithmic factors in the pair (d, n), and the constant c in this bound is some problem dependent constant that is strictly positive for any non-linear g^*. The analysis of Yang et al. [341] posits additional structure on the underlying parameter θ^* and improves the *rate* of the estimate (i.e., the dimension d in the bound (6.2) is replaced by a geometric quantity, but the $(\sigma^2 + c)$ term persists). Clearly, these bounds exhibit the same behavior for both large and small σ, and this is a limitation of these approaches that we would like to address. Adaptivity to noise variance is only achievable when we are able to drive the bias of the problem to zero at a faster rate by positing a good model for the function g^*.

- **Computational efficiency:** The SMLE, for instance, solves a non-convex problem to optimality and is NP-hard to compute for many nonparametric function classes. Variants of the SMLE are able to avoid some statistical issues with the SMLE, but they are still computationally intractable.
- **Dependence on the dimension:** Since a large portion of the semiparametric literature is classical, the dependence on the covariate dimension d is seldom made explicit. In many cases, this dependence is much worse than the linear dependence on d that we expect for parametric models.

Chapter-specific notation: Recall the notational convention introduced in Section 1.4. We complement this notation with a few other definitions that are used solely in this chapter and the corresponding technical proof section in Appendix B.2. We largely use capital letters X, Y, etc. to denote random variables/vectors, and small letters to denote their realizations, usually with the sample index x_i, y_i, etc. We reserve the notation Z for the standard Gaussian distribution, where the dimension can be inferred from context. Boldface capital letters $X, \mathbf{W}$, etc. are used to denote matrices; we let X^i denote the i-th column of X. We let $X^\dagger$ denote the Moore-Penrose pseudoinverse of a (tall) matrix X. For any positive integer p, we let $\mathbf{I}_p$ denote the $p \times p$ identity matrix. We deliberately eschew measure-theoretic considerations. Throughout, we write conditional expectations assuming that they exist. For a pair of continuous random variables (U, V) and a scalar u, we use the shorthand $\mathbb{E}[V|U = u]$ to denote the standard conditional expectation $\mathbb{E}[V|u]$.

6.2 Methodology and main result for general single index models

We now turn to the single-index model, which is the main focus of the paper. Throughout, we suppose that n samples drawn i.i.d. from the observation model

$$y_i = g^*(\langle x_i, \theta^* \rangle) + \epsilon_i, \tag{6.3}$$

once again assuming that $x_i \perp\!\!\!\perp \epsilon_i$, and that $x_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \mathbf{I}_d)$. We also assume that the noise ϵ_i is drawn from a σ-sub-Gaussian distribution and that the unknown parameter $\theta^* \in \mathbb{S}^{d-1}$ has unit norm. Assumptions on both the covariates and noise can be relaxed for subsets of our results, and we allude to this in Section 6.5. The univariate function is assumed to satisfy the inclusion $g^* \in \mathcal{G}$ for some nonparametric function class $\mathcal{G}$. Our procedure for parameter estimation in SIMs requires two natural *oracles*, which we introduce first.

6.2.1 Oracles: Labeling and Inverse Regression

As alluded to in the introduction, the first oracle that we require is a *labeling oracle*.

Labeling Oracle: Such an oracle outputs:

- A closed interval $\mathcal{I} \subseteq \mathbb{R}$, and a set of labeled samples $\mathcal{S} = \{i : \langle x_i, \theta^* \rangle \in \mathcal{I}\}$. Let W denote the truncation of the random variable $\langle X, \theta^* \rangle$ on this interval, having density

$$f_W(w) = \begin{cases} \frac{1}{\int_{x \in \mathcal{I}} \phi(x) dx} \phi(w) & \text{if } w \in \mathcal{I} \\ 0 & \text{otherwise,} \end{cases}$$

  where ϕ denotes the standard Gaussian density. Denote by $\mathcal{P}_Y$ the induced distribution on the response $Y = g^*(W) + \epsilon$, and let $\mathcal{Y}$ denote its sample space.

- A closed, convex[5] set $\mathcal{H}$ corresponding to the function class

$$\mathcal{H} \supseteq \left\{ y \mapsto \mathbb{E}[W|y] \;\middle|\; y = g(W) + \epsilon, g \in \mathcal{G} \right\}.$$

  In words, this contains functions mapping $\mathbb{R} \to \mathcal{I}$ that contains all conditional expectations under the "inverse" model. We use the shorthand h^* to denote the conditional expectation—which we refer to hereafter as the "inverse function"—formed when our observations are generated according to the link function g^*.

Note that in principle, outputting a set S via such a labeling oracle requires knowledge of the true parameter θ^*, which we are trying to estimate! However, in the sequel, we show an example of a class of single-index models for which this is also true. For now, assume that such a labeling oracle exists and let $N = |S|$ be the effective sample size that we work with. Note that N is, in principle, a random variable, but it will be helpful to think of it as a fixed integer for the rest of this section.

The "spirit" of the labeling oracle is to provide a region on which the "inverse" function is easy to reason about. With the labeling oracle in hand, note that the random variable W may be viewed as being generated according to the model

$$W = h^*(Y) + \xi, \tag{6.4}$$

where ξ is uncorrelated with $h^*(Y)$ by definition, and may be viewed as zero-mean noise. In the sequel, we use the convenient notation $\xi(y) = [W|Y = y] - \mathbb{E}[W|Y = y]$ to indicate that ξ depends on the realization y. The sample space of W is $\mathcal{I}$, and when the noise ϵ is supported on the entire real line, the sample space of Y is $\mathcal{Y} = \mathbb{R}$. We emphasize that in spite of how the labeling oracle above has been defined, we *do not* assume that we have access to realizations of the pair of random variables (W, ξ); one should view the observation model (6.4) simply as an analysis device.

The second oracle that we require is a nonparametric regression oracle over the function class $\mathcal{H}$.

[5] If the set $\mathcal{H}$ is not convex, then it suffices to work with its convex hull. More generally, we only require the set to be star-shaped around h^*, and if not, we can work with the star hull centered at h^*.

Inverse regression oracle: Our overall algorithm uses, as a black-box, an estimation procedure $\mathcal{A}$ over the function class $\mathcal{H}$. Given k i.i.d. samples drawn from a generic nonparametric regression model over the class $\mathcal{H}$, the procedure $\mathcal{A} : (\mathbb{R} \times \mathbb{R})^k \mapsto \mathcal{H}$ uses these samples to compute a function $\widehat{h} \in \mathcal{H}$ that optimizes some measure of fit to these samples. We place no restrictions (besides measurability) on such a procedure; our main result depends on the properties of the procedure through its "rate" function, introduced in Assumption 6.2.3.

With these two oracles in hand, we are now ready to present our procedure for parameter estimation in general SIMs. We denote the covariate distribution post-truncation (i.e., the distribution on the samples S) by $\mathcal{P}_X^{\mathcal{I}}$.

6.2.2 Reducing SIMs to regression: a meta-algorithm and its analysis

Our procedure is based on a natural alternating minimization principle applied iteratively for T steps. We begin by partitioning the N samples into $2T$ equal parts[6]. Denote such a partition by $\mathcal{D}_1, \ldots, \mathcal{D}_{2T}$; each of these sets has size $N/(2T)$ by construction. Our algorithm runs for T iterations; at each iteration, we use two of these data sets. Let us briefly describe iteration t of the algorithm, which uses the data sets $\mathcal{D}_{2t+1}$ and $\mathcal{D}_{2t+2}$.

On the first data set, we run the nonparametric procedure $\mathcal{A}$ on the set of pairs $(y_i, \langle x_i, \widehat{\theta}_t \rangle)_{i \in \mathcal{D}_{2t+1}}$, and form a function estimate $\widehat{h}_{t+1} \in \mathcal{H}$ such that $\widehat{h}_{t+1} = \mathcal{A}\left(\left(y_i, \langle x_i, \widehat{\theta}_t \rangle \right)_{i \in \mathcal{D}_{2t+1}} \right)$. In particular, we treat our *current* linear prediction $\langle x_i, \widehat{\theta}_t \rangle$ as a noisy observation of the true function evaluated at the point y_i. This is our minimization in the space of functions $\mathcal{H}$, through which we obtain an estimate of h^*. In order to intuitively reason about whether this step is sensible, consider the special case $\widehat{\theta}_t = \theta^*$. Here, the nonparametric procedure $\mathcal{A}$ obtains samples from the model $h^*(y_i) + \xi_i$ for each $i \in \mathcal{D}_{2t+1}$; these are simply noisy observations of the true function, and $\mathcal{A}$ is designed precisely to denoise these samples. On the other hand, if $\widehat{\theta}_t$ is close to θ^*, then we obtain samples from a similar model, but with some *additional noise*—our analysis will make this precise—that vanishes provided $\widehat{\theta}_t$ converges to θ^*.

With the function estimate $\widehat{h}_{t+1}$ in hand, we now turn to the second data set and run a linear regression. In particular, we regress $\left\{ \widehat{h}_{t+1}(y_i) \right\}_{i \in \mathcal{D}_{2t+2}}$ on the covariates $\{x_i\}_{i \in \mathcal{D}_{2t+2}}$ and obtain the linear parameter estimate $\widetilde{\theta}_{t+1}$. Finally, we output the normalized parameter estimate $\widehat{\theta}_{t+1} = \widetilde{\theta}_{t+1} / \|\widetilde{\theta}_{t+1}\|$ at the end of this iteration. Note that once again, one can reason about how sensible our linear regression step is by specializing to the case $\widehat{h}_{t+1} = h^*$; here, $h^*(y_i)$ is effectively a noisy sample of $\langle x_i, \theta^* \rangle$, and so we expect the linear regression to return an estimate that is close to θ^*. When $\widehat{h}_{t+1} \neq h^*$, this, once again, introduces additional noise in our observation process which vanishes when our function estimate $\widehat{h}_{t+1}$ converges to the true function h^*.

With this intuition—made concrete in the proof—we are then able to relate the error of parameter estimation at the next time step with the error at the current time step, and iterating this bound

[6] We assume that N is a multiple of $2T$ for simplicity.

allows us to improve upon the error of the initializer $\widehat{\theta}_0$. A formal description of the entire procedure is provided as Algorithm 4.

Algorithm 4: The LTI-SIM meta-algorithm with sample-splitting for the two regressions

```
Input: Data of $N$ samples $\{x_i, y_i\}_{i \in S}$ returned by the labeling oracle; nonparametric
       regression procedure $\mathcal{A}$; initial parameter $\widehat{\theta}_0$; number of iterations $T$.
Output: Final parameter estimate $\widehat{\theta}_T$.
Initialize $t \leftarrow 0$. Split the data into $2T$ equal portions indexed by $\mathcal{D}_1, \ldots, \mathcal{D}_{2T}$.
repeat
    Form the function estimate $\widehat{h}_{t+1} \in \mathcal{H}$ by computing
        $\widehat{h}_{t+1} = \mathcal{A}\left(\left(y_i, \langle x_i, \widehat{\theta}_t \rangle\right)_{i \in \mathcal{D}_{2t+1}}\right).$    (6.5)
    Letting $X_{t+1}$ denote the $\frac{N}{2T} \times d$ matrix with rows $\{x_i\}_{i \in \mathcal{D}_{2t+2}}$ and stacking up the
     responses $\left\{\widehat{h}_{t+1}(y_i)\right\}_{i \in \mathcal{D}_{2t+2}}$ in a vector $v$, compute
        $\widetilde{\theta}_{t+1} = X_{t+1}^{\dagger} v.$
    Compute the normalized parameter $\widehat{\theta}_{t+1} = \frac{\widetilde{\theta}_{t+1}}{\|\widetilde{\theta}_{t+1}\|}$.
until $t = T$;
Return $\widehat{\theta}_T$.
```

Note that we use two *separate* samples for the sub-steps of the algorithm in order to ensure that $\widehat{h}_{t+1}$ is independent of the samples used to perform the linear regression. In the full paper [243], we introduce and analyze a variant of the algorithm without sample-splitting in the special case $\sigma = 0$.

Remark 6.2.1 (LTI-SIM as alternating minimization)**.** *Note that for $X \sim \mathcal{P}_X^{\mathcal{I}}$, the observations obey the relation*

$$\langle \theta^*, X \rangle = h^*(Y) + \xi.$$

where ξ may be viewed as "noise" in the inverse problem. Thus, for a data set $\mathcal{D} \subseteq S$, it is reasonable to construct the loss function

$$\mathcal{L}_{\mathcal{D}}(\theta, h) := \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} (\langle \theta, x_i \rangle - h(y_i))^2,$$

and minimize it over the pair (θ, h) in order to obtain some measure of fit to the samples in the data set. However, this minimization is rendered non-convex by the constraint that the returned $\widehat{\theta}$

must be unit norm. Thus, step 2 of the LTI-SIM procedure may be viewed[7] *as minimizing this loss function over the function class $\mathcal{H}$, and steps 3 and 4 in conjunction as performing a minimization in parameter space.*

Remark 6.2.2 (Comparison with slicing estimators)**.** *Slicing estimators [8, 196] are based on the observation that for spherically symmetric distributions, the* conditional moments[8] $\mathbb{E}[X^{\otimes k}|Y]$ *capture properties of the true parameter θ^*. For instance, when $k = 1$, classical calculations show that under mild assumptions on g^*, the vector $\mathbb{E}[X|Y]$ aligns with the vector θ^* for almost every realization of Y. Thus, we may construct estimates of this conditional expectation from samples by* slicing *over y values, and this leads to a $\sqrt{n}$-consistent estimate for the parameter and is similar in many respects to the ADE procedure [45]. However, even when $\sigma = 0$, the randomness in the covariate X introduces noise in the empirical expectation, and so the error cannot decay at a rate faster than $\sqrt{n}$ even in this noiseless case.*

Algorithm 4 is also based on reasoning about a first-order conditional expectation, but relies on a model, provided by the labeling oracle, of further structure in the function $y \mapsto \mathbb{E}[W|Y = y]$. Intuitively, modeling this higher-order structure in conjunction with the first-order conditional expectation allows us to considerably refine the slicing estimate in an iterative fashion. The original slicing estimator can thus be used to provide a natural initialization $\widehat{\theta}_0$ for our procedure.

While our methodology is well-defined for any single index model in which we have access to a labeling oracle, our theoretical analysis of the algorithm requires the following assumptions.

Assumption 6.2.1. *The Gaussian volume of the set $\mathcal{I}$ is greater than κ, i.e.,* $\Pr\{Z \in \mathcal{I}\} \geq \kappa$ *for* $Z \in \mathcal{N}(0, 1)$.

Such an assumption is natural, and guarantees that we have a large enough "effective sample size", with N growing directly proportional to the true sample size n. For our next assumption, we require the following definition of a sub-Gaussian norm, which is a standard notion [313, 319].

Definition 6.2.1 (Sub-Gaussian norm)**.** *The L_2-Orlicz norm of a scalar random variable U is given by*

$$\|U\|_{\psi_2} = \inf\{t > 0 \mid \mathbb{E}[\exp(U^2/t^2)] \leq 2\}.$$

We also refer to this as the sub-Gaussian norm, and a random variable with sub-Gaussian norm bounded by σ is said to be σ-sub-Gaussian.

Assumption 6.2.2. *The noise of the inverse problem has sub-Gaussian norm ρ_σ uniformly for all $y \in \mathbb{R}$. Specifically, ρ_σ is a positive scalar such that*

$$\|\xi(y)\|_{\psi_2} \leq \rho_\sigma \qquad \textit{for all } y \in \mathcal{Y}.$$

[7] This is particularly true if the procedure $\mathcal{A}$ performs least squares, as in Theorem 6.4.1 to follow.

[8] The notation $v^{\otimes k}$ represents the tensor product of order k.

Remark 6.2.3. *Assumption 6.2.2 can be weakened in multiple ways. Firstly, the requirement that the noise be uniformly sub-Gaussian y-everywhere can be replaced with a requirement that it only holds over all y that can be realized with high probability. More generally, the sub-Gaussian assumption is not really required for our main result and can be weakened to allow for heavy-tailed noise—see the full paper [243] for such an extension to noise with bounded second moment.*

Finally, it can be verified that if the function g^* is invertible on the interval $\mathcal{I}$ and $\sigma = 0$, then Assumption 6.2.2 is trivially satisfied with $\rho_0 = 0$, since without noise, we have $\mathbb{E}[W|Y = y] = g^{-1}(y)$, and so $\xi(y) = 0$ almost surely. The next assumption requires that our inverse regression procedure output a useful function estimate.

Assumption 6.2.3. *Suppose we have k samples $\{y_i, w_i\}_{i=1}^k$ drawn i.i.d. from the observation model*

$$w_i = h^*(y_i) + \xi_i + z_i, \tag{6.6}$$

where the pair (y_i, ξ_i) is drawn from a joint distribution $\mathcal{P}_{Y,\xi}$ such that $\mathbb{E}[\xi|Y = y] = 0$ for each scalar y, the RV z_i is additional zero-mean, ρ-sub-Gaussian noise that is independent of the pair (y_i, ξ_i), and $h^* \in \mathcal{H}$ is an unknown function to be estimated. Suppose $\{\overline{y}_i\}_{i=1}^k$ are k fresh samples, each drawn i.i.d. from the distribution $\mathcal{P}_Y$. Then the procedure $\mathcal{A}\left((y_i, w_i)_{i=1}^k\right)$ returns a function $\widehat{h}$ satisfying

$$\frac{1}{k}\sum_{i=1}^{k}(\widehat{h}(\overline{y}_i) - h^*(\overline{y}_i))^2 \leq \overline{\mathcal{R}}_k^{\mathcal{A}}(h^*, \mathcal{P}_{Y,\xi}; \rho^2, \delta)$$

with probability greater than $1 - \delta$.

Through Assumption 6.2.3, we quantify the "quality" of the nonparametric procedure $\mathcal{A}$ through its population rate function $\overline{\mathcal{R}}_k^{\mathcal{A}}$. Indeed, computing these rate functions for specific nonparametric regression procedures is one of the principal goals of statistical learning theory [21, 302]. Note that unlike standard definitions of such a rate function, we allow the rate $\overline{\mathcal{R}}_k^{\mathcal{A}}$ to depend explicitly both on the underlying function h^*, and on the joint distribution of the noise and design points $\mathcal{P}_{Y,\xi}$. In the sequel, we visit settings in which the latter dependence can be removed if Assumption 6.2.2 also holds.

With these assumptions in place, we are now ready to state our main theorem. In the statement of the theorem, we track the error at time t by $\Delta_t = \sin^2 \angle\left(\widehat{\theta}_t, \theta^*\right)$; note that since each estimate $\widehat{\theta}_t$ has unit norm, there are absolute constants (c, C) such that

$$c\min\{\|\widehat{\theta}_t - \theta^*\|^2, \|\widehat{\theta}_t + \theta^*\|^2\} \leq \Delta_t \leq C\min\{\|\widehat{\theta}_t - \theta^*\|^2, \|\widehat{\theta}_t + \theta^*\|^2\},$$

whence Δ_t captures the squared ℓ_2 error of parameter estimation up to a sign[9]. We also use the shorthand $\overline{N} := N/2T$ and $\nu_t = \cos\angle\left(\widehat{\theta}_t, \theta^*\right) = \sqrt{1 - \Delta_t}$ for convenience, and denote by $\mathcal{P}_{Y,\xi}^*$

[9] While the sign ambiguity is inherent to even link functions g^*, it can otherwise be eliminated by assuming that $\widehat{\theta}_0$ forms an acute angle with θ^*.

the joint distribution of the random variables (Y, ξ) in the model (6.4). The shorthand $c \cdot \mathcal{P}^*_{Y,\xi}$ denotes the joint distribution of the scaled random variables $(cY, c\xi)$ in the model (6.4). Finally, recall the definition of the function h^* from the model (6.4).

Theorem 6.2.1. *Suppose that Assumptions 6.2.1, 6.2.2, and 6.2.3 hold, and that the iterates* $\widehat{\theta}_0, \ldots, \widehat{\theta}_T$ *are generated by Algorithm 4. Then there is a pair of absolute constants* (c_1, c_2) *such that for each* $t = 0, \ldots, T-1$, *if*

$$\Delta_t \leq \frac{99}{100}, \quad \overline{\mathcal{R}}^{\mathcal{A}}_{\overline{N}}\left(\nu_t h^*; \nu_t \mathcal{P}^*_{Y,\xi}, \Delta_t, \delta/3\right) + \rho_\sigma^2 \leq c_1 \kappa^2 \quad \textit{and} \quad \overline{N} \geq c_2 \max\left\{d, \kappa^{-2}\log^2(1/\kappa)\log\left(\frac{c_2}{\delta}\right)\right\}, \tag{6.7a}$$

then we have

$$\Delta_{t+1} \leq c_2 \Big\{\overline{\mathcal{R}}^{\mathcal{A}}_{\overline{N}}(\nu_t h^*; \nu_t \mathcal{P}^*_{Y,\xi}, \Delta_t, \delta/3) + \rho_\sigma^2\Big\} \cdot \left(\frac{d + \log(4/\delta)}{\overline{N}}\right) \tag{6.7b}$$

with probability exceeding $1-\delta$. *Moreover, on this event, if* $\angle(\widehat{\theta}_t, \theta^*) \leq \pi/2$, *then* $\angle(\widehat{\theta}_{t+1}, \theta^*) \leq \pi/2$.

The conditions (6.7a) present in the theorem warrant some discussion. The theorem requires that the iterate at time t satisfy $\Delta_t \leq 99/100$; the value of this constant is not important, and can be replaced with any other absolute constant[10] less than 1. The second condition

$$\overline{\mathcal{R}}^{\mathcal{A}}_{\overline{N}}\left(\nu_t h^*; \nu_t \mathcal{P}^*_{Y,\xi}, \Delta_t, \delta/3\right) + \rho_\sigma^2 \leq c_1$$

implies (qualitatively) that we are in the low noise regime with ρ_σ bounded above by an absolute constant. This is the regime in which we expect any gains to occur over classical semiparametric estimators, and in that sense, the condition should not be viewed as restrictive. Finally, the sample size condition $\overline{N} \geq c_2 \cdot d$ is also natural, and a consequence of the fact that we would like the linear regression step in the algorithm to return a unique solution. The accompanying technical condition $\overline{N} \geq c_2 \kappa^{-2} \log^2(1/\kappa) \log \frac{c_2}{\delta}$ ensures that the matrix X_{t+1} is well-conditioned.

Moving on to the theorem's conclusion, first note that it applies to *any* nonparametric estimation procedure that we use, and significant gains are obtained whenever the error rate $\overline{\mathcal{R}}^{\mathcal{A}}_{\overline{N}}\left(\nu_t h^*; \nu_t \mathcal{P}^*_{Y,\xi}, \Delta_t, \delta/3\right)$ is small. In particular, if $\overline{\mathcal{R}}^{\mathcal{A}}_{\overline{N}}\left(\nu_t h^*; \nu_t \mathcal{P}^*_{Y,\xi}, \Delta_t, \delta/3\right) = o(1)$, then running just *one step* of the procedure already obtains a better guarantee than that of classical estimators (cf. equation (6.2), with $n \equiv \overline{N}$). To obtain a final guarantee—which will typically be even sharper—the inequality needs to be applied iteratively; we do so in deriving Corollary 6.3.1 to follow. Finally, since the theorem applies to only one step of the iterative procedure, it is worth noting that the error Δ_t of previous step acts as the noise variance encountered by the nonparametric estimation procedure. This is what allows us to bootstrap the result and obtain a final rate. In Section 6.4.3, we derive a corollary of our main theorem when the procedure $\mathcal{A}$ is chosen to be the empirical risk minimizer. Let us now illustrate this guarantee on a specific subclass of monotone SIMs.

[10] It is also likely that this condition can be weakened to allow $\Delta_t \leq 1 - \mathcal{O}(d^{-1/2})$ (which would accommodate, say, a vector $\widehat{\theta}_t$ chosen uniformly at random from the unit sphere), but we do not concern ourselves with this extension since classical estimators can be used to guarantee that Δ_t is smaller than any pre-specified universal constant.

6.3 Consequences for monotone single index models

In this section, we apply the general result given by Theorem 6.4.1 to the case where the link function g^* is monotone. In this section, suppose that we have n i.i.d. samples drawn from the SIM (6.3) where the noise distribution is Gaussian of variance σ^2. We also make a further assumption on the link function g^*; we require some additional notation in order to state it. Let $\mathsf{c}_{n,\delta} = \sqrt{2\log(8n/\delta)}$, and recall that the set of sub-differentials of a function g at the point x are given by

$$\partial g(x) = \{y \in \mathbb{R} : g(z) \geq g(x) + y \cdot (z - x) \text{ for all } z \in \mathbb{R}\}.$$

For a pair of reals $a < b$, we say that $a \leq \partial g(x) \leq b$ if each element y in the set of sub-differentials obeys the inclusion $y \in [a, b]$. With these definitions in place, we make the following assumption on the link function g^*.

Assumption 6.3.1. *The function g^* is continuous with $0 < m \leq \partial g^*(z) \leq M < \infty$ for all $z \in [-\mathsf{c}_{n,\delta}, \mathsf{c}_{n,\delta}]$.*

Link functions employed in generalized linear models largely satisfy[11] Assumption 6.3.1, and more generally, the class of SIMs satisfying Assumption 6.3.1 has been extensively studied as a generalization of GLMs [161, 162]. Note that in contrast to general SIMs, the invertibility of the true function makes this class comparatively easier to handle. Let us now specify the two oracles that we require.

Labeling oracle: For this class of SIMs, the labeling oracle is trivial to implement. Simply output:

- The interval $\mathcal{I} = [-\mathsf{c}_{n,\delta}, \mathsf{c}_{n,\delta}]$,
- All n samples of the SIM, and
- The function class
$$\mathcal{H} = \left\{h : \mathbb{R} \mapsto \mathcal{I} \;\middle|\; h \text{ non-decreasing}\right\},$$
which is a convex set by definition. In Lemma 6.4.2 in the proof section, we show that this class contains, with high probability, all the appropriate conditional expectations that we hope to model.

Nonparametric inverse regression procedure: We let $\mathcal{A}$ correspond to the ERM procedure over the function class $\mathcal{H}$ defined above. In this special case, the algorithm can be implemented in near-linear time via the pool adjacent violators algorithm [19, 128].

[11] In some cases, it may be necessary to choose the tuple (m, M) to be functions of n and δ (e.g., for the logistic link function), but these will typically be functions that decrease/increase sub-polynomially in n.

With the labeling oracle and inverse regression procedure specified, it remains to verify the various technical assumptions required to apply Theorem 6.4.1. Let $\kappa_0 = \frac{M}{m}$ denote a natural notion of conditioning in the problem. In Lemma 6.4.3, we show that Assumption 6.2.2 holds with

$$\rho_\sigma \leq \rho_{\mathsf{mono}} := C\left(\sigma^2 \mathsf{c}_{n,\delta}\sqrt{\kappa_0^2 - 1} + \frac{\sigma}{m}\left\{\log(3\kappa_0) \vee \mathsf{c}_{n,\delta}\right\}\right). \tag{6.8}$$

When σ is small, i.e., in our regime of interest, we have $\rho_{\mathsf{mono}} \asymp \sigma \cdot \mathsf{c}_{n,\delta}$, where the $\asymp$ notation hides problem-dependent factors. Another special case is when $M = m$ and $g^*(z) = mz$ a.e.; here, we have $\rho_{\mathsf{mono}} = C\frac{\sigma}{m}\mathsf{c}_{n,\delta}$, and $\frac{\sigma}{m}$ is the right proxy for noise-to-signal ratio in linear models.

Bounds on the complexity terms are provided in Lemma 6.4.4, and Assumption 6.4.1 holds trivially with $b = \mathsf{c}_{n,\delta}$. We are thus led to the following corollary of Theorem 6.4.1, in which we use the shorthand $\overline{n} = n/2T$ for convenience.

Corollary 6.3.1. *Suppose that Assumption 6.3.1 holds, and that the labeling oracle and regression procedure are given by the discussion above. Then there is a tuple of absolute constants* (c_1, c_2, c_3, c_4) *such that for each* $t = 0, 1, \ldots, T-1$, *if*

$$\overline{n} \geq c_2 d, \qquad \Delta_t \leq \tfrac{99}{100}, \qquad \textit{and} \qquad \rho_{\mathsf{mono}} \leq c_1,$$

then

$$\Delta_{t+1} \leq c_2\left\{\left(\frac{\log \overline{n}}{\overline{n}}\right)^{2/3} + \left(\frac{\Delta_t + \rho_{\mathsf{mono}}^2}{\overline{n}}\right)^{2/3} + \rho_{\mathsf{mono}}^2\right\}\frac{d}{\overline{n}}\log\left(\frac{c_2}{\delta}\right)\cdot\log\left(\frac{c_2 n}{\delta}\right) \tag{6.9a}$$

with probability exceeding $1 - \delta$.
Consequently, if in addition we have $\overline{n} \geq c_2 d \log^2 n$, *then when* $c_3 \log(\log n) \leq T \leq c_4 \log(\log n)$, *we obtain*

$$\Delta_T \leq c_2 d \log n \cdot \left\{\left(\frac{\log n \cdot \log(\log n)}{n}\right)^{5/3} + \rho_{\mathsf{mono}}^2 \frac{\log n \cdot \log(\log n)}{n}\right\} \tag{6.9b}$$

with probability exceeding $1 - c_2 n^{-9}$.

Once again, a few comments are in order. First, note that by our discussion above, the bound (6.9b) recovers the correct behavior in a linear model up to a poly-logarithmic factor. Second, note the following consequence of the bound (6.9b) in order to facilitate a more transparent discussion. Assuming the initial angle made by θ_0 with θ^* is acute, we have

$$\|\widehat{\theta}_T - \theta^*\|^2 \lesssim \begin{cases} \sigma^2 \frac{d}{n} & \text{if } \sigma \geq n^{-1/3} \\ \frac{d}{n^{5/3}} & \text{otherwise,} \end{cases}, \tag{6.10}$$

where the $\lesssim$ notation above ignores both problem-dependent constants that depend on the pair (m, M), as well as logarithmic factors in n. Comparing the bounds (6.2) and (6.10), we see immediately that the estimation bias is significantly reduced, and this comparison helps explain the behavior seen in Figure 6.1.

While Corollary 6.3.1 clearly provides a guarantee that is significantly better than classical estimators when σ is small, it is worth noting that it is derived as a consequence of Theorem 6.4.1, which may not be the sharpest possible result obtainable when, for instance, $\sigma = 0$. In the full paper [243], we take a slightly different route towards understanding the zero-noise setting, by designing a slightly different procedure that is motivated by analysis considerations; we omit these details here.

6.4 Proofs of main results

In this section, we provide proofs of our main results. We begin by proving Theorem 6.4.1, and then derive the various corollaries stated in the main text. A few notes to the reader. Throughout our proofs, we assume that n is greater than some universal constant; the complementary case can be handled by appropriately modifying the constants in the proofs. Often, we work with the random variables defining a model—which we denote by capital letters—before instantiating the model on samples—which we denote using small letters. Finally, we use $c, c_1, c', \ldots$ to denote universal constants whose values may change from line to line.

6.4.1 Proof of Theorem 6.2.1

Each covariate is given by d i.i.d. random variables $X = (X_1, X_2, \ldots, X_d)$. Assume wlog by the rotational invariance of the Gaussian distribution that $\theta^* = e_1$, so that $\langle X, \theta^* \rangle = X_1$. Recall the random variable W given by the truncation of X_1 to the interval $\mathcal{I}$. Recall the function h^*, given by

$$h^*(y) = \mathbb{E}[W | Y = y] \qquad \text{for each } y \in \mathbb{R}.$$

Also recall the (unobservable) model (6.4) given by

$$W = h^*(y) + \xi(y),$$

for each fixed value of y, where $\xi(y) = [W|Y = y] - \mathbb{E}[W|Y = y]$ denotes noise that obeys $\mathbb{E}[\xi(y)] = 0$ for each y by definition, and is ρ_σ-sub-Gaussian for each $y \in \mathbb{R}$ by Assumption 6.2.2. Finally, recall that we denoted the covariate distribution post-truncation by $\mathcal{P}_X^{\mathcal{I}}$. Note that in each sample, we also observe $d - 1$ other covariates $X_2, \ldots, X_d$ each drawn from a standard Gaussian that is independent of everything else. Let $\alpha_t = \angle(\widehat{\theta}_t, \theta^*)$ and note that for $X \sim \mathcal{P}_X^{\mathcal{I}}$, we have

$$\langle X, \theta_t \rangle = \cos(\alpha_t) W + \sin(\alpha_t) \overline{X},$$

where $\overline{X} \sim \mathcal{N}(0, 1)$ is some linear combination of the random variables $X_2, \ldots, X_d$ (and therefore independent of W). Recall the shorthand $\nu_t = \cos(\alpha_t)$ and $\Delta_t = \sin^2(\alpha_t)$. Suppose for the rest of this proof that α_t is acute, so that $\sin(\alpha_t) = \sqrt{\Delta_t}$; the complementary case is similar, provided we work with the angle $\alpha_t = \angle(\widehat{\theta}_t, -\theta^*)$ instead. With this setup at hand, we are now ready to prove the theorem. We organize the proof by providing error guarantees for the two sub-steps of Algorithm 4, and then putting together the pieces.

Error due to nonparametric regression: The procedure $\mathcal{A}$ is given $\overline{N} = N/2T$ samples drawn from the observation model

$$\langle x_i,\, \widehat{\theta}_t\rangle = \nu_t \cdot h^*(y_i) + \nu_t \cdot \xi_i + \sqrt{\Delta_t}\overline{x}_i \qquad \text{for } i \in \mathcal{D}_{2t+1}.$$

By the star-shaped nature of $\mathcal{H}$, we have $\nu_t \cdot h^* \in \mathcal{H}$, so that this is now a nonparametric regression model where we observe $\overline{N}$ i.i.d. evaluations of the true function $\nu_t h^*$ corrupted by noise. In particular, comparing with the model (6.6), we have $\rho^2 = \Delta_t$. By Assumption 6.2.3, the procedure $\mathcal{A}$ uses these samples to then return a function $\widehat{h}_{t+1} \in \mathcal{H}$ that satisfies, for each $\delta \in (0, 1)$, the inequality

$$\frac{1}{\overline{N}} \sum_{i \in \mathcal{D}_{2t+2}} (\widehat{h}_{t+1}(y_i) - \nu_t h^*(y_i))^2 \leq \overline{\mathcal{R}}^{\mathcal{A}}_{\overline{N}}(\nu_t h^*; \nu_t \mathcal{P}^*_{Y,\xi}, \Delta_t, \delta)$$

with probability exceeding $1 - \delta$.

Error due to linear regression: We now show that performing the linear regression step leads to an error contraction by a multiplicative factor roughly $p/\overline{N}$. For this, we require the following lemma. For a vector v, we let v_i denote its i-th entry, and let $v_{\setminus i} = v - v_i \cdot e_i$ denote the vector with its i-th entry zeroed out.

Lemma 6.4.1. *Suppose we are given a matrix $X = (X^1, \ldots, X^d) \in \mathbb{R}^{n \times d}$, where the columns $X^2, \ldots, X^d \overset{i.i.d.}{\sim} \mathcal{N}(0, I_n)$. Also suppose we are given the n-dimensional vector $y = \tau X^1 + z$ for some scalar τ, and some vector $z \in \mathbb{R}^n$ that is fixed independently of the random vectors $X^2, \ldots, X^d$. Then there is an absolute constant c such that if $n \geq c \max\left\{d, \log\left(\frac{c}{\delta}\right)\right\}$, then the estimate $\beta = X^\dagger y$ obeys the inequalities*

$$\|\beta_{\setminus 1}\|^2 \leq 16 \cdot \left(\frac{d + \log(4/\delta)}{n^2}\right) \cdot \|z\|^2, \quad \textit{and} \tag{6.11a}$$

$$\beta_1^2 \geq \frac{\tau^2}{2} - 3\frac{\|z\|^2}{\|X^1\|^2}, \tag{6.11b}$$

with probability exceeding $1 - \frac{3\delta}{4}$. Moreover, on this event, if $\tau > \sqrt{\frac{3}{2}} \cdot \frac{\|z\|}{\|X^1\|}$, then $\beta_1 > 0$.

We prove this lemma at the end of the section. Let us now use it to provide an error guarantee on our problem. For a fresh draw of the pair (X, Y) with marginals $X \sim \mathcal{P}^{\mathcal{I}}_X$ and $Y \sim \mathcal{P}_Y$, we have

$$\nu_t W = \nu_t\langle X,\, \theta^*\rangle = \nu_t h^*(Y) + \nu_t \xi = \widehat{h}_{t+1}(Y) + \nu_t \xi + (\nu_t h^*(Y) - \widehat{h}_{t+1}(Y)),$$

so that rearranging yields

$$\widehat{h}_{t+1}(Y) = \nu_t\langle X,\, \theta^*\rangle - \nu_t \xi - (\nu_t h^*(Y) - \widehat{h}_{t+1}(Y)). \tag{6.12}$$

Notably, all random variables in the RHS are functions only of the tuple (W, ϵ) and hence independent of the random variables $X_2, \ldots, X_d$.

The linear regression step is performed on the samples $i \in \mathcal{D}_{2t+2}$. It is therefore helpful to instantiate the model (6.12) on these samples, and write

$$\widehat{h}_{t+1}(y_i) = \nu_t \langle x_i,\, \theta^* \rangle - \underbrace{\nu_t \xi_i - (\nu_t h^*(y_i) - \widehat{h}_{t+1}(y_i))}_{-\xi_i'} \text{ for } i \in \mathcal{D}_{2t+2};$$

crucially, due to sample-splitting across the two sub-steps of the algorithm, we have ensured that the function estimate $\widehat{h}_{t+1}$ can be regarded as fixed, since it is independent of the samples $i \in \mathcal{D}_{2t+2}$. Step 3 of the algorithm models the value $\widehat{h}_{t+1}(y_i)$ as a linear response to the covariates x_i. In particular, stack the covariates $\{x_i\}_{i \in \mathcal{D}_{2t+2}}$ in a matrix and the responses $\{\widehat{h}_{t+1}(y_i)\}_{i \in \mathcal{D}_{2t+2}}$ in a vector, and let $\xi' \in \mathbb{R}^{\overline{N}}$ denote a "noise" vector. Recall the condition $\overline{N} \geq c \max\{d, \log(c/\delta)\}$ assumed in Theorem 6.2.1, and recall that our regression estimate obtained as a result of step 3 of Algorithm 4 was denoted by $\widetilde{\theta}_{t+1}$. Applying Lemma 6.4.1 yields, with probability exceeding $1 - \frac{3\delta}{4}$, the implications

$$\|\widetilde{\theta}_{t+1}\|^2 \sin^2(\alpha_{t+1}) \leq 16 \cdot \|\xi'\|^2 \cdot \left(\frac{d + \log(4/\delta)}{\overline{N}^2} \right) \text{ and} \tag{6.13a}$$

$$\|\widetilde{\theta}_{t+1}\|^2 \cos^2(\alpha_{t+1}) \geq \frac{\nu_t^2}{2} - 3 \cdot \frac{\|\xi'\|^2}{\sum_{i \in \mathcal{D}_{2t+2}} \langle x_i,\, \theta^* \rangle^2}, \tag{6.13b}$$

where we have used the fact that the scalar τ in the lemma is equal to ν_t.

Putting together the pieces: We are finally ready to put together the pieces. Applying the Cauchy–Schwarz inequality yields

$$\|\xi'\|^2 \leq 2 \cdot \left(\sum_{i \in \mathcal{D}_{2t+2}} \nu_t^2 \xi_i^2 + \sum_{i \in \mathcal{D}_{2t+2}} (\nu_t h^*(y_i) - \widehat{h}_{t+1}(y_i))^2 \right).$$

By Assumption 6.2.2, each random variable ξ_i is ρ_σ-sub-Gaussian, so that

$$\Pr \left\{ \frac{1}{\overline{N}} \sum_{i \in \mathcal{D}_{2t+2}} \xi_i^2 \geq \rho_\sigma^2 \left(1 + \sqrt{\frac{\log(6/\delta)}{\overline{N}}} \right) \right\} \leq \delta/3 \tag{6.14}$$

for each $\overline{N} \geq \log(6/\delta)$. On the other hand, Assumption 6.2.3 guarantees that we have

$$\Pr \left\{ \frac{1}{\overline{N}} \sum_{i \in \mathcal{D}_{2t+2}} (\nu_t h^*(y_i) - \widehat{h}_{t+1}(y_i))^2 \geq \overline{\mathcal{R}}^{\mathcal{A}}_{\overline{N}}(\nu_t h^*; \nu_t \mathcal{P}^*_{Y,\xi}, \Delta_t, \delta/3) \right\} \leq \tfrac{\delta}{3}.$$

Putting together the pieces, we have

$$\frac{\|\xi'\|^2}{\overline{N}} \leq C \left(\nu_t^2 \rho_\sigma^2 + \overline{\mathcal{R}}_{\overline{N}}^{\mathcal{A}}(\nu_t h^*; \nu_t \mathcal{P}^*_{Y,\xi}, \Delta_t, \delta/3) \right) \tag{6.15a}$$

with probability greater than $1 - \frac{2\delta}{3}$. Additionally, Lemma B.2.2 from the appendix guarantees that provided $\overline{N} \geq c_1 \frac{\log^2(1/\kappa)}{\kappa^2} \log(4/\delta)$, we have

$$\sum_{i \in \mathcal{D}_{2t+2}} \langle x_i, \theta^* \rangle^2 = \sum_{i \in \mathcal{D}_{2t+2}} w_i^2 \geq \frac{1}{2} \kappa^2 \overline{N} \tag{6.15b}$$

with probability exceeding $1 - \delta/4$. Now note that we have $\Delta_t \leq 99/100$ by assumption, which guarantees the relation $\nu_t \geq 1/10$. Thus, on the intersection of the two events defined in inequality (6.15), inequality (6.13) yields

$$\tan^2(\alpha_{t+1}) \leq C \left(\rho_\sigma^2 + \overline{\mathcal{R}}_{\overline{N}}^{\mathcal{A}}(\nu_t h^*; \nu_t \mathcal{P}^*_{Y,\xi}, \Delta_t, \delta/3) \right) \cdot \left(\frac{d + \log(4/\delta)}{\overline{N}} \right),$$

where we have also used that the condition $\rho_\sigma^2 + \overline{\mathcal{R}}_{\overline{N}}^{\mathcal{A}}(\nu_t h^*; \nu_t \mathcal{P}^*_{Y,\xi}, \Delta_t, \delta/3) \leq c_1$ holds for a small enough constant c_1, to ensure that the RHS of inequality (6.13b) is bounded below by a universal positive constant. Finally, noting the elementary inequality $\sin^2 \alpha \leq \tan^2 \alpha$ concludes the proof. □

6.4.2 Proof of Lemma 6.4.1

Our proof of this lemma proceeds from first principles; we note that similar proofs are used to bound the variance inflation factor (VIF) in linear models (see, e.g., the book [235]). Use the more convenient notation $x = X^1$ and $\mathbf{W} = (X^2, \ldots, X^d)$, so that the matrix is given by $X = \begin{bmatrix} x & \mathbf{W} \end{bmatrix}$, and $X^\top X = \begin{bmatrix} \|x\|^2 & x^\top \mathbf{W} \\ \mathbf{W}^\top x & \mathbf{W}^\top \mathbf{W} \end{bmatrix}$. Note that for a general (invertible) symmetric matrix, the partial LDU decomposition can be written as

$$\begin{bmatrix} a & b^\top \\ b & \mathbf{C} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -\mathbf{C}^{-1} b & \mathbf{I} \end{bmatrix} \begin{bmatrix} (a - b^\top \mathbf{C}^{-1} b)^{-1} & 0 \\ 0 & \mathbf{C}^{-1} \end{bmatrix} \begin{bmatrix} 1 & -b^\top \mathbf{C}^{-1} \\ 0 & \mathbf{I} \end{bmatrix},$$

where $\mathbf{I}$ denotes the identity matrix of appropriate dimension. Applying this to the matrix $X^\top X$ and using the shorthand $\mathbf{P_W} = \mathbf{W}(\mathbf{W}^\top \mathbf{W})^{-1} \mathbf{W}^\top$ for the projection matrix onto the range of the matrix $\mathbf{W}$, we may write the pseudoinverse of X as

$$X^\dagger =$$
$$\begin{bmatrix} 1 & 0 \\ -(\mathbf{W}^\top \mathbf{W})^{-1} \mathbf{W}^\top x & \mathbf{I} \end{bmatrix} \begin{bmatrix} (\|x\|^2 - x^\top \mathbf{P_W} x)^{-1} & 0 \\ 0 & (\mathbf{W}^\top \mathbf{W})^{-1} \end{bmatrix} \begin{bmatrix} 1 & -x^\top \mathbf{W} (\mathbf{W}^\top \mathbf{W})^{-1} \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} x^\top \\ \mathbf{W}^\top \end{bmatrix}$$

Now for an arbitrary vector $v \in \mathbb{R}^n$, let $\langle x, v \rangle := x_v$; then we have

$$\begin{bmatrix} 1 & -x^\top \mathbf{W}(\mathbf{W}^\top \mathbf{W})^{-1} \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} x^\top \\ \mathbf{W}^\top \end{bmatrix} v = \begin{bmatrix} x_v - x^\top \mathbf{P_W} v \\ \mathbf{W}^\top v \end{bmatrix},$$

so that putting together the pieces yields

$$\begin{aligned} X^\dagger v &= \begin{bmatrix} 1 & 0 \\ -(\mathbf{W}^\top \mathbf{W})^{-1} \mathbf{W}^\top x & I \end{bmatrix} \begin{bmatrix} (\|x\|^2 - x^\top \mathbf{P_W} x)^{-1} & 0 \\ 0 & (\mathbf{W}^\top \mathbf{W})^{-1} \end{bmatrix} \begin{bmatrix} x_v - x^\top \mathbf{P_W} v \\ \mathbf{W}^\top v \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 \\ -(\mathbf{W}^\top \mathbf{W})^{-1} \mathbf{W}^\top x & \mathbf{I} \end{bmatrix} \begin{bmatrix} (\|x\|^2 - x^\top \mathbf{P_W} x)^{-1} \cdot (x_v - x^\top \mathbf{P_W} v) \\ \mathbf{W}^\dagger v \end{bmatrix}. \end{aligned}$$

Now using the shorthand $\mathbf{P}_\mathbf{W}^\perp = \mathbf{I} - \mathbf{P_W}$ for the projection matrix onto the orthogonal complement of $\mathbf{W}$, we have

$$\begin{aligned} X^\dagger v &= \begin{bmatrix} 1 & 0 \\ -(\mathbf{W}^\top \mathbf{W})^{-1} \mathbf{W}^\top x & \mathbf{I} \end{bmatrix} \begin{bmatrix} (x^\top \mathbf{P}_\mathbf{W}^\perp x)^{-1} \cdot (x^\top \mathbf{P}_\mathbf{W}^\perp v) \\ \mathbf{W}^\dagger v \end{bmatrix} \\ &= \begin{bmatrix} \tau_v \\ -\mathbf{W}^\dagger x \cdot \tau_v + \mathbf{W}^\dagger v \end{bmatrix}, \end{aligned}$$

where we have let $\tau_v := (x^\top \mathbf{P}_\mathbf{W}^\perp x)^{-1} \cdot (x^\top \mathbf{P}_\mathbf{W}^\perp v)$ for convenience.

Note that the above derivation holds for each $v \in \mathbb{R}^n$. We are interested in a vector that can be written as $v = \tau x + z$. In this case, we have

$$\begin{aligned} X^\dagger v &= X^\dagger(\tau x) + X^\dagger z \\ &= \tau e_1 + X^\dagger z \\ &= \begin{bmatrix} \tau_z + \tau \\ -\mathbf{W}^\dagger x \cdot \tau_z + \mathbf{W}^\dagger z \end{bmatrix}. \end{aligned}$$

Up to this point, all of our steps were deterministic; we now use the fact that $\mathbf{W}$ is a standard Gaussian random matrix. In particular, letting w_1 denote the first row of the matrix $\mathbf{W}$ and for any vector u fixed independently of $\mathbf{W}$, we have

$$\begin{aligned} \|\mathbf{W}^\dagger u\|^2 &\overset{d}{=} \|(\mathbf{W}^\top \mathbf{W})^{-1} \mathbf{W}^\top e_1\|^2 \|u\|^2 \\ &= \|(\mathbf{W}^\top \mathbf{W})^{-1} w_1\|^2 \|u\|^2 \\ &\le |||(\mathbf{W}^\top \mathbf{W})^{-1}|||_{\mathrm{op}}^2 \|w_1\|^2 \|u\|^2 \\ &\overset{\text{(i)}}{\le} \left(\frac{1}{\left(\sqrt{n} - \sqrt{d} - \sqrt{\log(4/\delta)}\right)^2} \right)^2 (d + \log(4/\delta)) \|u\|^2, \end{aligned}$$

where step (i) holds with probability exceeding $1 - \delta/4$ by tail bounds for χ^2 random variables, and the minimum singular value of a Gaussian random matrix [319].

We now use the assumption $n \geq c \max\{d, \log(4/\delta)\}$ for a large enough constant c to obtain the inequality

$$\|\mathbf{W}^{\dagger} u\|^2 \leq 2 \left(\frac{1}{n}\right)^2 \cdot \frac{\|u\|^2}{n} \left(d + \log\left(\frac{4}{\delta}\right)\right),$$

which holds for each fixed vector u with probability exceeding $1 - \delta/4$. Moreover, we have the equivalence $u^\top \mathbf{P}_{\mathbf{W}}^{\perp} u = \|\mathbf{P}_{\mathbf{W}}^{\perp} u\|^2$. Putting together the pieces and using the Cauchy–Schwarz inequality, the following sequence of bounds holds with probability exceeding $1 - \delta$:

$$\begin{aligned}
\frac{\|\beta_{\setminus 1}\|^2}{4} &\leq \left(\frac{d + \log(4/\delta)}{n^2}\right) \cdot (\tau_z^2 \|x\|^2 + \|z\|^2) \\
&= \left(\frac{d + \log(4/\delta)}{n^2}\right) \cdot \left(\frac{(x^\top \mathbf{P}_{\mathbf{W}}^{\perp} z)^2}{\|\mathbf{P}_{\mathbf{W}}^{\perp} x\|^4} \cdot \|x\|^2 + \|z\|^2\right) \\
&\overset{\text{(ii)}}{\leq} \left(\frac{d + \log(4/\delta)}{n^2}\right) \cdot \left(\frac{\|z\|^2}{\|\mathbf{P}_{\mathbf{W}}^{\perp} x\|^2} \cdot \|x\|^2 + \|z\|^2\right),
\end{aligned}$$

where step (ii) uses the Cauchy–Schwarz inequality and symmetry of the matrix $\mathbf{P}_{\mathbf{W}}^{\perp}$ to obtain $|x^\top \mathbf{P}_{\mathbf{W}}^{\perp} z| \leq \|\mathbf{P}_{\mathbf{W}}^{\perp} x\| \|z\|$.

Now note that we since $x \perp\!\!\!\perp \mathbf{W}$, we have $\frac{\|\mathbf{P}_{\mathbf{W}}^{\perp} x\|^2}{\|x\|^2} \overset{d}{=} \|\mathbf{P}_{\mathbf{W}}^{\perp} e_1\|^2$, which is the squared norm of a unit-norm n-dimensional vector projected onto a random $(n - d + 1)$-dimensional subspace. By well-known results (see, e.g, Dasgupta and Gupta [78]), this quantity is bounded above by $\frac{n - d + 1 + \log(4/\delta)}{n}$ with probability exceeding $1 - \delta/4$. Putting together the pieces once again with our assumption $n \geq c \max\{d, \log(4/\delta)\}$, we have

$$\frac{\|\beta_{\setminus 1}\|^2}{4} \leq \left(\frac{d + \log(4/\delta)}{n^2}\right) \cdot \left\{3 \cdot \|z\|^2 + \|z\|^2\right\},$$

with probability exceeding $1 - \frac{3\delta}{4}$.

To lower bound the signal term, we once again use the Cauchy–Schwarz inequality to obtain

$$\begin{aligned}
\beta_1^2 &\geq \frac{\tau^2}{2} - \tau_z^2 \\
&\geq \frac{\tau^2}{2} - 3 \cdot \frac{\|z\|^2}{\|x\|^2}.
\end{aligned}$$

This concludes the proof. □

Remark 6.4.1. *Lemma 6.4.1 illustrates the role of the approximation error in the problem, and can be used to reason about variants of classical semiparametric estimators. In particular, if $\mathbb{E}[g^*(Z)Z] = \mu \neq 0$, then we may write $g^*(X_1) = \mu X_1 + Z_{\pi,\sigma}$, where $Z_{\pi,\sigma}$ is uncorrelated with X_1 due to the orthogonality properties of Hermite polynomials [124]. Treating $\mathbb{E}|Z_{\pi,\sigma}|$ as the approximation error (which is a constant for any non-linear g^*), we see that Lemma 6.4.1 guarantees that regressing our observations $g(X_1) + \epsilon$ on $X_1, \ldots, X_n$ yields an estimate with error $\frac{d}{n}(\sigma^2 + \mathbb{E}|Z_{\pi,\sigma}|)$ (cf. equation (6.2)). The goal of first performing nonparametric regression to obtain a function estimate $\widehat{h}$ is to significantly reduce the approximation error of the problem.*

6.4.3 Implications for empirical risk minimization procedures

It is useful to particularize Theorem 6.2.1 to the case where $\mathcal{A}$ corresponds to the empirical risk minimization (ERM) algorithm over the function class $\mathcal{H}$. Since we are interested in performing ERM on i.i.d. samples drawn from the model (6.6), let us introduce it in this context. Given k i.i.d. samples $\{y_i, w_i\}_{i=1}^k$ drawn from this model, the ERM algorithm estimating the unknown nonparametric function $h^* \in \mathcal{H}$ returns the function

$$\widehat{h}_{\mathsf{ERM}} \in \operatorname*{argmin}_{h \in \mathcal{H}} \frac{1}{k} \sum_{i=1}^{k} (w_i - h(y_i))^2,$$

where we have chosen the squared loss given our assumption that the noise ϵ is sub-Gaussian. Note that this estimator exists since the function class $\mathcal{H}$ is closed and convex. The estimate is also random, due to both the randomness in the "design points" $y_1, \ldots, y_k$ and in the noise. Let us now discuss how one might bound the error rate of this algorithm with high probability.

A classical result in the study of the ERM algorithm [21, 173] is that the rate function is governed by the *local population Rademacher complexity* of the function class being estimated over. Let us first define a more general version of this quantity, valid for an arbitrary function class $\mathcal{F}$ mapping $\mathbb{R} \mapsto \mathbb{R}$ with k i.i.d. samples from our model (6.6). Let $y_1^k = (y_1, \ldots, y_k)$ denote the tuple of k i.i.d. design points drawn from the distribution $\mathcal{P}_Y$, and let $\eta = (\eta_1, \ldots, \eta_k)$ denote k i.i.d. Rademacher random variables drawn independently of everything else. Then the population Rademacher complexity of the function class is given by

$$\mathfrak{R}_k(\mathcal{F}) := \mathbb{E}_{\eta, y_1^k} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{k} \sum_{i=1}^{k} \eta_i f(y_i) \right| \right]. \tag{6.16a}$$

The Rademacher complexity defined in equation (6.16a) depends only on the function class $\mathcal{F}$ and design points $y_1, \ldots, y_k$, but not on the specific noise in the problem. In order to reason about how the noise affects the estimation procedure in our specific context, it is useful to also introduce another measure of complexity of the function class. Consider model (6.6), and denote by $\overline{\xi}_i := \overline{\rho}^{-1}(\xi_i + z_i)$ the rescaled noise in the i-th sample of our observations. Use the shorthand $\overline{\rho} := \sqrt{\rho_\sigma^2 + \rho^2}$, and let $\overline{\xi} := (\overline{\xi}_1, \ldots, \overline{\xi}_k)$. Then the *noise complexity* of the function class[12] $\mathcal{F}$ is defined as

$$\mathfrak{G}_k(\mathcal{F}; y_1^k) := \mathbb{E}_{\overline{\xi}} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{k} \sum_{i=1}^{k} \overline{\xi}_i \cdot f(y_i) \right| \right]. \tag{6.16b}$$

Note that in contrast to our definition of the population Rademacher complexity (6.16a), we no longer take an expectation over the random samples $y_1, \ldots, y_k$ in equation (6.16b), and so the noise complexity should be viewed as a random variable when the samples $y_1, \ldots, y_k$ are random.

[12] This is typically known as the Gaussian complexity when the noise is Gaussian, but we prefer the more general nomenclature at this stage of our development.

It is also useful to define the norms

$$\|f\|_2^2 := \mathbb{E}_{Y \sim \mathcal{P}_Y}[f^2(Y)] \qquad \text{and} \qquad \|f\|_k^2 := \sum_{i=1}^k f^2(y_i); \tag{6.17}$$

once again, the second norm should be viewed as random when the samples $y_1, \ldots, y_k$ are random. For either norm $\| \cdot \|$, let $\mathbb{B}(\| \cdot \|; t)$ denote the norm-ball of radius t centered at zero. Also define the shifted function class

$$\mathcal{H}_{h_0} = \{h - h_0 \mid h \in \mathcal{H}\},$$

where we have the equivalence $\mathcal{H} \equiv \mathcal{H}_0$.

With these definitions in place, analyses of the ERM algorithm rely on finding fixed points of certain local complexity measures, which we now define for our specific function class $\mathcal{H}_{h^*}$. For each positive integer k and pair of positive constants (γ_1, γ_2), define the quantities

$$\tau_k(h^*; \gamma_1) := \inf\left\{\tau > 0 : \mathfrak{R}_k(\mathcal{H}_{h^*} \cap \mathbb{B}(\| \cdot \|_2; \tau)) \leq \frac{\tau^2}{\gamma}\right\}, \text{ and} \tag{6.18a}$$

$$\mu_k(h^*, y_1^k; \gamma_2) := \inf\left\{\mu > 0 : \mathfrak{G}_k(\mathcal{H}_{h^*} \cap \mathbb{B}(\| \cdot \|_k; \mu); y_1^k) \leq \frac{\mu^2}{\gamma_2}\right\}. \tag{6.18b}$$

Note that the functional μ_k depends on the noise $\overline{\xi}$, while the functional τ_k does not. Let us provide some motivation for these complexity measures. A natural way to measure the error of the ERM is via its *fixed-design* loss

$$\|\widehat{h}_{\mathsf{ERM}} - h^*\|_k^2 := \frac{1}{k} \sum_{i=1}^k (\widehat{h}_{\mathsf{ERM}}(y_i) - h^*(y_i))^2, \tag{6.19a}$$

where $y_1, \ldots, y_k$ are precisely the k i.i.d. samples generated from the model (6.6) using which the ERM procedure is computed; consequently, the estimate $\widehat{h}_{\mathsf{ERM}}$ is not independent of the randomness in these samples. The noise complexity (6.16b) and associated critical inequality (6.18b) are useful in bounding this quantity. However, we are interested in controlling the error measured by the random variable

$$\frac{1}{k} \sum_{i=1}^k (\widehat{h}_{\mathsf{ERM}}(\overline{y}_i) - h^*(\overline{y}_i))^2, \tag{6.19b}$$

with *fresh* samples $\{\overline{y}_i\}_{i=1}^k$ drawn from the distribution $\mathcal{P}_Y$. A natural question is whether the error measures defined in equation (6.19) are close to each other. The functional τ_k defined in equation (6.18a) provides such a measure of closeness for an appropriate value of the scalar γ_1. In particular, a uniform law of large numbers holds in this problem, as a consequence of which both error metrics are in fact close to the *expected* error $\|\widehat{h}_{\mathsf{ERM}} - h^*\|_2^2$.

With this lengthy setup complete, we are finally ready to state our assumption on the function class $\mathcal{H}$. For simplicity, we assume that the function class is uniformly bounded.[13]

Assumption 6.4.1 (Bounded function class). *There is a positive constant b such that for all $h \in \mathcal{H}$, we have $\|h\|_\infty \leq b$.*

The following proposition states a bound on the rate of the ERM algorithm in terms of the complexity functions defined above, with the shorthand $\mathcal{H}^* \equiv \mathcal{H}_{h^*}$. Recall that the observations are corrupted by sub-Gaussian noise with parameter $\overline{\rho}$.

Proposition 6.4.1 (Theorems 14.1 and 13.5 of Wainwright [323]). *(a) Suppose that Assumption 6.4.1 holds, and that we observe k samples from the model* (6.6)*. Then there are absolute constants (c_1, c_2) such that for each scalar $u \geq \tau_k(h^*; b)$, we have*

$$\left| \|f\|_2^2 - \|f\|_k^2 \right| \leq \frac{1}{2}\|f\|_2^2 + \frac{1}{2}u^2$$

uniformly for all functions $f \in \mathcal{H}^*$, with probability exceeding $1 - c_2 \exp\left(-c_1 k \frac{u^2}{b^2}\right)$.

(b) Suppose that Assumption 6.2.2 holds. Then there are absolute constants (c_1, c_2) such that for each $u \geq \mu_k(h^*, y_1^k; 2\overline{\rho})$, the fixed-design loss (6.19a) *of the ERM algorithm run on k samples from the model* (6.6) *satisfies*

$$\Pr\left\{\|\widehat{h}_{\mathsf{ERM}} - h^*\|_k^2 \geq 16u\mu_k(h^*, y_1^k; 2\overline{\rho})\right\} \leq c_2 \exp\left(-c_1 k u \cdot \frac{\mu_k(h^*, y_1^k; 2\overline{\rho})}{\overline{\rho}^2}\right).$$

Applying this proposition leads to the following consequence of our main result Theorem 6.2.1 when the procedure $\mathcal{A}$ corresponds to the ERM algorithm. The proof follows straightforwardly by combining Theorem 6.2.1 and Proposition 6.4.1, but we provide it in Section 6.4.4 for completeness. We nevertheless state the result as a theorem for stylistic reasons. Recall the shorthand $\overline{N} := N/(2T)$ and $\rho_{\sigma,t} := \sqrt{\Delta_t^2 + \nu_t^2 \rho_\sigma^2}$, and let

$$\mathfrak{T}_t := \left(\tau_{\overline{N}}^2(\nu_t h^*; b) \vee \tfrac{b^2}{\overline{N}}\right) + \left(\mu_{\overline{N}}^2(\nu_t h^*, \{y_i\}_{i \in \mathcal{D}_{2t+1}}; 2\rho_{\sigma,t}) \vee \tfrac{\rho_{\sigma,t}^2}{\overline{N}}\right)$$

denote (what we will establish to be) an upper bound on the rate function used in Theorem 6.2.1.

Theorem 6.4.1. *Let the above notation prevail. Suppose that Assumptions 6.2.1, 6.2.2 and 6.4.1 hold. Also suppose that the iterates $\widehat{\theta}_0, \ldots, \widehat{\theta}_T$ are generated by running Algorithm 4 with procedure*

[13] This assumption can be relaxed to require that for some $p \geq q \geq 2$ and all $h \in \mathcal{H}$ with $\|h\|_2 \leq 1$ we have

$$\mathbb{E}[h^d(Y)] \leq b^{p-q}\mathbb{E}[h^q(Y)],$$

when $Y \sim \mathcal{P}_Y$. We do not pursue this extension here, and direct the reader to Wainwright [323] and Mendelson [222] for details.

$\mathcal{A}$ corresponding to the ERM algorithm. There are absolute positive constants (c_1, c_2) such that for each $t = 0, \ldots, T-1$, the following holds true: If

$$\Delta_t \leq 99/100, \qquad \mathfrak{T}_t \cdot \log\left(\tfrac{c_2}{\delta}\right) + \rho_\sigma^2 \leq c_1 \kappa^2, \quad \textit{and} \quad \overline{N} \geq c_2 \max\left\{p, \kappa^{-2}\log^2(1/\kappa)\log\left(\frac{c_2}{\delta}\right)\right\},$$

then

$$\Delta_{t+1} \leq c_2\Big(\mathfrak{T}_t \cdot \log\left(\tfrac{c_2}{\delta}\right) + \rho_\sigma^2\Big) \cdot \left(\frac{p + \log(4/\delta)}{\overline{N}}\right) \tag{6.20}$$

with probability exceeding $1 - \delta$.

It is worth making a few remarks on the theorem. Note that once again, we have stated the result only for one step of our iterative algorithm; in order to produce a guarantee on the 'final' iterate we will have to recurse this bound, and subsequently, bound (i) the number of iterations T required to reach a fixed point of the recursive relation (6.20), and (ii) the error of the fixed point of the recursion. To provide a qualitative answer to point (i), first note that for most nonparametric function classes, the RHS of equation (6.20) is always strictly positive for each finite $\overline{N}$, so that we can never hope to show an exact recovery guarantee for the algorithm. In other words, zero is not a fixed point of the error recursion. Typical arguments used in analyses of many nonparametric regression problems show that

$$\tau^2_{\overline{N}} \sim \left(\tfrac{C_1}{\overline{N}}\right)^{\lambda_1} \text{ and } \mu^2_{\overline{N}} \sim \left(\tfrac{C_2}{\overline{N}}\right)^{\lambda_2}, \tag{6.21}$$

where λ_1 and λ_2 are two fixed constants in the unit interval that depend on the regression problem, and the constants (C_1, C_2) depend on the remaining quantities that parameterize each of these complexity functions. In this case, it suffices to apply the error recursion for $T_0 = \mathcal{O}(\log\log(\overline{N}))$ iterations in order to arrive within a constant multiplicative factor of the fixed point[14], where the constants absorbed by this asymptotic notation depend on the other parameters of the problem, and the scalars (λ_1, λ_2). For a specific illustration of this phenomenon, see Corollary 6.3.1 to follow.

The abstract bound (6.21) also provides a qualitative answer to point (ii) above: taking $\overline{N} \to \infty$, we see immediately that the fixed point has error bounded by a quantity $(o(1) + \rho_\sigma^2)\frac{d}{\overline{N}}$. Comparing such an error bound with equation (6.2), we verify what was already alluded to after the statement of Theorem 6.2.1: when $\rho_\sigma \asymp \sigma$, using a consistent ERM estimator improves the rate of parameter estimation uniformly *for all* noise levels.

It is also helpful to state a consequence of the bound (6.20) when $\rho_\sigma = 0$; this is achieved for noiseless SIMs if the function g^* is invertible on the interval $\mathcal{I}$. Let $m^*_{\overline{N},\delta}$ denote the value of m satisfying the fixed point relation

$$m = c_2 \mu^2_{\overline{N}}\left(h^*, y_1^k; 2\sqrt{m}\right) \frac{d}{\overline{N}} \log\left(\tfrac{c_2}{\delta}\right).$$

[14] Since zero is not a fixed point, the number of iterations required to ensure convergence to within a multiplicative factor of the fixed point is finite; this is in contrast to problems for which we would like to guarantee exact recovery [236], and crucially, bounds the number of resampling steps required by the algorithm.

Then assuming that the error recursion converges to its fixed point, we have the bound

$$\Delta_T \leq C \left(\frac{d}{\overline{N}} \cdot \tau^2_{\overline{N}}(b) + m^*_{\overline{N},\delta} \right) \tag{6.22}$$

for the 'final' iterate of our algorithm. Once again, it is worth noting that the final error in the noiseless case is strictly better than the $d/\overline{N}$ rate if the complexity term $\tau_{\overline{N}}$ decays with $\overline{N}$; Corollary 6.3.1 provides an example of such a phenomenon.

Remark 6.4.2 (Sharpness in the noiseless regime)**.** *The bound* (6.22) *is unlikely to be the sharpest bound one can prove in general SIMs for the noiseless case. There are other analyses of the ERM tailored to capture the correct "version space" of the noiseless nonparametric regression problem, and these will be sharper than the bounds presented above. For a more in-depth discussion, see the full paper [243].*

6.4.4 Proof of Theorem 6.4.1

First, we use Proposition 6.4.1 to provide a bound on the rate function $\overline{\mathcal{R}}_k^{\mathsf{ERM}}(\nu_t h^*; \rho_{\sigma,t}, \delta/3)$. Using the notation of Assumption 6.2.3, let $\widehat{h}$ denote the function estimate obtained as a result of running the ERM on k samples from the model (6.6). For a sufficiently large constant c_2, set

$$u = c_2 \cdot \sqrt{\log\left(\tfrac{c_2}{\delta}\right)} \cdot \left(\tau_k(\nu_t h^*; b) \vee \tfrac{b}{\sqrt{k}}\right) \tag{6.23}$$

in Proposition 6.4.1(a), and consider the function $f := \widehat{h} - \nu_t h^* \in \mathcal{H}_{\nu_t h^*}$. This yields the bound

$$\|\widehat{h} - \nu_t h^*\|_2^2 \leq 2\|\widehat{h} - \nu_t h^*\|_k^2 + c_2 \cdot \left(\tau_k^2(\nu_t h^*; b) \vee \frac{b^2}{k}\right) \log\left(\tfrac{c_2}{\delta}\right)$$

with probability exceeding $1 - \delta/9$. Moreover, applying the same result to the set of fresh samples $\overline{y}_1, \ldots, \overline{y}_k$ (note that our definition of the norm, etc. would have to change, but the same result applies), we have

$$\frac{1}{k}\sum_{i=1}^{k} \left(\widehat{h}(\overline{y}_i) - \nu_t h^*(\overline{y}_i)\right)^2 \leq \frac{3}{2}\|\widehat{h} - \nu_t h^*\|_2^2 + \frac{1}{2}u^2;$$

choosing u according to equation (6.23) and putting together the pieces implies the bound

$$\frac{1}{k}\sum_{i=1}^{k} \left(\widehat{h}(\overline{y}_i) - \nu_t h^*(\overline{y}_i)\right)^2 \leq 3\|\widehat{h} - \nu_t h^*\|_k^2 + c_2 \cdot \left(\tau_k^2(\nu_t h^*; b) \vee \frac{b^2}{k}\right) \log\left(\tfrac{c_2}{\delta}\right)$$

with probability exceeding $1 - \frac{2\delta}{9}$.

Finally, we bound $\|\widehat{h} - \nu_t h^*\|_k^2$ using Proposition 6.4.1(b). Setting

$$u = c_2 \cdot \log\left(\tfrac{c_2}{\delta}\right) \cdot \left(\mu_k(\nu_t h^*, y_1^k; 2\rho_{\sigma,t}) \vee \tfrac{\rho_{\sigma,t}^2}{k \cdot \mu_k(\nu_t h^*, y_1^k; 2\rho_{\sigma,t})}\right)$$

and simplifying, we obtain

$$\|\widehat{h} - \nu_t h^*\|_k^2 \leq c_2 \left(\mu_k^2(\nu_t h^*, y_1^k; 2\rho_{\sigma,t}) \vee \tfrac{\rho_{\sigma,t}^2}{k} \right) \log \left(\tfrac{c_2}{\delta} \right)$$

with probability exceeding $1 - \frac{\delta}{9}$. Putting together the pieces, we have shown the bound

$$\overline{\mathcal{R}}_k^{\mathsf{ERM}}(\nu_t h^*; \rho_{\sigma,t}, \delta/3) \leq c_2 \left\{ \left(\tau_k^2(\nu_t h^*; b) \vee \tfrac{b^2}{k} \right) + \left(\mu_k^2(\nu_t h^*, y_1^k; 2\rho_{\sigma,t}) \vee \tfrac{\rho_{\sigma,t}^2}{k} \right) \right\} \log \left(\tfrac{c_2}{\delta} \right).$$

Substituting this expression into Theorem 6.2.1 by setting $k = \overline{N}$ completes the proof, since at iteration t of the iterative algorithm, we have $y_1^k = \{y_i\}_{i \in \mathcal{D}_{2t+1}}$. □

6.4.5 Proof of Corollary 6.3.1

The proof of this result makes crucial use of Theorem 6.4.1, and the reader is advised to scan the previous sections for its statement and proof. Given Theorem 6.4.1, it suffices—in addition to establishing Assumptions 6.2.1, 6.2.2, and 6.4.1—to bound the complexity functions τ_k and μ_k. All of these steps are presented in the following lemmas. Recall the value ρ_{mono} defined in equation (6.8), and our shorthand $\mathsf{c}_{n,\delta} = \sqrt{2 \log(2n/\delta)}$.

Lemma 6.4.2. *For a non-decreasing function $g : \mathbb{R} \to \mathbb{R}$, consider the observation model*

$$Y = g(X) + \sigma Z, \tag{6.24}$$

with $Z \sim \mathcal{N}(0, 1)$ and X drawn from some Lebesgue measurable distribution. Then the function $h(y) = \mathbb{E}[X | Y = y]$ exists a.e., and is non-decreasing.

Lemma 6.4.3. *Suppose that in the monotone single-index model* (6.24)*, the link function g satisfies Assumption 6.3.1, and the covariate distribution is given by a Gaussian truncated to the interval $[-\mathsf{c}_{n,\delta}, \mathsf{c}_{n,\delta}]$. Then, Assumption 6.2.2 holds with $\rho_\sigma \leq \rho_{\mathsf{mono}}$.*

The next two lemmas are stated assuming that the function class $\mathcal{H}$ is given by

$$\mathcal{H}(b) = \big\{ h : \mathbb{R} \mapsto [-b, b] \; \big| \; h \text{ non-decreasing} \big\} \tag{6.25}$$

for some positive real number b. Recall our notation for the shifted function class around h^*, given by $\mathcal{H}_{h^*} = \{h - h^* \mid h \in \mathcal{H}\}$. In the following lemmas, we also assume that $h^* \in \mathcal{H}$.

Lemma 6.4.4. *For each function $h^* \in \mathcal{H}$, integer k, sequence of samples y_1^k, and scalar γ, we have*

$$\tau_k^2(h^*; b) \leq c_2 b^2 \left(\frac{\log k}{k} \right)^{2/3}, \textit{ and} \tag{6.26a}$$

$$\mu_k^2(h^*, y_1^k; \gamma) \leq c_2 \left(\frac{\gamma^2 b \log k}{k} \right)^{2/3}. \tag{6.26b}$$

for a sufficiently large constant c_2.

Remark 6.4.3. *Note that we have not been particularly careful about the* exact *logarithmic factor in the bounds* (6.26)*, since there are other logarithmic terms present in the final bound of Corollary 6.3.1. However, we do note that it is likely that these bounds can be sharpened to remove the logarithmic factor appearing on the RHS.*

We prove these lemmas at the end of the section. Taking them as given for the moment, let us establish Corollary 6.3.1. Begin by defining the event

$$\mathcal{E} = \{|\langle x_i,\, \theta^* \rangle| \le \mathsf{c}_{n,\delta} \text{ for all } i \in [n]\},$$

and noting that $\Pr\{\mathcal{E}\} \ge 1 - \frac{\delta}{4}$ by standard Gaussian tail bounds. We work on this event for the rest of the proof, so that Lemma 6.4.2 guarantees the inclusion $\mathcal{H}(\mathsf{c}_{n,\delta}) \supseteq \{y \mapsto \mathbb{E}[\langle X,\, \theta^* \rangle | Y = y]\}$. Now, recall our shorthand $\rho_{\sigma,t} = \sqrt{\Delta_t + \nu_t \rho_\sigma^2}$, and suppose that the pair (Δ_t, δ) satisfies the inequalities

$$\Delta_t \le \frac{99}{100} \quad \text{and} \quad \left\{ \left(\tau^2_{\overline{n}}(\nu_t h^*; b) \vee \tfrac{b^2}{\overline{n}} \right) + \left(\mu^2_{\overline{n}}(\nu_t h^*, \{y_i\}_{i \in \mathcal{D}_{2t+1}}; 2\rho_{\sigma,t}) \vee \tfrac{\rho^2_{\sigma,t}}{\overline{n}} \right) \right\} \log\left(\tfrac{c_2}{\delta}\right) + \rho^2_{\mathsf{mono}} \le c_1. \tag{6.27}$$

The second condition is satisfied for large enough $\overline{n}$, and by the assumption that ρ_{mono} is bounded above by a small enough constant. The tuple $(\mu_{\overline{n}}, \tau_{\overline{n}}, b)$ is chosen according to Lemma 6.4.4. Then, applying Theorem 6.4.1 and substituting the bounds guaranteed by Lemmas 6.4.3 and 6.4.4 yields the guarantee

$$\begin{aligned} \Delta_{t+1} &\le c_2 \left\{ \mathsf{c}^2_{n,\delta} \left(\frac{\log \overline{n}}{\overline{n}} \right)^{2/3} + \left(\mathsf{c}_{n,\delta} \left(\Delta_t + \rho^2_{\mathsf{mono}} \right) \frac{\log \overline{n}}{\overline{n}} \right)^{2/3} + \rho^2_{\mathsf{mono}} \right\} \frac{d}{\overline{n}} \log\left(\frac{c_2}{\delta} \right) \\ &= c_2 \left\{ \left(\frac{\log \overline{n}}{\overline{n}} \right)^{2/3} + \left(\frac{\Delta_t + \rho^2_{\mathsf{mono}}}{\overline{n}} \right)^{2/3} + \rho^2_{\mathsf{mono}} \right\} \frac{d}{\overline{n}} \log\left(\frac{c_2}{\delta} \right) \cdot \mathsf{c}^2_{n,\delta} \end{aligned}$$

with probability at least $1 - \delta$, where the reader should recall that the values of the absolute constants may change from line to line. This establishes the bound (6.9a).

It remains to translate this guarantee into a bound on the final iterate (6.9b). Toward that end, set $\delta = n^{-10}$, and note that $\mathsf{c}^2_{n,\delta} \sim \log n$ to obtain the simplified one-step guarantee

$$\Delta_{t+1} \le c_2 \left\{ \left(\frac{\log \overline{n}}{\overline{n}} \right)^{2/3} + \left(\frac{\Delta_t + \rho^2_{\mathsf{mono}}}{\overline{n}} \right)^{2/3} + \rho^2_{\mathsf{mono}} \right\} \frac{d}{\overline{n}} \log^2 n,$$

which holds for each iteration t on the corresponding event $\mathcal{E}_t$. On $\mathcal{E}_t$ and under the assumption $\overline{n} \ge C\sigma^2(\kappa_0^2 - 1) d \log^2 n$, it can be verified that Δ_{t+1} satisfies condition (6.27) for a large enough constant C. Consequently, the argument can be applied iteratively; for an integer value T_0 to be determined shortly, condition on the event $\cap_{i=0}^{T_0} \mathcal{E}_i$. By the union bound, this event occurs with probability exceeding $1 - T_0 n^{-10}$. Abusing notation slightly, let ρ_{mono} now denote the same quantity but with this value of δ substituted.

Now choose an integer value T satisfying $C \log\log n \leq T \leq T_0$ for a large enough absolute constant C and any $T_0 \leq n$. Let us apply Lemma B.2.1 in the appendix with the substitutions

$$\begin{aligned} C_1 &= c_2 \frac{d}{\overline{n}} \log^2 n \left\{ \left(\frac{\log \overline{n}}{\overline{n}} \right)^{2/3} + \rho^2_{\mathsf{mono}} \right\}, \\ C_2 &= c_2 \frac{d}{\overline{n}} \log^2 n, \text{ and} \\ C_3 &= \rho^2_{\mathsf{mono}}, \end{aligned}$$

and note that $\gamma = 2/3$ and $\Delta_0 \leq 1$ by definition. Then by choosing C large enough, we can ensure that T is large enough to satisfy the condition required by Lemma B.2.1. Consequently, we have

$$\begin{aligned} \Delta_T &\leq c \cdot \left\{ \frac{d}{\overline{n}} \log^2 n \left\{ \left(\frac{\log \overline{n}}{\overline{n}} \right)^{2/3} + \rho^2_{\mathsf{mono}} \right\} + \frac{d}{\overline{n}} \log^2 n \cdot \left(\frac{\rho^2_{\mathsf{mono}}}{\overline{n}} \right)^{2/3} + \frac{p^3}{\overline{n}^5} \log^6 n \right\} \\ &\overset{\text{(i)}}{\leq} c \cdot \left\{ \underbrace{\frac{d \log^2 n}{\overline{n}} \left(\frac{\log \overline{n}}{\overline{n}} \right)^{2/3}}_{T_1} + \underbrace{\frac{d \log^2 n}{\overline{n}} \rho^2_{\mathsf{mono}}}_{T_2} + \underbrace{\frac{d \log^2 n}{\overline{n}} \cdot \left(\frac{\rho^2_{\mathsf{mono}}}{\overline{n}} \right)^{2/3}}_{T_3} \right\} \end{aligned} \tag{6.28}$$

where in step (i), we have used the condition $\overline{n} \gtrsim d$ to obtain the bound

$$\frac{p^3}{\overline{n}^5} \log^6 n \lesssim \frac{d \log^2 n}{\overline{n}} \left(\frac{\log \overline{n}}{\overline{n}} \right)^{2/3}.$$

Finally, some algebra reveals that if $T_2 \leq T_3$, then $T_3 \leq T_1$, and so we may drop the term T_3 from the bound by changing the absolute constant, and this concludes the proof. We note that the poly-logarithmic factors in the final bound have not been optimized. □

It remains to prove the various lemmas.

Proof of Lemma 6.4.2

We use f_X and f_Y to denote the marginal densities of the pair (X, Y). The notation $f_{X,Y}$ is used to denote their joint density, and let $f_{X|Y}$ denote the conditional density $X|Y$. We use $\phi(\cdot)$ to denote the standard Gaussian PDF, and $\mathcal{X}$ to denote the support of X. We have

$$\begin{aligned} h(y) := \mathbb{E}[X|Y = y] &= \int_{\mathcal{X}} x \frac{f_{X,Y}(x, y)}{f_Y(y)} dx \\ &= \frac{\int_{\mathcal{X}} x f_X(x) \phi(y - g(x)) dx}{f_Y(y)}. \end{aligned}$$

Now note that we have $\frac{d}{dy} f_{X,Y}(x,y) = -\left(\frac{y-g(x)}{\sigma^2}\right) f_{X,Y}(x,y)$ for each y; this follows by differentiating the Gaussian PDF. Further, note that $f_Y(y) = \int_{\mathcal{X}} f_{X,Y}(\widetilde{x}, y) d\widetilde{x}$, so we have

$$\begin{aligned}\sigma^2 h'(y) &= \frac{1}{(f_Y(y))^2} \cdot \Big(- \int_{\mathcal{X}} \int_{\mathcal{X}} x(y - g(x)) \cdot f_{X,Y}(\widetilde{x}, y) f_{X,Y}(x, y) dx d\widetilde{x} \\ &\qquad\qquad + \int_{\mathcal{X}} \int_{\mathcal{X}} x(y - g(\widetilde{x})) \cdot f_{X,Y}(\widetilde{x}, y) f_{X,Y}(x, y) dx d\widetilde{x} \Big) \\ &= \frac{1}{(f_Y(y))^2} \cdot \int_{\mathcal{X}} \int_{\mathcal{X}} x(g(x) - g(\widetilde{x})) f_{X,Y}(\widetilde{x}, y) f_{X,Y}(x, y) dx d\widetilde{x}. \end{aligned} \tag{6.29}$$

Since the same statement holds with the roles of x and $\widetilde{x}$ interchanged, we also have

$$\sigma^2 h'(y) = \frac{1}{(f_Y(y))^2} \cdot \int_{\mathcal{X}} \int_{\mathcal{X}} \widetilde{x}(g(\widetilde{x}) - g(x)) f_{X,Y}(\widetilde{x}, y) f_{X,Y}(x, y) dx d\widetilde{x}. \tag{6.30}$$

Summing equations (6.29) and (6.30) yields

$$2\sigma^2 h'(y) = \frac{1}{(f_Y(y))^2} \cdot \int_{\mathcal{X}} \int_{\mathcal{X}} (x - \widetilde{x})(g(x) - g(\widetilde{x})) f_{X,Y}(\widetilde{x}, y) f_{X,Y}(x, y) dx d\widetilde{x} \geq 0,$$

where the inequality follows from the monotonicity of g, which ensures that $(x - \widetilde{x})(g(x) - g(\widetilde{x}))$ is non-negative. □

Proof of Lemma 6.4.3

Recall that our (forward) observation model on the set of labeled samples is given by

$$Y = g^*(W) + \sigma Z,$$

where W is a standard Gaussian truncated to the interval $[-\mathsf{c}_{n,\delta}, \mathsf{c}_{n,\delta}]$, the link function $g^* : \mathbb{R} \to \mathbb{R}$ is monotone, and Z is a standard normal independent of everything else. In addition, Assumption 6.3.1 also implies that the bounds

$$m|a - b| \leq |g^*(a) - g^*(b)| \leq M|a - b| \tag{6.31}$$

hold for each pair of scalars (a, b).

We now split the rest of the proof into two cases.

Case $g^*(-\mathsf{c}_{n,\delta}) \leq y \leq g^*(\mathsf{c}_{n,\delta})$: In this case, note that the function g^{-1} is uniquely defined. Let us use the shorthand

$$\underline{\tau} \equiv \sqrt{1 + \sigma^2/M^2} \leq \overline{\tau} \equiv \sqrt{1 + \sigma^2/m^2},$$

and note that both of these quantities are equal to 1 when $\sigma = 0$. It also convenient to define

$$\overline{\sigma} = \sqrt{\frac{\sigma^2}{m^2+\sigma^2}}, \qquad\qquad \underline{\sigma} = \sqrt{\frac{\sigma^2}{M^2+\sigma^2}},$$
$$\overline{\mu}(y) = \frac{g^{-1}(y)}{1+\sigma^2/m^2}, \qquad \text{and} \qquad \underline{\mu}(y) = \frac{g^{-1}(y)}{1+\sigma^2/M^2}.$$

Once again, it is useful to keep in mind that we have $\overline{\sigma} \approx \sigma/m$ and $\overline{\mu}(y) \approx \underline{\mu}(y) \approx g^{-1}(y)$ in the small σ regime. Finally, let ϕ_τ denote the density of a zero-mean Gaussian with standard deviation τ, and let $\phi \equiv \phi_1$. Let Φ denote the CDF of the standard Gaussian.

We require the following lemma about the joint density of the pair (W, Y).

Lemma 6.4.5. *For each $g^*(-\mathsf{c}_{n,\delta}) \le y \le g^*(\mathsf{c}_{n,\delta})$, we have*

$$f_{W,Y}(w,y) \le \frac{\overline{\tau}\,\overline{\sigma}}{\sigma}\phi_{\overline{\tau}}\left(g^{-1}(y)\right)\phi_{\overline{\sigma}}\left(w-\overline{\mu}(y)\right)\kappa^{-1}\mathbf{1}\left\{w \in [-\mathsf{c}_{n,\delta}, \mathsf{c}_{n,\delta}]\right\} \textit{ and} \tag{6.32a}$$
$$f_Y(y) \ge \frac{\underline{\tau}\,\underline{\sigma}}{3\sigma}\,\phi_{\underline{\tau}}\left(g^{-1}(y)\right)\kappa^{-1}. \tag{6.32b}$$

The proof of Lemma 6.4.5 is postponed to the end of the subsection. Taking it as given for the moment, let us complete the proof of Lemma 6.4.3. Let us use the shorthand $W_y \equiv [W|Y=y]$. Lemma 6.4.5 yields the tail bound

$$\begin{aligned}
\Pr(|W_y - \overline{\mu}| \ge t\overline{\sigma}) &\le \frac{3\overline{\tau}\,\overline{\sigma}}{\underline{\tau}\,\underline{\sigma}} \cdot \frac{\phi_{\overline{\tau}}\left(g^{-1}(y)\right)}{\phi_{\underline{\tau}}\left(g^{-1}(y)\right)}\Phi(-t) \wedge 1 \\
&\le \frac{3\overline{\tau}\,\overline{\sigma}}{\underline{\tau}\,\underline{\sigma}} \cdot \frac{\phi_{\overline{\tau}}\left(g^{-1}(y)\right)}{\phi_{\underline{\tau}}\left(g^{-1}(y)\right)}\exp(-t^2/2) \wedge 1 \\
&= \frac{3\overline{\sigma}}{\underline{\sigma}}\exp\left(-\frac{\left(g^{-1}(y)\right)^2}{2}\cdot\left(\overline{\tau}^{-2}-\underline{\tau}^{-2}\right)\right)\exp(-t^2/2) \wedge 1 \\
&\overset{\text{(i)}}{\le} \frac{3M}{m}\exp\left(\frac{\left(g^{-1}(y)\right)^2}{2}\cdot\left(\overline{\tau}^{-2}-\underline{\tau}^{-2}\right)\right)\exp(-t^2/2) \wedge 1,
\end{aligned}$$

where in step (i), we have used the relation $\sqrt{\frac{\sigma^2+M^2}{\sigma^2+m^2}} \le \frac{M}{m}$. Further substituting the values of the pair $(\overline{\tau}, \underline{\tau})$, we have

$$\begin{aligned}
\Pr(|W_y - \overline{\mu}| > t\overline{\sigma}) &\le \frac{3M}{m}\exp\left(\frac{\left(g^{-1}(y)\right)^2}{2}\cdot\overline{\sigma}\cdot\underline{\sigma}\cdot Mm(M^2-m^2)\right)\exp(-t^2/2) \wedge 1 \\
&\overset{\text{(ii)}}{\le} \frac{3M}{m}\exp\left(\frac{\left(g^{-1}(y)\right)^2}{2}\cdot\sigma^2(M^2-m^2)\right)\exp(-t^2/2) \wedge 1,
\end{aligned}$$

where in step (ii), we have used the fact that when $\rho_{\mathsf{mono}} \leq c_1$, we have $\sigma \leq Cm$. We now make note of the following series of inequalities, which holds for each tuple of positive scalars (K, t) satisfying $K \geq 1$:

$$\begin{aligned} e^K e^{-t^2/2} \wedge 1 &\leq \exp\left\{K - t^2/2 \wedge 0\right\} \\ &= \exp\left\{K(1 - \tfrac{t^2}{2K} \wedge 0\right\} \\ &\overset{(\mathsf{i})}{\leq} \exp\left\{(1 - \tfrac{t^2}{2K}) \wedge 0\right\} \\ &\leq \exp\left\{1 - \tfrac{t^2}{2K}\right\}. \end{aligned}$$

where step (i) uses the fact that $K \geq 1$. Also define the positive scalar

$$\Gamma(y) := \sqrt{\frac{(g^{-1}(y))^2}{2} \cdot \sigma^2(M^2 - m^2) + \log\left(\tfrac{3M}{m}\right)}$$

for convenience, and note that we have

$$\sup_{y \in [g^*(-\mathsf{c}_{n,\delta}), g^*(\mathsf{c}_{n,\delta})]} \Gamma(y) \leq \Gamma := \sqrt{\frac{\mathsf{c}_{n,\delta}^2}{2} \cdot \sigma^2(M^2 - m^2) + \log\left(\tfrac{3M}{m}\right)}.$$

Putting together the pieces, we have

$$\Pr\left(\frac{|W_y - \overline{\mu}|}{\overline{\sigma}} \geq t\right) \leq \exp\left(1 - \frac{t^2}{2\Gamma^2(y)}\right),$$

and applying Lemma 5.5 of Vershynin [319] yields the inequality

$$\|\xi(y)\|_{\psi_2} \overset{(\mathsf{ii})}{\leq} C\|W_y - \overline{\mu}\|_{\psi_2} \leq C\overline{\sigma}\Gamma(y) \leq C\left(\frac{\sigma}{m} \wedge 1\right) \cdot \Gamma,$$

where step (ii) follows from the fact that centering does not change the sub-Gaussian constant by more than a constant factor (see, e.g., Lemma 2.6.8 of Vershynin [318]). Finally, using once again the fact that $\sigma \leq Cm$ and applying the elementary inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, which holds for any pair of positive reals (a, b) establishes the result for this case.

Case 2: When $y \notin [g^*(-\mathsf{c}_{n,\delta}), g^*(\mathsf{c}_{n,\delta})]$, we proceed by showing that the desired sub-Gaussian constant is still less than $C(\overline{\sigma}\Gamma \vee \overline{\sigma}\mathsf{c}_{n,\delta})$ for an absolute constant C. Let $\bar{y} := g^*(\mathsf{c}_{n,\delta})$, and first consider the case $y \geq \bar{y}$. Define the (non-negative) random variable $W' := \mathsf{c}_{n,\delta} - W_y$, and use the notation $W \equiv W_y$ for simplicity. First, note that it suffices to show a bound on the smallest positive γ such that

$$\mathbb{E}[e^{\gamma(W')^2} | Y = y] \leq 2,$$

since centering at the mean only affects the sub-Gaussian constant by a constant factor [318].

Then the ratio between the conditional densities when $Y = y$ and $Y = \overline{y}$ is given by

$$\begin{aligned}\frac{f_{W|Y}(w|y)}{f_{W|Y}(w|\overline{y})} &= \mathcal{C}_{y,\overline{y}}\frac{f_{W,Y}(w,y)}{f_{W,Y}(w,\overline{y})}\\ &= \mathcal{C}_{y,\overline{y}}\frac{\phi_\sigma(y-g^*(w))}{\phi_\sigma(\overline{y}-g^*(w))}\\ &= \mathcal{C}_{y,\overline{y}}\exp\left(-((y-g^*(w))^2-(\overline{y}-g^*(w))^2)/2\sigma^2)\right),\end{aligned}$$

where $\mathcal{C}_{y,\overline{y}}$ is a positive parameter that depends on the pair $(y, \overline{y})$, but is independent of the scalar w. This yields the further chain of bounds

$$\begin{aligned}\frac{f_{W|Y}(w|y)}{f_{W|Y}(w|\overline{y})} &= \mathcal{C}_{y,\overline{y}}\exp\left(-((y-\overline{y})(\frac{y+\overline{y}}{2}-g^*(w))/\sigma^2)\right)\\ &= \mathcal{C}_{y,\overline{y}}\exp\left(-((y-\overline{y})/\sigma^2)\right)^{\frac{y+\overline{y}}{2}-g^*(w)}\\ &= \mathcal{C}'_{y,\overline{y}}\exp\left((y-\overline{y})/\sigma^2\right)^{g^*(w)},\end{aligned}$$

for a different positive constant $\mathcal{C}'_{y,\overline{y}}$ that is independent of w. Since $y \geq \overline{y}$, the likelihood ratio is non-decreasing in w. Equivalently, the likelihood ratio decreases as the quantity $\mathsf{c}_{n,\delta} - w$ increases. Consequently, the random variables $(\mathsf{c}_{n,\delta} - W)^2$ and $\frac{f_{W|Y}(W|y)}{f_{W|Y}(W|\overline{y})}$ are negatively correlated, and for each $\gamma > 0$, we have

$$\begin{aligned}\mathbb{E}\left[e^{(W')^2/\gamma^2}|Y=y\right] &= \mathbb{E}\left[e^{\gamma^{-2}(\mathsf{c}_{n,\delta}-W)^2}\frac{f_{W|Y}(W|y)}{f_{W|Y}(W|\overline{y})}|Y=\overline{y}\right]\\ &\leq \mathbb{E}\left[e^{\gamma^{-2}(\mathsf{c}_{n,\delta}-W)^2}|Y=\overline{y}\right]\cdot\mathbb{E}\left[\frac{f_{W|Y}(W|y)}{f_{W|Y}(W|\overline{y})}|Y=\overline{y}\right]\\ &= \mathbb{E}\left[e^{\gamma^{-2}(\mathsf{c}_{n,\delta}-W)^2}|Y=\overline{y}\right]\\ &= \mathbb{E}\left[\exp\left(\gamma^{-2}(\mathsf{c}_{n,\delta}-\overline{\mu}(\overline{y}))^2+(\overline{\mu}(\overline{y})-W)^2\right)|Y=\overline{y})\right]\\ &= \exp\left(\gamma^{-2}(\mathsf{c}_{n,\delta}-\overline{\mu}(\overline{y}))^2\right)\mathbb{E}\left[\exp\left(\gamma^{-2}(\overline{\mu}(\overline{y})-W)^2\right)|Y=\overline{y})\right].\end{aligned}$$

Finally, note that we have $(\mathsf{c}_{n,\delta} - \overline{\mu}(\overline{y}))^2 = \mathsf{c}_{n,\delta}^2\frac{\sigma^2}{m^2+\sigma^2} = \mathsf{c}_{n,\delta}^2\overline{\sigma}^2$. On the other hand, by case 1 of the proof, the expectation term in the last display is bounded by a constant provided $\gamma^2 \geq \overline{\sigma}^2\Gamma^2$. By adjusting the constant factors, we can ensure that $\mathbb{E}\left[e^{\gamma^{-2}(W')^2}|Y=y\right] \leq 2$ provided $\gamma \geq C\overline{\sigma}^2(\Gamma^2 \vee \mathsf{c}_{n,\delta}^2)$, and this completes the proof for the case $y \geq \overline{y}$. An identical argument holds when $y \leq -\overline{y}$, and combining the two cases yields the lemma. □

Proof of Lemma 6.4.5: Recall the notation $\kappa = \Pr\{Z \in \mathcal{I}\}$, so that the density of the random variable W is given by

$$f_W(w) = \kappa^{-1}\phi(w)\mathbf{1}\{w \in \mathcal{I}\};$$

we use the shorthand $\kappa(w) := \kappa^{-1}\mathbf{1}\left\{w \in \mathcal{I}\right\}$ for convenience.

Let us begin by deriving the joint density. We have

$$\begin{aligned}
f_{W,Y}(w,y) &= f_{Y|W}(y|w) f_W(w) \\
&= \phi_\sigma(y - g(x))\ \phi(w)\kappa(w) \\
&\overset{\text{(i)}}{\leq} \phi_\sigma(m(g^{-1}(y) - w))\ \phi(w)\kappa(w) \\
&= \frac{1}{2\pi\sigma}\exp\left(-\frac{(g^{-1}(y) - w)^2}{2\sigma^2/m^2} - \frac{w^2}{2}\right)\kappa(w),
\end{aligned}$$

where step (i) follows from equation (6.31). Completing the squares and performing some more algebra leads to the relation

$$\begin{aligned}
f_{W,Y}(w,y) &\leq \frac{1}{2\pi\sigma}\exp\left(-\frac{g^{-1}(y)^2}{2(1+\sigma^2/m^2)}\right)\exp\left(-\frac{(1+\sigma^2/m^2)\left(w - \frac{g^{-1}(y)}{1+\sigma^2/m^2}\right)^2}{2\sigma^2/m^2}\right)\kappa(w) \\
&= \frac{1}{2\pi\sigma}\exp\left(-\frac{g^{-1}(y)^2}{2\overline{\tau}^2}\right)\exp\left(-\frac{(w - \overline{\mu}(y))^2}{2\overline{\sigma}^2}\right)\kappa(w) \\
&= \frac{\overline{\tau}\ \overline{\sigma}}{\sigma}\phi_{\overline{\tau}}\left(g^{-1}(y)\right)\phi_{\overline{\sigma}}\left(w - \overline{\mu}(y)\right)\kappa(w),
\end{aligned}$$

and this proves inequality (6.32a).

We now turn to establishing bound (6.32b). We have

$$\begin{aligned}
f_Y(y) &= \int_{\mathcal{I}} f_{W,Y}(w,y)dw \\
&= \int \phi_\sigma(y - g(x))\ \phi(w)\kappa(w)dw \\
&\overset{\text{(ii)}}{\geq} \int \phi_\sigma(M(g^{-1}(y) - w))\ \phi(w)\kappa(w)dw,
\end{aligned}$$

where step (ii) once again follows from equation (6.31). Completing the square similarly to above and performing some more algebra yields

$$\begin{aligned}
f_Y(y) &\geq \int \frac{1}{2\pi\sigma}\exp\left(-\frac{(g^{-1}(y))^2}{2(1+\sigma^2/M^2)}\right)\exp\left(-\frac{(1+\sigma^2/M^2)\left(w - \frac{g^{-1}(y)}{1+\sigma^2/M^2}\right)^2}{2\sigma^2/M^2}\right)\kappa(w)dw \\
&= \frac{\underline{\tau}\ \underline{\sigma}}{\sigma}\phi_{\underline{\tau}}\left(g^{-1}(y)\right)\int \phi_{\underline{\sigma}}\left(w - \underline{\mu}(y)\right)\kappa(w)dw.
\end{aligned}$$

The following sequence of relations then completes the proof.

$$\begin{aligned}\kappa \int \phi_{\underline{\sigma}}(w - \underline{\mu}(y))\kappa(w)dw &= \int_{-\mathsf{c}_{n,\delta}}^{\mathsf{c}_{n,\delta}} \phi_{\underline{\sigma}}(w - \underline{\mu}(y))dw \\ &\geq \int_{\underline{\mu}(y)}^{\mathsf{c}_{n,\delta}} \phi_{\underline{\sigma}}(w - \underline{\mu}(y))dw \\ &= \int_{0}^{\underline{\mu}(y)+\mathsf{c}_{n,\delta}} \phi_{\underline{\sigma}}(w)dw \\ &\geq \int_0^2 \phi(z)dz > 1/3,\end{aligned}$$

where we have used the inequalities $0 \leq \underline{\mu}(y) \leq \mathsf{c}_{n,\delta}$, $\mathsf{c}_{n,\delta} \geq 2$, and $\underline{\sigma} \leq 1$. □

Proof of Lemma 6.4.4

The proof of both claims in this lemma are based on the following result that bounds the expected supremum of the associated empirical process. In it, we let $\nu = (\nu_1, \dots, \nu_k)$ denote a sequence of i.i.d. 1-sub-Gaussian random variables that is independent of everything else.

Lemma 6.4.6. *For each function $h \in \mathcal{H}$, sequence of samples $y_1, \dots, y_k$, and scalar ϑ, we have*

$$\mathbb{E}_\nu \left[\sup_{\substack{h \in \mathcal{H} \\ \|h-h^*\|_k \leq \vartheta}} \left| \frac{1}{k} \sum_{i=1}^k \nu_i \cdot (h - h^*)(y_i) \right| \right] \leq c_2 \sqrt{\frac{b\vartheta\left(\log(b/\vartheta) \vee 1\right)}{k}}.$$

A variant of this claim can be found, for instance, in van de Geer [308], but we provide the proof at the end of this section for completeness. Taking the lemma as given for the moment, let us establish the two bounds. For convenience, we use the shorthand $\mathcal{H}^* \equiv \mathcal{H}_{h^*}$.

Proof of claim (6.26a)**:** We must establish a bound on the (localized) population Rademacher complexity of the function class $\mathcal{H}$, which contains functions that are uniformly bounded by b. It is helpful to work instead with the *empirical* Rademacher complexity, which, for an abstract function class $\mathcal{F}$ takes the form

$$\widehat{\mathfrak{R}}_k(\mathcal{F}) := \mathbb{E}_\eta \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{k} \sum_{i=1}^k \eta_i f(y_i) \right| \right]. \tag{6.33}$$

Note that we no longer take an expectation over the design points, and so this complexity measure should be viewed as a random variable when the samples $y_1, \dots, y_k$ are random. Recall the norm $\|\cdot\|_k$ defined in equation (6.17), and let $\widehat{\tau}_k(h^*; \gamma)$ denote the smallest positive solution to the (empirical) critical equality

$$\frac{\tau^2}{\gamma} = \widehat{\mathfrak{R}}_k(\mathcal{H}^* \cap \mathbb{B}(\|\cdot\|_k; \tau)) \tag{6.34}$$

for some positive scalar γ. Since the function class $\mathcal{H}^*$ is $2b$-bounded and star-shaped around 0, a slight modification of Proposition 14.25 of Wainwright (see also the discussion surrounding equations (14.6)-(14.8)) guarantees that there is an absolute constant c such that

$$\tau_k^2(h^*; b) \leq c \cdot \widehat{\tau}_k^2(h^*; b)$$

with probability exceeding $1 - \exp\{-c_1 k \tau_k^2(h^*; b)/b^2\}$. Consequently, it suffices to bound the (random) quantity $\widehat{\tau}_k$; we dedicate the rest of this proof to such a bound. Applying Lemma 6.4.6, we are looking for the smallest strictly positive solution to the inequality

$$\frac{\tau^2}{b} \leq c_2 \left(\frac{b\tau \log(b/\tau)}{k} \right)^{1/2},$$

and solving with equality yields the bound.

□

Proof of claim (6.26b)**:** By definition of the functional μ_k and Lemma 6.4.6, we are looking for the smallest strictly positive solution to the inequality

$$\frac{\mu^2}{\gamma} \leq c_2 \left(\frac{b\mu \log(b/\mu)}{k} \right)^{1/2},$$

and solving with equality yields the claimed bound. □

Proof of Lemma 6.4.6: Since we are interested in bounding the sub-Gaussian complexity over the class of bounded monotone functions, we appeal to arguments based on metric entropy bounds and chaining [313]. Let us provide some background for this method, starting with the definition of the covering number of a set in a metric space.

Definition 6.4.1 (Covering number)**.** *An ϵ-cover of a set $\mathbb{T}$ with respect to a metric ρ is a set $\{\theta^1, \theta^2, \ldots, \theta^N\} \subset \mathbb{T}$ such that for each $\theta \in \mathbb{T}$, there exists some $i \in [N]$ such that $\rho(\theta, \theta_i) \leq \epsilon$. The ϵ-covering number $N(\epsilon; \mathbb{T}, \rho)$ is the cardinality of the smallest ϵ-cover.*

The logarithm of the covering number is referred to as the metric entropy of a set. It is well known that the sub-Gaussian complexities of sets can be bounded via their metric entropies, and we employ this approach below. View the samples $y_1, \ldots, y_k$ as fixed in our particular problem, and use the shorthand $\mathbb{B}_n(\mu; \mathcal{H}^*)) := \{h \in \mathcal{H} \mid \|h - h^*\|_k \leq \mu\}$. Then we have the upper bound in terms of Dudley's entropy integral (see Theorem 5.22 of Wainwright [323]):

$$\mathbb{E}_\nu \left[\sup_{\substack{h \in \mathcal{H} \\ \|h - h^*\|_k \leq \vartheta}} \left| \frac{1}{k} \sum_{i=1}^{k} \nu_i \cdot (h - h^*)(y_i) \right| \right] \leq \frac{16}{\sqrt{k}} \int_0^\vartheta \sqrt{\log N(t; \mathbb{B}_n(\vartheta; \mathcal{H}^*), \| \cdot \|_k)} dt.$$

It is a classical fact that we have the bound

$$\log N(t; \mathbb{B}_n(\mu; \mathcal{H}^*), \|\cdot\|_k) \leq \frac{cb}{t} \max\{\log(b/t), 1\}$$

for some absolute constant c (see, e.g., Example 2.1(i) of van de Geer [308]), where b is the uniform bound on functions in $\mathcal{H}$. Substituting this bound and simplifying yields the claim. □

6.5 Summary and open questions

Our work provides a general-purpose method by which the bias in performing parameter estimation under the class of single-index models can be significantly reduced; crucially, this involves leveraging properties of the function class to which the nonparametric link function belongs. Our approach should be viewed as reduction based: given an appropriate labeling oracle, we are able to reduce the problem to performing nonparametric regression over a suitably defined inverse function class. Our analysis is black-box and also reduction-based, in that it allows any nonparametric function estimator, and derives a final rate of parameter estimation depending on the rate of the nonparametric estimator. We particularized this framework to the case where the nonparametric function estimator was given by least-squares, or empirical risk minimization.

To illustrate this general framework, we derived end-to-end parameter estimation guarantees for a sub-class of monotone single-index models, improving upon the rates of classical semiparametric estimators. Owing to the reduction in bias, this improvement is particularly stark as the noise level $\sigma \to 0$. In particular, when the model is noiseless, we showed a sharpened rate for the problem using a slightly different analysis method adapted to a natural variant of the procedure. In addition, we showed an information-theoretic identifiability limit for the problem of parameter estimation in monotone SIMs.

The generality of our framework raises many interesting questions. For instance, are there other classes of SIMs for which a labeling oracle is implementable in a computationally efficient manner? Another important assumption that was made in our paper was that of Gaussian covariates. Strictly speaking, this assumption can be weakened slightly provided the noise in the "nuisance" directions (corresponding to directions of covariate space that are orthogonal to $\langle X, \theta^* \rangle$) are well-behaved under conditioning. A rigorous extension to this class of covariates is an interesting open problem, and is likely to significantly broaden the scope of our results. Finally, there is the question—regarded as widely important in the statistical signal processing literature—of how these approaches should be modified when the true parameter $\theta^* \in \mathcal{K}$ for some (typically convex) set $\mathcal{K} \subseteq \mathbb{R}^d$. Is it sufficient to perform the linear regression step in our algorithms under this additional restriction? What are the rates achieved by such a procedure in the high signal regime?

In the broader context of this dissertation, let us emphasize the main takeaway of this chapter: that the alternating minimization methodology is both computationally tractable and adapts to the noise level in single-index models, resulting in significantly faster estimation rates in the low-noise (or high-signal) regime.

Part III

Reinforcement learning

What is the instance-specific complexity of estimating a value function?

Chapter 7

The policy evaluation problem

Reinforcement learning (RL) is a class of methods for the optimal control of dynamical systems [27–29, 292] that has begun to make inroads in a wide range of applied problem domains. This empirical research has revealed the limitations of our theoretical understanding of this class of methods—popular RL algorithms exhibit a variety of behavior across domains and problem instances, and existing theoretical bounds, which are generally based on worst-case assumptions, fail to capture this variety. An important theoretical goal is to develop *instance-specific* analyses that help to reveal what aspects of a given problem make it "easy" or "hard". In the context of this thesis, such a goal corresponds to the strongest form of adaptation guarantee, and allows distinctions to be drawn between ostensibly similar algorithms in terms of their performance profiles. In this portion of the thesis, we ask such precise questions for the *policy evaluation* problem. This chapter provides an introduction to this problem and equips the reader with the necessary background for Chapters 8 and 9 to follow.

7.1 Introduction

A variety of applications spanning science and engineering use Markov reward processes as models for real-world phenomena, including queuing systems, transportation networks, robotic exploration, game playing, and epidemiology. In some of these settings, the underlying parameters that govern the process are known to the modeler, but in others, these must be estimated from observed data. A salient example of the latter setting, which forms the main motivation for our work, is the policy evaluation problem encountered in Markov decision processes (MDPs) and reinforcement learning [27, 28, 292]. Here an agent operates in an environment whose dynamics are unknown: at each step, it observes the current state of the environment, and takes an action that changes its state according to some stochastic transition function determined by the environment. The goal is to evaluate the utility of some policy—that is, a mapping from states to actions, where utility is measured using rewards that the agent receives from the environment. These rewards are usually assumed to be additive over time, and since the policy determines the action to be taken at each state, the reward obtained at any time is simply a function of the current state of the agent.

Thus, this setting induces a Markov reward process (MRP) on the state space, in which both the underlying transitions and rewards are unknown to the agent. The agent only observes samples of state transitions and rewards.

Given these samples, the goal of the agent is to estimate the *value function* of the MRP. As noted above, in the context of Markov decision processes (MDPs), this problem is known as policy evaluation. The value function evaluated at a given state measures the expected long-term reward accumulated by starting at that state and running the underlying Markov chain. In applications, this value function encodes crucial information about the MRP. For example, there are MRPs in which the value function corresponds to the probability of a power grid failing [109] or the value of a board configuration in a game of Go [282]. Moreover, policy evaluation is an important component of many policy optimization algorithms for reinforcement learning, which use it as a sub-routine while searching for good policies to deploy in the environment.

The focus of the next two chapters is on understanding the policy evaluation problem in finite-state (or tabular) MRPs in an instance-dependent manner, focusing on the the generative setting in which the agent has access to a simulator that generates samples from the underlying MRP. In particular, we would like guarantees on the *sample complexity* of policy evaluation—defined as the number of samples required to obtain a value function estimate of some pre-specified error tolerance—as a function of the agent's environment, i.e., the transition and reward functions induced by the policy being evaluated. Local guarantees of this form provide more guidance for algorithm design in finite sample settings than their worst-case counterparts. Indeed, as sketched in the introductory chapter, this viewpoint underpins the important sub-field of local minimax complexity studied widely in the statistics and optimization literatures (e.g., [48, 354]), as well as in more recent work on online reinforcement learning algorithms [345].

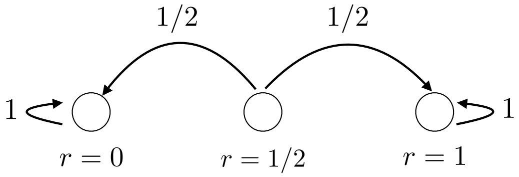


Figure 7.1: A simple 3-state Markov reward process.

The benefits of our instance-dependent guarantees are even evident in a model as simple as the 3-state MRP illustrated in Figure 7.1. Suppose that we observe noiseless rewards of this MRP and wish to compute its infinite-horizon value function with discount factor $\gamma \in (0, 1)$. Bounds based on the contractivity of the Bellman operator [160, 166, 324] imply that the ℓ_∞-error of the plug-in estimate scales proportionally to $1/(1-\gamma)^2$. The worst-case bounds of Azar et al. [7] imply a rate $1/(1-\gamma)^{3/2}$. But the optimal local result captured in Chapter 8 shows that the error is only proportional to $1/(1-\gamma)$. For a discount factor $\gamma = 0.99$, this improves the previous bounds by factors of 100 and 10, respectively, and consequently, the respective sample complexities by factors of 10^4 and 10^2. Instance-dependent results therefore allow us to differentiate problems that are "solvable" with finite samples from those that are not.

Before proceeding to our assessment of various algorithms, let us set down some important background for Chapters 8 and 9.

7.2 Background and problem formulation

In this section, we introduce the basic notation required to specify a Markov reward process, and formally define the problem of estimating value functions in the generative setting.

7.2.1 Markov reward processes and value functions

We study Markov reward processes defined on a finite set $\mathcal{X}$ of states, indexed as $\mathcal{X} = \{1, 2, \ldots, D\}$. The state evolution over time is determined by a set of transition functions $\{P(\,\cdot \mid x), \;\; x \in \mathcal{X}\}$, with the transition from state x to the next state being randomly chosen according to the distribution $P(\,\cdot \mid x)$. For notational convenience, we let $\mathbf{P} \in [0,1]^{D \times D}$ denote a row stochastic (Markov) transition matrix, where row j of this matrix—which we denote by p_j—collects the transition function of the j-th state. Also associated with an MRP is a *population* reward function $r : \mathcal{X} \mapsto \mathbb{R}$: transitioning from state x results in the reward $r(x)$. For convenience, we engage in a minor abuse of notation by letting r also denote a vector of length D, with r_j corresponding to the reward obtained at the j-th state.

In this part of the thesis, we consider the infinite-horizon, discounted reward as our notion for the long-term value of a state in the MRP. In particular, for a scalar discount factor $\gamma \in (0, 1)$, the long-term value of state x in the MRP is given by

$$\theta^*(x) := \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r(x_k) \;\middle|\; x_0 = x\right], \qquad \text{where } x_k \sim P(\,\cdot \mid x_{k-1}) \text{ for all } k \geq 1.$$

In words, this measures the expected discounted reward obtained by starting at the state x, where the expectation is taken with respect to the random transitions over states. Once again, we use θ^* to also denote a vector of length D, where θ^*_j corresponds to the value of the j-th state.

A note to the reader: in the sequel, we often reference a state simply by its index, and often refer to the state space $\mathcal{X} \equiv [D]$. Accordingly, we also use $P(\,\cdot \mid j)$ to denote the transition function corresponding to state $j \in [D]$.

7.2.2 Observation model

Given access to the true transition and reward functions, it is straightforward, at least in principle, to compute the value function. By definition, it is the unique solution of the Bellman fixed point relation

$$\theta^* = r + \gamma \mathbf{P} \theta^*. \tag{7.1}$$

In the learning setting, the pair $(\mathbf{P}, r)$ is unknown, and we instead assume access to a black box that generates samples from the transition and reward functions. In this part of the thesis, we operate under a setting known as the *synchronous or generative setting*; it is a stylized observation model that has been used extensively in the study of Markov decision processes (see Kearns and Singh [166] for an introduction). Let us introduce it in the context of MRPs: for a given sample index $k = 1, 2, \ldots, N$ and for each state $j \in [D]$, we observe a random next state $X_{k,j} \in [D]$ drawn according to the transition function $P(\,\cdot \mid j)$, and a random reward $R_{k,j}$ drawn from a conditional distribution $\mathcal{D}_r(\,\cdot \mid j)$. Throughout, we assume that the rewards are generated independently across states, with $\mathbb{E}[R_{k,j}] = r_j$. Letting $\rho(r)$ denote a non-negative vector indexed by the states $j \in [D]$, we assume the conditional distributions $\{\mathcal{D}_r(\,\cdot \mid j),\ j \in [D]\}$ are $\rho(r)$-sub-Gaussian, meaning that for each $j \in [D]$, we have

$$\mathbb{E}_{R\sim\mathcal{D}_r(\,\cdot\,|j)}\left[e^{\lambda(R-r_j)}\right] \leq e^{\frac{\lambda^2\rho_j^2(r)}{2}} \qquad \text{for all } \lambda \in \mathbb{R}. \tag{7.2}$$

With N such i.i.d. samples in hand, our goal is to estimate the value function θ^* in the ℓ_∞-error metric.

Such a goal is particularly relevant to the policy evaluation problem described in the introduction, since ℓ_∞-estimates of the value function can be used in conjunction with a policy improvement subroutine to eventually arrive at an optimal policy (see, e.g., Section 1.2.2. of the recent monograph [2]). We note in passing that bounds proved under the generative model may be translated into the more challenging online setting via the notion of Markov cover times (see, e.g., the papers [6, 95] for conversions of this type for Markov decision processes).

7.2.3 The plug-in estimator

A natural approach to this problem is use the observations to construct estimates $(\widehat{\mathbf{P}}, \widehat{r})$ of the pair $(\mathbf{P}, r)$, and then substitute or "plug in" these estimates into the Bellman equation, thereby obtaining the value function of the MRP having transition matrix $\widehat{\mathbf{P}}$ and reward vector $\widehat{r}$.

In order to define the plug-in estimator, let us introduce some helpful notation. For each time index k, we use the associated set of state samples $\{X_{k,j}, j \in [D]\}$ to form a random binary matrix $\mathbf{Z}_k \in \{0,1\}^{D\times D}$, in which row j has a single non-zero entry, determined by the sample $X_{k,j}$. Thus, the location of the non-zero entry in row j is drawn from the probability distribution defined by p_j, the j-th row of $\mathbf{P}$. Recall that our observations also include the stochastic reward vectors $\{R_k\}_{k=1}^N$ sampled from the reward distribution $\mathcal{D}_r$. Based on these observations, we define the sample means

$$\widehat{\mathbf{P}} = \frac{1}{N}\sum_{k=1}^{N}\mathbf{Z}_k \quad \text{and} \quad \widehat{r} = \frac{1}{N}\sum_{k=1}^{N} R_k, \tag{7.3}$$

which can be seen as unbiased estimates of the transition matrix $\mathbf{P}$ and the reward vector r, respectively.

The estimates $(\widehat{\mathbf{P}}, \widehat{r})$ define a new MRP, and its value function is given by the fixed point relation

$$\widehat{\theta}_{\mathsf{plug}} = \widehat{r} + \gamma\widehat{\mathbf{P}}\widehat{\theta}_{\mathsf{plug}}. \tag{7.4}$$

Solving this fixed point equation, we obtain the closed form expression $\widehat{\theta}_{\mathsf{plug}} = (\mathbf{I} - \gamma\widehat{\mathbf{P}})^{-1}\widehat{r}$ for the plug-in estimator. Note that the terminology "plug-in" arises the fact that $\widehat{\theta}_{\mathsf{plug}}$ is obtained by substituting the estimates $(\widehat{\mathbf{P}}, \widehat{r})$ into the original Bellman equation (7.1). We also note that in this special case—that is, the tabular setting without function approximation—the plug-in estimate is equivalent to the LSTD solution [39, 41].

In Chapter 8, we also introduce a close relative of the plug-in estimator that makes it significantly more "robust".

7.2.4 Operator view of stochastic approximation

Given these samples, define the k-th (noisy) linear operator $\widehat{\mathcal{T}}_k : \mathbb{R}^D \mapsto \mathbb{R}^D$ whose evaluation at the point θ is given by

$$\widehat{\mathcal{T}}_k(\theta) = R_k + \gamma \mathbf{Z}_k \theta. \tag{7.5}$$

The construction of these operators is inspired by the fact that we are interested in computing the fixed point of the population operator,

$$\mathcal{T} : \theta \mapsto r + \gamma \mathbf{P}\theta, \tag{7.6}$$

and a classical and natural way to do so is via a form of stochastic approximation known as temporal difference learning.

Classical temporal difference (TD) learning algorithms are parameterized by a sequence of stepsizes, $\{\alpha_k\}_{k\geq 1}$, with $\alpha_k \in (0, 1]$. Starting with an initial vector $\theta_1 \in \mathbb{R}^D$, the TD updates take the form

$$\theta_{k+1} = (1 - \alpha_k)\theta_k + \alpha_k \widehat{\mathcal{T}}_k(\theta_k) \quad \text{for } k = 1, 2, \ldots. \tag{7.7}$$

In the sequel, we discuss three popular stepsize choices:

$$\texttt{Constant stepsize:} \quad \alpha_k = \alpha, \qquad \text{where} \qquad 0 < \alpha \leq \alpha_{\max}. \tag{7.8a}$$

$$\texttt{Polynomial stepsize:} \quad \alpha_k = \frac{1}{k^\omega} \qquad \text{for some } \omega \in (0, 1). \tag{7.8b}$$

$$\texttt{Recentered-linear stepsize:} \quad \alpha_k = \frac{1}{1 + (1-\gamma)k}. \tag{7.8c}$$

In addition to the TD sequence (7.7), it is also natural to perform *Polyak-Ruppert averaging*, which produces a parallel sequence of averaged iterates

$$\widetilde{\theta}_k = \frac{1}{k}\sum_{j=1}^{k} \theta_j \quad \text{for } k = 1, 2, \ldots. \tag{7.9}$$

Such averaging schemes were introduced in the context of general stochastic approximation by Polyak [255] and Ruppert [270]. A large body of theoretical literature demonstrates that such an averaging scheme improves the rates of convergence of stochastic approximation when run with overly "aggressive" stepsizes.

7.2.5 Local complexity measures

In order to make our local guarantees precise, we require precise definitions of the "complexity" of an instance. First, define the span semi-norm of a value function θ as

$$\|\theta\|_{\text{span}} := \max_{x \in \mathcal{X}} \theta(x) - \min_{x \in \mathcal{X}} \theta(x).$$

Equivalently, the span semi-norm is equal to the variation of the vector $\theta \in \mathbb{R}^D$; see Puterman [258] for more details.

For each vector $\theta \in \mathbb{R}^D$, define the vector of empirical variances

$$\widehat{\sigma}^2(\theta) := \widehat{\mathbb{E}} \left| (\mathbf{Z} - \widehat{\mathbf{P}})\theta \right|^2,$$

where $\widehat{\mathbb{E}}$ denotes expectation over the empirical distribution (i.e., the random matrix $\mathbf{Z}$ is drawn uniformly at random from the set $\{\mathbf{Z}_k\}_{k=1}^N$). Note that given θ, this quantity is computable purely from the observed samples. Also define population variance vector

$$\sigma^2(\theta) := \mathbb{E} \left| (\mathbf{Z} - \mathbf{P})\theta \right|^2,$$

where in this case $\mathbf{Z}$ is drawn according to the population model $\mathbf{P}$.

Define the covariance matrix of the vector $(\mathbf{Z} - \mathbf{P})\theta$ as

$$\Sigma_{\mathbf{P}}(\theta) := \text{cov}_{\mathbf{Z} \sim \mathbf{P}} \left((\mathbf{Z} - \mathbf{P})\theta \right). \tag{7.10}$$

We often use $\Sigma(\theta)$ as a shorthand for $\Sigma_{\mathbf{P}}(\theta)$ when the underlying transition matrix $\mathbf{P}$ is clear from the context. With these definitions in hand, define the complexity measures

$$\nu(\mathbf{P}, \theta) := \max_{\ell \in [D]} \left(e_\ell^\top (\mathbf{I} - \gamma \mathbf{P})^{-1} \Sigma(\theta) (\mathbf{I} - \gamma \mathbf{P})^{-\top} e_\ell \right)^{1/2}, \quad \text{and} \tag{7.11a}$$

$$\rho(\mathbf{P}, r) := \sigma_r \left\| (\mathbf{I} - \gamma \mathbf{P})^{-1} \right\|_{2,\infty} \equiv \sigma_r \max_{\|u\|_2 = 1} \|(\mathbf{I} - \gamma \mathbf{P})^{-1} u\|_\infty. \tag{7.11b}$$

Note that $\nu(\mathbf{P}, \theta)$ corresponds to the maximal variance of the random vector $(\mathbf{I} - \gamma \mathbf{P})^{-1}(\mathbf{Z} - \mathbf{P})\theta$. We also use the convenient shorthand

$$b(\theta) := \frac{\|\theta\|_{\text{span}}}{1 - \gamma} \tag{7.11c}$$

in order to define our final measure of complexity.

7.3 A "toy" Markov reward process

In the next two chapters, we study the local behavior of various algorithms. It will be helpful to specialize these results to a toy MRP in which the instance-specific complexities can be changed in a transparent and straightforward fashion.

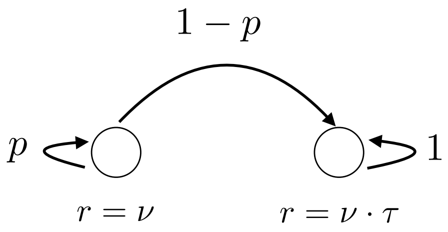


Figure 7.2: Illustration of the 2-state MRP used in the simulation. The triple of scalars (p, ν, τ), along with the discount factor γ, are parameters of the construction. The chain remains in state 1 with with probability p and transitions to state 2 with probability $1-p$; on the other hand, state 2 is absorbing. The rewards in states 1 and 2 are deterministic, specified by ν and $\nu\tau$, respectively.

This MRP consists of $D=2$ states, where state 1 stays fixed with probability p, transitions to state 2 with probability $1-p$, and state 2 is absorbing. The rewards in states 1 and 2 are given by ν and $\nu\tau$, respectively. Here the triple (p, ν, τ), along with the discount factor γ, are parameters of the construction. See Figure 7.2 for an illustration.

In some cases, it will be convenient to parameterize this MRP in a scalarized manner, in which case we vary the triple (p, ν, τ) in the following way. First, we fix a scalar λ in the unit interval $[0, 1]$, and then we set

$$p = \tfrac{4\gamma-1}{3\gamma}, \qquad \nu = 1, \quad \text{and} \quad \tau = 1-(1-\gamma)^{\lambda}. \tag{7.12}$$

Note that this sub-family of MRPs is fully parameterized by the pair (γ, λ). In particular, the local complexity quantities introduced in Section 7.2.5 can be explicitly calculated as functions of these two parameters. It can be shown via simple calculations that the underlying MRP satisfies

$$\nu(\mathbf{P}, \theta^*) \sim \left(\frac{1}{1-\gamma}\right)^{1.5-\lambda}, \quad \rho(\mathbf{P}, r) = 0 \quad \text{and} \quad b(\theta^*) \sim \left(\frac{1}{1-\gamma}\right)^{2-\lambda}, \tag{7.13a}$$

and furthermore, that

$$\|\sigma(\theta^*)\|_\infty \sim \left(\frac{1}{1-\gamma}\right)^{0.5-\lambda}, \; \|(\mathbf{I}-\gamma\mathbf{P})^{-1}\sigma(\theta^*)\|_\infty \sim \left(\frac{1}{1-\gamma}\right)^{1.5-\lambda}, \; \text{and} \; \|\theta^*\|_{\text{span}} \sim \left(\frac{1}{1-\gamma}\right)^{1-\lambda}, \tag{7.13b}$$

provided $N \gtrsim 1/(1-\gamma)$ and $\gamma \geq \frac{1}{2}$. These calculations are presented in Appendix C.1, but the bounds (7.13) will be useful both in our lower bound constructions and simulations to follow.

Having formally set up the necessary background, we now turn to formal guarantees on these procedures and their variants in Chapters 8 and 9.

Chapter 8

Instance-dependent adaptation of plug-in estimator

As a natural first step towards providing local guarantees for the policy evaluation problem, we analyze the plug-in estimator for the problem, which estimates the underlying transition and reward functions from the samples, and outputs the value function of the MRP in which these estimates correspond to the ground truth parameters. We also analyze a robust variant of this approach, and provide minimax lower bounds that hold over subsets of the parameter space.

Related work: Markov reward processes have a rich history originating in the theory of Markov chains and renewal processes; we refer the reader to the classical books [99] and [90] for introductions to the subject. The policy evaluation problem has seen considerable interest in the stochastic control and reinforcement learning communities, and various algorithms have been analyzed in both asymptotic [36, 295] and non-asymptotic [183, 289] settings. Chapter 3 of the monograph by Szepesvari [294] provides a brief introduction to these methods, and the recent survey by Dann et al. [77] focuses on methods based on temporal differences [293].

In the language of temporal difference (TD) algorithms, the plug-in approach that we analyze corresponds to the least squares temporal difference (LSTD) solution [41] in the tabular setting, without function approximation. While TD algorithms for policy evaluation have been analyzed by many previous papers, their focus is typically either on (i) how function approximation affects the algorithm [301], (ii) asymptotic convergence guarantees [36, 295] or (iii) establishing convergence rates in metrics of the ℓ_2-type [183, 289, 295]. Since ℓ_2-type metrics can be associated with an inner product, many specialized analyses can be ported over from the literature on stochastic optimization (e.g., [9, 232]).[1] On the other hand, our focus is on providing non-asymptotic guarantees in the ℓ_∞-error metric, since these are particularly compatible with policy iteration methods. In particular, policy iteration can be shown to converge at a geometric rate when combined with policy evaluation methods that are accurate in ℓ_∞-norm (e.g., see the books [2, 29]). Also, given that we

[1] Here we have only referenced some representative papers; see the references in Szepesvari [294] for a broader overview.

Problem	Algorithm	Paper	Model	Sample-size	Guarantee	Technique
State-action value estimation in MDPs	Plug-in	[166], [160]	Synchronous	Non-asymptotic	Global, ℓ_∞	Hoeffding
		[7]	Synchronous	Non-asymptotic	Global, ℓ_∞	Bernstein
	Stochastic approximation: Q-learning & variants	[37]	Synchronous	Asymptotic	Global, conv. in dist.	ODE method
		[84]	Synchronous	Asymptotic	Local, conv. in dist.	Asymptotic normality
		[324], [65]	Synchronous	Non-asymptotic	Local, ℓ_∞	Bernstein, Moreau envelope
		[6], [280], [325]	Synchronous	Non-asymptotic	Global, ℓ_∞	Bernstein, variance reduction
Optimal value estimation in MDPs	Plug-in	[7]	Synchronous	Non-asymptotic	Global, ℓ_∞	Bernstein
		[3]	Synchronous	Non-asymptotic	Global, ℓ_∞	Bernstein + decoupling
	Stochastic approximation	[281]	Synchonous	Non-asymptotic	Global, ℓ_∞	Bernstein + variance reduction
Policy evaluation in MRPs	**Plug-in**	**This chapter**	**Synchronous**	**Non-asymptotic**	**Local, ℓ_∞**	**Bernstein + leave-one-out**
	Stochastic approximation: TD-learning	[295], [255], [150], [37], [84]	Synchronous, trajectories	Asymptotic	Local, ℓ_2 and conv. in dist.	Averaging, ODE method
		[183], [30], [289]	Synchronous, trajectories	Non-asymptotic	Global, ℓ_2	Averaging, martingales
	TD-learning with function approximation	[301], [306]	Trajectories	Asymptotic	Global oracle inequality Local, conv. in dist.	Asymptotic normality
		[30], [86], [72]	Synchronous, trajectories	Non-asymptotic	Global and local, ℓ_2	Population to sample
	Median of means	**This chapter**	**Synchronous**	**Non-asymptotic**	**Local, ℓ_∞**	**Robustness**

Table 8.1: A subset of results in the tabular and infinite-horizon discounted setting, both for policy evaluation in MRPs and policy optimization in MDPs. For a broader overview of results, see Gosavi [123] for the setting of infinite-horizon average reward, and Dann and Brunskill [76] for the episodic setting. The "technique" vertical of the table is only meant to showcase a representative subset of those employed. Our contributions are highlighted in red.

are interested in fine-grained, instance-dependent guarantees, we first study the problem without function approximation.

As briefly alluded to before, there has also been some recent focus on obtaining instance-dependent guarantees in online reinforcement learning settings [209, 283]. These analyses have led to more practically applicable algorithms that provide, for instance, horizon-independent regret bounds for certain episodic MDPs [154, 345], thereby improving upon worst-case bounds. Recent work has also established some instance-dependent bounds for the problem of state-action value function estimation in Markov decision processes, for both ordinary Q-learning [324] and a variance-reduced improvement [325]. However, we currently lack the localized lower bounds that would allow us to understand the fundamental limits of the problem in a more local sense, except in some special cases for asymptotic settings; for instance, see Ueno et al. [306] and Devraj and Meyn [84] for bounds of this type for LSTD and stochastic approximation, respectively. We hope that our analysis of the simpler policy evaluation problem will be useful in broadening the scope of such guarantees.

Portions of our analysis exploit a decoupling that is induced by a leave-one-out technique. We note that leave-one-out techniques are frequently used in probabilistic analysis (e.g., [81]). In the context of Markov processes, arguments that are related to but distinct from those appearing in this chapter have been used in analyzing estimates of the stationary distribution of a Markov chain [63], and for analyzing optimal policies in reinforcement learning [3].

For the reader's convenience, we have collected many of the relevant results both in policy optimization and evaluation in Table 8.1, along with the settings and sample-size regimes in which they apply, the nature of the guarantee, and the salient techniques used.

Contributions: We study the problem of estimating the infinite-horizon, discounted value function of a tabular MRP in ℓ_∞-norm, assuming access to state transitions and reward samples under the generative model. Our first main result, Theorem 8.1.1, analyzes the plug-in estimator, showing two types of guarantees: on one hand, we derive high-probability upper bounds on the error that can be computed based on the observed data, and on the other, we show upper bounds that depend on the underlying (unknown) population transition matrix and reward function. The latter result is achieved via a decoupling argument that we expect to be more broadly applicable to problems of this type.

Corollary 8.2.1 then specializes the population-based result in Theorem 8.1.1 to natural sub-classes of MRPs. Theorem 8.2.1 provides minimax lower bounds for these sub-classes, showing—in conjunction with Corollary 8.2.1—that the plug-in approach is minimax optimal over the class of MRPs with uniformly bounded reward functions. However, these results suggest that the plug-in approach is *not* minimax-optimal over the class of MRPs having value functions with bounded variance under the transition model. Consequently, we analyze an approach based on the median-of-means device and show that this modified estimator is minimax optimal over the class of MRPs having value functions with bounded variance.

Chapter-specific notation: Recall the notational convention introduced in Section 1.4. We complement this notation with a few other definitions that are used solely in this chapter and the corresponding technical proof section in Appendix C.2. We let $\|\mathbf{M}\|_{1,\infty}$ denote the maximum ℓ_1-norm of the rows of a matrix $\mathbf{M}$, and refer to it as the $(1, \infty)$-operator norm of a matrix.

8.1 Guarantees for the plug-in approach

Recall the local complexity measures defined in Section 7.2.5 of Chapter 7. With this notation, we are ready to state two results for the plug-in estimator.

Theorem 8.1.1. *There is a pair of universal constants (c_1, c_2) such that if $N \geq c_1 \frac{\gamma^2}{(1-\gamma)^2} \log(8D/\delta)$, then each of the following statements holds with probability at least $1 - \delta$.*

(a) We have

$$\|\widehat{\theta}_{\mathsf{plug}} - \theta^*\|_\infty \leq c_2 \left\{ \sqrt{\frac{\log(8D/\delta)}{N}} \left(\gamma \|(\mathbf{I} - \gamma\widehat{\mathbf{P}})^{-1}\widehat{\sigma}(\widehat{\theta}_{\mathsf{plug}})\|_\infty + \frac{\|\rho(r)\|_\infty}{1-\gamma} \right) + \frac{\log(8D/\delta)}{N} \cdot \gamma b(\widehat{\theta}_{\mathsf{plug}}) \right\}. \tag{8.1a}$$

(b) We have

$$\|\widehat{\theta}_{\mathsf{plug}} - \theta^*\|_\infty \leq c_2 \left\{ \sqrt{\frac{\log(8D/\delta)}{N}} \left(\gamma \left\|(\mathbf{I} - \gamma\mathbf{P})^{-1}\sigma(\theta^*)\right\|_\infty + \frac{\|\rho(r)\|_\infty}{1-\gamma} \right) + \frac{\log(8D/\delta)}{N} \cdot \gamma b(\theta^*) \right\}. \tag{8.1b}$$

It is worth making a few comments on this theorem, which provides two instance-dependent upper bounds on the error of the plug-in approach. Assuming for simplicity of discussion[2] that the maximum noise reward parameter $\|\rho(r)\|_\infty$ is known, then part (a) of the theorem provides a bound that can be evaluated based on the observed data; bounds of this form are especially useful in downstream analyses. For instance, a central consideration in policy iteration methods is to obtain "good enough" value function estimates $\widehat{\theta}$ for fixed policies, in that we have $\|\widehat{\theta} - \theta^*\|_\infty \leq \epsilon$ for some prescribed tolerance ϵ. Theorem 8.1.1(a) provides a method by which such a bound may be verified for the plug-in approach: compute the statistic on the RHS of bound (8.1a); if this is less than ϵ, then the bound $\|\widehat{\theta}_{\mathsf{plug}} - \theta^*\|_\infty \leq \epsilon$ holds with probability exceeding $1 - \delta$.

On the other hand, Theorem 8.1.1(b) provides a guarantee that depends on the unknown problem instance. From the perspective of the analysis, this is the more difficult bound to establish, since it requires a leave-one-out technique to decouple dependencies between the estimate $\widehat{\theta}_{\mathsf{plug}}$ and the matrix $\widehat{\mathbf{P}}$. We expect our technique—presented in full in Section 8.4.2—and its variants to be more

[2] We note that when $\rho(r)$ is *not* known but the reward distribution is (say) Gaussian, it is straightforward to provide an entry-wise upper bound for it by computing the empirical standard deviation of rewards from samples, and using this to define a high-probability and data-dependent bound on the sub-Gaussian parameter.

broadly useful in analyzing other problems in reinforcement learning besides the policy evaluation problem considered here.

Third, note that our lower bound on the sample size—which evaluates to $N \geq \frac{c_1}{(1-\gamma)^2} \log(8D/\delta)$ for any strictly positive discount factor—is unavoidable in general. In particular, for any fixed reward-noise parameter $\|\rho(r)\|_\infty > 0$, this condition is required in order to obtain a consistent estimate of the value function.[3] On the other hand, in the special case of deterministic rewards ($\|\rho(r)\|_\infty = 0$), we suspect that this condition can be weakened, but leave this for future work.

Finally, it is worth noting that there are two terms in the bounds of Theorem 8.1.1: the first term corresponds to a notion of standard deviations of the estimated/true value function and reward, and the second depends on the span semi-norm of the value function. Are both of these terms necessary? What is the optimal rate at which any value function can be estimated?

8.2 Assessing optimality

The questions asked above motivate the analysis to be presented in this section, which has two parts.

8.2.1 Lower bounds on the local minimax risk

In order to study the question of optimality in this chapter, we adopt the notion of *local minimax risk*, in which the performance of an estimator is measured in a worst-case sense locally over natural subsets of the parameter space. In the next chapter, we revisit this question and provide even more refined lower bounds that show that the plug-in estimator is in fact optimal when viewed from the local asymptotic minimax framework (see Section 9.1).

Our upper bounds in the previous section depend on the problem instance via the standard deviation function $\sigma(\theta^*)$, the reward standard deviation $\rho(r)$, and the span semi-norm of θ^*. Accordingly, we define the following subsets[4] of Markov reward processes (MRPs):

$$\mathcal{M}_{\mathsf{var}}(\vartheta, \varrho) := \Big\{\text{set of all MRPs s.t. } \|\sigma(\theta^*)\|_\infty \leq \vartheta \text{ and } \|\rho(r)\|_\infty \leq \varrho\Big\}, \tag{8.2a}$$

$$\mathcal{M}_{\mathsf{vfun}}(\zeta, \varrho) := \Big\{\text{set of all MRPs s.t. } \|\theta^*\|_{\mathrm{span}} \leq \zeta \text{ and } \|\rho(r)\|_\infty \leq \varrho\Big\}, \qquad \text{and} \tag{8.2b}$$

$$\mathcal{M}_{\mathsf{rew}}(r_{\max}, \varrho) := \Big\{\text{set of all MRPs s.t. } \|r\|_\infty \leq r_{\max} \text{ and } \|\rho(r)\|_\infty \leq \varrho\Big\}. \tag{8.2c}$$

Letting $\mathcal{M}$ be any one of these sets, we use the shorthand $\theta \in \mathcal{M}$ to mean that θ is the value function of some MRP in the set $\mathcal{M}$. Each choice of the set $\mathcal{M}$ defines the local minimax risk given

[3] For instance, even with known transition dynamics, estimating the value function of a single state to within additive error ϵ requires $\Omega\left(\frac{1}{(1-\gamma)^2\epsilon^2}\right)$ samples of the noisy reward.

[4] The following mnemonic device may help the reader appreciate and remember notation: the symbol ϑ, or "vartheta", stands for a measure of the variability in the value function θ; the symbol ϱ, or "varrho", represents the variability in reward samples, and $r_{\max}$ represents the maximum absolute reward mean.

by

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \mathcal{M}} \mathbb{E}\left[\|\widehat{\theta} - \theta^*\|_\infty\right],$$

where the infimum ranges over all measurable functions $\widehat{\theta}$ of N observations from the generative model. With this set-up, we can now state some lower bounds in terms of such local minimax risks:

Theorem 8.2.1. *There is a pair of absolute constants (c_1, c_2) such that for all $\gamma \in [\frac{1}{2}, 1)$ and sample sizes $N \geq \frac{c_1}{1-\gamma} \log(D/2)$, the following statements hold.*

(a) For each triple of positive scalars $(\vartheta, \zeta, \varrho)$ satisfying[5] *$\vartheta \leq \zeta\sqrt{1-\gamma}$, we have*

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \mathcal{M}_{\mathrm{var}}(\vartheta,\varrho) \cap \mathcal{M}_{\mathsf{vfun}}(\zeta,\varrho)} \mathbb{E}\left[\|\widehat{\theta} - \theta^*\|_\infty\right] \geq \frac{c_2}{1-\gamma}\sqrt{\frac{\log(D/2)}{N}}\,(\vartheta + \varrho)\,. \tag{8.3a}$$

(b) For each pair of positive scalars $(r_{\max}, \varrho)$ satisfying $r_{\max} \geq \varrho\sqrt{\frac{\log D}{N}}$, we have

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \mathcal{M}_{\mathsf{rew}}(r_{\max},\varrho)} \mathbb{E}\left[\|\widehat{\theta} - \theta^*\|_\infty\right] \geq \frac{c_2}{1-\gamma}\sqrt{\frac{\log(D/2)}{N}}\left(\frac{r_{\max}}{(1-\gamma)^{1/2}} + \varrho\right). \tag{8.3b}$$

Equipped with these lower bounds, we can now assess the local minimax optimality of the plug-in estimator. In order to facilitate this comparison, let us state a corollary of Theorem 8.1.1 that provides bounds on the worst-case error of the plug-in estimator over particular subsets of the parameter space. In order to further simplify the comparison, we restrict our attention to the range $\gamma \in [\frac{1}{2}, 1)$ covered by the lower bounds.

Corollary 8.2.1. *There are absolute constants (c_3, c_4) such that for all $\gamma \in [\frac{1}{2}, 1)$ and sample sizes*[6] *$N \geq \frac{c_3}{(1-\gamma)^2} \log(8D/\delta)$, the following statements hold.*

(a) Consider a triple of positive scalars $(\vartheta, \zeta, \varrho)$ such that[7] *$\vartheta \leq \zeta$. Then for any value function $\theta^* \in \mathcal{M}_{\mathsf{var}}(\vartheta, \varrho) \cap \mathcal{M}_{\mathsf{vfun}}(\zeta, \varrho)$, we have*

$$\|\widehat{\theta}_{\mathsf{plug}} - \theta^*\|_\infty \leq \frac{c_4}{1-\gamma}\left\{\sqrt{\frac{\log(8D/\delta)}{N}}\,(\vartheta + \varrho) + \frac{\log(8D/\delta)}{N} \cdot \zeta\right\} \tag{8.4a}$$

with probability at least $1 - \delta$.

[5] We conjecture that this lower bound can be proved under the weaker condition $\vartheta \leq \zeta$, thereby matching the condition present in Corollary 8.2.1(a).

[6] As shown in the proof, part (a) of the corollary holds without this assumption on the sample size, but we state it here to facilitate a direct derivation of Corollary 8.2.1 from Theorem 8.1.1.

[7] It is worth noting that the condition $\vartheta \leq \zeta$ in part (a) of the corollary does not entail any loss of generality, since we always have $\|\sigma(\theta^*)\|_\infty \leq \|\theta^*\|_{\mathrm{span}}$. Indeed, for MRPs in which $\|\sigma(\theta^*)\|_\infty \ll \|\theta^*\|_{\mathrm{span}}$, the second term on the RHS of inequality (8.4a) will dominate the bound unless the sample size N is large.

(b) *Consider an arbitrary pair of positive scalars* $(r_{\max}, \varrho)$. *Then for any value function* $\theta^* \in \mathcal{M}_{\mathsf{rew}}(r_{\max}, \varrho)$, *we have*

$$\|\widehat{\theta}_{\mathsf{plug}} - \theta^*\|_\infty \leq \frac{c_4}{1-\gamma}\sqrt{\frac{\log(8D/\delta)}{N}}\left(\frac{r_{\max}}{(1-\gamma)^{1/2}} + \varrho\right) \tag{8.4b}$$

with probability at least $1 - \delta$.

By comparing Corollary 8.2.1(b) with Theorem 8.2.1(b), we see that the plug-in estimator is minimax optimal (up to constant factors) over the class $\mathcal{M}_{\mathsf{rew}}(r_{\max}, \varrho)$. This conclusion parallels that of Azar et al. [7] for the related problem of optimal state-value function estimation in MDPs. (In our notation, their work applies to the special case of $\varrho = 0$, but their analysis can easily be extended to this more general setting.)

A comparison of part (a) of the two results is more interesting. Here we see that the first term in the upper bound (8.4a) matches the lower bound (8.3a) up to a constant factor. The second term of inequality (8.4a), however, does not have an analogous component in the lower bound, and this leads us to the interesting question of whether the analysis of the plug-in estimator can be sharpened so as to remove the dependence of the error on the span semi-norm $\|\theta^*\|_{\text{span}}$. Proposition C.2.1, presented in Appendix C.2, shows that this is impossible in general, and that there are MRPs in which the ℓ_∞ error can be *lower bounded* by a term that is proportional to the span semi-norm.

This raises another natural question: Is there a different estimator whose error can be bounded independently of the span semi-norm $\|\theta^*\|_{\text{span}}$, and which is able achieve the lower bound (8.3a)? In the next section, we introduce such an estimator via a median-of-means device.

8.2.2 Closing the gap via the median-of-means method

In many situations, the span semi-norm of a value function θ^* may be much larger its variance $\sigma(\theta^*)$ under the transition model. Such a discrepancy arises when there are states with extremely large positive (or negative) rewards that are visited with very low probability. In such cases, the second terms in the bounds (8.1) dominate the first. It is thus of interest to derive bounds that are purely "variance-dependent" and independent of the span norm. In order to do so, we analyze a slight variant of the plug-in approach. In particular, we analyze the *median-of-means* estimator, which is a standard robust alternative to the sample mean in other scenarios [188, 234]. In the context of reinforcement learning, Pazis et al. [250] made use of it for online policy optimization in MDPs.

In our setting, we only employ median-of-means to obtain a better estimate of term depending on the transition matrix; we still use the estimate $\widehat{r}$ defined in equation (7.3) as our estimate of the reward function.[8] Given the data set $\{\mathbf{Z}_k\}_{k=1}^N$ and some vector $\theta \in \mathbb{R}^D$, the median-of-means estimate $\widehat{\mathcal{M}}(\theta)$ of the population expectation $\mathbf{P}\theta$ is given by the following nonlinear operation:

[8] In principle, one could run a median-of-means estimate on the combination of reward and transition, but this is not necessary in our setting due to the sub-Gaussian assumption on the reward noise (7.2). Slight modifications of our techniques also yield bounds for the combined median-of-means estimate assuming only that the standard deviation of the reward noise is bounded entry-wise by the vector $\rho(r)$.

- First, split the data set into K equal parts denoted $\{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$, where each subset $\mathcal{D}_i$ has size $m = \lfloor N/K \rfloor$.
- Second, compute the empirical mean $\widehat{\mu}_i(\theta) := \frac{1}{m} \sum_{k \in \mathcal{D}_i} \mathbf{Z}_k \theta$ for each $i \in [K]$.
- Finally, return the quantity $\widehat{\mathcal{M}}(\theta) := \mathrm{med}(\widehat{\mu}_1(\theta), \ldots, \widehat{\mu}_K(\theta))$, where the median—defined for convenience as the $\lfloor K/2 \rfloor$-th order statistic—is taken entry-wise.

The random operator $\widehat{\mathcal{M}}$ defines the *median-of-means empirical Bellman operator*, given by

$$\widehat{\mathcal{T}}_N^{\mathsf{MoM}}(\theta) := \widehat{r} + \gamma \widehat{\mathcal{M}}(\theta). \tag{8.5}$$

As shown in Lemma 8.4.6 (see Section 8.4), this operator is γ-contractive in the ℓ_∞-norm. Consequently, it has a unique fixed point, which we term the *median-of-means value function estimate*, denoted by $\widehat{\theta}_{\mathsf{MoM}}$.

In practice, the estimate $\widehat{\theta}_{\mathsf{MoM}}$ can be found by starting at an arbitrary initialization and repeatedly applying the γ-contractive operator $\widehat{\mathcal{T}}_N^{\mathsf{MoM}}$ until convergence.[9] The following theorem provides a population-based guarantee on the error of this estimator.

Theorem 8.2.2. *Suppose that the median-of-means operator $\widehat{\mathcal{M}}$ is constructed with the parameter choice $K = 8 \log(4D/\delta)$. Then there is a universal constant c such that we have*

$$\|\widehat{\theta}_{\mathsf{MoM}} - \theta^*\|_\infty \leq \frac{c}{1-\gamma} \sqrt{\frac{\log(8D/\delta)}{N}} \Big(\gamma \|\sigma(\theta^*)\|_\infty + \|\rho(r)\|_\infty \Big) \tag{8.6}$$

with probability exceeding $1 - \delta$.

We have thus achieved our goal of obtaining a purely variance-dependent bound. Indeed, for each pair of positive scalars (ϑ, ϱ), any value function $\theta^* \in \mathcal{M}_{\mathrm{var}}(\vartheta, \varrho)$, and reward distribution satisfying $\|\rho(r)\|_\infty \leq \varrho$, we have

$$\|\widehat{\theta}_{\mathsf{MoM}} - \theta^*\|_\infty \leq \frac{c}{1-\gamma} \sqrt{\frac{\log(8D/\delta)}{N}} \left(\vartheta + \varrho \right),$$

with probability exceeding $1 - \delta$. Integrating this tail bound yields an analogous upper bound on the expected error, which matches the lower bound (8.3a) on the expected error up to a constant factor. As a corollary, we conclude that the minimax risk over the class $\mathcal{M}_{\mathrm{var}}(\vartheta, \varrho)$ scales as

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \mathcal{M}_{\mathrm{var}}(\vartheta, \varrho)} \mathbb{E}\left[\|\widehat{\theta} - \theta^*\|_\infty \right] \asymp \frac{1}{1-\gamma} \sqrt{\frac{\log(D)}{N}} \left(\vartheta + \varrho \right), \tag{8.7}$$

and is achieved (up to constant factors) by the estimator $\widehat{\theta}_{\mathsf{MoM}}$.

[9] Since the operator is γ-contractive, it suffices to run this iterative algorithm for $\log_\gamma \epsilon$ to obtain an ϵ-approximate fixed point in an additive sense.

However, our results fall short of showing that the estimator $\widehat{\theta}_{\mathsf{MoM}}$ is minimax optimal over the class $\mathcal{M}_{\mathsf{rew}}(r_{\max}, \varrho)$ of MRPs with bounded rewards. Indeed, for any value function θ^* in the class $\mathcal{M}_{\mathsf{rew}}(r_{\max}, \varrho)$, Theorem 8.2.2 yields the corollary

$$\|\widehat{\theta}_{\mathsf{MoM}} - \theta^*\|_\infty \leq \frac{c}{1-\gamma}\sqrt{\frac{\log(8D/\delta)}{N}}\left(\gamma\frac{r_{\max}}{1-\gamma} + \varrho\right)$$

with probability exceeding $1-\delta$. Comparing inequality (8.3b) with this bound, we see that our upper bound on the median-of-means estimator is sub-optimal by a factor $(1-\gamma)^{-\frac{1}{2}}$ in the discount complexity. From a technical standpoint, this is due to the fact that our upper bound in Theorem 8.2.2 involves the functional $\frac{1}{1-\gamma}\|\sigma(\theta^*)\|_\infty$ and not the sharper functional $\|(\mathbf{I}-\gamma\mathbf{P})^{-1}\sigma(\theta^*)\|_\infty$ present in Theorem 8.1.1(b). We believe that this gap is not intrinsic to the MoM method, and conjecture that an upper bound depending on the latter functional can be proved for the estimator $\widehat{\theta}_{\mathsf{MoM}}$; this would guarantee that the median-of-means estimator is also minimax optimal over the class $\mathcal{M}_{\mathsf{rew}}(r_{\max}, \varrho)$.

8.3 Numerical experiments

In this section, we explore the sharpness of our theoretical predictions, for both the plug-in and the median-of-means (MoM) estimator. Our bounds predict a range of behaviors depending on the scaling of the maximum standard deviation $\|\sigma(\theta^*)\|_\infty$, and the span semi-norm (for the plug-in estimator). Let us verify these scalings via some simple experiments.

8.3.1 Behavior on the "hard" example used for the lower bound

First, we use a simple variant of our lower bound construction illustrated in Figure 7.2 of Chapter 7, where we additionally choose the scalar parameterization (7.12). Recall the bound (7.13b) on the local complexity measures of this class of MRPs.

Consequently, by the bound (8.6) from Theorem 8.2.2, for a fixed sample size N, the MoM estimator should have ℓ_∞-norm scaling as $\left(\frac{1}{1-\gamma}\right)^{1.5-\lambda}$. The same prediction also holds for the plug-in estimator, assuming that $N \gtrsim \frac{1}{(1-\gamma)}$.

In order to test this prediction, we fixed the parameter $\lambda \in [0,1]$, and generated a range of MRPs with different values of the discount factor γ. For each such MRP, we drew $N = 10^4$ samples from the generative observation model and computed both the plug-in and median-of-means estimators, where the latter estimator was run with the choice $K = 20$. While the plug-in estimator has a simple closed-form expression, the MoM estimator was obtained by running the median-of-means Bellman operator $\widehat{\mathcal{T}}_N^{\mathsf{MoM}}$ iteratively until it converged to its fixed point; we declared that convergence had occurred when the ℓ_∞-norm of the difference between successive iterates fell below 10^{-8}.

In Figure 8.1, we plot the ℓ_∞-error, of both the plug-in approach as well as the median-of-means estimator, as a function of γ. The plot shows the behavior for three distinct values $\lambda = \{0, 0.5, 1\}$. Each point on each curve is obtained by averaging 1000 Monte Carlo trials of the experiment. Note that on this log-log plot, we see a linear relationship between the log ℓ_∞-error and log discount

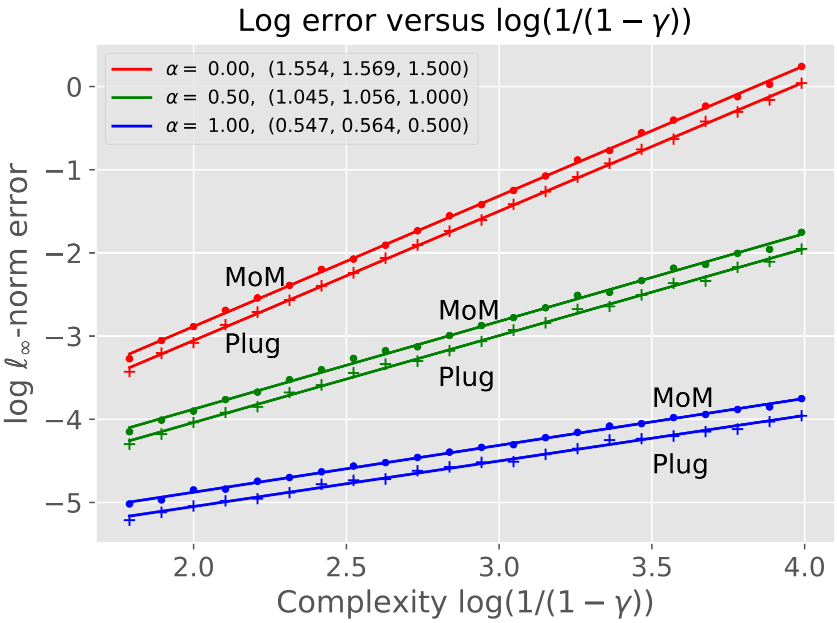


Figure 8.1: Log-log plot of the ℓ_∞-error versus the discount complexity parameter $1/(1-\gamma)$ for both the plug-in estimator (in $+$ markers) and median-of-means estimator (in $\bullet$ markers) computed on the toy MRP from Figure 7.2 with the scalar parameterization (7.12). Errors are averaged over $T = 1000$ trials with $N = 10^4$ samples each. We have also plotted the least-squares fits through these points, and the slopes of these lines are provided in the legend. In particular, the legend contains the tuple of slopes $(\widehat{\beta}_{\text{plug}}, \widehat{\beta}_{\text{MoM}}, \beta^*)$ for each value of λ. Logarithms are to the natural base.

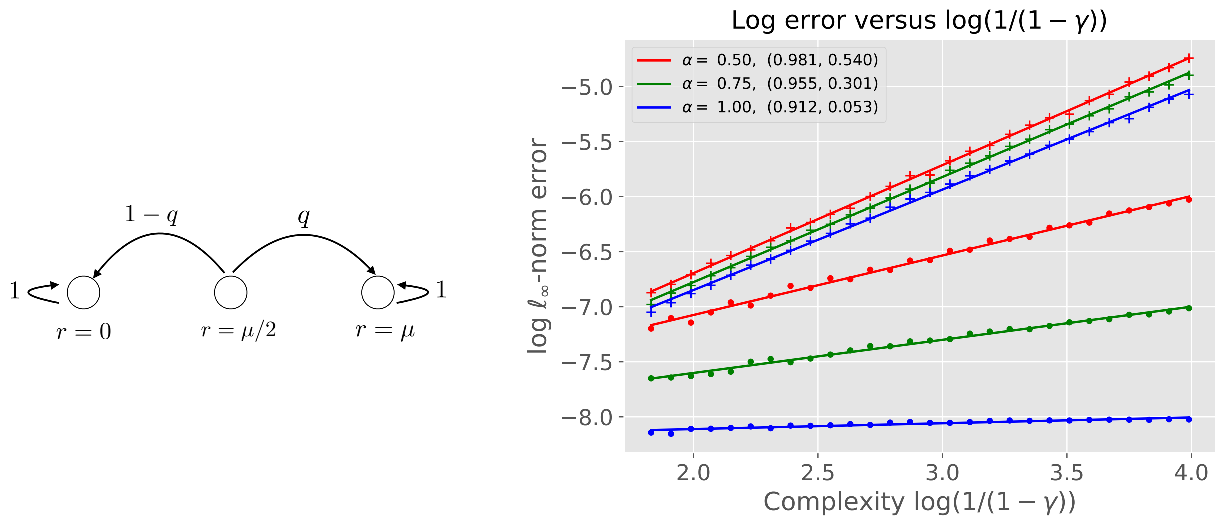
complexity, with the slopes depending on the value of λ. More precisely, from our calculations above, our theory predicts that the log ℓ_∞-error should be related to the log complexity $\log\left(\frac{1}{1-\gamma}\right)$ in a linear fashion with slope

$$\beta^* = 1.5 - \lambda.$$

Consequently, for both the plug-in and MoM estimators, we performed a linear regression to estimate these slopes, denoted by $\widehat{\beta}_{\text{plug}}$ and $\widehat{\beta}_{\text{MoM}}$ respectively. The plot legend reports the triple $(\widehat{\beta}_{\text{plug}}, \widehat{\beta}_{\text{MoM}}, \beta^*)$, and for each we see good agreement between the theoretical prediction β^* and its empirical counterparts.

8.3.2 When does the MoM estimator perform better than plug-in?

Our theoretical results predict that the MoM estimator should outperform the plug-in approach when the span semi-norm of the value function $\|\theta^*\|_{\text{span}}$ is much larger than its maximum standard deviation $\|\sigma(\theta^*)\|_\infty$. Indeed, Proposition C.2.1 in Appendix C.2 demonstrates that there are MRPs on which the ℓ_∞-error of the plug-in estimator grows with the span semi-norm of the optimal value function. Let us now simulate the behavior of both the plug-in and MoM approach on this MRP, constructed by taking $D/3$ copies of the 3-state MRP in Figure 8.2(a).



(a) (b)

Figure 8.2: (a) Illustration of the MRP $\mathcal{R}_{\mathbf{1}}(q, \mu)$. In the simulation as well as in the lower bound construction of Proposition C.2.1, we concatenate $D/3$ such MRPs to produce an MRP on D states. For the simulation, we choose $D = 3\left\lfloor \left(\frac{1}{1-\gamma}\right)^{\lambda} \right\rfloor$ and set $q = \frac{10}{ND}$ and $\mu = 1$. (b) Log-log plot of the ℓ_∞-error versus the discount complexity parameter $1/(1-\gamma)$ for both the plug-in estimator (in $+$ markers) and median-of-means estimator (in $\bullet$ markers) averaged over $T = 1000$ trials with $N = 10^4$ samples each. We have also plotted the least-squares fits through these points, and the slopes of these lines are provided in the legend. In particular, the legend contains the tuple of slopes $(\widehat{\beta}_{\text{plug}}, \widehat{\beta}_{\text{MoM}})$ for each value of λ. Logarithms are to the natural base.

Our simulation is carried out on $N = 10^4$ samples from this D-state MRP, with noiseless observations of the reward. In order to parameterize the MRP via the discount factor alone, we fix the pair (q, D) in the following way. First, we fix a scalar λ in the unit interval $[0, 1]$, and then set

$$D = 3\left\lfloor \left(\frac{1}{1-\gamma}\right)^{\lambda} \right\rfloor \quad \text{and} \quad q = \frac{10}{ND}.$$

Note that this sub-family of MRPs is fully parameterized by the pair (γ, λ). The construction also ensures that

$$\frac{\|\theta^*\|_\infty}{N} \gg \frac{\|\sigma(\theta^*)\|_\infty}{\sqrt{N}}, \tag{8.8}$$

for this chosen parameterization, and furthermore, that the ratio of the LHS and RHS of inequality (8.8) increases as the dimension D increases (see the proof of Proposition C.2.1).

As shown in Proposition C.2.1 in Appendix C.2, the ℓ_∞ error of the plug-in estimator for this family of MRPs can be *lower bounded* by $\|\theta^*\|_\infty/N$. It is also straightforward to show that the error of the MoM estimator is *upper bounded* by the quantity $\frac{\|\sigma(\theta^*)\|_\infty}{\sqrt{N}}$. Now increasing the value of λ increases the dimension D, and so the MoM estimator should behave better and better for larger values of λ. In particular, this behavior can be captured in the log-log plot of the error against $1/(1-\gamma)$, which is presented in Figure 8.2(b).

The plot shows the behavior for three distinct values $\lambda = \{0.5, 0.75, 1\}$. Each point on each curve is obtained by averaging 1000 Monte Carlo trials of the experiment. As expected, the MoM estimator consistently outperforms the plug-in estimator for each value of λ. Moreover, on this log-log plot, we see a linear relationship between the log ℓ_∞-error and log discount complexity, with the slopes depending on the value of λ. For both the plug-in and MoM estimators, we performed a linear regression to estimate these slopes, denoted by $\widehat{\beta}_{\text{plug}}$ and $\widehat{\beta}_{\text{MoM}}$ respectively. The plot legend reports the pair $(\widehat{\beta}_{\text{plug}}, \widehat{\beta}_{\text{MoM}})$, and we see that the gap between the slopes increases as λ increases.

8.4 Proofs of main results

We now turn to the proofs of our main results. Throughout our proofs, the reader should recall that the values of absolute constants may change from line-to-line. We also use the following facts repeatedly. First, for a row stochastic matrix $\mathbf{M}$ with non-negative entries and any scalar $\gamma \in [0, 1)$, we have the infinite series

$$(\mathbf{I} - \gamma \mathbf{M})^{-1} = \sum_{t=0}^{\infty} (\gamma \mathbf{M})^t, \tag{8.9a}$$

which implies that the entries of $(\mathbf{I} - \gamma \mathbf{M})^{-1}$ are all non-negative. Second, for any such matrix, we also have the bound $\|(\mathbf{I} - \gamma \mathbf{M})^{-1}\|_{1,\infty} \leq \frac{1}{1-\gamma}$. Finally, for any matrix $\mathbf{A}$ with positive entries and a vector v of compatible dimension, we have the elementwise inequality

$$|\mathbf{A} v| \preceq \mathbf{A} |v|. \tag{8.9b}$$

8.4.1 Proof of Theorem 8.1.1, part (a)

Throughout this proof, we adopt the convenient shorthand $\widehat{\theta} \equiv \widehat{\theta}_{\text{plug}}$ for notational convenience. By the Bellman equations (7.1) and (7.4) for θ^* and $\widehat{\theta}$, respectively, we have

$$\widehat{\theta} - \theta^* = \gamma \left\{ \widehat{\mathbf{P}} \widehat{\theta} - \mathbf{P} \theta^* \right\} + (\widehat{r} - r) \;=\; \gamma \widehat{\mathbf{P}} (\widehat{\theta} - \theta^*) + \gamma (\widehat{\mathbf{P}} - \mathbf{P}) \theta^* + (\widehat{r} - r).$$

Introducing the shorthand $\widehat{\Delta} := \widehat{\theta} - \theta^*$ and re-arranging implies the relation

$$\widehat{\Delta} = \gamma (\mathbf{I} - \gamma \widehat{\mathbf{P}})^{-1} (\widehat{\mathbf{P}} - \mathbf{P}) \theta^* + (\mathbf{I} - \gamma \widehat{\mathbf{P}})^{-1} (\widehat{r} - r), \tag{8.10}$$

and consequently, the elementwise inequality

$$|\widehat{\Delta}| \preceq \gamma(\mathbf{I} - \gamma\widehat{\mathbf{P}})^{-1}|(\widehat{\mathbf{P}} - \mathbf{P})\theta^*| + (\mathbf{I} - \gamma\widehat{\mathbf{P}})^{-1}|(\widehat{r} - r)|, \tag{8.11}$$

where we have used the relation (8.9b) with the matrix $\mathbf{A} = (\mathbf{I} - \gamma\widehat{\mathbf{P}})^{-1}$. Given the sub-Gaussian condition on the stochastic rewards, we can apply Hoeffding's inequality combined with the union bound to obtain the elementwise inequality $|\widehat{r} - r| \preceq c\sqrt{\frac{\log(8D/\delta)}{N}} \cdot \rho(r)$, which holds with probability at least $1 - \frac{\delta}{4}$. Since the matrix $(\mathbf{I} - \gamma\widehat{\mathbf{P}})^{-1}$ has non-negative entries and $(1, \infty)$-norm at most $\frac{1}{1-\gamma}$, we have

$$(\mathbf{I} - \gamma\widehat{\mathbf{P}})^{-1}|\widehat{r} - r| \preceq \frac{c}{1-\gamma}\|\rho(r)\|_\infty\sqrt{\frac{\log(8D/\delta)}{N}}\mathbf{1}. \tag{8.12a}$$

with the same probability. On the other hand, by Bernstein's inequality, we have

$$|(\widehat{\mathbf{P}} - \mathbf{P})\theta^*| \preceq c\left\{\sqrt{\frac{\log(8D/\delta)}{N}} \cdot \sigma(\theta^*) + \|\theta^*\|_{\text{span}}\frac{\log(8D/\delta)}{N} \cdot \mathbf{1}\right\}$$

with probability at least $1 - \frac{\delta}{4}$, and hence

$$(\mathbf{I} - \gamma\widehat{\mathbf{P}})^{-1}|(\widehat{\mathbf{P}} - \mathbf{P})\theta^*| \preceq c\left\{\sqrt{\frac{\log(8D/\delta)}{N}} \cdot \|(\mathbf{I} - \gamma\widehat{\mathbf{P}})^{-1}\sigma(\theta^*)\|_\infty + \frac{\|\theta^*\|_{\text{span}}}{1-\gamma}\frac{\log(8D/\delta)}{N}\right\} \cdot \mathbf{1}. \tag{8.12b}$$

Substituting the bounds (8.12a) and (8.12b) into the elementwise inequality (8.11), we find that

$$|\widehat{\Delta}| \preceq c\left\{\sqrt{\frac{\log(8D/\delta)}{N}} \cdot \left(\gamma\|(\mathbf{I} - \gamma\widehat{\mathbf{P}})^{-1}\sigma(\theta^*)\|_\infty + \frac{\|\rho(r)\|_\infty}{1-\gamma}\right) + \frac{\gamma\|\theta^*\|_{\text{span}}}{1-\gamma}\frac{\log(8D/\delta)}{N}\right\} \cdot \mathbf{1} \tag{8.13}$$

with probability at least $1 - \frac{\delta}{2}$.

Our next step is to relate the pair of population quantities $(\sigma(\theta^*), \|\theta^*\|_{\text{span}})$ to their empirical analogues $(\widehat{\sigma}(\widehat{\theta}), \|\widehat{\theta}\|_{\text{span}})$. The following lemma provides such a bound.

Lemma 8.4.1 (Population to empirical variance)**.** *We have the element-wise inequality*

$$\sigma(\theta^*) \preceq 2\widehat{\sigma}(\widehat{\theta}) + 2|\widehat{\Delta}| + c'\|\theta^*\|_{\text{span}}\sqrt{\frac{\log(8D/\delta)}{N}} \cdot \mathbf{1} \tag{8.14}$$

with probability at least $1 - \delta/2$.

Taking this lemma as given for the moment, let us complete the proof.

Since the matrix $(\mathbf{I} - \gamma\widehat{\mathbf{P}})^{-1}$ has non-negative entries, we can multiply both sides of the elementwise inequality (8.14) by it; doing so and taking the ℓ_∞-norm yields

$$\|(\mathbf{I} - \gamma\widehat{\mathbf{P}})^{-1}\sigma(\theta^*)\|_\infty \leq 2\|(\mathbf{I} - \gamma\widehat{\mathbf{P}})^{-1}\widehat{\sigma}(\widehat{\theta})\|_\infty + \frac{2\|\widehat{\Delta}\|_\infty}{1-\gamma} + \frac{c'\|\theta^*\|_{\text{span}}}{1-\gamma}\sqrt{\frac{\log(8D/\delta)}{N}}.$$

Substituting back into the elementwise inequality (8.13) and taking ℓ_∞-norms of both sides, we find that

$$\|\widehat{\Delta}\|_\infty \leq c\left\{\sqrt{\frac{\log(8D/\delta)}{N}}\left(\gamma\|(\mathbf{I} - \gamma\widehat{\mathbf{P}})^{-1}\widehat{\sigma}(\widehat{\theta})\|_\infty + \frac{\|\rho(r)\|_\infty}{1-\gamma}\right) + \frac{\gamma\|\theta^*\|_{\text{span}}}{1-\gamma}\frac{\log(8D/\delta)}{N}\right\} + \frac{2c\gamma}{1-\gamma}\sqrt{\frac{\log(8D/\delta)}{N}}\|\widehat{\Delta}\|_\infty.$$

Since the span semi-norm satisfies the triangle inequality, we have

$$\|\theta^*\|_{\text{span}} \leq \|\widehat{\theta}\|_{\text{span}} + \|\widehat{\Delta}\|_{\text{span}} \leq \|\widehat{\theta}\|_{\text{span}} + 2\|\widehat{\Delta}\|_\infty.$$

Substituting this bound and re-arranging yields

$$\kappa\|\widehat{\theta} - \theta^*\|_\infty \leq c\left\{\sqrt{\frac{\log(8D/\delta)}{N}}\left(\gamma\|(\mathbf{I} - \gamma\widehat{\mathbf{P}})^{-1}\widehat{\sigma}(\widehat{\theta})\|_\infty + \frac{\|\rho(r)\|_\infty}{1-\gamma}\right) + \frac{\gamma\|\widehat{\theta}\|_{\text{span}}}{1-\gamma}\frac{\log(8D/\delta)}{N}\right\}.$$

where we have introduced the shorthand $\kappa := 1 - \frac{2c\gamma}{1-\gamma}\left(\sqrt{\frac{\log(8D/\delta)}{N}} + \frac{\log(8D/\delta)}{N}\right)$. Finally, by choosing the pre-factor c_1 in the lower bound $N \geq c_1\gamma^2\frac{\log(8D/\delta)}{(1-\gamma)^2}$ large enough, we can ensure that $\kappa \geq \frac{1}{2}$, thereby completing the proof of Theorem 8.1.1(a).

Proof of Lemma 8.4.1

We now turn to the proof of the auxiliary result in Lemma 8.4.1. We begin by noting that the statement is trivially true when $N \leq \log(8D/\delta)$, since we have

$$\sigma(\theta^*) \preceq \|\theta^*\|_{\text{span}}\mathbf{1}.$$

Thus, by adjusting the constant factors in the statement of the lemma, it suffices to prove the lemma under the assumption $N \geq c\log(8D/\delta)$ for a sufficiently large absolute constant c. Accordingly, we make this assumption for the rest of the proof.

We use the following convenient notation for expectations. Let $\mathbb{E}$ denote the vector expectation operator, with the convention that $\mathbb{E}[v] = \mathbf{P}v$. Similarly, let $\widehat{\mathbb{E}}$ denote the vector empirical expectation operator, given by $\widehat{\mathbb{E}}[v] = \widehat{\mathbf{P}}v$. These operators are applied elementwise by definition, and we let $\mathbb{E}_i$ and $\widehat{\mathbb{E}}_i$ denote the i-th entry of each operator, respectively.

With this notation, we have

$$\begin{aligned}
\sigma^2(\theta^*) &= \mathbb{E}\,|\theta^* - \mathbb{E}[\theta^*]|^2 \\
&= (\mathbb{E} - \widehat{\mathbb{E}})\Big|\theta^* - \mathbb{E}[\theta^*]\Big|^2 + \widehat{\mathbb{E}}\,|\theta^* - \mathbb{E}[\theta^*]|^2 \\
&\preceq (\mathbb{E} - \widehat{\mathbb{E}})\Big|\theta^* - \mathbb{E}[\theta^*]\Big|^2 + 2\left|\widehat{\mathbb{E}}[\theta^*] - \mathbb{E}[\theta^*]\right|^2 + 2\widehat{\mathbb{E}}\left|\theta^* - \widehat{\mathbb{E}}[\theta^*]\right|^2 \\
&= \underbrace{(\mathbb{E} - \widehat{\mathbb{E}})\Big|\theta^* - \mathbb{E}[\theta^*]\Big|^2}_{T_1} + \underbrace{2\left|\widehat{\mathbb{E}}[\theta^*] - \mathbb{E}[\theta^*]\right|^2}_{T_2} + 2\widehat{\sigma}^2(\theta^*). \qquad (8.15)
\end{aligned}$$

We claim that the terms T_1 and T_2 are bounded as follows:

$$T_1 \preceq \frac{\sigma^2(\theta^*)}{4} + c\|\theta^*\|_{\text{span}}^2 \frac{\log(8D/\delta)}{N} \cdot \mathbf{1}, \quad \text{and} \qquad (8.16a)$$

$$T_2 \preceq c\left\{\frac{\log(8D/\delta)}{N} \cdot \sigma^2(\theta^*) + \left(\|\theta^*\|_{\text{span}} \frac{\log(8D/\delta)}{N}\right)^2 \cdot \mathbf{1}\right\}, \qquad (8.16b)$$

where each bound holds with probability at least $1 - \frac{\delta}{4}$. Taking these bounds as given for the moment, as long as $N \geq c' \log(8D/\delta)$ for a sufficiently large constant c', we can ensure that

$$T_1 + T_2 \preceq \frac{\sigma^2(\theta^*)}{2} + c\|\theta^*\|_{\text{span}}^2 \frac{\log(8D/\delta)}{N} \cdot \mathbf{1},$$

Substituting back into our earlier bound (8.15), we find that

$$\frac{\sigma^2(\theta^*)}{2} \preceq 2\widehat{\sigma}^2(\theta^*) + c'\|\theta^*\|_{\text{span}}^2 \frac{\log(8D/\delta)}{N} \cdot \mathbf{1}.$$

Rearranging and taking square roots entry-wise, we find that

$$\sigma(\theta^*) \preceq \sqrt{4\widehat{\sigma}^2(\theta^*) + 2c'\|\theta^*\|_{\text{span}}^2 \frac{\log(8D/\delta)}{N} \cdot \mathbf{1}} \;\preceq\; 2\widehat{\sigma}(\theta^*) + c'\|\theta^*\|_{\text{span}} \sqrt{\frac{\log(8D/\delta)}{N}} \cdot \mathbf{1}.$$

Finally noting that we have the entry-wise inequality $\widehat{\sigma}(\theta^*) \preceq \widehat{\sigma}(\widehat{\theta}) + |\widehat{\theta} - \theta^*|$ establishes the claim of Lemma 8.4.1.
It remains to prove the bounds (8.16a) and (8.16b).

Proof of bound (8.16a)**:** For each index $i \in [D]$, define the random variable $Y_i := \big(\theta^*_J - \mathbb{E}_i[\theta^*]\big)^2$, where J is an index chosen at random from the distribution p_i. By definition, each random variable Y_i is non-negative, and so with $\mathbb{E}$ now denoting the regular expectation of a scalar random variable, we have lower tail bound (Proposition 2.14, [323])

$$\mathbb{P}\left[\mathbb{E}[Y_i] - Y_i \geq s\right] \leq \exp\left(-\frac{ns^2}{2\mathbb{E}[Y_i^2]}\right) \qquad \text{for all } s > 0.$$

Moreover, we have $Y_i \leq \|\theta^*\|_{\mathrm{span}}^2$ almost surely, from which we obtain

$$\mathbb{E}[Y_i^2] \leq \|\theta^*\|_{\mathrm{span}}^2 \mathbb{E}_i\left[(\theta^* - \mathbb{E}[\theta^*])^2\right] \;=\; \|\theta^*\|_{\mathrm{span}}^2 \sigma_i^2(\theta^*).$$

Putting together the pieces yields the elementwise inequality

$$T_1 \preceq c\|\theta^*\|_{\mathrm{span}} \sqrt{\frac{\log(8D/\delta)}{N}} \cdot \sigma(\theta^*) \overset{\text{(i)}}{\preceq} \frac{\sigma^2(\theta^*)}{8} + c'\|\theta^*\|_{\mathrm{span}}^2 \frac{\log(8D/\delta)}{N},$$

with probability at least $1 - \delta/3$, where in step (i), we have used the inequality $2ab \leq \nu a^2 + \nu^{-1} b^2$, which holds for any triple of positive scalars (a, b, ν).

Proof of the bound (8.16b)**:** From Bernstein's inequality, we have the element-wise bound

$$\left|\widehat{\mathbb{E}}[\theta^*] - \mathbb{E}[\theta^*]\right| \preceq c\left\{\sqrt{\frac{\log(8D/\delta)}{N}} \cdot \sigma(\theta^*) + \|\theta^*\|_{\mathrm{span}} \frac{\log(8D/\delta)}{N} \cdot \mathbf{1}\right\}$$

with probability at least $1 - \delta/4$, and hence

$$T_2 \preceq c\left\{\frac{\log(8D/\delta)}{N} \cdot \sigma^2(\theta^*) + \left(\|\theta^*\|_{\mathrm{span}} \frac{\log(8D/\delta)}{N}\right)^2 \cdot \mathbf{1}\right\},$$

as claimed.

8.4.2 Proof of Theorem 8.1.1, part (b)

Once again, we employ the shorthand $\widehat{\theta} \equiv \widehat{\theta}_{\mathsf{plug}}$ for notational convenience, and also the shorthand $\widehat{\Delta} = \widehat{\theta} - \theta^*$. Note that it suffices to show the inequality

$$\Pr\left\{\|\widehat{\theta} - \theta^*\|_\infty \geq c\gamma \left\|(\mathbf{I} - \gamma\mathbf{P})^{-1}|(\widehat{\mathbf{P}} - \mathbf{P})\theta^*|\right\|_\infty + c(1-\gamma)^{-1}\|\widehat{r} - r\|_\infty\right\} \leq \frac{\delta}{2}, \tag{8.17}$$

from which the theorem follows by application of a Bernstein bound to the first term and Hoeffding bound to the second, in a similar fashion to the inequalities (8.12). We therefore dedicate the rest of the proof to establishing inequality (8.17).

Proving the bound (8.17)

We have

$$\widehat{\Delta} = \widehat{\theta} - \theta^* = \gamma\widehat{\mathbf{P}}\widehat{\theta} - \gamma\mathbf{P}\theta^* + (\widehat{r} - r) = \gamma(\widehat{\mathbf{P}} - \mathbf{P})\widehat{\theta} + \gamma\mathbf{P}\widehat{\Delta} + (\widehat{r} - r),$$

which implies that

$$\begin{aligned}\widehat{\Delta} - (\mathbf{I} - \gamma\mathbf{P})^{-1}(\widehat{r} - r) &= \gamma(\mathbf{I} - \gamma\mathbf{P})^{-1}(\widehat{\mathbf{P}} - \mathbf{P})\widehat{\theta} \\ &= \gamma(\mathbf{I} - \gamma\mathbf{P})^{-1}(\widehat{\mathbf{P}} - \mathbf{P})\widehat{\Delta} + \gamma(\mathbf{I} - \gamma\mathbf{P})^{-1}(\widehat{\mathbf{P}} - \mathbf{P})\theta^*.\end{aligned} \tag{8.18}$$

Since all entries of $(\mathbf{I} - \gamma\mathbf{P})^{-1}$ are non-negative, we have the element-wise inequalities

$$\begin{aligned} |\widehat{\Delta}| &\preceq \gamma(\mathbf{I} - \gamma\mathbf{P})^{-1}|(\widehat{\mathbf{P}} - \mathbf{P})\widehat{\Delta}| + \gamma(\mathbf{I} - \gamma\mathbf{P})^{-1}|(\widehat{\mathbf{P}} - \mathbf{P})\theta^*| + (\mathbf{I} - \gamma\mathbf{P})^{-1}|\widehat{r} - r| \\ &\preceq \gamma(\mathbf{I} - \gamma\mathbf{P})^{-1}|(\widehat{\mathbf{P}} - \mathbf{P})\widehat{\Delta}| + \gamma(\mathbf{I} - \gamma\mathbf{P})^{-1}|(\widehat{\mathbf{P}} - \mathbf{P})\theta^*| + \frac{1}{1-\gamma}\|\widehat{r} - r\|_\infty \cdot \mathbf{1}. \end{aligned} \tag{8.19}$$

The second and third terms are already in terms of the desired population-level functionals in equation (8.17). It remains to bound the first term.

Note that the key difficulty here is the fact that the two matrices $\widehat{\mathbf{P}} - \mathbf{P}$ and $\widehat{\Delta}$ are not independent. As a first attempt to address this dependence, one is tempted to use the fact that provided N is large enough, each row of $\widehat{\mathbf{P}} - \mathbf{P}$ has small ℓ_1-norm; for instance, see Weissman et al. [330] for sharp bounds of this type. In particular, this would allow us to work with the entry-wise bounds

$$|(\widehat{\mathbf{P}} - \mathbf{P})\widehat{\Delta}| \preceq \|\widehat{\mathbf{P}} - \mathbf{P}\|_{1,\infty}\|\widehat{\Delta}\|_\infty \cdot \mathbf{1} \precsim C\sqrt{\frac{D}{N}}\|\widehat{\Delta}\|_\infty \cdot \mathbf{1},$$

where the final relation hides logarithmic factors in the pair (D, δ). Proceeding in this fashion, we would then bound each entry in the first term of equation (8.19) by $\gamma(1-\gamma)^{-1}\sqrt{\frac{D}{N}}\|\widehat{\Delta}\|_\infty$; then choosing N large enough such that $\gamma(1-\gamma)^{-1}\sqrt{\frac{D}{N}} \leq 1/2$ suffices to establish bound (8.17). However, this requires a sample size $N \gtrsim \frac{\gamma^2}{(1-\gamma)^2}D$, while we wish to obtain the bound (8.17) with the sample size $N \gtrsim \frac{\gamma^2}{(1-\gamma)^2}$. This requires a more delicate analysis.

Our analysis instead proceeds entry-by-entry, and uses a leave-one-out sequence to carefully decouple the dependence between $\widehat{\mathbf{P}} - \mathbf{P}$ and $\widehat{\Delta}$. Let us introduce some notation to make this precise. For each $i \in [D]$, recall that we used $\widehat{p}_i$ and p_i to denote row i of the matrices $\widehat{\mathbf{P}}$ and $\mathbf{P}$, respectively. Let $\widehat{\mathbf{P}}^{(i)}$ denote the i-th leave-one-out transition matrix, which is identical to $\widehat{\mathbf{P}}$ except with row i replaced by the population vector p_i. Let $\widehat{\theta}^{(i)} := (\mathbf{I} - \gamma\widehat{\mathbf{P}}^{(i)})^{-1}r$ be the value function estimate based on $\widehat{\mathbf{P}}^{(i)}$ and the *true* reward vector r, and denote the associated difference vector by $\widehat{\Delta}^{(i)} := \widehat{\theta}^{(i)} - \theta^*$.

Now note that we have

$$\Big[(\widehat{\mathbf{P}} - \mathbf{P})\widehat{\Delta}\Big]_i = \langle \widehat{p}_i - p_i,\, \widehat{\Delta}\rangle = \langle \widehat{p}_i - p_i,\, \widehat{\Delta}^{(i)}\rangle + \langle \widehat{p}_i - p_i,\, \widehat{\theta} - \widehat{\theta}^{(i)}\rangle.$$

This decomposition is helpful because, now, the vectors $\widehat{p}_i - p_i$ and $\widehat{\Delta}^{(i)}$ are independent by construction, so that standard tail bounds can be used on the first term. For the second term, we use the fact that $\widehat{\theta} \approx \widehat{\theta}^{(i)}$, since the latter is obtained by replacing just one row of the estimated transition matrix. Formally, this closeness will be argued by using the matrix inversion formula. We collect these two results in the following lemma.

Lemma 8.4.2. *Suppose that the sample size is lower bounded as $N \geq c'\gamma^2 \frac{\log(8D/\delta)}{(1-\gamma)^2}$. Then with probability at least $1 - \frac{\delta}{2D}$ and for each $i \in [D]$, we have*

$$\gamma |\langle \widehat{p}_i - p_i,\, \widehat{\Delta}^{(i)} \rangle| \leq c \left\{ \gamma \|\widehat{\Delta}\|_\infty \sqrt{\frac{\log(8D/\delta)}{N}} + \gamma \, |\langle \widehat{p}_i - p_i,\, \theta^* \rangle| + \|r - \widehat{r}\|_\infty \right\} \quad \textit{and} \tag{8.20a}$$

$$\gamma |\langle \widehat{p}_i - p_i,\, \widehat{\theta}^{(i)} - \widehat{\theta} \rangle| \leq c \left\{ \gamma \|\widehat{\Delta}\|_\infty \sqrt{\frac{\log(8D/\delta)}{N}} + \gamma \, |\langle \widehat{p}_i - p_i,\, \theta^* \rangle| + \|r - \widehat{r}\|_\infty \right\}. \tag{8.20b}$$

With this lemma in hand, let us complete the proof. Combining the bounds of Lemma 8.4.2 with a union bound over all D entries yields the elementwise inequality

$$\gamma \left| (\widehat{\mathbf{P}} - \mathbf{P}) \widehat{\Delta} \right| \preceq c\gamma \left| (\widehat{\mathbf{P}} - \mathbf{P}) \theta^* \right| + c \left\{ \gamma \|\widehat{\Delta}\|_\infty \sqrt{\frac{\log(8D/\delta)}{N}} + \|\widehat{r} - r\|_\infty \right\} \mathbb{1}$$

with probability at least $1 - \delta/2$. Since the entries of $(\mathbf{I} - \gamma \mathbf{P})^{-1}$ are non-negative, we can multiply both sides of this inequality by it, thereby obtaining

$$\begin{aligned} &\gamma (\mathbf{I} - \gamma \mathbf{P})^{-1} \left| (\widehat{\mathbf{P}} - \mathbf{P}) \widehat{\Delta} \right| \\ &\qquad \preceq c\gamma (\mathbf{I} - \gamma \mathbf{P})^{-1} \left| (\widehat{\mathbf{P}} - \mathbf{P}) \theta^* \right| + \frac{c}{1-\gamma} \left\{ \gamma \|\widehat{\Delta}\|_\infty \sqrt{\frac{\log(8D/\delta)}{N}} + \|\widehat{r} - r\|_\infty \right\} \mathbf{1}. \end{aligned}$$

Returning to the upper bound (8.19), we have shown that

$$\|\widehat{\Delta}\|_\infty \leq c\gamma \frac{\|\widehat{\Delta}\|_\infty}{1-\gamma} \sqrt{\frac{\log(8D/\delta)}{N}} + c'\gamma \left\| (\mathbf{I} - \gamma \mathbf{P})^{-1} |(\widehat{\mathbf{P}} - \mathbf{P}) \theta^*| \right\|_\infty + \frac{c}{1-\gamma} \|r - \widehat{r}\|_\infty.$$

Under the assumed lower bound on the sample size $N \geq c'\gamma^2 \frac{\log(8D/\delta)}{(1-\gamma)^2}$, this inequality implies that

$$\|\widehat{\Delta}\|_\infty \leq c'\gamma \left\| (\mathbf{I} - \gamma \mathbf{P})^{-1} |(\widehat{\mathbf{P}} - \mathbf{P}) \theta^*| \right\|_\infty + \frac{c}{1-\gamma} \|r - \widehat{r}\|_\infty,$$

as claimed (8.17). □

We now proceed to a proof of Lemma 8.4.2, which uses the following structural lemma relating the quantities $\widehat{\Delta}^{(i)}$ and $\widehat{\Delta}$.

Lemma 8.4.3. *Suppose that the sample size is lower bounded as $N \geq c'\gamma^2 \frac{\log(8D/\delta)}{(1-\gamma)^2}$. Then with probability at least $1 - \frac{\delta}{4D}$ and for each $i \in [D]$, we have*

$$\|\widehat{\Delta}^{(i)}\|_\infty \leq c \|\widehat{\Delta}\|_\infty + \frac{c}{1-\gamma} \Big\{ \gamma \, |\langle \widehat{p}_i - p_i,\, \theta^* \rangle| + \|\widehat{r} - r\|_\infty \Big\}. \tag{8.21}$$

This lemma is proved in Section 8.4.2 to follow.

Proof of Lemma 8.4.2

We prove the two bounds in turn.

Proof of inequality (8.20a)**:** Note that $\widehat{p}_i - p_i$ and $\widehat{\Delta}^{(i)}$ are independent by construction, so that the Hoeffding inequality yields

$$|\langle \widehat{p}_i - p_i,\, \widehat{\Delta}^{(i)}\rangle| \leq c\|\widehat{\Delta}^{(i)}\|_\infty \sqrt{\frac{\log(8D/\delta)}{N}} \tag{8.22}$$

with probability at least $1 - \delta/(4D)$.

Using this in conjunction with inequality (8.21) from Lemma 8.4.3 yields the bound

$$\begin{aligned}\gamma|\langle \widehat{p}_i - p_i,\, \widehat{\Delta}^{(i)}\rangle| &\leq c\gamma\|\widehat{\Delta}\|_\infty \sqrt{\frac{\log(8D/\delta)}{N}} + \frac{c\gamma}{1-\gamma}\sqrt{\frac{\log(8D/\delta)}{N}}\Big\{\gamma\,|\langle \widehat{p}_i - p_i,\, \theta^*\rangle| + \|\widehat{r} - r\|_\infty\Big\} \\ &\overset{\text{(i)}}{\leq} c\gamma\|\widehat{\Delta}\|_\infty \sqrt{\frac{\log(8D/\delta)}{N}} + c\gamma\,|\langle \widehat{p}_i - p_i,\, \theta^*\rangle| + c\|\widehat{r} - r\|_\infty,\end{aligned}$$

where in step (i), we have used the lower bound on the sample size $N \geq c' \frac{\gamma^2}{(1-\gamma)^2} \log(8D/\delta)$.

Proof of inequality (8.20b)**:** The proof of this claim is more involved. Using the relation (8.18) (with suitable modifications of terms), we have

$$\begin{aligned}\widehat{\theta}^{(i)} - \widehat{\theta} &= \gamma(\mathbf{I} - \gamma\widehat{\mathbf{P}})^{-1}(\widehat{\mathbf{P}}^{(i)} - \widehat{\mathbf{P}})\widehat{\theta}^{(i)} + (\mathbf{I} - \gamma\widehat{\mathbf{P}})^{-1}(r - \widehat{r}) \\ &= -\gamma(\mathbf{I} - \gamma\widehat{\mathbf{P}})^{-1}e_i\left(\langle \widehat{p}_i - p_i,\, \widehat{\theta}^{(i)}\rangle\right) + (\mathbf{I} - \gamma\widehat{\mathbf{P}})^{-1}(r - \widehat{r}).\end{aligned} \tag{8.23}$$

Moreover, the Woodbury matrix identity [144] yields

$$\mathbf{M} := \left(\mathbf{I} - \gamma\widehat{\mathbf{P}}\right)^{-1} - \left(\mathbf{I} - \gamma\widehat{\mathbf{P}}^{(i)}\right)^{-1} = -\gamma\frac{(\mathbf{I} - \gamma\widehat{\mathbf{P}}^{(i)})^{-1}e_i(\widehat{p}_i - p_i)^T(\mathbf{I} - \gamma\widehat{\mathbf{P}}^{(i)})^{-1}}{1 - \gamma(\widehat{p}_i - p_i)^T(\mathbf{I} - \gamma\widehat{\mathbf{P}}^{(i)})^{-1}e_i}.$$

Consequently,

$$\begin{aligned}&\langle \widehat{p}_i - p_i,\, \widehat{\theta}^{(i)} - \widehat{\theta}\rangle \\ &\quad= -\gamma(\widehat{p}_i - p_i)^\top(\mathbf{I} - \gamma\widehat{\mathbf{P}})^{-1}e_i\left(\langle \widehat{p}_i - p_i,\, \widehat{\theta}^{(i)}\rangle\right) + (\widehat{p}_i - p_i)^\top(\mathbf{I} - \gamma\widehat{\mathbf{P}})^{-1}(r - \widehat{r}) \\ &\quad= -\gamma(\widehat{p}_i - p_i)^\top(\mathbf{I} - \gamma\widehat{\mathbf{P}}^{(i)})^{-1}e_i\left(\langle \widehat{p}_i - p_i,\, \widehat{\theta}^{(i)}\rangle\right) + (\widehat{p}_i - p_i)^\top(\mathbf{I} - \gamma\widehat{\mathbf{P}}^{(i)})^{-1}(r - \widehat{r}) \\ &\qquad\quad - \gamma(\widehat{p}_i - p_i)^\top\mathbf{M}e_i\left(\langle \widehat{p}_i - p_i,\, \widehat{\theta}^{(i)}\rangle\right) + (\widehat{p}_i - p_i)^\top\mathbf{M}(r - \widehat{r}) \\ &\quad= \left(\langle \widehat{p}_i - p_i,\, \widehat{\theta}^{(i)}\rangle\right)\cdot\frac{2Z_i^2 - Z_i}{1 - Z_i} + T_i\cdot\frac{1 - 2Z_i}{1 - Z_i},\end{aligned} \tag{8.24}$$

where we have defined, for convenience, the random variables

$$Z_i := \gamma(\widehat{p}_i - p_i)^\top(\mathbf{I} - \gamma\widehat{\mathbf{P}}^{(i)})^{-1}e_i \quad \text{and} \quad T_i := (\widehat{p}_i - p_i)^\top(\mathbf{I} - \gamma\widehat{\mathbf{P}}^{(i)})^{-1}(r - \widehat{r}).$$

Since $\widehat{p}_i - p_i$ is independent of the vector $(\mathbf{I} - \gamma\widehat{\mathbf{P}}^{(i)})^{-1}(r - \widehat{r})$, applying the Hoeffding bound yields the inequality

$$|T_i| \leq \frac{c}{1-\gamma}\|r - \widehat{r}\|_\infty\sqrt{\frac{\log(8D/\delta)}{N}}$$

with probability exceeding $1 - \delta/(4D)$.

On the other hand, exploiting independence between the vectors $\widehat{p}_i - p_i$ and $(\mathbf{I} - \gamma\widehat{\mathbf{P}}^{(i)})^{-1}e_i$ and applying the Hoeffding bound, we also have

$$|Z_i| \leq \frac{c\gamma}{1-\gamma}\sqrt{\frac{\log(8D/\delta)}{N}}$$

with probability least $1 - \delta/(4D)$. Taking $N \geq c'\frac{\gamma^2}{(1-\gamma)^2}\log(8D/\delta)$ for a sufficiently large constant c' ensures that $\gamma|T_i| \leq \|r - \widehat{r}\|_\infty$ and $|Z_i| \leq 1/4$, so that with probability exceeding $1 - \delta/(2D)$, inequality (8.24) yields

$$\begin{aligned}\gamma|\langle\widehat{p}_i - p_i,\, \widehat{\theta}^{(i)} - \widehat{\theta}\rangle| &\leq c\Big\{\gamma|\langle\widehat{p}_i - p_i,\, \widehat{\theta}^{(i)}\rangle| + \|r - \widehat{r}\|_\infty\Big\} \\ &\leq c\Big\{\gamma|\langle\widehat{p}_i - p_i,\, \widehat{\Delta}^{(i)}\rangle| + \gamma|\langle\widehat{p}_i - p_i,\, \theta^*\rangle| + \|r - \widehat{r}\|_\infty\Big\}.\end{aligned}$$

Finally, applying part (a) of Lemma 8.4.2 completes the proof. □

Proof of Lemma 8.4.3

Recall our leave-one-out matrix $\widehat{\mathbf{P}}^{(i)}$, and the explicit bound (8.22). We have

$$\Big|\langle\widehat{p}_i - p_i,\, \widehat{\theta}^{(i)}\rangle\Big| \leq \Big|\langle\widehat{p}_i - p_i,\, \widehat{\Delta}^{(i)}\rangle\Big| + |\langle\widehat{p}_i - p_i,\, \theta^*\rangle| \;\leq\; c\|\widehat{\Delta}^{(i)}\|_\infty\sqrt{\frac{\log(8D/\delta)}{N}} + |\langle\widehat{p}_i - p_i,\, \theta^*\rangle| \tag{8.25}$$

with probability at least $1 - \delta/(4D)$. Substituting inequality (8.25) into the bound (8.23), we find that

$$\|\widehat{\theta}^{(i)} - \widehat{\theta}\|_\infty \leq \frac{c}{1-\gamma}\left\{\gamma\|\widehat{\Delta}^{(i)}\|_\infty \cdot \sqrt{\frac{\log(8D/\delta)}{N}} + \gamma\,|\langle\widehat{p}_i - p_i,\, \theta^*\rangle| + \|r - \widehat{r}\|_\infty\right\}. \tag{8.26}$$

Finally, the triangle inequality yields

$$\begin{aligned}\|\widehat{\Delta}^{(i)}\|_\infty &\leq \|\widehat{\Delta}\|_\infty + \|\widehat{\theta}^{(i)} - \widehat{\theta}\|_\infty \\ &\leq\; \|\widehat{\Delta}\|_\infty + \frac{c}{1-\gamma}\left\{\gamma\|\widehat{\Delta}^{(i)}\|_\infty \cdot \sqrt{\frac{\log(8D/\delta)}{N}} + \gamma\,|\langle\widehat{p}_i - p_i,\, \theta^*\rangle| + \|r - \widehat{r}\|_\infty\right\}.\end{aligned}$$

For $N \geq c'\gamma^2 \frac{\log(8D/\delta)}{(1-\gamma)^2}$ with c' sufficiently large, we have

$$\|\widehat{\Delta}^{(i)}\|_\infty \leq c\|\widehat{\Delta}\|_\infty + \frac{c}{1-\gamma}\Big\{\gamma\,|\langle \widehat{p}_i - p_i,\, \theta^*\rangle| + \|\widehat{r} - r\|_\infty\Big\}$$

with probability at least $1 - \frac{\delta}{4D}$, which completes the proof of Lemma 8.4.3. □

Since Corollary 8.2.1 follows from Theorem 8.1.1, we prove it first before moving to a proof of Theorem 8.2.1 in Section 8.4.4.

8.4.3 Proof of Corollary 8.2.1

In order to prove part (a), consider inequality (8.10) and further use the fact that we have the ℓ_∞-bound $\|(\mathbf{I} - \gamma\widehat{\mathbf{P}})^{-1}\|_{1,\infty} \leq \frac{1}{1-\gamma}$ to obtain the element-wise bound

$$|\widehat{\theta} - \theta^*| \preceq \frac{\gamma}{1-\gamma}\|(\widehat{\mathbf{P}} - \mathbf{P})\theta^*\|_\infty \mathbf{1} + \frac{\|\widehat{r} - r\|_\infty}{1-\gamma} \cdot \mathbf{1}.$$

Applying Bernstein's bound to the first term and Hoeffding's bound to the second completes the proof.

In order to prove part (b) of the corollary, we apply Lemma 7 of Azar et al. [7]—in particular, equation (17) of that paper. Tailored to this setting, their result leads to the point-wise bound

$$\|(\mathbf{I} - \gamma\mathbf{P})^{-1}\sigma(\theta^*)\|_\infty \leq c\frac{r_{\max}}{(1-\gamma)^{3/2}}.$$

We also have the bound

$$\|\theta^*\|_{\text{span}} \leq 2\|\theta^*\|_\infty = 2\|(\mathbf{I} - \gamma\mathbf{P})^{-1}r\|_\infty \leq \frac{2r_{\max}}{1-\gamma},$$

so that combining the pieces and applying Theorem 8.1.1(b), we obtain

$$\|\widehat{\theta} - \theta^*\|_\infty \leq \frac{c}{(1-\gamma)}\left\{\sqrt{\frac{\log(8D/\delta)}{N}}\left(\gamma\frac{r_{\max}}{(1-\gamma)^{1/2}} + \|\rho(r)\|_\infty\right) + \gamma \cdot \frac{\log(8D/\delta)}{N}\frac{r_{\max}}{1-\gamma}\right\}.$$

Finally, when $N \geq c_1\frac{\log(8D/\delta)}{1-\gamma}$ for a sufficiently large constant c_1, we have

$$\frac{\log(8D/\delta)}{N}\frac{r_{\max}}{1-\gamma} \leq c\sqrt{\frac{\log(8D/\delta)}{N}}\frac{r_{\max}}{(1-\gamma)^{1/2}},$$

thereby establishing the claim. □

8.4.4 Proof of Theorem 8.2.1

For all of our lower bounds, we assume that the reward distribution takes the Gaussian form

$$\mathcal{D}_r(\cdot \mid j) = \mathcal{N}(r_j, \varrho^2) \tag{8.27}$$

for each state j. Note that this reward distribution satisfies $\|\rho(r)\|_\infty = \varrho$ by construction.

Let us begin with a short overview of our proof, which proceeds in two steps. First, we suppose that the transition matrix $\mathbf{P}$ is known exactly, and the hardness of the estimation problem is due to noisy observations of the reward function. In particular, letting $\mathcal{M}_{\mathbf{I}}(r_{\max}, \varrho)$ denote the class of all MRPs with the specific reward observation model (8.27), and for which the transition matrix is the identity matrix $\mathbf{I}$ and the rewards are uniformly bounded as $\|r\|_\infty \leq r_{\max}$, we show that

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \mathcal{M}_{\mathbf{I}}(r_{\max}, \varrho)} \mathbb{E}\|\widehat{\theta} - \theta^*\|_\infty \geq c \left\{ \frac{\varrho}{1-\gamma} \cdot \sqrt{\frac{\log(D)}{N}} \wedge \frac{r_{\max}}{1-\gamma} \right\}. \tag{8.28}$$

Note that for each pair of positive scalars $(\vartheta, r_{\max})$ we have the inclusions

$$\mathcal{M}_{\mathbf{I}}(r_{\max}, \varrho) \subseteq \mathcal{M}_{\mathrm{var}}(\vartheta, \varrho) \quad \text{and} \quad \mathcal{M}_{\mathbf{I}}(r_{\max}, \varrho) \subseteq \mathcal{M}_{\mathsf{rew}}(r_{\max}, \varrho),$$

and so that the lower bound (8.28) carries over to the classes $\mathcal{M}_{\mathrm{var}}(\vartheta, \varrho)$ and $\mathcal{M}_{\mathsf{rew}}(r_{\max}, \varrho)$.

Next, we suppose that the population reward function r is known exactly ($\varrho = 0$), and the hardness of the estimation problem is only due to uncertainty in the transitions. Under this setting, we prove the lower bounds

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \mathcal{M}_{\mathrm{var}}(\vartheta, 0)} \mathbb{E}\|\widehat{\theta} - \theta^*\|_\infty \geq c \frac{\vartheta}{1-\gamma} \cdot \sqrt{\frac{\log(D/2)}{N}}, \qquad \text{and} \tag{8.29a}$$

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \mathcal{M}_{\mathsf{rew}}(r_{\max}, 0)} \mathbb{E}\|\widehat{\theta} - \theta^*\|_\infty \geq c \frac{r_{\max}}{(1-\gamma)^{3/2}} \cdot \sqrt{\frac{\log(D/2)}{N}}. \tag{8.29b}$$

Since $\mathcal{M}_{\mathrm{var}}(\vartheta, 0) \subset \mathcal{M}_{\mathrm{var}}(\vartheta, \varrho)$ for any $\varrho > 0$, these lower bounds also carry over to the more general setting. The minimax lower bounds of Theorem 8.2.1 are obtained by taking the maximum of the bounds (8.28) and (8.29). Let us now establish the two previously claimed bounds.

Proof of claim (8.28)

For some positive scalar λ to be chosen shortly, consider D distinct reward vectors $\{r^{(1)}, \ldots, r^{(D)}\}$, where the vector $r^{(i)} \in \mathbb{R}^D$ has entries

$$r_j^{(i)} := \begin{cases} \lambda & \text{if } i = j \\ 0 & \text{otherwise,} \end{cases} \qquad \text{for all } j \in [D].$$

Denote by $\mathcal{R}^{(i)}$ the MRP with reward function $r^{(i)}$; and transition matrix $\mathbf{I}$. Thus, the i-th value function is given by the vector $(\theta^*)^{(i)} := \frac{1}{1-\gamma} r^{(i)}$.

By construction, we have $\|(\theta^*)^{(i)} - (\theta^*)^{(j)}\|_\infty = \lambda/(1-\gamma)$ for each pair of distinct indices (i, j). Furthermore, the KL divergence between Gaussians of variance ϱ^2 centered at $r^{(i)}$ and $r^{(j)}$ is given by

$$\mathsf{KL}\left(\mathcal{N}(r^{(i)}, \varrho^2 \mathbf{I}) \parallel \mathcal{N}(r^{(j)}, \varrho^2 \mathbf{I})\right) = \frac{\|r^{(i)} - r^{(j)}\|_2^2}{\varrho^2} = \frac{2\lambda^2}{\varrho^2}.$$

Thus, applying the local packing version of Fano's method (§15.3.3, [323]), we have

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \{\mathcal{M}^{(i)}\}_{i\in[D]}} \mathbb{E}\|\widehat{\theta} - \theta^*\|_\infty \geq c \frac{\lambda}{1-\gamma}\left(1 - \frac{2\frac{\lambda^2}{\varrho^2}N + \log 2}{\log D}\right).$$

Setting $\lambda = \varrho\sqrt{\frac{\log D}{6N}} \wedge r_{\max}$ yields the claimed lower bound.

Proof of claim (8.29)

This lower bound is based on a modification of constructions used by Lattimore and Hutter [184] and Azar et al. [7]. Our proof, however, is tailored to the generative observation model.

Our proof is structured as follows. First, we construct a family of "hard" MRPs and prove a minimax lower bound as a function of parameters used to define this family. Constructing this family of hard instances requires us to first define a basic building block: a two-state MRP that was illustrated in Figure 7.2. After obtaining this general lower bound, we then set the scalars that parameterize the hard class MRP appropriately to obtain the two claimed bounds.

We now describe the two-state MRP in more detail. For a pair of parameters (p, τ), each in the unit interval $[0, 1]$, and a positive scalar ν, consider the two-state Markov reward process $\mathcal{R}_\mathbf{0}(p, \nu, \tau)$, with transition matrix and reward vector given by

$$\mathbf{P_0} = \begin{bmatrix} p & 1-p \\ 0 & 1 \end{bmatrix} \qquad \text{and} \qquad r_\mathbf{0} = \begin{bmatrix} \nu \\ \nu \cdot \tau \end{bmatrix},$$

respectively. See Figure 7.2 in Chapter 7 for an illustration of this MRP.

A straightforward calculation yields that it has value function and corresponding standard deviation vector given by

$$\theta^*(p, \nu, \tau) = \nu \begin{bmatrix} \frac{1-\gamma+\gamma\tau(1-p)}{(1-\gamma p)(1-\gamma)} \\ \frac{\tau}{1-\gamma} \end{bmatrix} \qquad \text{and} \qquad \sigma(\theta^*) = \nu \begin{bmatrix} \frac{(1-\tau)\sqrt{p(1-p)}}{1-\gamma p} \\ 0 \end{bmatrix}, \tag{8.30}$$

respectively, where we have used the shorthand $\theta^* \equiv \theta^*(p, \nu, \tau)$. We also have $\|\theta^*\|_{\text{span}} = \frac{\nu(1-\tau)}{1-\gamma p}$; the two scalars (ν, τ) allow us to control the quantities $\|\sigma(\theta^*)\|_\infty$ and $\|\theta^*\|_{\text{span}}$. Index the states of this MRP by the set $\{0, 1\}$, and consider now a sample drawn from this MRP under the generative model. We see a pair of states drawn according to the respective rows of the transition matrix $\mathbf{P_0}$; the first state is drawn according to the Bernoulli distribution Ber(p), and the second state is deterministic and equal to 1. For convenience, we use $\mathbb{P}(p) = (\text{Ber}(p), 1)$ to denote the distribution of this pair of states.

Our hard class of instances is based in part on the difficulty of distinguishing two such MRPs that are close in a specific sense. Let us make this intuition precise. For two scalar values $0 \leq p_2 \leq p_1 \leq 1$, some algebra yields the relation

$$\|\theta^*(p_1, \nu, \tau) - \theta^*(p_2, \nu, \tau)\|_\infty = \nu \cdot \frac{(p_1 - p_2)(1 - \tau)}{(1 - \gamma p_1)(1 - \gamma p_2)}. \tag{8.31}$$

In the sequel, we work with the choices

$$p_1 = \frac{4\gamma - 1}{3\gamma} \quad \text{and} \quad p_2 = p_1 - \frac{1}{8}\sqrt{\frac{p_1(1-p_1)}{N}\log(D/2)},$$

which, under the assumed lower bound on the sample size N, are both scalars in the range $\left[\frac{1}{2}, 1\right)$ for all discount factors $\gamma \in \left[\frac{1}{2}, 1\right)$. Moreover, it is worth noting the relations

$$\begin{aligned} 1 - p_1 &= \frac{1-\gamma}{3\gamma}, & & c_1\frac{1-\gamma}{3\gamma} \leq 1 - p_2 \leq c_2\frac{1-\gamma}{3\gamma} \\ 1 - \gamma p_1 &= \frac{4}{3}(1-\gamma), & \text{and} \qquad & c_1(1-\gamma) \leq 1 - \gamma p_2 \leq c_2(1-\gamma), \end{aligned} \tag{8.32}$$

where the inequalities on the right hold provided $N \geq \frac{c\gamma}{1-\gamma}\log(D/2)$ for a sufficiently large constant c. Here the pair of constants (c_1, c_2) are universal, depend only on c, and may change from line to line.

We also require the following lemma, proved in Section 8.4.4 to follow, which provides a useful bound on the KL divergence between $\mathbb{P}(p_1)$ and $\mathbb{P}(p_2)$.

Lemma 8.4.4. *For each pair* $p, q \in [1/2, 1)$*, we have*

$$\mathsf{KL}\left(\mathbb{P}(p) \| \mathbb{P}(q)\right) \leq \frac{(p-q)^2}{(p \vee q)(1 - (p \vee q))}.$$

We are now in a position to describe the hard family of MRPs over which we prove a general lower bound. Suppose that D is even for convenience, and consider a set of $D/2$ "master" MRPs $\bar{\mathcal{M}} := \{\mathcal{R}_1, \ldots, \mathcal{R}_{D/2}\}$ each on D states[10] constructed as follows. Decompose each master MRP into $D/2$ sub-MRPs of two states each; index the k-th sub-MRP in the j-th master MRP by $\mathcal{R}_{j,k}$. For each pair $j, k \in [D/2]$, set

$$\mathcal{R}_{j,k} = \begin{cases} \mathcal{R}_\mathbf{0}(p_1, \nu, \tau) & \text{if } j \neq k \\ \mathcal{R}_\mathbf{0}(p_2, \nu, \tau) & \text{otherwise.} \end{cases}$$

[10] Note that this step is only required in order to "tensorize" the construction in order to obtain the optimal dependence on the dimension. If, instead of the ℓ_∞ error, one was interested in estimating the value function at a fixed state of the MRP, then this tensorization is no longer needed.

Let θ_j^* denote the value function corresponding to MRP $\mathcal{R}_j$, and let $\mathbb{P}_j^N$ denote the distribution of state transitions observed from the MRP $\mathcal{R}_j$ under the generative model. Also note that for each $i \in [D/2]$, we have

$$\|\sigma(\theta_i^*)\|_\infty = \nu \frac{(1-\tau)\sqrt{p_1(1-p_1)}}{(1-\gamma p_1)}. \tag{8.33}$$

Lower bounding the minimax risk over this class: We again use the local packing form of Fano's method (§15.3.3, [323]) to establish a lower bound. Choose some index J uniformly at random from the set $[D/2]$, and suppose that we draw N i.i.d. samples $Y^N := (Y_1, \dots, Y_N)$ from the MRP $\mathcal{R}_J$ under the generative model. Here each $Y_i \in \mathcal{X}^D$ represents a random set of D states, and the goal of the estimator is to identify the random index J and, consequently, to estimate the value function θ_J^*. Let us now lower bound the expected error incurred in this $(D/2)$-ary hypothesis testing problem. Fano's inequality yields the bound

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \bar{\mathcal{M}}} \mathbb{E}\|\widehat{\theta} - \theta^*\|_\infty \geq \frac{1}{2} \min_{j \neq k} \|\theta_j^* - \theta_k^*\|_\infty \left(1 - \frac{I(J; Y^N) + \log 2}{\log(D/2)}\right), \tag{8.34}$$

where $I(J; Y^N)$ denotes the mutual information between J and Y^N.

Let us now bound the two terms that appear in inequality (8.34). By equation (8.31), we have

$$\|\theta_j^* - \theta_k^*\|_\infty = \nu \cdot \frac{(p_1 - p_2)(1-\tau)}{(1-\gamma p_1)(1-\gamma p_2)} \qquad \text{for all } 1 \leq j \neq k \leq D/2.$$

Furthermore, since the samples $Y_1, \dots, Y_N$ are i.i.d., the chain rule of mutual information yields

$$\begin{aligned} \frac{1}{N} I(J; Y^N) = I(J; Y_1) &\leq \max_{j \neq k} \mathsf{KL}(\mathbb{P}_j \| \mathbb{P}_k) \\ &\overset{\text{(i)}}{=} \mathsf{KL}(\mathbb{P}(p_1) \| \mathbb{P}(p_2)) + \mathsf{KL}(\mathbb{P}(p_2) \| \mathbb{P}(p_1)) \\ &\overset{\text{(ii)}}{\leq} 2 \frac{(p_1 - p_2)^2}{p_1(1-p_1)}, \end{aligned}$$

where step (i) is a consequence of the construction, which ensures that the distributions $\mathbb{P}_j$ and $\mathbb{P}_k$ coincide on all but the j-th and k-th sub-MRPs. On the other hand, step (ii) follows from Lemma 8.4.4, and the fact that $p_2 \leq p_1$.

Putting together the pieces, we now have

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \bar{\mathcal{M}}} \mathbb{E}\|\widehat{\theta} - \theta^*\|_\infty \geq \frac{\nu}{2} \cdot \frac{(p_1 - p_2)(1-\tau)}{(1-\gamma p_1)(1-\gamma p_2)} \left(1 - \frac{2N \frac{(p_1 - p_2)^2}{p_1(1-p_1)} + \log 2}{\log(D/2)}\right).$$

Recall the choice $p_1 - p_2 = \frac{1}{8}\sqrt{\frac{p_1(1-p_1)}{N} \log(D/2)}$. For $D \geq 8$, this ensures, for a small enough positive constant c, the bound

$$\inf_{\widehat{\theta}} \sup_{\theta^* \in \bar{\mathcal{M}}} \mathbb{E}\|\widehat{\theta} - \theta^*\|_\infty \geq c\nu \frac{(1-\tau)\sqrt{p_1(1-p_1)}}{(1-\gamma p_1)} \cdot \sqrt{\frac{\log(D/2)}{N}} \frac{1}{1-\gamma p_2}. \tag{8.35}$$

With the relation (8.35) at hand, we now turn to proving the two sub-claims in equation (8.29).

Proof of claim (8.29a)**:** Recall equation (8.33); for $i \in [D/2]$, we have

$$\|\sigma(\theta_i^*)\|_\infty = \nu \frac{(1-\tau)\sqrt{p_1(1-p_1)}}{(1-\gamma p_1)} \quad \text{and} \quad \|\theta_i^*\|_{\text{span}} = \nu \frac{(1-\tau)}{(1-\gamma p_1)}.$$

Now for every pair of scalars (ϑ, ζ) satisfying $\vartheta = \zeta\sqrt{1-\gamma}$, set $\tau = 1/2$ and $\nu = 2\zeta(1-\gamma p_1)$. With this choice of parameters, we have the inclusion $\mathcal{M}_{\mathsf{var}}(\vartheta, 0) \cap \mathcal{M}_{\mathsf{vfun}}(\zeta, 0) \subseteq \bar{\mathcal{M}}$, and evaluating the bound (8.35) yields

$$\begin{aligned} \inf_{\widehat{\theta}} \sup_{\theta^* \in \mathcal{M}_{\mathsf{var}}(\vartheta,0)\cap\mathcal{M}_{\mathsf{vfun}}(\zeta,0)} \mathbb{E}\|\widehat{\theta} - \theta^*\|_\infty &\geq c\vartheta\sqrt{\frac{\log(D/2)}{N}}\frac{1}{1-\gamma p_2} \\ &\overset{\text{(ii)}}{=} c\vartheta\sqrt{\frac{\log(D/2)}{N}}\frac{1}{1-\gamma}, \end{aligned}$$

where in step (ii), we have used inequality (8.32). The same lower bound clearly also extends to the set $\mathcal{M}_{\mathsf{var}}(\vartheta, 0) \cap \mathcal{M}_{\mathsf{vfun}}(\zeta, 0)$ for $\zeta \geq \vartheta(1-\gamma)^{-1/2}$; this establishes part (a) of the theorem.

Proof of claim (8.29b)**:** Given a value $r_{\max}$, set $\tau = 0$ and $\nu = r_{\max}$ and note that the rewards of all the MRPs in the set $\bar{\mathcal{M}}$ satisfy $\|r\|_\infty \leq \nu$. Hence, we have $\mathcal{M}_{\mathsf{rew}}(r_{\max}, 0) \subseteq \bar{\mathcal{M}}$ for this choice of parameters. Using inequality (8.35) and recalling the bounds (8.32) once again, we have

$$\begin{aligned} \inf_{\widehat{\theta}} \sup_{\theta^* \in \mathcal{M}_{\mathsf{rew}}(r_{\max},0)} \mathbb{E}\|\widehat{\theta} - \theta^*\|_\infty &\geq c r_{\max}\frac{\sqrt{p_1(1-p_1)}}{(1-\gamma p_1)} \cdot \sqrt{\frac{\log(D/2)}{N}}\frac{1}{1-\gamma p_2} \\ &\geq c\frac{r_{\max}}{(1-\gamma)^{3/2}}\sqrt{\frac{\log(D/2)}{N}}. \end{aligned}$$

Proof of Lemma 8.4.4

By construction, the second state of the Markov chain is absorbing, so it suffices to consider the KL divergence between the first components of the distributions $\mathbb{P}(p)$ and $\mathbb{P}(q)$. These are Bernoulli random variables Ber(p) and Ber(q), and the following calculation bounds their KL divergence:

$$\begin{aligned} \mathsf{KL}\left(\text{Ber}(p)\|\text{Ber}(q)\right) &= p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q} \\ &\overset{\text{(ii)}}{\leq} p \cdot \frac{p-q}{q} + (1-p)\cdot\frac{q-p}{1-q} \\ &= \frac{(p-q)^2}{q(1-q)}, \end{aligned}$$

where step (ii) uses the inequality $\log(1+x) \leq x$, which is valid for all $x > -1$. A similar inequality holds with the roles of p and q reversed, and the denominator of the expression is lower for the larger value $p \vee q$. This completes the proof.

8.4.5 Proof of Theorem 8.2.2

Note that the median-of-means operator is applied elementwise; denote the i-th such operator by $\widehat{\mathcal{M}}_i$. Let $\widehat{\mathcal{M}} - \mathbf{P}$ denote the elementwise difference of operator $\widehat{\mathcal{M}}$ and the linear operator $\mathbf{P}$; its i-th component is given by the operator $\widehat{\mathcal{M}}_i(\cdot) - \langle p_i, \cdot \rangle$.

We require two technical lemmas in the proof. The power of the median-of-means device is clarified by the first lemma, which is an adaptation of classical results (see, e.g., [153, 234]).

Lemma 8.4.5. *Suppose that $K = 8\log(4D/\delta)$ and $m = \lfloor N/K \rfloor$. Then there is a universal constant c such that for each index $i \in [D]$ and each fixed vector $\theta \in \mathbb{R}^D$, we have*

$$\Pr\left\{ |(\widehat{\mathcal{M}}_i - p_i)(\theta)| \geq c\, \sigma_i(\theta) \sqrt{\frac{\log(8D/\delta)}{N}} \right\} \leq \frac{\delta}{4D}.$$

Comparing this lemma to the Bernstein bound (cf. equation (8.12b)), we see that we no longer pay in the span semi-norm $\|\theta^*\|_{\mathrm{span}}$, and this is what enables us to establish the solely variance-dependent bound (8.6).

We also require the following lemma that guarantees that the median-of-means Bellman operator is contractive.

Lemma 8.4.6. *The median-of-means operator is 1-Lipschitz in the ℓ_∞-norm, and satisfies*

$$|\widehat{\mathcal{M}}(\theta_1) - \widehat{\mathcal{M}}(\theta_2)| \leq \|\theta_1 - \theta_2\|_\infty \qquad \textit{for all vectors } \theta_1, \theta_2 \in \mathbb{R}^D.$$

Consequently, the empirical operator $\widehat{\mathcal{T}}_N^{\mathsf{MoM}}$ is γ-contractive in ℓ_∞-norm and satisfies

$$|\widehat{\mathcal{T}}_N^{\mathsf{MoM}}(\theta_1) - \widehat{\mathcal{T}}_N^{\mathsf{MoM}}(\theta_2)| \leq \gamma \|\theta_1 - \theta_2\|_\infty \qquad \textit{for all pairs of value functions } (\theta_1, \theta_2).$$

See Section 8.4.5 for the proof of Lemma 8.4.6.

We are now in a position to establish the theorem, where we now use the shorthand $\widehat{\theta} \equiv \widehat{\theta}_{\mathsf{MoM}}$ for convenience. Note that the vectors θ^* and $\widehat{\theta}$ satisfy the fixed point relations

$$\theta^* = r + \gamma \mathbf{P} \theta^*, \quad \text{and} \quad \widehat{\theta} = \widehat{r} + \gamma \widehat{\mathcal{M}}(\widehat{\theta}),$$

respectively. Taking differences, the error vector $\widehat{\Delta} = \widehat{\theta} - \theta^*$ satisfies the relation

$$\begin{aligned} \widehat{\theta} - \theta^* &= \gamma(\widehat{\mathcal{M}}(\widehat{\Delta} + \theta^*) - \mathbf{P}\theta^*) + \widehat{r} - r \\ &= \gamma(\widehat{\mathcal{M}}(\widehat{\Delta} + \theta^*) - \widehat{\mathcal{M}}(\theta^*)) + \gamma(\widehat{\mathcal{M}} - \mathbf{P})(\theta^*) + (\widehat{r} - r). \end{aligned}$$

Taking ℓ_∞-norms on both sides and using the triangle inequality, we have

$$\begin{aligned} \|\widehat{\Delta}\|_\infty &\leq \gamma \|\widehat{\mathcal{M}}(\theta^* + \widehat{\Delta}) - \widehat{\mathcal{M}}(\theta^*)\|_\infty + \gamma |(\widehat{\mathcal{M}} - \mathbf{P})(\theta^*)| + \|\widehat{r} - r\|_\infty \\ &\overset{\text{(i)}}{\leq} \gamma \|\widehat{\Delta}\|_\infty + \gamma |(\widehat{\mathcal{M}} - \mathbf{P})(\theta^*)| + \|\widehat{r} - r\|_\infty, \end{aligned}$$

where step (i) is a result of Lemma 8.4.6. Finally, applying Lemma 8.4.5 in conjunction with the Hoeffding inequality and a union bound over all D indices completes the proof.

Proof of Lemma 8.4.6

The second claim follows directly from the first by noting that

$$|\widehat{\mathcal{T}}_N^{\mathsf{MoM}}(\theta_1) - \widehat{\mathcal{T}}_N^{\mathsf{MoM}}(\theta_2)| = \gamma|\widehat{\mathcal{M}}(\theta_1) - \widehat{\mathcal{M}}(\theta_2)|.$$

In order to prove the first claim, recall that for each $\theta \in \mathbb{R}^D$, we have $\widehat{\mathcal{M}}(\theta) = \mathrm{med}(\widehat{\mu}_1(\theta), \ldots, \widehat{\mu}_K(\theta))$, where the median—defined as the $\lfloor K/2 \rfloor$-th order statistic—is taken entry-wise. By definition, for each $i \in [K]$, we have

$$\begin{aligned}\|\widehat{\mu}_i(\theta_1) - \widehat{\mu}_i(\theta_2)\|_\infty &= \left\| \left(\frac{1}{m} \sum_{k \in \mathcal{D}_i} \mathbf{Z}_k \right) (\theta_1 - \theta_2) \right\|_\infty \\ &\leq \left\| \frac{1}{m} \sum_{k \in \mathcal{D}_i} \mathbf{Z}_k \right\|_{1,\infty} \|\theta_1 - \theta_2\|_\infty \\ &\overset{\text{(i)}}{=} \|\theta_1 - \theta_2\|_\infty,\end{aligned}$$

where step (i) is a result of the fact that $\frac{1}{m} \sum_{k \in \mathcal{D}_i} \mathbf{Z}_k$ is a row stochastic matrix with non-negative entries. Finally, we have the entry-wise bound

$$\begin{aligned}|\widehat{\mathcal{M}}(\theta_1) - \widehat{\mathcal{M}}(\theta_2)| &= |\,\mathrm{med}(\widehat{\mu}_1(\theta_1), \ldots, \widehat{\mu}_K(\theta_1)) - \mathrm{med}(\widehat{\mu}_1(\theta_2), \ldots, \widehat{\mu}_K(\theta_2))| \\ &\overset{\text{(ii)}}{\preceq} \|\theta_1 - \theta_2\|_\infty \cdot \mathbf{1},\end{aligned}$$

where step (ii) follows from our definition of the median as the $\lfloor K/2 \rfloor$-th order statistic, and Lemma 8.4.7 to follow. This completes the proof of Lemma 8.4.6. □

Lemma 8.4.7. *For each pair of vectors (u, v) of dimension D and each index $i \in [D]$, we have*

$$|u_{(i)} - v_{(i)}| \leq \|u - v\|_\infty.$$

Proof. Assume without loss of generality that the entries of u are sorted in increasing order (so that $u_1 \leq u_2 \leq \ldots \leq u_D$), and let w denote a vector containing the entries of v sorted in increasing order. We then have

$$|u_{(i)} - v_{(i)}| = |u_i - w_i| \leq \|u - w\|_\infty \overset{\text{(i)}}{\leq} \|u - v\|_\infty,$$

where step (i) follows from the rearrangement inequality applied to the ℓ_∞-norm [320]. □

8.5 Summary and open questions

Our work investigates the local minimax complexity of value function estimation in Markov reward processes. Our upper bounds are instance-dependent, and we also provide minimax lower bounds that hold over natural subsets of the parameter space. The plug-in approach is shown to be optimal over the class of MRPs with bounded rewards, and a variant based on the median-of-means device achieves optimality over the class of MRPs having value functions with bounded variance.

Our results also leave a few interesting questions unresolved. Is Corollary 8.2.1(a) sharp, say up to a logarithmic factor in the dimension? Is the median-of-means approach minimax-optimal over the class of MRPs having bounded rewards? We conjecture that both of these questions can be answered in the affirmative. Are our results also sharp under alternative local minimax parameterizations (say in terms of the functional $\|(\mathbf{I}-\gamma\mathbf{P})^{-1}\sigma(\theta^*)\|_\infty$)? Is there a more fine-grained lower bound analysis that shows the (sub)-optimality of these approaches, and are there better adaptive procedures for this problem? There is also the related question of whether a minimax lower bound can be proved over a local neighborhood of every point θ^*. These questions are answered in the affirmative in Chapter 9 to follow.

In a complementary direction, another interesting question is to ask how function approximation affects these bounds. Our techniques should be useful in answering some of these questions, and also more broadly in proving analogous guarantees in the more challenging policy optimization setting.

There is the question of removing our assumption on the generative model: How does the plug-in estimator behave when it is computed on a sampled *trajectory* of the system? A classical solution is the blocking method of simulating the generative model from such samples [344]: given a sampled trajectory, chop it into pieces of length (roughly) equal to the mixing time of the Markov chain, and to treat the respective first sample from each of these pieces as (approximately) independent. But clearly, this approach is somewhat wasteful, and there have been recent refinements in related problems when the mixing time can become arbitrarily large [284]. It would be interesting to explore these approaches and derive instance-dependent guarantees in the L^2_μ-norm, where μ is the stationary distribution of the Markov chain.

It would also be interesting to analyze other policy evaluation algorithms with this local perspective. In Chapter 9 to follow, we provide non-asymptotic guarantees on the ℓ_∞-error for the popular family of temporal-difference learning algorithms that are based on stochastic approximation.

Chapter 9

Is stochastic approximation instance-optimal?

In this chapter, we study a class of stochastic approximation algorithms for policy evaluation, with a focus on developing instance-dependent bounds. Our results complement those of Chapter 8, in which we analyzed the plug-in estimator and its robust variant.

Related work

We begin with a broad overview of related work on stochastic approximation for policy evaluation, which goes by the same of *temporal difference* (TD) learning. For a broader discussion of related work on learning in Markov reward processes, see Chapter 8.

Asymptotic theory: The TD update was originally proposed by Sutton [293], and is typically used in conjunction with an appropriate parameterization of value functions; see [77] for a comprehensive survey. Classical results on the algorithm are typically asymptotic, and include both convergence guarantees [37] and examples of divergence [10]; see the seminal work [301] for conditions that guarantee asymptotic convergence.

It is worth noting that the TD algorithm is a form of linear stochastic approximation, and can be fruitfully combined with the iterate-averaging procedure put forth independently by Polyak [255] and Ruppert [270]. In this context, the work of Polyak and Juditsky [255] deserves special mention, since it shows that under fairly mild conditions, the TD algorithm converges when combined with Polyak-Ruppert iterate averaging. To be clear, in the specific context of the policy evaluation problem, the results in the Polyak-Juditsky paper [255] allow noise only in the observations of rewards (i.e., the transition function is assumed to be known). However, the underlying techniques can be extended to derive results in the setting in which we only observe samples of transitions; for instance, see the work of Tadic [295] for results of this type.

Non-asymptotic theory: Recent years have witnessed significant interest in understanding TD-type algorithms from the non-asymptotic standpoint. Bhandari et al. [30] focus on proving ℓ_2-

guarantees for the TD algorithm when combined with Polyak-Ruppert iterate averaging. They consider both the generative model as well as the Markovian noise model, and provide non-asymptotic guarantees on the expected error. Their results also extend to analyses of the popular TD(λ) variant of the algorithm, as well as to Q-learning in specific MDP instances. Also noteworthy is the analysis of Lakshminarayanan and Szepesvari [183], carried out in parallel with Bhandari et al. [30]; it provides similar guarantees on the TD(0) algorithm with constant stepsize and averaging. Note that both of these analyses focus on ℓ_2-guarantees (equipped with an associated inner product), and thus can directly leverage proof techniques for stochastic optimization [9, 232].

Other related results[1] include those of Dalal et al. [72], Doan et al. [86], Korda and La [175], and also more contemporary papers [321, 337]. The latter three of these papers introduce a variance-reduced form of temporal difference learning, a variant of which we analyze in this paper.

Instance-dependent results: The focus on instance-dependent guarantees for TD algorithms is recent, and results are available both in the ℓ_2-norm setting [30, 72, 183, 337] and the ℓ_∞-norm settings [246] (see Chapter 8). In general, however, the guarantees provided by work to date are not sharp. For instance, the bounds in [72] scale exponentially in relevant parameters of the problem, whereas the papers [30, 183, 337] do not capture the correct "variance" of the problem instance at hand. In the previous chapter, we derived ℓ_∞ bounds on policy evaluation for the plug-in estimator. These results were shown to be locally minimax optimal in certain regions of the parameter space. There has also been some recent focus on obtaining instance-dependent guarantees in online reinforcement learning settings [209]. This has resulted in more practically useful algorithms that provide, for instance, horizon-independent regret bounds for certain episodic MDPs [154, 345]. Recent work has also established some instance-dependent bounds, albeit not sharp over the whole parameter space, for the problem of state-action value function estimation in Markov decision processes, for both ordinary Q-learning [324] and a variance-reduced improvement [325].

Contributions

In this paper, we study stochastic approximation algorithms for evaluating the value function of a Markov reward process in the discounted setting. Our goal is to provide a sharp characterization of performance in the ℓ_∞-norm, for procedures that are given access to state transitions and reward samples under the generative model. In practice, temporal difference learning is typically applied with an additional layer of (linear) function approximation. In the current paper, so as to bring the instance dependence into sharp focus, we study the algorithms without this function approximation step. In this context, we tell a story with three parts, as detailed below:

Local minimax lower bounds: Global minimax analysis provides bounds that hold uniformly over large classes of models. In this paper, we seek to gain a more refined understanding of how the difficulty of policy evaluation varies as a function of the instance. In order to do so, we undertake

[1] There were some errors in the results of Korda and La [175] that were pointed out by both Lakshminarayanan and Szepesvari [183] and Xu et al. [337].

an analysis of the local minimax risk associated with a problem. We first prove an asymptotic statement (Proposition 9.1.1) that characterizes the local minimax risk up to a logarithmic factor; it reveals the relevance of two functionals of the instance that we define. In proving this result, we make use of the classical asymptotic minimax theorem [135, 186, 187]. We then refine this analysis by deriving a *non-asymptotic* local minimax bound, as stated in Theorem 9.1.1, which is derived using the non-asymptotic local minimax framework of Cai and Low [48], an approach that builds upon the seminal concept of hardest local alternatives that can be traced back to Stein [290].

Non-asymptotic suboptimality of iterate averaging: Our local minimax lower bounds raise a natural question: Do standard procedures for policy evaluation achieve these instance-specific bounds? In Section 9.2, we address this question for the TD(0) algorithm with iterate averaging. Via a careful simulation study, we show that for many popular stepsize choices, the algorithm *fails* to achieve the correct instance-dependent rate in the non-asymptotic setting, even when the sample size is quite large. This is true for both the constant stepsize, as well as polynomial stepsizes of various orders. Notably, the algorithm with polynomial stepsizes of certain orders achieves the local risk in the asymptotic setting (see Theorem 1).

Non-asymptotic optimality of variance reduction: In order to remedy this issue with iterate averaging, we propose and analyze a variant of TD learning with variance reduction, showing both through theoretical (see Theorem 2) and numerical results (see Figure 9.2) that this algorithm achieves the correct instance-dependent rate provided the sample size is larger than an explicit threshold. Thus, this algorithm is provably better than TD(0) with iterate averaging.

Chapter-specific notation: Recall the notational convention introduced in Section 1.4. We complement this notation with a few other definitions that are used solely in this chapter and the corresponding technical proof section in Appendix C.3. Recall from Chapter 8 that we let $\|\mathbf{M}\|_{1,\infty}$ denote the maximum ℓ_1-norm of the rows of a matrix $\mathbf{M}$, and refer to it as the $(1, \infty)$-operator norm of a matrix. More generally, for scalars $q, p \geq 1$, we define $|\!|\!|\mathbf{M}|\!|\!|_{p,q} \overset{\mathsf{def}}{=} \sup_{\|x\|_p \leq 1} \|\mathbf{M}x\|_q$. We let $\mathbf{M}^\dagger$ denote the Moore-Penrose pseudoinverse of a matrix $\mathbf{M}$.

We turn to the statements of our main results and discussion of their consequences. All of our statements involve certain measures of the local complexity of a given problem, which we introduce first. We then turn to the statement of lower bounds on the ℓ_∞-norm error in policy evaluation. In Section 9.1, we prove two lower bounds. Our first result, stated as Proposition 9.1.1, is asymptotic in nature (holding as the sample size $N \to +\infty$). Our second lower bound, stated as Theorem 9.1.1, provides a result that holds for a range of finite sample sizes. Given these lower bounds, it is then natural to wonder about known algorithms that achieve them. Concretely, does the TD(0) algorithm combined with Polyak-Ruppert averaging achieve these instance-dependent bounds? In Section 9.2, we undertake a careful empirical study of this question, and show that in the non-asymptotic setting, this algorithm fails to match the instance-dependent bounds. This finding sets up the analysis in Section 9.3, where we introduce a variance-reduced version of TD(0), and prove that it does achieve the instance-dependent lower bounds from Theorem 9.1.1 up to a logarithmic factor in dimension.

9.1 Local minimax lower bound

Throughout this section, we use the letter $\mathcal{P}$ to denote an individual problem instance, $\mathcal{P} = (\mathbf{P}, r)$, and use $\theta(\mathcal{P}) := \theta^* = (\mathbf{I} - \gamma \mathbf{P})^{-1} r$ to denote the *target* of interest. The aim of this section is to establish *instance-specific* lower bounds for estimating $\theta(\mathcal{P})$ under the generative observation model. In order to do so, we adopt a local minimax approach.

The remainder of this the section is organized as follows. In Section 9.1, we prove an asymptotic local minimax lower bound, valid as the sample size N tends to infinity. It gives an explicit Gaussian limit for the rescaled error that can be achieved by any procedure. The asymptotic covariance in this limit law depends on the problem instance, and is very closely related to the functionals $\nu(\mathbf{P}, \theta^*)$ and $\rho(\mathbf{P}, r)$ that we have defined. Moreover, we show that this limit can be achieved—in the asymptotic limit—by the TD algorithm combined with Polyak-Ruppert averaging. While this provides a useful sanity check, in practice we implement estimators using a finite number of samples N, so it is important to obtain non-asymptotic lower bounds for a full understanding. With this motivation, Section 9.1 provides a new, *non-asymptotic* instance-specific lower bound for the policy evaluation problem. We show that the quantities $\nu(\mathbf{P}, \theta^*)$ and $\rho(\mathbf{P}, r)$ also cover the instance-specific complexity in the finite-sample setting. In proving this non-asymptotic lower bound, we build upon techniques in the statistical literature based on constructing hardest one-dimensional alternatives [34, 49, 87, 88, 290]. As we shall see in later sections, while the TD algorithm with averaging is instance-specific optimal in the asymptotic setting, it *fails* to achieve our non-asymptotic lower bound.

Asymptotic local minimax lower bound

Our first approach towards an instance-specific lower bound is an asymptotic one, based on classical local asymptotic minimax theory. For regular and parametric families, the Hájek–Le Cam local asymptotic minimax theorem [135, 186, 187] shows that the Fisher information—an instance-specific functional—characterizes a fundamental asymptotic limit. Our model class is both parametric and regular, and so this classical theory applies to yield an asymptotic local minimax bound. Some additional work is needed to relate this statement to the more transparent complexity measures $\nu(\mathbf{P}, \theta^*)$ and $\rho(\mathbf{P}, r)$ that we have defined.

In order to state our result, we require some additional notation. Fix an instance $\mathcal{P} = (\mathbf{P}, r)$. For any $\epsilon > 0$, we define an ϵ-neighborhood of problem instances by

$$\mathfrak{N}(\mathcal{P}; \epsilon) = \left\{\mathcal{P}' = (\mathbf{P}', r') : \|\mathbf{P} - \mathbf{P}'\|_F + \|r - r'\|_2 \leq \epsilon\right\}.$$

Adopting the ℓ_∞-norm as the loss function, the *local asymptotic minimax risk* is given by

$$\mathfrak{M}_\infty(\mathcal{P}) \equiv \mathfrak{M}_\infty(\mathcal{P}; \|\cdot\|_\infty) = \lim_{c \to \infty} \lim_{N \to \infty} \inf_{\hat{\theta}_N} \sup_{\mathcal{Q} \in \mathfrak{N}(\mathcal{P}; c/\sqrt{N})} \mathbb{E}_{\mathcal{Q}} \left[\sqrt{N} \left\|\hat{\theta}_N - \theta(\mathcal{Q})\right\|_\infty\right]. \tag{9.1}$$

Here the infimum is taken over all estimators $\widehat{\theta}_N$ that are measurable functions of N i.i.d. observations drawn according to the generative observation model.

Our first main result characterizes the local asymptotic risk $\mathfrak{M}_\infty(\mathcal{P})$ exactly, and shows that it is attained by stochastic approximation with averaging. Recall the Polyak-Ruppert (PR) sequence $\{\widetilde{\theta}_k\}_{k\geq 1}$ defined in Eq. (7.9), and let $\{\widetilde{\theta}_k^\omega\}_{k\geq 1}$ denote this sequence when the underlying SA algorithm is the TD update with the polynomial stepsize sequence (7.8b) with exponent ω.

Proposition 9.1.1. *Let $Z \in \mathbb{R}^D$ be a multivariate Gaussian with zero mean and covariance matrix*

$$(\mathbf{I} - \gamma\mathbf{P})^{-1}(\gamma^2\Sigma_{\mathbf{P}}(\theta(\mathcal{P})) + \sigma_r^2\mathbf{I})(\mathbf{I} - \gamma\mathbf{P})^{-\top}. \tag{9.2a}$$

Then the local asymptotic minimax risk at problem instance $\mathcal{P}$ is given by

$$\mathfrak{M}_\infty(\mathcal{P}) = \mathbb{E}[\|Z\|_\infty]. \tag{9.2b}$$

Furthermore, for each problem instance $\mathcal{P}$ and scalar $\omega \in (1/2, 1)$, this limit is achieved by the TD algorithm with an ω-polynomial stepsize and PR-averaging:

$$\lim_{N\to\infty} \sqrt{N} \cdot \mathbb{E}\left[\|\widetilde{\theta}_N^\omega - \theta(\mathcal{P})\|_\infty\right] = \mathbb{E}[\|Z\|_\infty]. \tag{9.2c}$$

With the convention that $\theta^* \equiv \theta(\mathcal{P})$, a short calculation bounding the maximum absolute value of sub-Gaussian random variables (see, e.g., Ex. 2.11 in Wainwright [323]) yields the sandwich relation

$$\gamma\nu(\mathbf{P}, \theta^*) + \rho(\mathbf{P}, r) \leq \mathbb{E}[\|Z\|_\infty] \leq \sqrt{2\log D} \cdot (\gamma\nu(\mathbf{P}, \theta^*) + \rho(\mathbf{P}, r)),$$

so that Proposition 9.1.1 shows that, up to a logarithmic factor in dimension D, the local asymptotic minimax risk is entirely characterized by the functional $\gamma\nu(\mathbf{P}, \theta^*) + \rho(\mathbf{P}, r)$.

It should be noted that lower bounds similar to Eq. (9.2b) have been shown for specific classes of stochastic approximation algorithms [303]. However, to the best of our knowledge, a local minimax lower bound—one applying to any procedure that is a measurable function of the observations—is not available in the existing literature.

Furthermore, Eq. (9.2c) shows that stochastic approximation with polynomial stepsizes and averaging attains the exact local asymptotic risk. Our proof of this result essentially mirrors that of Polyak and Juditsky [255], and amounts to verifying their assumptions under the policy evaluation setting. Given this result, it is natural to ask if averaging is optimal also in the non-asymptotic setting; answering this question is the focus of the next two sections of the paper.

Non-asymptotic local minimax lower bound

Proposition 9.1.1 provides an instance-specific lower bound on $\theta(\mathcal{P})$ that holds asymptotically. In order to obtain a non-asymptotic guarantee, we borrow ideas from the non-asymptotic framework introduced by Cai and Low [49] for nonparametric shape-constrained inference. Adapting their definition of local minimax risk to our problem setting, given the loss function $L(\theta-\theta^*) = \|\theta-\theta^*\|_\infty$, the (normalized) *local non-asymptotic minimax risk* for $\theta(\cdot)$ at instance $\mathcal{P} = (\mathbf{P}, r)$ is given by

$$\mathfrak{M}_N(\mathcal{P}) = \sup_{\mathcal{P}'} \inf_{\hat{\theta}_N} \max_{\mathcal{Q}\in\{\mathcal{P},\mathcal{P}'\}} \sqrt{N} \cdot \mathbb{E}_{\mathcal{Q}}\left[\|\widehat{\theta}_N - \theta(\mathcal{Q})\|_\infty\right]. \tag{9.3}$$

Here the infimum is taken over all estimators $\widehat{\theta}_N$ that are measurable functions of N i.i.d. observations drawn according to the generative observation model, and the normalization by $\sqrt{N}$ is for convenience. The definition (9.3) is motivated by the notion of the hardest one-dimensional alternative [312, Ch. 25]. Indeed, given an instance $\mathcal{P}$, the local non-asymptotic risk $\mathfrak{M}_N(\mathcal{P})$ first looks for the hardest alternative $\mathcal{P}'$ against $\mathcal{P}$ (which should be local around $\mathcal{P}$), then measures the worst-case risk over $\mathcal{P}$ and its (local) hardest alternative $\mathcal{P}'$.

With this definition in hand, we lower bound the local non-asymptotic minimax risk using the complexity measures $\nu(\mathbf{P}, \theta^*)$ and $\rho(\mathbf{P}, r)$ defined in Eq. (7.11):

Theorem 9.1.1. *There exists a universal constant $c > 0$ such that for any instance $\mathcal{P} = (\mathbf{P}, r)$, the local non-asymptotic minimax risk is lower bounded as*

$$\mathfrak{M}_N(\mathcal{P}) \geq c\Big(\gamma\nu(\mathbf{P}, \theta^*) + \rho(\mathbf{P}, r)\Big). \tag{9.4}$$

This bound is valid for all sample sizes N that satisfy

$$N \geq N_0 := \max\left\{\frac{\gamma^2}{(1-\gamma)^2}, \frac{b^2(\theta^*)}{\nu^2(\mathbf{P}, \theta^*)}\right\}. \tag{9.5}$$

A few comments are in order. First, it is natural to wonder about the necessity of condition (9.5) on the sample size N in our lower bound. Chapter 8 upper bounds on the ℓ_∞-error of the plugin estimator, and these results also require a bound of this type. In fact, when the rewards are observed with noise (i.e., for any $\sigma_r > 0$), the condition $N \gtrsim \frac{\gamma^2}{(1-\gamma)^2}$ is natural, since it is necessary in order to obtain an estimate of the value function with $\mathcal{O}(1)$ error. On the other hand, in the special case of deterministic rewards ($\sigma_r = 0$), it is interesting to ask how the fundamental limits of the problem behave in the absence of this condition.

Second, note that the non-asymptotic lower bound (9.4) is closely connected to the asymptotic local minimax bound from Proposition 9.1.1. In particular, for any sample size N satisfying the lower bound (9.5), our non-asymptotic lower bound (9.4) coincides with the asymptotic lower bound (9.2b) up to a constant factor. Thus, it cannot be substantially sharpened. The finite-sample nature of the lower bound (9.4) is a powerful tool for assessing optimality of procedures: it provides a performance benchmark that holds over a large range of finite sample sizes N. Indeed, in the next section, we study the performance of the TD learning algorithm with Polyak-Ruppert averaging. While this procedure achieves the local minimax lower bound asymptotically, as guaranteed by Eq. (9.2c) in Proposition 9.1.1, it falls short of doing so in natural *finite-sample* scenarios.

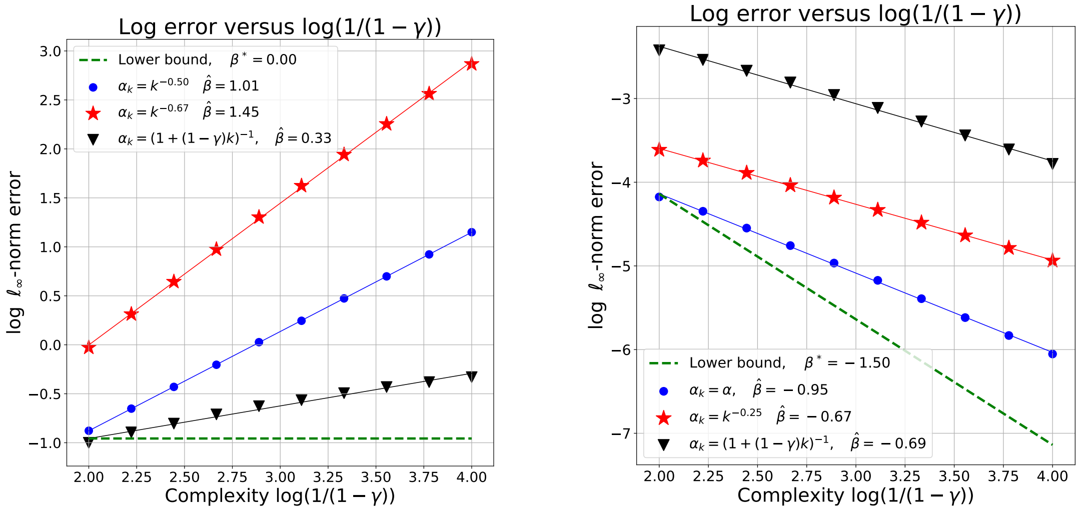
9.2 Suboptimality of averaging

Polyak and Juditsky [255] provide a general set of conditions under which a given stochastic-approximation (SA) algorithm, when combined with Polyak-Ruppert averaging, is guaranteed to have asymptotically optimal behavior. For the current problem, the bound (9.2c) in Proposition 9.1.1, which is proved using the Polyak-Juditsky framework, shows that SA with polynomial stepsizes and averaging have this favorable asymptotic property.

However, asymptotic theory of this type gives no guarantees in the finite-sample setting. In particular, suppose that we are given a sample size N that scales as $(1-\gamma)^{-2}$, as specified in our lower bounds. Does the averaged TD(0) algorithm exhibit optimal behavior in this non-asymptotic setting? In this section, we answer this question in the negative. More precisely, we use the parameterized family of Markov reward processes described in Figure 7.2 of Chapter 7, and provide careful simulations that reveal the suboptimality of TD without averaging.

A simulation study

In order to compare the behavior of averaged TD with the lower bound, we performed a series of experiments of the following type. For a fixed parameter λ in the range $[0, 1.5]$, we generated a range of MRPs with different values of the discount factor γ. For each value of the discount parameter γ, we consider the problem of estimating θ^* using a sample size N set to be one of two possible values: namely, $N \in \left\{ \lceil \frac{8}{(1-\gamma)^2} \rceil, \lceil \frac{8}{(1-\gamma)^3} \rceil \right\}$.



(a) $\lambda = 0.5, \quad N = \lceil \frac{8}{(1-\gamma)^2} \rceil$.

(b) $\lambda = 1.5, \quad N = \lceil \frac{8}{(1-\gamma)^3} \rceil$.

Figure 9.1: Log-log plots of the ℓ_∞-error versus the discount complexity parameter $1/(1-\gamma)$ for various algorithms. Each point represents an average over 1000 trials, with each trial simulations are for the 2-state MRP depicted in Figure 7.2 with the parameter choices $p = \frac{4\gamma-1}{3\gamma}$, $\nu = 1$ and $\tau = 1 - (1-\gamma)^\lambda$. We have also plotted the least-squares fits through these points, and the slopes of these lines are provided in the legend. In particular, the legend contains the stepsize choice for averaged SA (denoted as α_k), the slope $\hat{\beta}$ of the least-squares line, and the ideal value β^* of the slope computed in equation 9.6. We also include the lower bound predicted by Theorem 9.1.1 for these examples as a dotted line for comparison purposes. Logarithms are to the natural base.

In Figure 9.1, we plot the ℓ_∞-error of the averaged SA, for constant stepsize (7.8a), polynomial-decay stepsize (7.8b) and recentered linear stepsize (7.8c), as a function of γ. The plots show the behavior for $\lambda \in \{0.5, 1.5\}$. Each point on each curve is obtained by averaging 1000 Monte Carlo trials of the experiment. Using our lower bound calculations above in conjunction with the bound (7.13a), the log ℓ_∞-error is related to the complexity $\log\left(\frac{1}{1-\gamma}\right)$ in a linear fashion; we use β^* to denote the slope of this idealized line. Simple algebra yields

$$\beta^* = \frac{1}{2} - \lambda \quad \text{for} \quad N = \frac{1}{(1-\gamma)^2}, \quad \text{and} \quad \beta^* = -\lambda \qquad \text{for} \quad N = \frac{1}{(1-\gamma)^3}. \tag{9.6}$$

In other words, for an algorithm which achieves the lower bound predicted by our theory, we expect a linear relationship between the log ℓ_∞-error and log discount complexity $\log\left(\frac{1}{1-\gamma}\right)$, with the slope β^*.

Accordingly, for the averaged SA estimators with the stepsize choices in (7.8a)-(7.8c), we performed a linear regression to estimate the slopes between the log ℓ_∞-error and the log discount-complexity $\log\left(\frac{1}{1-\gamma}\right)$. The plot legend reports the stepsize choices α_k and the slope $\widehat{\beta}$ of the fitted regression line. We also include the lower bound in the plots, as a dotted line along with its slope, for a visual comparison. We see that the slopes corresponding to the averaged SA algorithm are higher compared to the ideal slopes of the dotted lines. Stated differently, this means that the averaged SA algorithm does not achieve the lower bound with either the constant step or the polynomial-decay step. Overall, the simulations provided in this section demonstrate that the averaged SA algorithm, although guaranteed to be asymptotically optimal by Eq. (9.2c) in Proposition 9.1.1, does not yield the ideal non-asymptotic behavior.

9.3 Variance-reduced policy evaluation

In this section, we propose and analyze a variance-reduced version of the TD learning algorithm. As in standard variance-reduction schemes, such as SVRG [156], our algorithm proceeds in epochs. In each epoch, we run a standard stochastic approximation scheme, but we recenter our updates in order to reduce their variance. The recentering uses an empirical approximation to the population Bellman operator $\mathcal{T}$.

We describe the behavior of the algorithm over epochs by a sequence of operators, $\{\mathcal{V}_m\}_{m\geq 1}$, which we define as follows. At epoch m, the method uses a vector $\overline{\theta}_m$ in order to recenter the update, where the vector $\overline{\theta}_m$ should be understood as the best current approximation to the unknown vector θ^*. In the ideal scenario, such a recentering would involve the quantity $\mathcal{T}(\overline{\theta}_m)$, where $\mathcal{T}$ denotes the population operator previously defined in Eq. (7.6). Since we lack direct access to the population operator $\mathcal{T}$, however, we use the Monte Carlo approximation

$$\widetilde{\mathcal{T}}_{N_m}(\overline{\theta}_m) := \frac{1}{N_m} \sum_{i \in \mathfrak{D}_m} \widehat{\mathcal{T}}_i(\overline{\theta}_m), \tag{9.7}$$

where the empirical operator $\widehat{\mathcal{T}}_i$ is defined in Eq. (7.5). Here the set $\mathfrak{D}_m$ is a collection of N_m i.i.d. samples, independent of all other randomness.

Given the pair $(\overline{\theta}_m, \widetilde{\mathcal{T}}_{N_m}(\overline{\theta}_m))$ and a stepsize $\alpha \in (0,1)$, we define the operator $\mathcal{V}_\ell$ on $\mathbb{R}^d$ as follows:

$$\theta \mapsto \mathcal{V}_k\left(\theta; \alpha, \overline{\theta}_m, \widetilde{\mathcal{T}}_{N_m}\right) := (1-\alpha)\theta + \alpha\left\{\widehat{\mathcal{T}}_k(\theta) - \widehat{\mathcal{T}}_k(\overline{\theta}_m) + \widetilde{\mathcal{T}}_{N_m}(\overline{\theta}_m)\right\}. \tag{9.8}$$

As defined in Eq. (7.5), the quantity $\widehat{\mathcal{T}}_\ell$ is a stochastic operator, where the randomness is independent of the set of samples $\mathfrak{D}_m$ used to define $\widetilde{\mathcal{T}}_{N_m}$. Consequently, the stochastic operator $\widehat{\mathcal{T}}_\ell$ is independent of the recentering vector $\widetilde{\mathcal{T}}_{N_m}(\overline{\theta}_m)$. Moreover, by construction, for each $\theta \in \mathbb{R}^D$, we have

$$\mathbb{E}\left[\widehat{\mathcal{T}}_k(\theta) - \widehat{\mathcal{T}}_k(\overline{\theta}_m) + \widetilde{\mathcal{T}}_{N_m}(\overline{\theta}_m)\right] = \mathcal{T}(\theta).$$

Thus, we see that $\mathcal{V}_k$ can be seen as an unbiased stochastic approximation of the population-level Bellman operator. As will be clarified in the analysis, the key effect of the recentering steps is to reduce its associated variance.

A single epoch

Based on the variance-reduced policy evaluation update defined in Eq. (9.8), we are now ready to define a single epoch of the overall algorithm. We index epochs using the integers $m = 1, 2, \ldots, M$, where M corresponds to the total number of epochs to be run. Epoch m requires as inputs the following quantities:

- a vector $\overline{\theta}$, which is chosen to be the output of the previous epoch,
- a positive integer K denoting the number of steps within the given epoch,
- a positive integer N_m denoting the number of samples used to calculate the Monte Carlo update (9.7),
- a sequence of stepsizes $\{\alpha_k\}_{k\geq 1}^K$ with $\alpha_k \in (0,1)$, and
- a set of fresh samples $\{\widehat{\mathcal{T}}_i\}_{i\in\mathfrak{E}_m}$, with $|\mathfrak{E}_m| = N_m + K$. The first N_m samples are used to define the dataset $\mathfrak{D}_m$ that underlies the Monte Carlo update (9.7), whereas the remaining K samples are used in the K steps within each epoch.

We summarize the operations within a single epoch in Algorithm 5.

The choice of the stepsize sequence $\{\alpha_k\}_{k\geq 1}$ is crucial, and it also determines the epoch length K. Roughly speaking, it is sufficient to choose a large enough epoch length to ensure that the error is reduced by a constant factor in each epoch. In Section 9.3 to follow, we study three popular stepsize choices—the constant stepsize (7.8a), the polynomial stepsize (7.8b) and the recentered linear stepsize (7.8c)—and provide lower bounds on the requisite epoch length in each case.

Algorithm 5: RunEpoch $(\overline{\theta}; K, N_m, \{\alpha_k\}_{k=1}^{K}, \{\widehat{\mathcal{T}}_i\}_{i\in\mathfrak{E}_m})$

```
Given (a) Epoch length $K$ , (b) Recentering vector $\overline{\theta}$ , (c) Recentering sample size $N_m$,
    (d) Stepsize sequence $\{\alpha_k\}_{k\geq 1}^{K}$, (e) Samples $\{\widehat{\mathcal{T}}_i\}_{i\in\mathfrak{E}_m}$
Compute the recentering quantity $\widetilde{\mathcal{T}}_{N_m}(\overline{\theta}) := \frac{1}{N_m} \sum_{i\in\mathfrak{D}_m} \widehat{\mathcal{T}}_i(\overline{\theta})$
Initialize $\theta_1 = \overline{\theta}$
for $k = 1, 2, \ldots, K$ do
    Compute the variance-reduced update:
        $$\theta_{k+1} = \mathcal{V}_k\left(\theta_k; \alpha_k, \overline{\theta}, \widetilde{\mathcal{T}}_{N_m}\right)$$
end for
```

Overall algorithm

We are now ready to specify our variance-reduced policy-evaluation (VRPE) algorithm. The overall algorithm has five inputs: (a) an integer M, denoting the number of epochs to be run, (b) an integer K, denoting the length of each epoch, (c) a sequence of sample sizes $\{N_m\}_{m=1}^{M}$ denoting the number of samples used for recentering, (d) Sample batches $\{\{\widehat{\mathcal{T}}_i\}_{i\in\mathfrak{E}_m}\}_{m=1}^{M}$ to be used in m epochs, and (e) a sequence of stepsize $\{\alpha_k\}_{k\geq 1}$ to be used in each epoch. Given these five inputs, we summarize the overall procedure in Algorithm 6:

Algorithm 6: Variance-reduced policy evaluation (VRPE)

```
Given (a) Number of epochs $M$, (b) Epoch length $K$ , (c) Recentering sample sizes
    $\{N_m\}_{m=1}^{M}$, (d) Sample batches $\{\widehat{\mathcal{T}}_i\}_{i\in\mathfrak{E}_m}$, for $m = 1, \ldots, M$, (e) Stepsize $\{\alpha_k\}_{k=1}^{K}$
Initialize at $\overline{\theta}_1$
for $m = 1, 2, \ldots, M$ do
    $\overline{\theta}_{m+1} = \mathrm{RunEpoch}\left(\overline{\theta}_m; K, N_m, \{\alpha\}_{k=1}^{K}, \{\widehat{\mathcal{T}}_i\}_{i\in\mathfrak{E}_m}\right)$
end for
Return $\overline{\theta}_{M+1}$ as the final estimate
```

In the next section, we provide a detailed description on how to choose these input parameters for three popular stepsize choices (7.8a)–(7.8c). Finally, we reiterate that at epoch m, the algorithm uses $N_m + K$ new samples, and the samples used in the epochs are independent of each other. Accordingly, the total number of samples used in M epochs is given by $KM + \sum_{m=1}^{M} N_m$.

Instance-dependent guarantees

Given a desired failure probability, $\delta \in (0, 1)$, and a total sample size N, we specify the following choices of parameters in Algorithm 6:

$$\texttt{Number of epochs :} \quad M := \log_2 \left(\frac{N(1-\gamma)^2}{8 \log((8D/\delta) \cdot \log N)} \right) \tag{9.9a}$$

$$\texttt{Recentering sample sizes :} \quad N_m := 2^m \frac{4^2 \cdot 9^2 \cdot \log(8MD/\delta)}{(1-\gamma)^2} \qquad \text{for } m = 1, \ldots, M \tag{9.9b}$$

$$\texttt{Sample batches:} \quad \text{Partition the } N \text{ samples to obtain } \{\widehat{\mathcal{T}}_i\}_{i \in \mathfrak{C}_m} \text{ for } m = 1, \ldots M \tag{9.9c}$$

$$\texttt{Epoch length:} \quad K = \frac{N}{2M} \tag{9.9d}$$

In the following theorem statement, we use (c_1, c_2, c_3, c_4) to denote universal constants.

Theorem 9.3.1. *(a) Suppose that the input parameters of Algorithm 6 are chosen according to Eq.* (9.9)*. Furthermore, suppose that the sample size N satisfies one of the following three stepsize-dependent lower bounds:*

(a) $\frac{N}{M} \geq c_1 \frac{\log(8ND/\delta)}{(1-\gamma)^3}$ *for recentered linear stepsize* $\alpha_k = \frac{1}{1+(1-\gamma)k}$,

(b) $\frac{N}{M} \geq c_2 \log(8ND/\delta) \cdot \left(\frac{1}{1-\gamma}\right)^{\left(\frac{1}{1-\omega} \vee \frac{2}{\omega}\right)}$ *for polynomial stepsize* $\alpha_k = \frac{1}{k^\omega}$ *with* $0 < \omega < 1$,

(c) $\frac{N}{M} \geq \frac{c_3}{\log\left(\frac{1}{1-\alpha(1-\gamma)}\right)}$ *for constant stepsize* $\alpha_k = \alpha \leq \frac{1}{5^2 \cdot 32^2} \cdot \frac{(1-\gamma)^2}{\log(8ND/\delta)}$.

Then for any initialization $\overline{\theta}_1$, the output $\overline{\theta}_{M+1}$ satisfies

$$\begin{aligned} \|\overline{\theta}_{M+1} - \theta^*\|_\infty &\leq c_4 \cdot \left\|\overline{\theta}_1 - \theta^*\right\|_\infty \cdot \frac{\log^2((8D/\delta) \cdot \log N)}{N^2(1-\gamma)^4} \\ &\quad + c_4 \cdot \left\{ \sqrt{\frac{\log(8DM/\delta)}{N}} \cdot \Big(\gamma \cdot \nu(\mathbf{P}, \theta^*) + \rho(\mathbf{P}, r)\Big) + \frac{\log(8DM/\delta)}{N} \cdot b(\theta^*) \right\}, \end{aligned} \tag{9.10}$$

with probability exceeding $1 - \delta$.

See Section 9.4.3 for the proof of this theorem.

A few comments on the upper bound provided in Theorem 9.3.1 are in order. In order to facilitate a transparent discussion in this section, we use the notation $\gtrsim$ in order to denote a relation that holds up to logarithmic factors in the tuple $(N, D, (1-\gamma)^{-1})$.

Initialization dependence: The first term on the right-hand side of the upper bound (9.10) depends on the initialization $\overline{\theta}_1$. It should be noted that when viewed as a function of the sample size N, this initialization-dependent term decays at a faster rate compared to the other two terms. This indicates that the performance of Algorithm 6 does not depend on the initialization $\overline{\theta}_1$ in a significant way. A careful look at the proof (cf. Section 9.4.3) reveals that the coefficient of $\|\overline{\theta}_1 - \theta^*\|_\infty$ in the bound (9.10) can be made significantly smaller. In particular, for any $p \geq 1$ the first term in the right-hand side of bound (9.10) can be replaced by

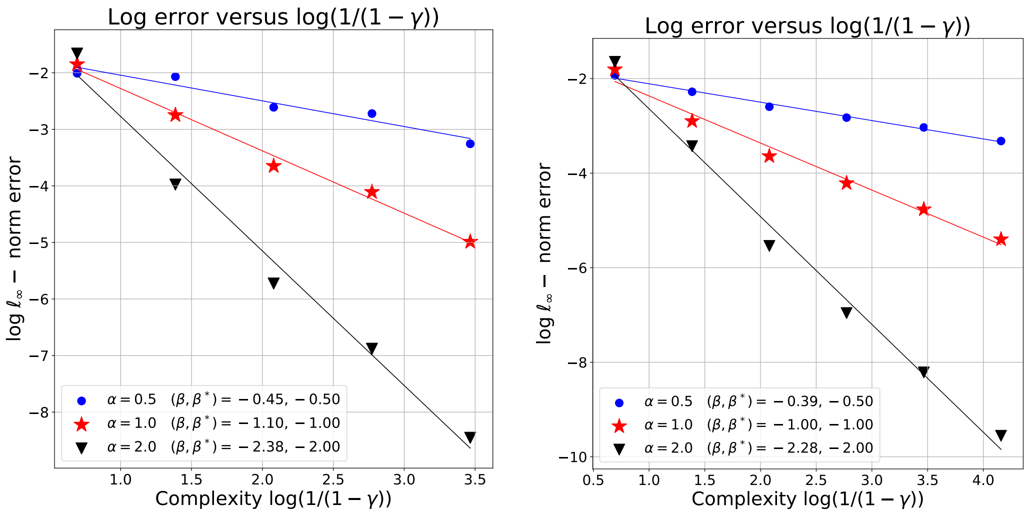
$$c_4 \cdot \frac{\|\overline{\theta}_1 - \theta^*\|_\infty}{N^p} \cdot \frac{\log^p((8D/\delta) \cdot \log N)}{(1-\gamma)^{2p}},$$

by increasing the recentering sample size (9.9b) by a constant factor and changing the values of the absolute constants (c_1, c_2, c_3, c_4), with these values depending only on the value of p. We have stated and proved a version for $p = 2$. Assuming the number of samples N satisfies $N \geq (1-\gamma)^{-(2+\Delta)}$ for some $\Delta > 0$, the first term on the right-hand side of bound (9.10) can always be made smaller than the other two terms. In the sequel we show that each of the lower bound conditions (a)-(c) in the statement of Theorem 9.3.1 requires a lower bound condition $N \gtrsim (1-\gamma)^{-3}$.

Comparing the upper and lower bounds: The second and the third terms in (9.10) show the instance-dependent nature of the upper bound, and they are the dominating terms. Furthermore, assuming that the minimum sample size requirements from Theorems 9.1.1 and 9.3.1 are met, we find that the upper bound (9.10) matches the lower bound (9.4) up to logarithmic terms.

It is worthwhile to explicitly compute the minimum sample size requirements in Theorems 9.1.1 and 9.3.1. Ignoring the logarithmic terms and constant factors for the moment, unwrapping the lower bound conditions (a)-(c) in Theorem 9.3.1, we see that for both the constant stepsize and the recentered linear stepsize the sample size needs to satisfy $N \gtrsim (1-\gamma)^{-3}$. For the polynomial stepsize $\alpha_k = \frac{1}{k^\omega}$, the sample size has to be at least $(1-\gamma)^{-\left(\frac{1}{1-\omega} \vee \frac{2}{\omega}\right)}$. Minimizing the last bound for different values of $\omega \in (0, 1)$, we see that the minimum value is attained at $\omega = 2/3$, and in that case the bound (9.10) is valid when $N \gtrsim (1-\gamma)^{-3}$. Overall, for all the three stepsize choices discussed in Theorem 9.3.1 we require $N \gtrsim (1-\gamma)^{-3}$ in order to certify the upper bound. Returning to Theorem 9.1.1, from assumption (9.5) we see that in the best case scenario, Theorem 9.1.1 is valid as soon as $N \gtrsim (1-\gamma)^{-2}$. Putting together the pieces we find that the sample size requirement for Theorem 9.3.1 is more stringent than that of Theorem 9.1.1. Currently we do not know whether the minimum sample size requirements in Theorems 9.1.1 and 9.3.1 are necessary; answering this question is an interesting future research direction.

Simulation study: It is interesting to demonstrate the sharpness of our bounds via a simulation study, using the same scheme as our previous study of TD(0) with averaging. In Figure 9.2, we report the results of this study; see the figure caption for further details. At a high level, we see that the VRPE algorithm, with either the recentered linear stepsize (panel (a)) or the polynomial stepsize $t^{-2/3}$, produces errors that decay with the exponents predicted by our instance-dependent theory for $\lambda \in \{0.5, 1.0, 2.0\}$. See the figure caption for further details.



(a) Recentered-linear stepsize (b) Polynomial stepsize

Figure 9.2: Log-log plots of the ℓ_∞-error versus the discount complexity parameter $1/(1-\gamma)$ for the VRPE algorithm. Each point is computed from an average over 1000 trials. Each trial entails drawing $N = \lceil \frac{8}{(1-\gamma)^3} \rceil$ samples from the 2-state MRP in Figure 7.2 with the parameter choices $p = \frac{4\gamma - 1}{3\gamma}$, $\nu = 1$ and $\tau = 1 - (1-\gamma)^\lambda$. Each line on each plot represents a different value of λ, as labeled in the legend. We have also plotted the least-squares fits through these points, and the slopes of these lines are also provided in the legend. We also report the pair $(\widehat{\beta}, \beta^*)$, where the coefficient $\widehat{\beta}$ denotes the slope of the least-squares fit and β^* denotes the slope predicted from the lower bound calculation (9.6). (a) Performance of VRPE for the recentered linear stepsize (7.8c). (b) Performance of VRPRE with polynomially decaying stepsizes (7.8b) with $\omega = 2/3$.

9.4 Proofs of main results

We now turn to the proofs of our main results.

9.4.1 Proof of Proposition 9.1.1

Recall the definition of the matrix $\Sigma_{\mathbf{P}}(\theta)$ from Eq. (7.10), and define the covariance matrix

$$V_{\mathcal{P}} = (\mathbf{I} - \gamma \mathbf{P})^{-1} (\gamma^2 \Sigma_{\mathbf{P}}(\theta) + \sigma_r^2 \mathbf{I}) (\mathbf{I} - \gamma \mathbf{P})^{-T}. \tag{9.11}$$

Recall that we use Z to denote a multivariate Gaussian random vector $Z \sim \mathcal{N}(0, V_{\mathcal{P}})$, and that the sequence $\{\widetilde{\theta}_k^\omega\}_{k \geq 1}$ is generated by averaging the iterates of stochastic approximation with polynomial stepsizes (7.8b) with exponent ω. With this notation, the two claims of the theorem are:

$$\mathfrak{M}_\infty(\mathcal{P}) = \mathbb{E}[\|Z\|_\infty], \quad \text{and} \tag{9.12a}$$

$$\lim_{N \to \infty} \mathbb{E}\left[\sqrt{N} \cdot \|\widetilde{\theta}_N^\omega - \theta^*\|_\infty\right] = \mathbb{E}[\|Z\|_\infty]. \tag{9.12b}$$

We now prove each of these claims separately.

Proof of Eq. (9.12a)

For the reader's convenience, let us state a version of the Hájek–Le Cam local asymptotic minimax theorem:

Theorem 9.4.1. *Let $\{P_{\vartheta'}\}_{\vartheta' \in \Theta}$ be a family of parametric models, quadratically mean differentiable with Fisher information matrices $J_{\vartheta'}$. Fix some parameter $\vartheta \in \Theta$, and consider a function $\psi : \Theta \to \mathbb{R}^D$ that is differentiable at ϑ. Then for any quasi-convex loss $L : \mathbb{R}^D \to \mathbb{R}$, we have:*

$$\lim_{c \to \infty} \lim_{N \to \infty} \inf_{\hat{\vartheta}_N} \sup_{\substack{\vartheta' \\ \|\vartheta' - \vartheta\|_2 \leq c/\sqrt{N}}} \mathbb{E}_{\vartheta'} \Big[L\big(\sqrt{N} \cdot (\hat{\vartheta}_N - \vartheta')\big) \Big] = \mathbb{E}[L(Z)], \tag{9.13}$$

where the infimum is taken over all estimators $\hat{\vartheta}_N$ that are measurable functions of N i.i.d. data points drawn from P_ϑ, and the expectation is taken over a multivariate Gaussian $Z \sim \mathcal{N}(0, \nabla\psi(\vartheta)^T J_\vartheta^\dagger \nabla\psi(\vartheta))$.

Returning to the problem at hand, let $\vartheta = (\mathbf{P}, r)$ denote the unknown parameters of the model and let $\psi(\vartheta) = \theta(\mathcal{P}) = (\mathbf{I} - \gamma\mathbf{P})^{-1} r$ denote the target vector. A direct application of Theorem 9.4.1 shows that

$$\mathfrak{M}_\infty(\mathcal{P}) = \mathbb{E}[\|Z\|_\infty] \text{ where } Z = \mathcal{N}(0, \nabla\psi(\vartheta)^T J_\vartheta^\dagger \nabla\psi(\vartheta)), \tag{9.14}$$

where J_ϑ is the Fisher information at ϑ. The following result provides a more explicit form of the covariance of Z:

Lemma 9.4.1. *We have the identity*

$$\nabla\psi(\vartheta)^T J_\vartheta^\dagger \nabla\psi(\vartheta) = (\mathbf{I} - \gamma\mathbf{P})^{-1}(\gamma^2 \Sigma_{\mathbf{P}}(\theta) + \sigma_r^2 \mathbf{I})(\mathbf{I} - \gamma\mathbf{P})^{-T}. \tag{9.15}$$

Although the proof of this claim is relatively straightforward, it involves some lengthy and somewhat tedious calculations; we refer the reader to Appendix C.3.1 for the proof.

Given the result from Lemma 9.4.1, the claim (9.12a) follows by substituting the relation (9.15) into (9.14).

Proof of Eq. (9.12b)

The proof of this claim follows from the results of Polyak and Juditsky [255, Theorem 1], once their assumptions are verified for TD(0) with polynomial stepsizes. Recall that the TD iterates in Eq. (7.7) are given by the sequence $\{\theta_k\}_{k \geq 1}$, and that $\widetilde{\theta}_k^\omega$ denotes the k-th iterate generated by averaging.

For each $k \geq 1$, note the following equivalence between the notation of our paper and that of Polyak and Juditsky [255], or PJ for short:

$$x_k \equiv \theta_k, \qquad \gamma_k \equiv \alpha_k, \qquad \mathbf{A} \equiv \mathbf{I} - \gamma \mathbf{P}, \quad \text{and} \quad \xi_k = (R_k - r) + (\mathbf{Z}_k - \mathbf{P})\theta_k.$$

Let us now verify the various assumptions in the PJ paper. Assumption 2.1 in the PJ paper holds by definition, since the matrix $\mathbf{I} - \gamma \mathbf{P}$ is Hurwitz. Assumption 2.2 in the PJ paper is also satisfied by the polynomial stepsize sequence for any exponent $\omega \in (0, 1)$.

It remains to verify the assumptions that must be satisfied by the noise sequence $\{\xi_k\}_{k \geq 1}$. In order to do so, write the k-th such iterate as

$$\xi_k = (R_k - r) + (\mathbf{Z}_k - \mathbf{P})\theta^* + (\mathbf{Z}_k - \mathbf{P})(\theta_k - \theta^*).$$

Since $\mathbf{Z}_k$ is independent of the sequence $\{\theta_i\}_{i=1}^k$, it follows that the condition

$$\lim_{N \to \infty} \mathbb{E}\left[\|\theta_N - \theta^*\|_2^2\right] \tag{9.16}$$

suffices to guarantee that Assumptions 2.3–2.5 in the PJ paper are satisfied. We now claim that for each $\omega \in (1/2, 1]$, condition (9.16) is satisfied by the TD iterates. Taking this claim as given for the moment, note that applying Theorem 1 of Polyak and Juditsky [255] establishes claim (9.12b), for any exponent $\omega \in (1/2, 1)$.

It remains to establish condition (9.16). For any $\omega \in (1/2, 1]$, the sequence of stepsizes $\{\alpha_k\}_{k \geq 1}$ satisfies the conditions

$$\sum_{k=1}^{\infty} \alpha_k = \infty \quad \text{and} \quad \sum_{k=1}^{\infty} \alpha_k^2 < \infty.$$

Consequently, classical results due to Robbins and Munro [263, Theorem 2] guarantee ℓ^2-convergence of θ_N to θ^*.

9.4.2 Proof of Theorem 9.1.1

Throughout the proof, we use the notation $\mathcal{P} = (\mathbf{P}, r)$ and $\mathcal{P}' = (\mathbf{P}', r')$ to denote, respectively, the problem instance at hand and its alternative. Moreover, we use $\theta^* \equiv \theta(\mathcal{P})$ and $\theta(\mathcal{P}')$ to denote the associated target parameters for each of the two problems $\mathcal{P}$ and $\mathcal{P}'$. We use $\Delta_{\mathbf{P}} = \mathbf{P} - \mathbf{P}'$ and $\Delta_r = r - r'$ to denote the differences of the parameters. For probability distributions, we use P and P' to denote the marginal distribution of a single observation under $\mathcal{P}$ and $\mathcal{P}'$, and use P^N and $(P')^N$ to denote the distribution of N i.i.d observations drawn from P or P', respectively.

Proof structure

We introduce two special classes of alternatives of interest, denoted as $\mathcal{S}_1$ and $\mathcal{S}_2$ respectively:

$$\mathcal{S}_1 = \{\mathcal{P}' = (\mathbf{P}', r') \mid r' = r\}, \quad \text{and} \quad \mathcal{S}_2 = \{\mathcal{P}' = (\mathbf{P}', r') \mid \mathbf{P}' = \mathbf{P}\}.$$

In words, the class $\mathcal{S}_1$ consists of alternatives $\mathcal{P}'$ that have the same reward vector r as $\mathcal{P}$, but a different transition matrix $\mathbf{P}'$. Similarly, the class $\mathcal{S}_2$ consists of alternatives $\mathcal{P}'$ with the same transition matrix $\mathbf{P}$, but a different reward vector. By restricting the alternative $\mathcal{P}'$ within class $\mathcal{S}_1$ and $\mathcal{S}_2$, we can define *restricted versions* of the local minimax risk, namely

$$\mathfrak{M}_N(\mathcal{P};\mathcal{S}_1) \equiv \sup_{\mathcal{P}'\in\mathcal{S}_1} \inf_{\hat{\theta}_N} \max_{\mathcal{P}\in\{\mathcal{P},\mathcal{P}'\}} \mathbb{E}_{\mathcal{P}}\left[\sqrt{N}\cdot\left\|\hat{\theta}_N-\theta(\mathcal{P})\right\|_\infty\right], \quad \text{and} \tag{9.17a}$$

$$\mathfrak{M}_N(\mathcal{P};\mathcal{S}_2) \equiv \sup_{\mathcal{P}'\in\mathcal{S}_2} \inf_{\hat{\theta}_N} \max_{\mathcal{P}\in\{\mathcal{P},\mathcal{P}'\}} \mathbb{E}_{\mathcal{P}}\left[\sqrt{N}\cdot\left\|\hat{\theta}_N-\theta(\mathcal{P})\right\|_\infty\right]. \tag{9.17b}$$

The main part of the proof involves showing that there is a universal constant $c > 0$ such that the lower bounds

$$\mathfrak{M}_N(\mathcal{P};\mathcal{S}_1) \geq c\cdot\gamma\nu(\mathbf{P},\theta^*), \quad \text{and} \tag{9.18a}$$

$$\mathfrak{M}_N(\mathcal{P};\mathcal{S}_2) \geq c\cdot\rho(\mathbf{P},r) \tag{9.18b}$$

both hold (assuming that the sample size N is sufficiently large to satisfy the condition (9.5)). Since we have $\mathfrak{M}_N(\mathcal{P}) \geq \max\{\mathfrak{M}_N(\mathcal{P};\mathcal{S}_1),\mathfrak{M}_N(\mathcal{P};\mathcal{S}_2)\}$, these lower bounds in conjunction imply the claim Theorem 9.1.1. The next section shows how to prove these two bounds.

Proof of the lower bounds (9.18a) **and** (9.18b)**:**

Our first step is to lower bound the local minimax risk for each problem class in terms of a modulus of continuity between the Hellinger distance and the ℓ_∞-norm.

Lemma 9.4.2. *For each $\mathcal{S} \in \{\mathcal{S}_1,\mathcal{S}_2\}$, we have the lower bound $\mathfrak{M}_N(\mathcal{P};\mathcal{S}) \geq \frac{1}{8}\cdot\underline{\mathfrak{M}}_N(\mathcal{P};\mathcal{S})$, where we define*

$$\underline{\mathfrak{M}}_N(\mathcal{P};\mathcal{S}) \overset{\text{def}}{=} \sup_{\mathcal{P}'\in\mathcal{S}} \left\{\sqrt{N}\cdot\left\|\theta(\mathcal{P})-\theta(\mathcal{P}')\right\|_\infty \mid d_{\text{hel}}(P,P') \leq \frac{1}{2\sqrt{N}}\right\}. \tag{9.19}$$

The proof of Lemma 9.4.2 follows a relatively standard argument, one which reduces estimation to testing; see Appendix C.3.2 for details.

This lemma allows us to focus our remaining attention on lower bounding the quantity $\underline{\mathfrak{M}}_N(\mathcal{P};\mathcal{S})$. In order to do so, we need both a lower bound on the ℓ_∞-norm $\left\|\theta(\mathcal{P})-\theta(\mathcal{P}')\right\|_\infty$ and an upper bound on the Hellinger distance $d_{\text{hel}}(P,P')$. These two types of bounds are provided in the following two lemmas. We begin with lower bounds on the ℓ_∞-norm:

Lemma 9.4.3. *(a) For any $\mathcal{P}$ and for all $\mathcal{P}' \in \mathcal{S}_1$, we have*

$$\|\theta(\mathcal{P})-\theta(\mathcal{P}')\|_\infty \geq \left(1-\frac{\gamma}{1-\gamma}\|\Delta_{\mathbf{P}}\|_\infty\right)_+ \cdot \left\|\gamma(\mathbf{I}-\gamma\mathbf{P})^{-1}\Delta_{\mathbf{P}}\theta^*\right\|_\infty. \tag{9.20a}$$

(b) For any $\mathcal{P}$ and for all $\mathcal{P}' \in \mathcal{S}_2$, we have

$$\|\theta(\mathcal{P})-\theta(\mathcal{P}')\|_\infty \geq \left\|(\mathbf{I}-\gamma\mathbf{P})^{-1}\Delta_r\right\|_\infty. \tag{9.20b}$$

See Appendix C.3.2 for the proof of this claim.

Next, we require upper bounds on the Hellinger distance:

Lemma 9.4.4. *(a) For each $\mathcal{P}$ and for all $\mathcal{P}' \in \mathcal{S}_1$, we have*

$$d_{\mathrm{hel}}(P, P')^2 \leq \frac{1}{2}\sum_{i,j} \frac{\left((\Delta_{\mathbf{P}})_{i,j}\right)^2}{\mathbf{P}_{i,j}}. \tag{9.21a}$$

(b) For each $\mathcal{P}$ and for all $\mathcal{P}' \in \mathcal{S}_2$, we have

$$d_{\mathrm{hel}}(P, P')^2 \leq \frac{1}{2\sigma_r^2} \left\|r_1 - r_2\right\|_2^2. \tag{9.21b}$$

See Appendix C.3.2 for the proof of this upper bound.

Using Lemmas 9.4.3 and 9.4.4, we can derive two different lower bounds. First, we have the lower bound $\underline{\mathfrak{M}}_N(\mathcal{P}; \mathcal{S}_1) \geq \underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_1)$, where

$$\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_1) \equiv \sup_{\mathcal{P}' \in \mathcal{S}_1} \left\{ \sqrt{N} \cdot \left(1 - \frac{\gamma \left\|\Delta_{\mathbf{P}}\right\|_\infty}{1-\gamma}\right)_+ \cdot \left\|\gamma(\mathbf{I} - \gamma\mathbf{P})^{-1}\Delta_{\mathbf{P}}\theta^*\right\|_\infty \;\mid\; \sum_{i,j} \frac{\left((\Delta_{\mathbf{P}})_{i,j}\right)^2}{\mathbf{P}_{i,j}} \leq \frac{1}{2N} \right\}. \tag{9.22a}$$

Second, we have the lower bound $\underline{\mathfrak{M}}_N(\mathcal{P}; \mathcal{S}_2) \geq \underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_2)$, where

$$\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_2) \equiv \sup_{\mathcal{P}' \in \mathcal{S}_2} \left\{ \sqrt{N} \cdot \left\| \left(\mathbf{I} - \gamma\mathbf{P}\right)^{-1} \Delta_r \right\|_\infty \;\; \frac{1}{\sigma_r^2} \left\|r_1 - r_2\right\|_2 \leq \frac{1}{2N} \right\}. \tag{9.22b}$$

In order to complete the proofs of the two lower bounds (9.18a) and (9.18b), it suffices to show that

$$\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_2) \geq \frac{1}{\sqrt{2}} \cdot \rho(\mathbf{P}, r), \quad \text{and} \tag{9.23a}$$

$$\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_1) \geq \frac{1}{2\sqrt{2}} \cdot \gamma\nu(\mathbf{P}, \theta^*). \tag{9.23b}$$

Proof of the bound (9.23a)**:** This lower bound is easy to show—it follows from the definition:

$$\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_2) = \frac{\sigma_r}{\sqrt{2}} \left\|(\mathbf{I} - \gamma\mathbf{P})^{-1}\Delta_r\right\|_\infty = \frac{1}{\sqrt{2}}\rho(\mathbf{P}, r).$$

Proof of the bound (9.23b)**:** The proof of this claim is much more delicate. Our strategy is to construct a special "hard" alternative $\overline{\mathcal{P}} \in \mathcal{S}_1$, that leads to a good lower bound on $\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_1)$. Lemma 9.4.5 below is the main technical result that we require:

Lemma 9.4.5. *There exists some probability transition matrix* $\bar{\mathbf{P}}$ *with the following properties:*

(a) It satisfies the constraint $\sum_{i,j} \frac{\big((\bar{\mathbf{P}}-\mathbf{P})_{i,j}\big)^2}{\mathbf{P}_{i,j}} \leq \frac{1}{2N}$.

(b) It satisfies the inequalities

$$\big\|\bar{\mathbf{P}} - \mathbf{P}\big\|_\infty \leq \frac{1}{\sqrt{2N}}, \quad \textit{and} \quad \big\|\gamma(\mathbf{I} - \gamma\mathbf{P})^{-1}(\bar{\mathbf{P}} - \mathbf{P})\theta^*\big\|_\infty \geq \frac{\gamma}{\sqrt{2N}} \cdot \nu(\mathbf{P}, \theta^*).$$

See Appendix C.3.2 for the proof of this claim.

Given the matrix $\bar{\mathbf{P}}$ guaranteed by this lemma, we consider the "hard" problem $\overline{\mathcal{P}} := (\bar{\mathbf{P}}, r) \in \mathcal{S}_1$. From the definition of $\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_1)$ in Eq. (9.22a), we have that

$$\begin{aligned}\underline{\mathfrak{M}}'_N(\mathcal{P}; \mathcal{S}_1) &\geq \sqrt{N} \cdot \left(1 - \frac{\gamma}{1-\gamma}\big\|\mathbf{P} - \bar{\mathbf{P}}\big\|_\infty\right)_+ \cdot \big\|\gamma(\mathbf{I} - \gamma\bar{\mathbf{P}})^{-1}(\mathbf{P} - \bar{\mathbf{P}})\theta^*\big\|_\infty \\ &\geq \sqrt{N} \cdot \left(1 - \frac{\gamma}{1-\gamma} \cdot \frac{1}{\sqrt{2N}}\right)_+ \cdot \frac{\gamma}{\sqrt{2N}} \cdot \nu(\mathbf{P}, \theta^*) \geq \frac{1}{2\sqrt{2}} \cdot \gamma\nu(\mathbf{P}, \theta^*),\end{aligned}$$

where the last inequality follows by the assumed lower bound $N \geq \frac{4\gamma^2}{(1-\gamma)^2}$. This completes the proof of the lower bound (9.23b).

9.4.3 Proof of Theorem 9.3.1

This section is devoted to the proof of Theorem 9.3.1, which provides the achievability results for variance-reduced policy evaluation.

Proof of part (a):

We begin with a lemma that characterizes the progress of Algorithm 6 over epochs:

Lemma 9.4.6. *Under the assumptions of Theorem 9.3.1 (a), there is an absolute constant* c *such that for each epoch* $m = 1, \ldots, M$*, we have:*

$$\begin{aligned}\big\|\overline{\theta}_{m+1} - \theta^*\big\|_\infty &\leq \frac{\|\overline{\theta}_m - \theta^*\|_\infty}{4} \\ &\quad + c\left\{\sqrt{\frac{\log(8DM/\delta)}{N_m}}\Big(\gamma \cdot \nu(\mathbf{P}, \theta^*) + \rho(\mathbf{P}, r)\Big) + \frac{\log(8DM/\delta)}{N_m} \cdot b(\theta^*)\right\},\end{aligned} \tag{9.24}$$

with probability exceeding $1 - \frac{\delta}{M}$.

Taking this lemma as given for the moment, let us complete the proof. We use the shorthand

$$\tau_m := \sqrt{\frac{\log(8DM/\delta)}{N_m}}\Big(\gamma \cdot \nu(\mathbf{P}, \theta^*) + \rho(\mathbf{P}, r)\Big) \quad \text{and} \quad \eta_m := \frac{\log(8DM/\delta)}{N_m} \cdot b(\theta^*) \tag{9.25}$$

to ease notation, and note that $\frac{\tau_m}{\sqrt{2}} \leq \tau_{m+1}$ and $\frac{\eta_m}{2} \leq \eta_{m+1}$, for each $m \geq 1$. Using this notation and unwrapping the recursion relation from Lemma 9.4.6, we have

$$\begin{aligned} \left\|\bar{\theta}_{M+1} - \theta^*\right\|_\infty &\leq \frac{\left\|\bar{\theta}_M - \theta^*\right\|_\infty}{4} + c(\tau_M + \eta_M) \\ &\overset{(i)}{\leq} \frac{\left\|\bar{\theta}_{M-1} - \theta^*\right\|_\infty}{4^2} + \frac{c}{2}\left(\tau_M + \eta_M\right) + c(\tau_M + \eta_M) \\ &\overset{(ii)}{\leq} \frac{\left\|\bar{\theta}_1 - \theta^*\right\|_\infty}{4^M} + 2c(\tau_M + \eta_M). \end{aligned}$$

Here, step (i) follows by applying the one-step application of the recursion (9.24), and by using the upper bounds $\frac{\tau_m}{\sqrt{2}} \leq \tau_{m+1}$ and $\frac{\eta_m}{2} \leq \eta_{m+1}$. Step (ii) follows by repeated application of the recursion (9.24). The last inequality holds with probability at least $1 - \delta$ by a union bound over the M epochs.

It remains to express the quantities 4^M, τ_M and η_M—all of which are controlled by the recentering sample size N_M—in terms of the total number of available samples N. Towards this end, observe that the total number of samples used for recentering at M epochs is given by

$$\sum_{m=1}^{M} N_m \asymp 2^M \cdot \frac{\log(8MD/\delta)}{(1-\gamma)^2}.$$

Substituting the value of $M = \log_2\left(\frac{N(1-\gamma)^2}{8\log((8D/\delta)\cdot\log N)}\right)$ we have

$$c_1 N \leq N_M \asymp \sum_{m=1}^{M} N_m \leq \frac{N}{2},$$

where c_1 is a universal constant. Consequently, the total number of samples used by Algorithm 6 is given by

$$MK + \sum_{m=1}^{M} N_m \leq \frac{N}{2} + \frac{N}{2} = N,$$

where in the last equation we have used the fact that $MK = \frac{N}{2}$. Finally, using $M = \log_2\left(\frac{N(1-\gamma)^2}{8\log((8D/\delta)\cdot\log N)}\right)$ we have the following relation for some universal constant c:

$$4^M = c \cdot \frac{N^2(1-\gamma)^4}{\log^2((8D/\delta)\cdot\log N)}$$

Putting together the pieces, we conclude that

$$\|\overline{\theta}_{M+1} - \theta^*\|_\infty \leq c_2 \left\|\overline{\theta}_1 - \theta^*\right\|_\infty \cdot \frac{\log^2((8D/\delta) \cdot \log N)}{N^2(1-\gamma)^4} \\ + c_2 \left\{ \sqrt{\frac{\log(8DM/\delta)}{N}} \Big(\gamma \cdot \nu(\mathbf{P}, \theta^*) + \rho(\mathbf{P}, r) \Big) + \frac{\log(8DM/\delta)}{N} \cdot b(\theta^*) \right\},$$

for a suitable universal constant c_2. The last bound is valid with probability exceeding $1-\delta$ via the union bound. In order to complete the proof, it remains to prove Lemma 9.4.6, which we do in the following subsection.

Proof of Lemma 9.4.6

We now turn to the proof of the key lemma within the argument. We begin with a high-level overview in order to provide intuition. In the m-th epoch that updates the estimate from $\overline{\theta}_m$ to $\overline{\theta}_{m+1}$, the vector $\overline{\theta} \equiv \overline{\theta}_m$ is used to recenter the updates. Our analysis of the m-th epoch is based on a sequence of recentered operators $\{\mathcal{J}_k^m\}_{k \geq 1}$ and their population analogs $\mathcal{J}^m(\theta)$. The action of these operators on a point θ is given by the relations

$$\mathcal{J}_k^m(\theta) := \widehat{\mathcal{T}}_k(\theta) - \widehat{\mathcal{T}}_k(\overline{\theta}_m) + \widetilde{\mathcal{T}}_N(\overline{\theta}_m), \quad \text{and} \quad \mathcal{J}^m(\theta) := \mathcal{T}(\theta) - \mathcal{T}(\overline{\theta}_m) + \widetilde{\mathcal{T}}_N(\overline{\theta}_m). \tag{9.26a}$$

By definition, the updates within epoch m can be written as

$$\theta_{k+1} = (1-\alpha_k)\,\theta_k + \alpha_k \mathcal{J}_k^m(\theta_k)\,. \tag{9.26b}$$

Note that the operator $\mathcal{J}^m$ is γ-contractive in $\|\cdot\|_\infty$-norm, and as a result it has a unique fixed point, which we denote by $\widehat{\theta}_m$. Since $\mathcal{J}^m(\theta) = \mathbb{E}\left[\mathcal{J}_k^m(\theta)\right]$ by construction, when studying epoch m, it is natural to analyze the convergence of the sequence $\{\theta_k\}_{k\geq 1}$ to $\widehat{\theta}_m$.

Suppose that we have taken K steps within epoch m. Applying the triangle inequality yields the bound

$$\|\overline{\theta}_{m+1} - \theta^*\|_\infty = \|\theta_{K+1} - \theta^*\|_\infty \leq \left\|\theta_{K+1} - \widehat{\theta}_m\right\|_\infty + \left\|\widehat{\theta}_m - \theta^*\right\|_\infty. \tag{9.26c}$$

With this decomposition, our proof of Lemma 9.4.6 is based on two auxiliary lemmas that provide high-probability upper bounds on the two terms on the right-hand side of inequality (9.26c).

Lemma 9.4.7. *Let (c_1, c_2, c_3) be positive numerical constants, and suppose that the epoch length K satisfies one the following three stepsize-dependent lower bounds:*

(a) $K \geq c_1 \frac{\log(8KMD/\delta)}{(1-\gamma)^3}$ for recentered linear stepsize $\alpha_k = \frac{1}{1+(1-\gamma)k}$,

(b) $K \geq c_2 \log(8KMD/\delta) \cdot \left(\frac{1}{1-\gamma}\right)^{\left(\frac{1}{1-\omega} \vee \frac{2}{\omega}\right)}$ for polynomial stepsize $\alpha_k = \frac{1}{k^\omega}$ with $0 < \omega < 1$,

(c) $K \geq \frac{c_3}{\log\left(\frac{1}{1-\alpha(1-\gamma)}\right)}$ *for constant stepsize* $\alpha_k = \alpha \leq \frac{(1-\gamma)^2}{\log(8KMD/\delta)} \cdot \frac{1}{5^2 \cdot 32^2}$.

Then after K *update steps with epoch* m*, the iterate* θ_{K+1} *satisfies the bound*

$$\|\theta_{K+1} - \widehat{\theta}_m\|_\infty \leq \frac{1}{8}\|\overline{\theta}_m - \theta^*\|_\infty + \frac{1}{8}\|\widehat{\theta}_m - \theta^*\|_\infty \quad \textit{with probability at least } 1 - \tfrac{\delta}{2M}. \tag{9.27}$$

See Appendix C.3.3 for the proof of this claim.

Our next auxiliary result provides a high-probability bound on the difference $\|\widehat{\theta}_m - \theta^*\|_\infty$.

Lemma 9.4.8. *There is an absolute constant* c_4 *such that for any recentering sample size satisfying* $N_m \geq 4^2 \cdot 9^2 \cdot \frac{\log(MD/\delta)}{(1-\gamma)^2}$*, we have*

$$\|\widehat{\theta}_m - \theta^*\|_\infty \leq \tfrac{1}{9}\|\overline{\theta}_m - \theta^*\|_\infty + c_4 \left\{ \sqrt{\frac{\log(8DM/\delta)}{N_m}} \Big(\gamma \cdot \nu(\mathbf{P}, \theta^*) + \rho(\mathbf{P}, r)\Big) + \frac{\log(8DM/\delta)}{N_m} \cdot b(\theta^*) \right\},$$

with probability exceeding $1 - \frac{\delta}{2M}$.

See Appendix C.3.3 for the proof of this claim.

With Lemmas 9.4.7 and 9.4.8 in hand, the remainder of the proof is straightforward. Recall from Eq. (9.25) the shorthand notation τ_m and η_m. Using our earlier bound (9.26c), we have that at the end of epoch m (which is also the starting point of epoch $m+1$),

$$\begin{aligned}
\left\|\overline{\theta}_{m+1} - \theta^*\right\|_\infty &\leq \|\theta_{K+1} - \widehat{\theta}_m\|_\infty + \|\widehat{\theta}_m - \theta^*\|_\infty \\
&\overset{(i)}{\leq} \left\{ \frac{\|\overline{\theta}_m - \theta^*\|_\infty}{8} + \frac{1}{8}\left\|\widehat{\theta}_m - \theta^*\right\|_\infty \right\} + \left\|\widehat{\theta}_m - \theta^*\right\|_\infty \\
&= \frac{\|\overline{\theta}_m - \theta^*\|_\infty}{8} + \frac{9}{8} \cdot \left\|\widehat{\theta}_m - \theta^*\right\|_\infty \\
&\overset{(ii)}{\leq} \frac{\|\overline{\theta}_m - \theta^*\|_\infty}{8} + \frac{1}{8}\left\{\|\overline{\theta}_m - \theta^*\|_\infty + c_4(\tau_m + \eta_m)\right\} \\
&\leq \frac{\|\overline{\theta}_m - \theta^*\|_\infty}{4} + c_4(\tau_m + \eta_m),
\end{aligned}$$

where inequality (i) follows from Lemma 9.4.7(a), and inequality (ii) from Lemma 9.4.8. Finally, the sequence of inequalities above holds with probability at least $1 - \frac{\delta}{M}$ via a union bound. This completes the proof of Lemma 9.4.6.

Proof of Theorem 9.3.1, parts (b) and (c)

The proofs of Theorem 9.3.1 parts (b) and (c) require versions of Lemma 9.4.6 for the polynomial stepsize (7.8b) and constant stepsize (7.8a), respectively. These two versions of Lemma 9.4.6 can be obtained by simply replacing Lemma 9.4.7, part (a), by Lemma 9.4.7, parts (b) and (c), respectively, in the proof of Lemma 9.4.6.

9.5 Summary and open questions

In this chapter, we have discussed the problem of policy evaluation in discounted Markov decision processes via stochastic approximation. Our contribution is three-fold. First, we provided a non-asymptotic instance-dependent local-minimax bound on the ℓ_∞-error for the policy evaluation problem under the generative model. Next, via careful simulations, we showed that the standard TD-learning algorithm—even when combined with Polyak-Rupert iterate averaging—does not yield ideal non-asymptotic behavior as captured by our lower bound. In order to remedy this difficulty, we introduced and analyzed a variance-reduced version of the standard TD-learning algorithm which achieves our non-asymptotic instance-dependent lower bound up to logarithmic factors. Both the upper and lower bounds discussed in this paper hold when the sample size is bigger than an explicit threshold; relaxing this minimum sample size requirement is an interesting future research direction. Finally, we point out that although we have focused on the tabular policy evaluation problem, the variance-reduced algorithm discussed in this paper can be applied in more generality, and it would be interesting to explore applications of this algorithm to non-tabular settings.

In the broader context of this dissertation, this chapter (and this part of the thesis more generally) focused on the adaptation question in policy evaluation. This perspective both asks and answers interesting questions and has helped guide our algorithmic developments. We expect that demanding instance-specific adaptation from other algorithms in reinforcement learning can lead to interesting statistical and algorithmic insights that a focus solely on worst-case optimality cannot.

Part IV

Appendices

Technical material

Appendix A

Technical material for part I

A.1 Technical lemmas used in Chapter 2

In this section, we provide statements and proofs of the technical lemmas used in the proofs of our main theorems.

A.1.1 Supporting proofs for Theorem 2.3.1: $d = 1$ case

We provide a proof of Lemma 2.4.1 in this section; see Section 2.4.1 for the proof of Theorem 2.3.1 (case $d = 1$) given Lemma 2.4.1. We begin by restating the lemma for convenience.

Lemma 2.4.1. *For $d = 1$ and any two permutation matrices Π and Π^*, and provided $\frac{\|x^*\|_2^2}{\sigma^2} > 1$, we have*

$$\Pr\{\Delta(\Pi, \Pi^*) \leq 0\} \leq c' \exp\left(-c\, \mathsf{d_H}(\Pi, \Pi^*) \log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right)\right).$$

Proof of Lemma 2.4.1

Before the proof, we establish notation. For each $\delta > 0$, define the events

$$\mathcal{F}_1(\delta) = \left\{ |\|P_{\Pi^*}^{\perp} y\|_2^2 - \|P_{\Pi}^{\perp} w\|_2^2| \geq \delta \right\}, \text{ and} \tag{A.1a}$$

$$\mathcal{F}_2(\delta) = \left\{ \|P_{\Pi}^{\perp} y\|_2^2 - \|P_{\Pi}^{\perp} w\|_2^2 \leq 2\delta \right\}. \tag{A.1b}$$

Evidently,

$$\{\Delta(\Pi, \Pi^*) \leq 0\} \subseteq \mathcal{F}_1(\delta) \cup \mathcal{F}_2(\delta). \tag{A.2}$$

Indeed, if neither $\mathcal{F}_1(\delta)$ nor $\mathcal{F}_2(\delta)$ occurs

$$\Delta(\Pi, \Pi^*) = \left(\|P_{\Pi}^{\perp} y\|_2^2 - \|P_{\Pi}^{\perp} w\|_2^2\right) - \left(\|P_{\Pi^*}^{\perp} y\|_2^2 - \|P_{\Pi}^{\perp} w\|_2^2\right) > 2\delta - \delta = \delta.$$

Thus, to prove Lemma 2.4.1, we shall bound the probability of the two events $\mathcal{F}_1(\delta)$ and $\mathcal{F}_2(\delta)$ individually, and then invoke the union bound. Note that inequality (A.2) holds for all values of $\delta > 0$; it is convenient to choose $\delta^* := \frac{1}{3}\|P_{\Pi}^{\perp}\Pi^* A x^*\|_2^2$. With this choice, the following lemma bounds the probabilities of the individual events over randomness in w conditioned on a given A. Its proof is postponed to the end of the section.

Lemma A.1.1. *For any $\delta > 0$ and with $\delta^* = \frac{1}{3}\|P_{\Pi}^{\perp}\Pi^* A x^*\|_2^2$, we have*

$$\Pr_w\{\mathcal{F}_1(\delta)\} \le c' \exp\left(-c\frac{\delta}{\sigma^2}\right), \textit{ and} \tag{A.3a}$$

$$\Pr_w\{\mathcal{F}_2(\delta^*)\} \le c' \exp\left(-c\frac{\delta^*}{\sigma^2}\right). \tag{A.3b}$$

The next lemma, proved in Section A.1.1, is needed in order to incorporate the randomness in A into the required tail bound. It is convenient to introduce the shorthand $T_\Pi := \|P_{\Pi}^{\perp}\Pi^* A x^*\|_2^2$.

Lemma A.1.2. *For $d = 1$ and any two permutation matrices Π and Π^* at Hamming distance h, we have*

$$\Pr_A\{T_\Pi \le t\|x^*\|_2^2\} \le 6\exp\left(-\frac{h}{10}\left[\log\frac{h}{t} + \frac{t}{h} - 1\right]\right) \tag{A.4}$$

for all $t \in [0, h]$.

We now have all the ingredients to prove Lemma 2.4.1.

Proof of Lemma 2.4.1. Applying Lemma A.1.1 and using the union bound yields

$$\begin{aligned}\Pr_w\{\Delta(\Pi,\Pi^*) \le 0\} &\le \Pr_w\{\mathcal{F}_1(\delta^*)\} + \Pr_w\{\mathcal{F}_2(\delta^*)\} \\ &\le c' \exp\left(-c\frac{T_\Pi}{\sigma^2}\right).\end{aligned} \tag{A.5}$$

Combining bound (A.5) with Lemma A.1.2 yields

$$\begin{aligned}\Pr\{\Delta(\Pi,\Pi^*) \le 0\} &\le c' \exp\left(-c\frac{t\|x^*\|_2^2}{\sigma^2}\right)\Pr_A\{T_\Pi \ge t\|x^*\|_2^2\} + \Pr_A\{T_\Pi \le t\|x^*\|_2^2\} \\ &\le c' \exp\left(-c\frac{t\|x^*\|_2^2}{\sigma^2}\right) + 6\exp\left(-\frac{h}{10}\left[\log\frac{h}{t} + \frac{t}{h} - 1\right]\right),\end{aligned} \tag{A.6}$$

where the last inequality holds provided that $t \in [0, h]$, and the probability in the LHS is now taken over randomness in both w *and* A.

Using the shorthand $\mathsf{snr} := \frac{\|x^*\|_2^2}{\sigma^2}$, setting $t = h\frac{\log \mathsf{snr}}{\mathsf{snr}}$, and noting that $t \in [0, h]$ since $\mathsf{snr} > 1$, we have

$$\Pr\{\Delta(\Pi,\Pi^*) \le 0\} \le c' \exp\left(-ch\log\mathsf{snr}\right) + 6\exp\left(-\frac{h}{10}\left[\log\left(\frac{\mathsf{snr}}{\log\mathsf{snr}}\right) + \frac{\log\mathsf{snr}}{\mathsf{snr}} - 1\right]\right).$$

It is easily verified that for all $\mathsf{snr} > 1$, we have

$$\log\left(\frac{\mathsf{snr}}{\log \mathsf{snr}}\right) + \frac{\log \mathsf{snr}}{\mathsf{snr}} - 1 > \frac{\log \mathsf{snr}}{4}. \tag{A.7}$$

Hence, after substituting for snr, we have

$$\Pr\{\Delta(\Pi, \Pi^*) \leq 0\} \leq c' \exp\left(-ch \log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right)\right). \tag{A.8}$$

□

Proof of Lemma A.1.1

We prove each claim of the lemma separately.

Proof of claim (A.3a) To start, note that by definition of the linear model, we have $\|P_{\Pi^*}^{\perp} y\|_2^2 = \|P_{\Pi^*}^{\perp} w\|_2^2$. Letting Z_ℓ denote a χ^2 random variable with ℓ degrees of freedom, we claim that

$$\|P_{\Pi^*}^{\perp} w\|_2^2 - \|P_{\Pi}^{\perp} w\|_2^2 = Z_k - \tilde{Z}_k,$$

where $k := \min(d, \mathsf{d_H}(\Pi, \Pi^*))$.

For the rest of the proof, we adopt the shorthand $\Pi \setminus \Pi' := \mathsf{range}(\Pi A) \setminus \mathsf{range}(\Pi' A)$, and $\Pi \cap \Pi' := \mathsf{range}(\Pi A) \cap \mathsf{range}(\Pi' A)$. Now, by the Pythagorean theorem, we have

$$\|P_{\Pi^*}^{\perp} w\|_2^2 - \|P_{\Pi}^{\perp} w\|_2^2 = \|P_{\Pi} w\|_2^2 - \|P_{\Pi^*} w\|_2^2.$$

Splitting it up further, we can then write

$$\|P_{\Pi} w\|_2^2 = \|P_{\Pi \cap \Pi^*} w\|_2^2 + \|(P_{\Pi} - P_{\Pi \cap \Pi^*}) w\|_2^2,$$

where we have used the fact that $P_{\Pi \cap \Pi^*} P_{\Pi} = P_{\Pi \cap \Pi^*} = P_{\Pi \cap \Pi^*} P_{\Pi^*}$.

Similarly for the second term, we have $\|P_{\Pi^*} w\|_2^2 = \|P_{\Pi \cap \Pi^*} w\|_2^2 + \|(P_{\Pi^*} - P_{\Pi \cap \Pi^*}) w\|_2^2$, and hence,

$$\|P_{\Pi} w\|_2^2 - \|P_{\Pi^*} w\|_2^2 = \|(P_{\Pi} - P_{\Pi \cap \Pi^*}) w\|_2^2 - \|(P_{\Pi^*} - P_{\Pi \cap \Pi^*}) w\|_2^2.$$

Now each of the two projection matrices above has rank[1] $\mathsf{dim}(\Pi \setminus \Pi^*) = k$, which completes the proof of the claim. To prove the lemma, note that for any $\delta > 0$, we can write

$$\Pr\{\mathcal{F}_1(\delta)\} \leq \Pr\{|Z_k - k| \geq \delta/2\} + \Pr\{|\tilde{Z}_k - k| \geq \delta/2\}.$$

Using the sub-exponential tail-bound on χ^2 random variables (see Lemma A.2.2 in Appendix A.2.2) completes the proof. □

[1] With probability 1

Proof of claim (A.3b) We begin by writing

$$\Pr_w\{\mathcal{F}_2(\delta)\} = \Pr_w\left\{\underbrace{\|P_\Pi^\perp \Pi^* A x^*\|_2^2 + 2\langle P_\Pi^\perp \Pi^* A x^*, P_\Pi^\perp w\rangle}_{R(A,w)} \leq 2\delta\right\}.$$

We see that conditioned on A, the random variable $R(A, w)$ is distributed as $\mathcal{N}(T_\Pi, 4\sigma^2 T_\Pi)$, where we have used the shorthand $T_\Pi := \|P_\Pi^\perp \Pi^* A x^*\|_2^2$.

So applying standard Gaussian tail bounds (see, for example, Boucheron et al. [38]), we have

$$\Pr_w\{\mathcal{F}_2(\delta)\} \leq \exp\left(-\frac{(T_\Pi - 2\delta)^2}{8\sigma^2 T_\Pi}\right).$$

Setting $\delta = \delta^* := \frac{1}{3}T_\Pi$ completes the proof. □

Proof of Lemma A.1.2

In the case $d = 1$, the matrix A is composed of a single vector $a \in \mathbb{R}^n$. Recalling the random variable $T_\Pi = \|P_\Pi^\perp \Pi^* A x^*\|_2^2$, we have

$$\begin{aligned} T_\Pi &= (x^*)^2 \left(\|a\|_2^2 - \frac{1}{\|a\|_2^2}\langle a_\Pi, a\rangle^2\right) \\ &\overset{\text{(i)}}{\geq} (x^*)^2 \left(\|a\|_2^2 - |\langle a, a_\Pi\rangle|\right) \\ &= \frac{(x^*)^2}{2} \min\left(\|a - a_\Pi\|_2^2, \|a + a_\Pi\|_2^2\right), \end{aligned}$$

where step (i) follows from the Cauchy Schwarz inequality. Applying a union bound then yields

$$\Pr\{T_\Pi \leq t(x^*)^2\} \leq \Pr\{\|a - a_\Pi\|_2^2 \leq 2t\} + \Pr\{\|a + a_\Pi\|_2^2 \leq 2t\}.$$

Let Z_ℓ and $\tilde{Z}_\ell$ denote (not necessarily independent) χ^2 random variables with ℓ degrees of freedom. We split the analysis into two cases.

Case $h \geq 3$ Lemma A.2.1 from Appendix A.2.1 guarantees that

$$\frac{\|a - a_\Pi\|_2^2}{2} \overset{d}{=} Z_{h_1} + Z_{h_2} + Z_{h_3}, \text{ and} \tag{A.9a}$$

$$\frac{\|a + a_\Pi\|_2^2}{2} \overset{d}{=} \widetilde{Z}_{h_1} + \widetilde{Z}_{h_2} + \widetilde{Z}_{h_3} + \widetilde{Z}_{n-h}, \tag{A.9b}$$

where $\overset{d}{=}$ denotes equality in distribution and $h_1, h_2, h_3 \geq \frac{h}{5}$ with $h_1 + h_2 + h_3 = h$. An application of the union bound then yields

$$\Pr\{\|a - a_\Pi\|_2^2 \leq 2t\} \leq \sum_{i=1}^{3} \Pr\left\{Z_{h_i} \leq t\frac{h_i}{h}\right\}.$$

Similarly, provided that $h \geq 3$, we have

$$\begin{aligned}\Pr\{\|a + a_\Pi\|_2^2 \leq 2t\} &\leq \Pr\{\widetilde{Z}_{h_1} + \widetilde{Z}_{h_2} + \widetilde{Z}_{h_3} + \widetilde{Z}_{n-h} \leq t\} \\ &\overset{\text{(ii)}}{\leq} \Pr\{\widetilde{Z}_{h_1} + \widetilde{Z}_{h_2} + \widetilde{Z}_{h_3} \leq t\} \\ &\overset{\text{(iii)}}{\leq} \sum_{i=1}^{3} \Pr\left\{\widetilde{Z}_{h_i} \leq t\frac{h_i}{h}\right\},\end{aligned}$$

where inequality (ii) follows from the non-negativity of Z_{n-h}, and the monotonicity of the CDF; and inequality (iii) from the union bound. Finally, bounds on the lower tails of χ^2 random variables (see Lemma B.1.6 in Appendix B.1.4) yield

$$\begin{aligned}\Pr\left\{Z_{h_i} \leq t\frac{h_i}{h}\right\} &= \Pr\left\{\widetilde{Z}_{h_i} \leq t\frac{h_i}{h}\right\} \\ &\overset{\text{(iv)}}{\leq} \left(\frac{t}{h}\exp\left(1 - \frac{t}{h}\right)\right)^{h_i/2} \\ &\overset{\text{(v)}}{\leq} \left(\frac{t}{h}\exp\left(1 - \frac{t}{h}\right)\right)^{h/10}.\end{aligned}$$

Here, inequality (iv) is valid provided $\frac{th_i}{h} \leq h_i$, or equivalently, if $t \leq h$, whereas inequality (v) follows since $h_i \geq h/5$ and the function $xe^{1-x} \in [0, 1]$ for all $x \in [0, 1]$. Combining the pieces proves Lemma A.1.2 for $h \geq 3$.

Case $h = 2$ In this case, we have

$$\frac{\|a - a_\Pi\|_2^2}{2} \overset{d}{=} 2Z_1, \quad \text{and} \quad \frac{\|a + a_\Pi\|_2^2}{2} \overset{d}{=} 2\widetilde{Z}_1 + \widetilde{Z}_{n-2}.$$

Proceeding as before by applying the union bound and Lemma B.1.6, we have that for $t \leq 2$, the random variable T_Π obeys the tail bound

$$\begin{aligned}\Pr\{T_\Pi \leq t(x^*)^2\} &\leq 2\left(\frac{t}{2}\exp\left(1 - \frac{t}{2}\right)\right)^{1/2} \\ &\leq 6\left(\frac{t}{h}\exp\left(1 - \frac{t}{h}\right)\right)^{h/10}, \text{ for } h = 2.\end{aligned}$$

□

A.1.2 Supporting proofs for Theorem 2.3.1: $d > 1$ case

We now provide a proof of Lemma 2.4.2. See Section 2.4.2 for the proof of Theorem 2.3.1 (case $d > 1$) given Lemma 2.4.2. We begin by restating the lemma for convenience.

Lemma 2.4.2. *For any* $1 < d < n$*, any two permutation matrices* Π *and* Π^* *at Hamming distance* h*, and provided* $\left(\frac{\|x^*\|_2^2}{\sigma^2}\right) n^{-\frac{2n}{n-d}} > \frac{5}{4}$*, we have*

$$\Pr\{\Delta(\Pi, \Pi^*) \leq 0\} \leq c' \max\left[\exp\left(-n \log \frac{n}{2}\right), \exp\left(ch\left(\log\left(\frac{\|x^*\|_2^2}{\sigma^2}\right) - \frac{2n}{n-d}\log n\right)\right)\right]. \tag{2.11}$$

Proof of Lemma 2.4.2

The first part of the proof is exactly the same as that of Lemma 2.4.1. In particular, Lemma A.1.1 applies without modification to yield a bound identical to the inequality (A.5), given by

$$\Pr_w\{\Delta(\Pi, \Pi^*) \leq 0\} \leq c' \exp\left(-c\frac{T_\Pi}{\sigma^2}\right), \tag{A.10}$$

where $T_\Pi = \|P_\Pi^\perp \Pi^* A x^*\|_2^2$, as before.

The major difference from the $d = 1$ case is in the random variable T_Π. Accordingly, we state the following parallel lemma to Lemma A.1.2.

Lemma A.1.3. *For* $1 < d < n$*, any two permutation matrices* Π *and* Π^* *at Hamming distance* h*, and* $t \leq hn^{-\frac{2n}{n-d}}$*, we have*

$$\Pr_A\{T_\Pi \leq t\|x^*\|_2^2\} \leq 2\max\{T_1, T_2\}, \tag{A.11}$$

where

$$T_1 = \exp\left(-n\log\frac{n}{2}\right), \text{ and}$$
$$T_2 = 6\exp\left(-\frac{h}{10}\left[\log\left(\frac{h}{tn^{\frac{2n}{n-d}}}\right) + \frac{tn^{\frac{2n}{n-d}}}{h} - 1\right]\right).$$

The proof of Lemma A.1.3 appears in Section A.1.2. We are now ready to prove Lemma 2.4.2.

Proof of Lemma 2.4.2. We prove Lemma 2.4.2 from Lemma A.1.3 and equation (A.10) by an argument similar to the one before. In particular, in a similar vein to the steps leading up to equation (A.6), we have

$$\Pr\{\Delta(\Pi, \Pi^*) \leq 0\} \leq c' \exp\left(-c\frac{t\|x^*\|_2^2}{\sigma^2}\right) + \Pr_A\{T_\Pi \leq t\|x^*\|_2^2\}. \tag{A.12}$$

We now use the shorthand $\mathsf{snr} := \frac{\|x^*\|_2^2}{\sigma^2}$ and let $t^* = h\frac{\log\left(\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}\right)}{\mathsf{snr}}$. Noting that $\mathsf{snr} \cdot n^{-\frac{2n}{n-d}} > 5/4$ yields $t^* \leq hn^{-\frac{2n}{n-d}}$, we set $t = t^*$ in inequality (A.12) to obtain

$$\begin{aligned} &\Pr\{\Delta(\Pi, \Pi^*) \leq 0\} \\ &\quad \leq c' \exp\left(-ch\log sn^{-\frac{2n}{n-d}}\right) + \Pr_A\{T_\Pi \leq t^*\|x^*\|_2^2\}. \end{aligned} \tag{A.13}$$

Since $\Pr_A\{T_\Pi \leq t^* \|x^*\|_2^2\}$ can be bounded by a maximum of two terms (A.11), we now split the analysis into two cases depending on which term attains the maximum.

Case 1 First, suppose that the second term attains the maximum in inequality (A.11), i.e., $\Pr_A\{T_\Pi \leq t^*\|x^*\|_2^2\} \leq 12\exp\left(-\frac{h}{10}\left[\log\left(\frac{h}{t^* n^{\frac{2n}{n-d}}}\right) + \frac{t^* n^{\frac{2n}{n-d}}}{h} - 1\right]\right)$. Substituting for t^*, we have

$$\begin{aligned}
&\Pr_A\{T_\Pi \leq t^*\|x^*\|_2^2\} \\
&\leq 12\exp\left(-\frac{h}{10}\log\left(\frac{\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}}{\log\left(\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}\right)}\right)\right)\cdot\exp\left(-\frac{h}{10}\left[\frac{\log\left(\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}\right)}{\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}} - 1\right]\right).
\end{aligned}$$

We have $\mathsf{snr}\cdot n^{-\frac{2n}{n-d}} > \frac{5}{4}$, a condition which leads to the following pair of easily verifiable inequalities:

$$\log\left(\frac{\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}}{\log\left(\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}\right)}\right) + \frac{\log\left(\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}\right)}{\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}} - 1 \geq \frac{\log \mathsf{snr}\cdot n^{-\frac{2n}{n-d}}}{4}, \text{ and} \tag{A.14a}$$

$$\log\left(\frac{\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}}{\log\left(\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}\right)}\right) + \frac{\log\left(\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}\right)}{\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}} - 1 \leq 5\log\left(\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}\right). \tag{A.14b}$$

Using inequality (A.14a), we have

$$\Pr_A\{T_\Pi \leq t^*\|x^*\|_2^2\} \leq 12\exp\left(-ch\log\left(\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}\right)\right). \tag{A.15}$$

Inequality (A.14b) will be useful in the second case to follow. Now using inequalities (A.15) and (A.13) together yields

$$\Pr\{\Delta(\Pi,\Pi^*) \leq 0\} \leq c'\exp\left(-ch\log\left(\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}\right)\right). \tag{A.16}$$

It remains to handle the second case.

Case 2 Suppose now that $\Pr_A\{T_\Pi \leq t^*\|x^*\|_2^2\} \leq 2\exp\left(-n\log\frac{n}{2}\right)$, i.e., that the first term in RHS of inequality (A.11) attains the maximum when $t = t^*$. In this case, we have

$$\begin{aligned}
\exp\left(-n\log\frac{n}{2}\right) &\geq 6\exp\left(-\frac{h}{10}\left[\log\left(\frac{h}{t^* n^{\frac{2n}{n-d}}}\right) + \frac{t^* n^{\frac{2n}{n-d}}}{h} - 1\right]\right) \\
&\overset{\text{(i)}}{\geq} c'\exp\left(-ch\log\left(\mathsf{snr}\cdot n^{-\frac{2n}{n-d}}\right)\right),
\end{aligned}$$

where step (i) follows from the right inequality (A.14b). Now substituting into inequality (A.13), we have

$$\begin{aligned}\Pr\{\Delta(\Pi, \Pi^*) \leq 0\} &\leq c' \exp\left(-ch \log\left(\mathsf{snr} \cdot n^{-\frac{2n}{n-d}}\right)\right) + 2\exp\left(-n \log \frac{n}{2}\right) \\ &\leq c' \exp\left(-n \log \frac{n}{2}\right). \qquad \text{(A.17)}\end{aligned}$$

Combining equations (A.16) and (A.17) completes the proof of Lemma 2.4.2. □

Proof of Lemma A.1.3

We begin by reducing the problem to the case $x^* = e_1 \|x^*\|_2$, where e_1 represents the first standard basis vector in $\mathbb{R}^d$. In particular, if $Wx^* = e_1\|x^*\|_2$ for a $d \times d$ unitary matrix W and writing $A = \widetilde{A}W$, we have by rotation invariance of the Gaussian distribution that the the entries of $\widetilde{A}$ are distributed as i.i.d. standard Gaussians. It can be verified that $T_\Pi = \|I - \Pi\widetilde{A}(\widetilde{A}^\top\widetilde{A})^{-1}(\Pi\widetilde{A})^\top\Pi^*\widetilde{A}e_1\|_2^2\|x^*\|_2^2$. Since $\widetilde{A} \overset{d}{=} A$, the reduction is complete.

In order to keep the notation uncluttered, we denote the first column of A by a. We also denote the span of the first column of ΠA by S_1 and that of the last $d-1$ columns of ΠA by S_{-1}. Denote their respective orthogonal complements by $S_1^\perp$ and $S_{-1}^\perp$. We then have

$$\begin{aligned}T_\Pi &= \|x^*\|_2^2 \|P_\Pi^\perp a\|_2^2 \\ &= \|x^*\|_2^2 \|P_{S_{-1}^\perp \cap S_1^\perp} a\|_2^2 \\ &= \|x^*\|_2^2 \|P_{S_{-1}^\perp \cap S_1^\perp} P_{S_1^\perp} a\|_2^2.\end{aligned}$$

We now condition on a. Consequently, the subspace $S_1^\perp$ is a *fixed* $(n-1)$-dimensional subspace. Additionally, $S_{-1}^\perp \cap S_1^\perp$ is the intersection of a uniformly random $(n-(d-1))$-dimensional subspace with a fixed $(n-1)$-dimensional subspace, and is therefore a uniformly random $(n-d)$-dimensional subspace within $S_1^\perp$. Writing $u = \frac{P_{S_1^\perp} a}{\|P_{S_1^\perp} a\|_2}$, we have

$$T_\Pi \overset{d}{=} \|x^*\|_2^2 \|P_{S_{-1}^\perp \cap S_1^\perp} u\|_2^2 \|P_{S_1^\perp} a\|_2^2.$$

Now since $u \in S_1^\perp$, note that $\|P_{S_{-1}^\perp \cap S_1^\perp} u\|_2^2$ is the squared length of a projection of an $(n-1)$-dimensional unit vector onto a uniformly chosen $(n-d)$-dimensional subspace. In other words, denoting a uniformly random projection from m dimensions to k dimensions by P_k^m and noting that u is a unit vector, we have

$$\|P_{S_{-1}^\perp \cap S_1^\perp} u\|_2^2 \overset{d}{=} \|P_{n-d}^{n-1} v_1\|_2^2 \overset{\text{(i)}}{=} 1 - \|P_{d-1}^{n-1} v_1\|_2^2,$$

where v_1 represents a fixed standard basis vector in $n-1$ dimensions. The quantities P_{n-d}^{n-1} and P_{d-1}^{n-1} are projections onto orthogonal subspaces, and step (i) is a consequence of the Pythagorean theorem.

Now removing the conditioning on a, we see that the term for $d > 1$ can be lower bounded by the corresponding T_Π for $d = 1$, but scaled by a random factor – the norm of a random projection. Using $T^1_\Pi := \|P^\perp_{S_1} a\|_2^2 \|x^*\|_2^2$ to denote T_Π when $d = 1$, we have

$$T_\Pi = (1 - X_{d-1}) T^1_\Pi, \tag{A.18}$$

where we have introduced the shorthand $X_{d-1} = \|P^{n-1}_{d-1} v_1\|_2^2$.

We first handle the random projection term in equation (A.18) using Lemma A.2.3 in Appendix A.2.2. In particular, substituting $\beta = (1 - z)\frac{n-1}{d-1}$ in inequality (A.26) yields

$$\begin{aligned}
\Pr\{1 - X_{d-1} \le z\} &\le \left(\frac{n-1}{d-1}\right)^{(d-1)/2} \left(\frac{z(n-1)}{n-d}\right)^{(n-d)/2} \\
&\overset{\text{(i)}}{\le} \sqrt{\binom{n-1}{d-1}} \sqrt{\binom{n-1}{n-d}} z^{\frac{n-d}{2}} \\
&= \binom{n-1}{d-1} z^{\frac{n-d}{2}} \\
&\overset{\text{(ii)}}{\le} 2^{n-1} z^{\frac{n-d}{2}},
\end{aligned}$$

where in steps (i) and (ii), we have used the standard inequality $2^n \ge \binom{n}{r} \ge \left(\frac{n}{r}\right)^r$. Now setting $z = n^{\frac{-2n}{n-d}}$, which ensures that $(1 - z)\frac{n-1}{d-1} > 1$ for all $d < n$ and large enough n, we have

$$\Pr\{1 - X_{d-1} \le n^{\frac{-2n}{n-d}}\} \le \exp\left(-n \log \frac{n}{2}\right). \tag{A.19}$$

Applying the union bound then yields

$$\begin{aligned}
&\Pr\{T_\Pi \le t\|x^*\|_2^2\} \\
&\quad \le \Pr\{1 - X_{d-1} \le n^{\frac{-2n}{n-d}}\} + \Pr\{T^1_\Pi \le t n^{\frac{2n}{n-d}} \|x^*\|_2^2\}.
\end{aligned} \tag{A.20}$$

We have already computed an upper bound on $\Pr\{T^1_\Pi \le t n^{\frac{2n}{n-d}} \|x^*\|_2^2\}$ in Lemma A.1.2. Applying it yields that provided $t \le h n^{-\frac{2n}{n-d}}$, we have

$$\Pr\{T^1_\Pi \le t n^{\frac{2n}{n-d}} \|x^*\|_2^2\} \le 6 \left(\frac{t n^{\frac{2n}{n-d}}}{h} \exp\left(1 - \frac{t n^{\frac{2n}{n-d}}}{h}\right)\right)^{h/10}. \tag{A.21}$$

Combining equations (A.21) and (A.19) with the union bound (A.20) and performing some algebraic manipulation then completes the proof of Lemma A.1.3. □

A.1.3 Supporting proofs for Proposition 2.3.1

The following technical lemma was used in the proof of Proposition 2.3.1. Recall that the observation vector y is drawn from the mixture distribution

$$\mathbb{M}(\bar{h}) = \frac{1}{|\mathbb{B}_{\mathsf{H}}(\bar{h})|} \sum_{\Pi \in \mathfrak{S}_n} \mathbb{P}_{\Pi},$$

where $\mathbb{P}_\Pi$ denotes the Gaussian distribution $\mathcal{N}(\Pi A x^*, \sigma^2 I_n)$.

Lemma A.1.4. *For y drawn according to the mixture distribution $\mathbb{M}(\bar{h})$, we have*

$$\det \mathbb{E}\left[yy^\top\right] \leq (\sigma^2 + \|x^*\|_2^2)^n \,(1+n) \left(\frac{\bar{h}}{n}\right)^{n-1}.$$

Proof of Lemma A.1.4. We first explicitly calculate the matrix $Y := \mathbb{E}\left[yy^\top\right]$. Note that the diagonal entries take the form

$$Y_{ii} = (x^*)^\top \mathbb{E}\left[a_{\pi_i} a_{\pi_i}^\top\right] x^* + \mathbb{E}\left[w_{\pi_i}^2\right] = \|x^*\|_2^2 + \sigma^2.$$

The off-diagonal entries can be evaluated as

$$\begin{aligned} Y_{ij} &= (x^*)^\top \mathbb{E}\left[a_{\pi_i} a_{\pi_j}^\top\right] x^* + \mathbb{E}\left[w_{\pi_i} w_{\pi_j}\right] \\ &\overset{(\mathsf{i})}{=} \left(\frac{n-\bar{h}}{n} + \frac{\bar{h}}{n^2}\right)\left(\|x^*\|_2^2 + \sigma^2\right), \text{ for } i \neq j, \end{aligned}$$

where step (i) follows since

$$\pi_i = \pi_j \text{ with probability } \frac{n-\bar{h}}{n} + \frac{\bar{h}}{n^2}. \tag{A.22}$$

Equation (A.22) is a consequence of the fact that a uniform permutation over $\mathbb{B}_{\mathsf{H}}(\bar{h})$ can be generated by first picking $\bar{h}$ positions (the permutation set) out of $[n]$ uniformly at random, and then uniformly permuting those $\bar{h}$ positions. The probability that $\pi_i = \pi_j$ is equal to the probability that $\pi_i = i$, an event that occurs if:
(a) position i is not chosen in the permutation set, which happens with probability $\frac{n-\bar{h}}{n}$, or if
(b) position i is in the permutation set but the permutation maps i to itself, which happens with probability $\frac{\bar{h}}{n}\frac{1}{n}$.

Hence, the determinant of Y is given by $\det Y = (\|x^*\|_2^2 + \sigma^2)^n \det \overline{Y}$, where we have defined $\overline{Y} := \frac{1}{\|x^*\|_2^2 + \sigma^2} Y$. Note that $\overline{Y}$ is a highly structured matrix, and so its determinant can be computed exactly. In particular, letting the scalar $\beta = 1$ denote the identical diagonal entries of $\overline{Y}$ and the scalar γ denote its identical off-diagonal entries, it is easy to verify that the all ones vector $\mathbf{1}$ is an eigenvector of $\overline{Y}$, with corresponding eigenvalue $\beta + (n-1)\gamma$. Additionally, for any vector v that obeys $\mathbf{1}^\top v = 0$, we have

$$\overline{Y} v = (\beta - \gamma) v + \gamma (v^\top \mathbf{1}) \mathbf{1} = (\beta - \gamma) v,$$

and so the remaining $n-1$ eigenvalues are identically $\beta - \gamma$.

Substituting for β and γ, the eigenvalues of $\overline{Y}$ are given by

$$\lambda_1(\overline{Y}) = 1 + \frac{(n-\bar{h})(n-1)}{n} + \frac{\bar{h}(n-1)}{n^2}, \text{ and}$$

$$\lambda_2(\overline{Y}) = \lambda_3(\overline{Y}) = \cdots = \lambda_n(\overline{Y}) = \frac{\bar{h}}{n} - \frac{\bar{h}}{n^2}.$$

Hence, we have

$$\begin{aligned}\det \overline{Y} &= \left(1 + \frac{(n-\bar{h})(n-1)}{n} + \frac{\bar{h}(n-1)}{n^2}\right)\left(\frac{\bar{h}}{n} - \frac{\bar{h}}{n^2}\right)^{n-1} \\ &\le \left(1 + n - \bar{h} + \frac{\bar{h}}{n}\right)\left(\frac{\bar{h}}{n}\right)^{n-1} \\ &\le (1+n)\left(\frac{\bar{h}}{n}\right)^{n-1},\end{aligned}$$

where in the last step, we have used the fact that $0 < \bar{h} \le n$. This completes the proof. □

A.1.4 Supporting proofs for Theorem 2.3.3

The following technical lemma was used in the proof of Theorem 2.3.3, and we recall the setting for convenience. For any estimator $\widehat{\Pi}$, we denote by the indicator random variable $E(\widehat{\Pi}, D)$ whether or not the $\widehat{\Pi}$ has acceptable distortion, i.e., $E(\widehat{\Pi}, D) = \mathbb{I}[\mathsf{d_H}(\widehat{\Pi}, \Pi^*) \ge D]$, with $E = 1$ representing the error event. Assume Π^* is picked uniformly at random in $\mathfrak{S}_n$.

Lemma A.1.5. *The probability of error is lower bounded as*

$$\Pr\{E(\widehat{\Pi}, D) = 1\} \ge 1 - \frac{I(\Pi^*; y, A) + \log 2}{\log n! - \log \frac{n!}{(n-D+1)!}}. \tag{A.23}$$

Proof of Lemma A.1.5. We use the shorthand $E := E(\widehat{\Pi}, D)$ in this proof to simplify notation. Proceeding by the usual proof of Fano's inequality, we begin by expanding $H(E, \Pi^* | y, A = a, \widehat{\Pi})$ in two ways:

$$H(E, \Pi^* | y, A, \widehat{\Pi}) = H(\Pi^* | y, A, \widehat{\Pi}) + H(E | \Pi^*, y, A, \widehat{\Pi}) \tag{A.24a}$$

$$= H(E | y, A, \widehat{\Pi}) + H(\Pi^* | E, y, A, \widehat{\Pi}). \tag{A.24b}$$

Since $\Pi^* \to (y, A) \to \widehat{\Pi}$ forms a Markov chain, we have $H(\Pi^*|y, A, \widehat{\Pi}) = H(\Pi^*|y, A)$. Non-negativity of entropy yields $H(E|\Pi^*, y, A, \widehat{\Pi}) \geq 0$. Since conditioning cannot increase entropy, we have $H(E|y, A, \widehat{\Pi}) \leq H(E) \leq \log 2$, and $H(\Pi^*|E, y, A, \widehat{\Pi}) \leq H(\Pi^*|E, \widehat{\Pi})$. Combining all of this with the pair of equations (A.24) yields

$$\begin{aligned} H(\Pi^*|y, A) &\leq H(\Pi^*|E, \widehat{\Pi}) + \log 2 \\ &= \Pr\{E = 1\} H(\Pi^*|E = 1, \widehat{\Pi}) \\ &\quad + (1 - \Pr\{E = 1\})\, H(\Pi^*|E = 0, \widehat{\Pi}) + \log 2. \end{aligned} \tag{A.25}$$

We now use the fact that uniform distributions maximize entropy to bound the two terms as $H(\Pi^*|E = 1, \widehat{\Pi}) \leq H(\Pi^*) = \log n!$, and $H(\Pi^*|E = 0, \widehat{\Pi}) \leq \log \frac{n!}{(n-D+1)!}$, where the last inequality follows since $E = 0$ reveals that Π^* is within a Hamming ball of radius $D - 1$ around $\widehat{\Pi}$, and the cardinality of that Hamming ball is $\frac{n!}{(n-D+1)!}$.

Substituting back into inequality (A.25) yields

$$\begin{aligned} &\Pr\{E = 1\} \left(\log n! - \log \frac{n!}{(n - D + 1)!} \right) + H(\Pi^*) \\ &\qquad \geq H(\Pi^*|y, A) - \log 2 - \log \frac{n!}{(n - D + 1)!} + \log n!, \end{aligned}$$

where we have added the term $H(\Pi^*) = \log n!$ to both sides. Simplifying then yields inequality (A.23). □

A.2 Auxiliary results for Chapter 2

In this section, we prove a preliminary lemma about permutations that is useful in many of our proofs. We also derive tight bounds on the lower tails of χ^2-random variables and state an existing result on tail bounds for random projections.

A.2.1 Independent sets of permutations

In this section, we prove a combinatorial lemma about permutations. Given a Gaussian random vector $Z \in \mathbb{R}^n$, we use this lemma to characterize the distribution of $Z \pm \Pi Z$ as a function of the permutation Π. In order to state the lemma, we need to set up some additional notation. For a permutation π on k objects, let G_π denote the corresponding undirected incidence graph, i.e., $V(G_\pi) = [k]$, and $(i, j) \in E(G_\pi)$ iff $j = \pi(i)$ or $i = \pi(j)$.

Lemma A.2.1. *Let π be a permutation on $k \geq 3$ objects such that $\mathsf{d_H}(\pi, I) = k$. Then the vertices of G_π can be partitioned into three sets V_1, V_2, V_3 such that each is an independent set, and $|V_1|, |V_2|, |V_3| \geq \lfloor \frac{k}{3} \rfloor \geq \frac{k}{5}$.*

Proof. Note that for any permutation π, the corresponding graph G_π is composed of cycles, and the vertices in each cycle together form an independent set. Consider one such cycle. We can go through the vertices in the order induced by the cycle, and alternate placing them in each of the 3 partitions. Clearly, this produces independent sets, and furthermore, having 3 partitions ensures that the last vertex in the cycle has some partition with which it does not share edges. If the cycle length $C \equiv 0 \pmod 3$, then each partition gets $C/3$ vertices, otherwise the smallest partition has $\lfloor C/3 \rfloor$ vertices. The partitions generated from the different cycles can then be combined (with relabeling, if required) to ensure that the largest partition has cardinality at most 1 more than that of the smallest partition. □

A.2.2 Tail bounds on random projections

We state the following lemma for general sub-exponential random variables (see, e.g., Boucheron et al. [38]). We use it in the context of χ^2 random variables.

Lemma A.2.2. *Let X be a sub-exponential random variable. Then for all $t > 0$, we have*

$$\Pr\{|X - \mathbb{E}[X]| \geq t\} \leq c' e^{-ct}.$$

Lastly, we require tail bounds on the norms of random projections, a problem that has been studied extensively in the literature on dimensionality reduction. The following lemma, a consequence of the Chernoff bound, is taken from Dasgupta and Gupta [78, Lemma 2.2b].

Lemma A.2.3 ([78])**.** *Let x be a fixed n-dimensional vector, and let P_d^n be a projection matrix from n-dimensional space to a uniformly randomly chosen d-dimensional subspace, where $d \leq n$. Then we have for every $\beta > 1$ that*

$$\Pr\{\|P_d^n x\|_2^2 \geq \frac{\beta d}{n}\|x\|_2^2\} \leq \beta^{d/2}\left(1 + \frac{(1-\beta)d}{n-d}\right)^{(n-d)/2}. \tag{A.26}$$

A.2.3 Strong converse for Gaussian channel capacity

The following result due to Shannon [278] provides a strong converse for the Gaussian channel. The non-asymptotic version as stated here was also derived by Yoshihara [343].

Lemma A.2.4 ([343])**.** *Consider a vector Gaussian channel on n coordinates with message power P and noise power σ^2, whose capacity is given by $\overline{R} = \log\left(1 + \frac{P}{\sigma^2}\right)$. For any codebook $\mathcal{C}$ with $|\mathcal{C}| = 2^{nR}$, if for some $\epsilon > 0$ we have*

$$R > (1+\epsilon)\overline{R},$$

then the probability of error $p_e \geq 1 - 2 \cdot 2^{-n\epsilon}$ for n large enough.

A.3 Preliminary lemmas for Chapter 3

Before proceeding to the proofs of our main results, we provide two lemmas that underlie many of our arguments. The proofs of these lemmas can be found in Sections A.3 and A.3, respectively. Let us denote the ℓ_∞ norm of A by $\|A\|_\infty = \max_{i\in[n_1],\, j\in[n_2]} |A_{i,j}|$.

The first lemma establishes concentration of a linear form of observations $Y_{i,j}$ around its mean.

Lemma A.3.1. *For any fixed matrix $A \in \mathbb{R}^{n_1\times n_2}$ and scalar $u \geq 0$, and under our observation model* (3.1)*, we have*

$$|\langle Y - M^*, A\rangle| \leq 2(\zeta+1)\Big(\sqrt{e-1}\,\|A\|_F\sqrt{\frac{n_1n_2}{N}u} + \|A\|_\infty\frac{n_1n_2}{N}u\Big)$$

with probability at least $1-4e^{-u}$.

Consequently, for any nonempty subset $\mathcal{S} \subset [n_1]\times[n_2]$, it holds that

$$\Big|\sum_{(i,j)\in\mathcal{S}}(Y_{i,j} - M^*_{i,j})\Big| \leq 8(\zeta+1)\bigg(\sqrt{\frac{|\mathcal{S}|n_1n_2}{N}\log(n_1n_2)} + \frac{n_1n_2}{N}\log(n_1n_2)\bigg)$$

with probability at least $1-4(n_1n_2)^{-4}$.

The next lemma generalizes Theorem 5 of Shah et al. [274] to any model in which the noise satisfies a "mixed tail" assumption. More precisely, a random matrix $W \in \mathbb{R}^{n_1\times n_2}$ is said to have an (α,β)-mixed tail if there exist (possibly (n_1,n_2)-dependent) positive scalars α and $\beta \leq n_1^2$ such that for any fixed matrix $A \in \mathbb{R}^{n_1\times n_2}$ and $u \geq 0$, we have

$$\Pr\Big\{\big|\langle W, A\rangle\big| \geq \big(\alpha\|A\|_F\sqrt{u} + \beta\|A\|_\infty u\big)\Big\} \leq 2e^{-u}. \tag{A.27}$$

It is worth mentioning that similar (but less general) lemmas characterizing the estimation error for a bivariate isotonic matrix were also proved in prior work [55, 56].

Lemma A.3.2. *Let $n_2 \leq n_1$, and consider the observation model $Y = M^* + W$. Suppose that the noise matrix W satisfies the (α,β)-mixed tail condition* (A.27)*.*
(a) There is an absolute constant c such that for all $M^* \in \mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}}$, the least squares estimator $\widehat{M}_{\mathsf{LS}}(\mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}}, Y)$ satisfies

$$\begin{aligned}\big\|\widehat{M}_{\mathsf{LS}}(\mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}}, Y) - M^*\big\|_F^2 \leq c\bigg\{&\alpha^2 n_1\log n_1 + \beta n_2(\log n_1)^2 + \beta n_1\log n_1 \\ &+ \Big[\alpha\sqrt{n_1n_2}(\log n_1)^2\Big] \wedge \Big[\alpha^2 n_1n_2\log(n_1/\alpha + e)\Big] \wedge \Big[\alpha^{4/3}n_1^{1/3}n_2(\log n_1)^{2/3}\Big]\bigg\}\end{aligned}$$

with probability at least $1 - n_1^{-3n_1}$.

(b) There is an absolute constant c such that for all $M^* \in \mathbb{C}_{\mathsf{BISO}}$, the least squares estimator $\widehat{M}_{\mathsf{LS}}(\mathbb{C}_{\mathsf{BISO}}, Y)$ satisfies

$$\big\|\widehat{M}_{\mathsf{LS}}(\mathbb{C}_{\mathsf{BISO}}, Y) - M^*\big\|_F^2 \le c \bigg\{ \beta n_2 (\log n_1)^2 \\ + \Big[\alpha \sqrt{n_1 n_2} (\log n_1)^2\Big] \wedge \Big[\alpha^2 n_1 n_2 \log(n_1/\alpha + e)\Big] \wedge \Big[\alpha^{4/3} n_1^{1/3} n_2 (\log n_1)^{2/3}\Big] \bigg\}$$

with probability at least $1 - n_1^{-3n_1}$.

Proof of Lemma A.3.1

Recall that the observation matrix Y is defined by

$$Y_{i,j} = \frac{n_1 n_2}{N} \sum_{\ell=1}^{N'} y_\ell \, \mathbf{1}\{X_\ell = E^{(i,j)}\}.$$

Let $\{I_{i,j}\}_{i \in [n_1], j \in [n_2]}$ be the partition of $[N']$ defined so that $\ell \in I_{i,j}$ if and only if $X_\ell = E^{(i,j)}$. In other words, the observation y_ℓ is a noisy version of entry $M^*_{i,j}$ for each $\ell \in I_{i,j}$. Then, we have

$$Y_{i,j} = \frac{n_1 n_2}{N} \sum_{\ell \in I_{i,j}} y_\ell.$$

By Poissonization $N' \sim \mathsf{Poi}(N)$, we know that the random variables $|I_{i,j}|$ are i.i.d. $\mathsf{Poi}(\frac{N}{n_1 n_2})$, and that the quantities $Y_{i,j}$ are independent for $(i, j) \in [n_1] \times [n_2]$. Moreover, we may write

$$Y_{i,j} - M^*_{i,j} = \frac{n_1 n_2}{N} \sum_{\ell \in I_{i,j}} (y_\ell - M^*_{i,j}) + M^*_{i,j} \frac{n_1 n_2}{N} \Big(|I_{i,j}| - \frac{N}{n_1 n_2} \Big).$$

As a result, it holds that

$$\langle Y - M^*, A \rangle = Z_1 + Z_2,$$

where we define

$$Z_1 = \sum_{\substack{i \in [n_1] \\ j \in [n_2]}} A_{i,j} \frac{n_1 n_2}{N} \sum_{\ell \in I_{i,j}} (y_\ell - M^*_{i,j}) \quad \text{and}$$

$$Z_2 = \sum_{\substack{i \in [n_1] \\ j \in [n_2]}} A_{i,j} M^*_{i,j} \frac{n_1 n_2}{N} \Big(|I_{i,j}| - \frac{N}{n_1 n_2} \Big).$$

We now control Z_1 and Z_2 separately.

First, for $s \in \mathbb{R}$, we compute

$$\begin{aligned}\mathbb{E}\exp(sZ_1) &= \mathbb{E}_{I_{i,j}}\left[\mathbb{E}_{y_\ell}\left[\exp\left(s\sum_{i,j} A_{i,j}\frac{n_1n_2}{N}\sum_{\ell\in I_{i,j}}(y_\ell - M^*_{i,j})\right)\,\Big|\, I_{i,j}\right]\right]\\ &= \mathbb{E}_{I_{i,j}}\left[\prod_{i,j}\prod_{\ell\in I_{i,j}}\mathbb{E}_{y_\ell}\left[\exp\left(s\frac{n_1n_2}{N}A_{i,j}(y_\ell - M^*_{i,j})\right)\,\Big|\, I_{i,j}\right]\right]\\ &\leq \mathbb{E}_{I_{i,j}}\left[\prod_{i,j}\prod_{\ell\in I_{i,j}}\exp\left(\left(\zeta s\frac{n_1n_2}{N}A_{i,j}\right)^2\right)\right] \quad \text{if} \quad |s| \leq \frac{N}{n_1n_2\|A\|_\infty \zeta},\end{aligned} \tag{A.28}$$

where inequality(A.28) holds since the quantity $y_\ell - M^*_{i,j}$ is sub-exponential with parameter ζ. It then follows that for $|s| \leq \frac{N}{n_1n_2\|A\|_\infty\zeta}$, we have

$$\begin{aligned}\mathbb{E}\exp(sZ_1) &\leq \mathbb{E}\left[\prod_{i,j}\exp\left(\left(\zeta s\frac{n_1n_2}{N}A_{i,j}\right)^2|I_{i,j}|\right)\right]\\ &= \prod_{i,j}\mathbb{E}\left[\exp\left(\left(\zeta s\frac{n_1n_2}{N}A_{i,j}\right)^2|I_{i,j}|\right)\right]\\ &\overset{\text{(i)}}{=} \prod_{i,j}\exp\left\{\frac{N}{n_1n_2}\left[\exp\left(\left(\zeta s\frac{n_1n_2}{N}A_{i,j}\right)^2\right) - 1\right]\right\},\end{aligned}$$

where step (i) follows by explicit computation since $|I_{i,j}|$ are i.i.d. $\mathsf{Poi}(\frac{N}{n_1n_2})$ random variables. We may now apply the inequality $e^x - 1 \leq (e-1)x$ valid for $x \in [0,1]$, with the substitution

$$x = \left(\zeta s\frac{n_1n_2}{N}A_{i,j}\right)^2;$$

note that this lies in the range $[0,1]$ for all $|s| \leq \frac{N}{n_1n_2\|A\|_\infty\zeta}$. Thus, for s in this range, we have

$$\begin{aligned}\mathbb{E}\exp(sZ_1) &\leq \prod_{i,j}\exp\left\{(e-1)\frac{n_1n_2}{N}\left(\zeta sA_{i,j}\right)^2\right\}\\ &= \exp\left\{(e-1)\frac{n_1n_2}{N}\zeta^2s^2\|A\|_F^2\right\}.\end{aligned}$$

Using the Chernoff bound, we obtain for $|s| \leq \frac{N}{n_1n_2\|A\|_\infty\zeta}$ and all $t \geq 0$ the tail bound

$$\Pr\{Z_1 \geq t\} \leq e^{-st}\,\mathbb{E}\exp(sZ_1) \leq \exp\left\{(e-1)\frac{n_1n_2}{N}\zeta^2s^2\|A\|_F^2 - st\right\}.$$

The optimal choice $s = \frac{Nt}{2(e-1)n_1n_2\zeta^2\|A\|_F^2} \wedge \frac{N}{n_1n_2\|A\|_\infty\zeta}$ yields the bound

$$\Pr\{Z_1 \geq t\} \leq \exp\left(-\frac{N}{n_1n_2}\left(\frac{t^2}{4(e-1)\zeta^2\|A\|_F^2} \wedge \frac{t}{2\zeta\|A\|_\infty}\right)\right). \tag{A.29}$$

The lower tail bound is obtained analogously.

Let us now turn to the noise term Z_2. For $s \in \mathbb{R}$, we have

$$\begin{aligned}
\mathbb{E}\exp(sZ_2) &= \mathbb{E}\Big[\exp\Big(s\sum_{i,j} A_{i,j}M^*_{i,j}\frac{n_1n_2}{N}\Big(|I_{i,j}| - \frac{N}{n_1n_2}\Big)\Big)\Big] \\
&= \prod_{i,j}\mathbb{E}\Big[\exp\Big(sA_{i,j}M^*_{i,j}\frac{n_1n_2}{N}\Big(|I_{i,j}| - \frac{N}{n_1n_2}\Big)\Big)\Big] \\
&= \prod_{i,j}\exp\Big\{\frac{N}{n_1n_2}\Big[\exp\Big(sA_{i,j}M^*_{i,j}\frac{n_1n_2}{N}\Big) - sA_{i,j}M^*_{i,j}\frac{n_1n_2}{N} - 1\Big]\Big\},
\end{aligned}$$

where the last step uses the explicit MGF of $|I_{i,j}|$. We may now apply the inequality $e^x - x - 1 \le (e-2)x^2$ valid for $x \in [0,1]$, with the substitution

$$x = sA_{i,j}M^*_{i,j}\frac{n_1n_2}{N};$$

note that this quantity is in the range $[0,1]$ provided $|s| \le \frac{N}{n_1n_2\|A\|_\infty}$.

Thus, for all s in this range, we have

$$\begin{aligned}
\mathbb{E}\exp(sZ_2) &\le \prod_{i,j}\exp\Big\{(e-2)\frac{n_1n_2}{N}\big(sA_{i,j}M^*_{i,j}\big)^2\Big\} \\
&\le \exp\Big\{(e-2)\frac{n_1n_2}{N}s^2\|A\|_F^2\Big\}.
\end{aligned}$$

A similar Chernoff bound argument then yields, for each $t \ge 0$, the bound

$$\Pr\{Z_2 \ge t\} \le \exp\Big(-\frac{N}{n_1n_2}\Big(\frac{t^2}{4(e-2)\|A\|_F^2} \wedge \frac{t}{2\|A\|_\infty}\Big)\Big), \tag{A.30}$$

and the lower tail bound holds analogously.

Combining the tail bounds (A.29) on Z_1 and (A.30) on Z_2, we obtain

$$\big|\langle Y - M^*, A\rangle\big| \le |Z_1| + |Z_2| \le 2(\zeta+1)\Big(\sqrt{e-1}\|A\|_F\sqrt{\frac{n_1n_2}{N}u} + \|A\|_\infty\frac{n_1n_2}{N}u\Big)$$

with probability at least $1 - 4e^{-u}$, for each $u \ge 0$.

The second consequence of the lemma follows immediately from the first assertion by taking A to be the indicator of the subset $\mathcal{S}$, and some algebraic manipulation. □

Proof of Lemma A.3.2

We first state several lemmas that will be used in the proof. The following variational formula due to Chatterjee [58] is convenient for controlling the performance of a least squares estimator over any closed (not necessarily convex) set.

Lemma A.3.3 (Chatterjee's variational formula)**.** *Let $\mathcal{C}$ be a closed subset of $\mathbb{R}^{n_1 \times n_2}$. Suppose that $Y = M^* + W$ where $M^* \in \mathcal{C}$ and $W \in \mathbb{R}^{n_1 \times n_2}$. Let $\widehat{M}_{\mathsf{LS}}(\mathcal{C}, Y)$ denote the least squares estimator, that is, the projection of Y onto $\mathcal{C}$. Define a function $f_{M^*} : \mathbb{R}_+ \to \mathbb{R}$ by*

$$f_{M^*}(t) = \sup_{\substack{M \in \mathcal{C} \\ \|M - M^*\|_2 \le t}} \langle W, M - M^* \rangle - \frac{t^2}{2}.$$

If there exists $t^* > 0$ such that $f_{M*}(t) < 0$ for all $t \ge t^*$, then we have $\|\widehat{M}_{\mathsf{LS}}(\mathcal{C}, Y) - M^*\|_2 \le t^*$.

This deterministic form is proved in [104, Lemma 6.1]. Here we simply state the result in matrix form for convenient application.

The following chaining tail bound due to Dirksen [85, Theorem 3.5] is tailored for bounding the supremum of an empirical process with a mixed tail. We state a version specialized to our setup. For each positive scalar c, let $c\|\cdot\|$ denote the norm $\|\cdot\|$ scaled by c. Let $\gamma_p(\mathcal{C}, \|\cdot\|)$ and $\mathsf{diam}(\mathcal{C}, \|\cdot\|)$ denote Talagrand's γ_p functional (see, e.g., [85]) and the diameter of the set $\mathcal{C}$ in the distance induced by the norm $\|\cdot\|$, respectively.

Lemma A.3.4 (Generic chaining tail bounds)**.** *Let W be a random matrix in $\mathbb{R}^{n_1 \times n_2}$ satisfying the (α, β)-mixed tail condition* (A.27)*. Let $\mathcal{C}$ be a subset of $\mathbb{R}^{n_1 \times n_2}$ and let $M^* \in \mathcal{C}$. Then there exists a universal positive constant c such that for any $u \ge 1$, we have*

$$\Pr\Big\{ \sup_{M \in \mathcal{C}} \big|\langle W, M - M^* \rangle\big| \ge c\Big(\gamma_2(\mathcal{C}, \alpha\|\cdot\|_F) + \gamma_1(\mathcal{C}, \beta\|\cdot\|_\infty) \\ + \sqrt{u}\,\mathsf{diam}(\mathcal{C}, \alpha\|\cdot\|_F) + u\,\mathsf{diam}(\mathcal{C}, \beta\|\cdot\|_\infty)\Big)\Big\} \le e^{-u}.$$

The final lemma bounds the metric entropy of the set $\mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}}$ in the ℓ_2 or ℓ_∞ norm. The ℓ_2 entropy bound is known [56, 274]. For $\epsilon > 0$ and a set $\mathcal{C}$ equipped with a norm $\|\cdot\|$, let $N(\epsilon; \mathcal{C}, \|\cdot\|)$ denote the ϵ-metric entropy of $\mathcal{C}$ in the norm $\|\cdot\|$.

Lemma A.3.5. *There is an absolute positive constant c such that for any $\epsilon \le \sqrt{n_1 n_2}$, we have*

$$\log N(\epsilon; \mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}}, \|\cdot\|_F) \le 2n_1 \log n_1 + \\ \Big[c\frac{n_1 n_2}{\epsilon^2}\Big(\log\frac{\sqrt{n_1 n_2}}{\epsilon}\Big)^2\Big] \wedge \Big(c\, n_1 n_2 \log\frac{\sqrt{n_1 n_2}}{\epsilon}\Big) \wedge \Big(c\frac{\sqrt{n_1 n_2}}{\epsilon} n_2 \log n_1\Big); \tag{A.31a}$$

$$\log N(\epsilon; \mathbb{C}_{\mathsf{BISO}}, \|\cdot\|_F) \le \\ \Big[c\frac{n_1 n_2}{\epsilon^2}\Big(\log\frac{\sqrt{n_1 n_2}}{\epsilon}\Big)^2\Big] \wedge \Big(c\, n_1 n_2 \log\frac{\sqrt{n_1 n_2}}{\epsilon}\Big) \wedge \Big(c\frac{\sqrt{n_1 n_2}}{\epsilon} n_2 \log n_1\Big), \tag{A.31b}$$

and the metric entropies are zero if $\epsilon > \sqrt{n_1 n_2}$. We also have, for each $\epsilon \le 1$, the bounds

$$\log N(\epsilon; \mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}}, \|\cdot\|_\infty) \le \Big[\frac{n_2}{\epsilon}\log(en_1)\Big] \wedge \Big(n_1 n_2 \log\frac{e}{\epsilon}\Big) + 2n_1 \log n_1; \tag{A.32a}$$

$$\log N(\epsilon; \mathbb{C}_{\mathsf{BISO}}, \|\cdot\|_\infty) \le \Big[\frac{n_2}{\epsilon}\log(en_1)\Big] \wedge \Big(n_1 n_2 \log\frac{e}{\epsilon}\Big), \tag{A.32b}$$

and the metric entropies are zero if $\epsilon > 1$.

The proof of this lemma is provided at the end of the section.

Taking these lemmas as given, we are ready to prove Lemma A.3.2. We only provide the proof for the class $\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}}$; the proof for the class $\mathbb{C}_{\mathsf{BISO}}$ is analogous. Let $\mathcal{B}_{M^*}(t)$ denote the ball of radius t in the Frobenius norm centered at M^*. To apply Lemma A.3.3, we define

$$g(t) = \sup_{M \in \mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}} \cap \mathcal{B}_{M^*}(t)} \langle W, M - M^* \rangle \quad \text{and} \quad f(t) = g(t) - \frac{t^2}{2}.$$

The key is to bound this supremum $g(t)$. By the assumption on the noise matrix W, Lemma A.3.4 immediately implies that with probability $1 - e^{-u}$, we have

$$\begin{aligned} g(t) &\lesssim \gamma_2(\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}} \cap \mathcal{B}_{M^*}(t), \alpha \| \cdot \|_F) + \gamma_1(\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}} \cap \mathcal{B}_{M^*}(t), \beta \| \cdot \|_\infty) \\ &\quad + \sqrt{u}\, \mathsf{diam}(\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}} \cap \mathcal{B}_{M^*}(t), \alpha \| \cdot \|_F) + u\, \mathsf{diam}(\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}} \cap \mathcal{B}_{M^*}(t), \beta \| \cdot \|_\infty). \end{aligned} \tag{A.33}$$

The diameters are, in turn, bounded as

$$\begin{aligned} &\mathsf{diam}(\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}} \cap \mathcal{B}_{M^*}(t), \alpha \| \cdot \|_F) \le \alpha\big(t \wedge \sqrt{n_1 n_2}\big) \quad \text{and} \\ &\mathsf{diam}(\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}} \cap \mathcal{B}_{M^*}(t), \beta \| \cdot \|_\infty) \le \beta(t \wedge 1) \end{aligned}$$

since each entry of $M \in \mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}}$ is in $[0, 1]$.

It remains to bound the γ_1 and γ_2 functionals of the set $\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}} \cap \mathcal{B}_{M^*}(t)$. These functionals can be bounded by entropy integral bound (see equation (2.3) of [85])

$$\gamma_p(\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}} \cap \mathcal{B}_{M^*}(t), \| \cdot \|) \le c_p \int_0^\infty \big[\log N\big(\epsilon, \mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}} \cap \mathcal{B}_{M^*}(t), \| \cdot \|\big)\big]^{1/p} d\epsilon,$$

valid for a constant $c_p > 0$ and any norm $\| \cdot \|$. We use this bound for $p = 1$ and $p = 2$, and the metric entropy bounds established in Lemma A.3.5.

Let us begin by establishing a bound on the γ_2 functional by writing

$$\gamma_2(\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}} \cap \mathcal{B}_{M^*}(t), \alpha \| \cdot \|_F) \le c_2 \int_0^{\alpha(t \wedge \sqrt{n_1 n_2})} \big[\log N\big(\epsilon, \mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}} \cap \mathcal{B}_{M^*}(t), \alpha \| \cdot \|_F\big)\big]^{1/2} d\epsilon.$$

We now make two observations about the metric entropy on the RHS. First, note that scaling the Frobenius norm by a factor α amounts to replacing ϵ by ϵ/α in (A.31a). Second, notice that the metric entropy is expressed as a minimum of three terms; we provide a bound for each of the terms separately, and then obtain the final bound as a minimum of the three cases.

For the first term, notice that the minimum is never attained when $\epsilon \le \alpha n_1^{-5}$, so that we have

$$\begin{aligned} \gamma_2(\mathbb{C}_{\mathsf{Perm}}^{\mathsf{r,c}} \cap \mathcal{B}_{M^*}(t), \alpha \| \cdot \|_F) &\lesssim \int_{\alpha n_1^{-5}}^{\alpha(t \wedge \sqrt{n_1 n_2})} \left\{ \frac{\alpha^2 n_1 n_2}{\epsilon^2} \Big(\log \frac{\alpha \sqrt{n_1 n_2}}{\epsilon} \Big)^2 + 2 n_1 \log n_1 \right\}^{1/2} d\epsilon \\ &\lesssim \int_{\alpha n_1^{-5}}^{\alpha \sqrt{n_1 n_2}} \frac{\alpha \sqrt{n_1 n_2}}{\epsilon} \log \frac{\alpha \sqrt{n_1 n_2}}{\epsilon} d\epsilon + \alpha t \sqrt{n_1 \log n_1} \\ &\lesssim \alpha \sqrt{n_1 n_2} (\log n_1)^2 + \alpha t \sqrt{n_1 \log n_1}. \end{aligned}$$

Now consider the second term of the bound (A.31a); we have

$$\begin{aligned}\gamma_2(\mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}} \cap \mathcal{B}_{M^*}(t), \alpha\|\cdot\|_F) &\lesssim \int_0^{\alpha(t\wedge\sqrt{n_1n_2})} \Big\{\Big(n_1n_2 \log \frac{\alpha\sqrt{n_1n_2}}{\epsilon}\Big) + 2n_1 \log n_1\Big\}^{1/2} d\epsilon \\ &\lesssim \alpha t\sqrt{n_1n_2\log(n_1/t+e)} + \alpha t\sqrt{n_1\log n_1}.\end{aligned}$$

Finally, the third term of bound (A.31a) can be used to obtain

$$\begin{aligned}\gamma_2(\mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}} \cap \mathcal{B}_{M^*}(t), \alpha\|\cdot\|_F) &\lesssim \int_0^{\alpha t} \Big\{\Big(\frac{\alpha\sqrt{n_1n_2}}{\epsilon} n_2 \log n_1\Big) + 2n_1 \log n_1\Big\}^{1/2} d\epsilon \\ &\lesssim \alpha t^{1/2} n_1^{1/4} n_2^{3/4}\sqrt{\log n_1} + \alpha t\sqrt{n_1\log n_1}.\end{aligned}$$

With the three bounds combined, we have

$$\begin{aligned}&\gamma_2(\mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}} \cap \mathcal{B}_{M^*}(t), \alpha\|\cdot\|_F) \lesssim \alpha t\sqrt{n_1\log n_1} + \\ &\Big[\alpha\sqrt{n_1n_2}(\log n_1)^2\Big] \wedge \Big[\alpha t\sqrt{n_1n_2\log(n_1/t+e)}\Big] \wedge \Big(\alpha t^{1/2}n_1^{1/4}n_2^{3/4}\sqrt{\log n_1}\Big).\end{aligned}$$

Let us now turn to bounding the γ_1 functional in ℓ_∞ norm. Scaling the $\|\cdot\|_\infty$ norm by a factor β amounts to replacing ϵ by ϵ/β in the metric entropy bound (A.32a), so we have

$$\begin{aligned}&\gamma_1(\mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}} \cap \mathcal{B}_{M^*}(t), \beta\|\cdot\|_\infty) \le c_1 \int_0^{\beta(t\wedge 1)} \log N\big(\epsilon, \mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}} \cap \mathcal{B}_{M^*}(t), \beta\|\cdot\|_\infty\big) d\epsilon \\ &\le c_1 \int_0^{\beta} \Big\{\Big[\frac{\beta n_2}{\epsilon}\log(en_1)\Big] \wedge \Big(n_1n_2\log\frac{e\beta}{\epsilon}\Big) + 2n_1\log n_1\Big\} d\epsilon \\ &\lesssim \int_0^{\beta n_1^{-5}} n_1n_2 \log\frac{\beta}{\epsilon} d\epsilon + \int_{\beta n_1^{-5}}^{\beta} \frac{\beta n_2}{\epsilon}\log n_1 d\epsilon + \beta n_1 \log n_1 \\ &\lesssim \beta n_2(\log n_1)^2 + \beta n_1\log n_1.\end{aligned}$$

Putting together the pieces with the bound (A.33), we obtain

$$\begin{aligned}g(t) &\lesssim \Big[\alpha\sqrt{n_1n_2}(\log n_1)^2\Big] \wedge \Big[\alpha t\sqrt{n_1n_2\log(n_1/t+e)}\Big] \wedge \Big(\alpha t^{1/2}n_1^{1/4}n_2^{3/4}\sqrt{\log n_1}\Big) \\ &\quad + \alpha t\sqrt{n_1\log n_1} + \beta n_2(\log n_1)^2 + \beta n_1\log n_1 + \alpha t\sqrt{u} + \beta u.\end{aligned}$$

As a result, there is a universal positive constant c such that choosing

$$\begin{aligned}t \ge t^* = c_3\Big\{\Big[\sqrt{\alpha}(n_1n_2)^{1/4}\log n_1\Big] \wedge \Big[\alpha\sqrt{n_1n_2\log(n_1/\alpha+e)}\Big] \wedge \Big[\alpha^{2/3}n_1^{1/6}n_2^{1/2}(\log n_1)^{1/3}\Big] \\ + \alpha\sqrt{n_1\log n_1} + \sqrt{\beta n_2}\log n_1 + \sqrt{\beta n_1\log n_1} + \alpha\sqrt{u} + \sqrt{\beta u}\Big\}\end{aligned} \tag{A.34}$$

yields the bound $g(t) < t^2/8$. This holds for each *individual* $t \geq t^*$ with probability $1 - e^{-u}$. We now prove that $f(t) < 0$ *simultaneously* for all $t \geq t^*$ with high probability. Note that typically, the star-shaped property of the set suffices to provide such a bound, but we include the full proof for completeness.

To this end, we first note that by assumption,

$$\Pr\{|W_{i,j}| \geq \alpha\sqrt{u} + \beta u\} \leq 2e^{-u}.$$

A union bound then implies that

$$\Pr\left\{\|W\|_F \geq (\alpha\sqrt{u} + \beta u)\sqrt{n_1 n_2}\right\} \leq 2n_1 n_2 e^{-u}.$$

Therefore, we have with probability at least $1 - 2n_1 n_2 e^{-u}$ that

$$g(t) \leq t\|W\|_F \leq t(\alpha\sqrt{u} + \beta u)\sqrt{n_1 n_2}$$

simultaneously for all $t \geq 0$. On this event, it holds that $f(t) < 0$ for all $t \geq t^{\#} = 3(\alpha\sqrt{u} + \beta u)\sqrt{n_1 n_2}$.

For $t \in [t^*, t^{\#}]$, we employ a discretization argument (clearly, we can assume $t^{\#} \geq t^*$ without loss of generality). Let $T = \{t_1, \ldots, t_k\}$ be a discretization of the interval $[t^*, t^{\#}]$ such that $t^* = t_1 < \cdots < t_k = t^{\#}$ and $2t_1 \geq t_2$. Note that T can be chosen so that

$$|T| = k \leq \log_2 \frac{t^{\#}}{t^*} + 1 \leq \log_2\left((3 + 3\sqrt{\beta u})\sqrt{n_1 n_2}\right) + 1 \leq 7\log(n_1 u),$$

where we used the assumption that $\beta \leq n_1^2$. Using the high probability bound $g(t) < t^2/8$ for each individual $t \geq t^*$ and a union bound over T, we obtain that with probability at least $1 - 7\log(n_1 u)e^{-u}$,

$$\max_{t \in T} g(t) - t^2/8 < 0.$$

On this event, we use the fact that $g(t)$ is non-decreasing and that $t_i \geq t_{i+1}/2$ to conclude that for each $t \in [t_i, t_{i+1}]$ where $i \in [k-1]$, we have

$$f(t) = g(t) - t^2/2 \leq g(t_{i+1}) - t_i^2/2 \leq g(t_{i+1}) - t_{i+1}^2/8 \leq \max_{t \in T} g(t) - t^2/8 < 0.$$

In summary, we obtain that $f(t) < 0$ for all $t \geq t^*$ simultaneously with probability at least $1 - 2n_1 n_2 e^{-u} - 7\log(n_1 u)e^{-u}$. Choosing $u = 4n_1 \log n_1$, recalling the definition of t^* in (A.34) and applying Lemma A.3.3, we conclude that with probability at least $1 - n_1^{-3n_1}$,

$$\begin{aligned}&\left\|\widehat{M}_{\mathsf{LS}}(\mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}}, Y) - M^*\right\|_F^2 \leq (t^*)^2 \lesssim \\ &\left[\alpha\sqrt{n_1 n_2}(\log n_1)^2\right] \wedge \left[\alpha^2 n_1 n_2 \log(n_1/\alpha + e)\right] \wedge \left[\alpha^{4/3} n_1^{1/3} n_2 (\log n_1)^{2/3}\right] \\ &+ \alpha^2 n_1 \log n_1 + \beta n_2 (\log n_1)^2 + \beta n_1 \log n_1.\end{aligned}$$

The entire argument can be repeated for the class $\mathbb{C}_{\mathsf{BISO}}$, in which case terms of the order $n_1 \log n_1$ disappear as there is no latent permutation. Since the argument is analogous, we omit the details. This completes the proof of the lemma. □

Proof of Lemma A.3.5 Note that $\sqrt{n_1 n_2}$ and 1 are the diameters of $\mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}}$ in ℓ_2 and ℓ_∞ norms respectively, so we can assume $\epsilon \leq \sqrt{n_1 n_2}$ or 1 in the two cases.

For the ℓ_2 metric entropy, Lemma 3.4 of [56] yields

$$\log N(\epsilon; \mathbb{C}_{\mathsf{BISO}}, \|\cdot\|_F) \leq c_1 \frac{n_1 n_2}{\epsilon^2} \Big(\log \frac{\sqrt{n_1 n_2}}{\epsilon} \Big)^2,$$

which is the first term of (A.31a). In addition, since $\mathbb{C}_{\mathsf{BISO}}$ is contained in the ball in $\mathbb{R}^{n_1 \times n_2}$ of radius $\sqrt{n_1 n_2}$ centered at zero, we have the simple bound

$$\log N(\epsilon; \mathbb{C}_{\mathsf{BISO}}, \|\cdot\|_F) \leq c_2 n_1 n_2 \log \frac{\sqrt{n_1 n_2}}{\epsilon},$$

which is the second term of (A.31a).

Moreover, any matrix in $\mathbb{C}_{\mathsf{BISO}}$ has Frobenius norm bounded by $\sqrt{n_1 n_2}$ and has non-decreasing columns so $\mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}}$ is a subset of the class of matrices considered in Lemma 6.7 of the paper [104] with $A = 0$ and $t = \sqrt{n_1 n_2}$ (see also equation (6.9) of the paper for the notation). The aforementioned lemma yields the bound

$$\log N(\epsilon; \mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}}, \|\cdot\|_F) \leq c_3 \frac{\sqrt{n_1 n_2}}{\epsilon} n_2 \log n_1.$$

Taking the minimum of the three bounds above yields the ℓ_2 metric entropy bound (A.31b) on the class of bivariate isotonic matrices. Since $\mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}}$ is a union of $n_1! n_2!$ permuted versions of $\mathbb{C}_{\mathsf{BISO}}$, combining this bound with an additive term $2n_1 \log n_1$ provides the bound for the class $\mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}}$.

For the ℓ_∞ metric entropy, we again start with $\mathbb{C}_{\mathsf{BISO}}$. Let us define a discretization $I_\epsilon = \{0, \epsilon, 2\epsilon, \dots, \lfloor 1/\epsilon \rfloor \epsilon\}$ and a set of matrices

$$\mathcal{Q} = \big\{ M \in \mathbb{C}_{\mathsf{BISO}} : M_{i,j} \in I_\epsilon \big\},$$

which is a discretized version of $\mathbb{C}_{\mathsf{BISO}}$. We claim that $\mathcal{Q}$ is an ϵ-net of $\mathbb{C}_{\mathsf{BISO}}$ in the ℓ_∞ norm. Indeed, for any $M \in \mathbb{C}_{\mathsf{BISO}}$, we can define a matrix M' by setting

$$M'_{i,j} = \operatorname*{argmin}_{a \in I_\epsilon} |a - M_{i,j}|$$

with the convention that if $M_{i,j} = (k + 0.5)\epsilon$ for an integer $0 \leq k < \lfloor 1/\epsilon \rfloor$, then we set $M'_{i,j} = (k+1)\epsilon$. It is not hard to see that $M' \in \mathcal{Q}$ and moreover $\|M' - M\|_\infty \leq \epsilon$. Therefore the claim is established.

It remains to bound the cardinality of $\mathcal{Q}$. Since each column of $\mathcal{Q}$ is non-decreasing and takes values in I_ϵ having cardinality $\lfloor 1/\epsilon \rfloor + 1$, it is well known (by a "stars and bars" argument) that the number of possible choices for each column of a matrix in $\mathcal{Q}$ can be bounded as

$$\binom{n_1 + \lfloor 1/\epsilon \rfloor}{\lfloor 1/\epsilon \rfloor} \leq \Big(e \frac{n_1 + \lfloor 1/\epsilon \rfloor}{\lfloor 1/\epsilon \rfloor} \Big)^{\lfloor 1/\epsilon \rfloor} \wedge \Big(e \frac{n_1 + \lfloor 1/\epsilon \rfloor}{n_1} \Big)^{n_1},$$

where we used the bound $\binom{n}{k} \leq (\frac{en}{k})^k$ for any $0 \leq k \leq n$. Since a matrix in $\mathcal{Q}$ has n_2 columns, we obtain

$$\log|\mathcal{Q}| \leq n_2 \log\binom{n_1 + \lfloor 1/\epsilon \rfloor}{\lfloor 1/\epsilon \rfloor} \leq \Big[\frac{n_2}{\epsilon}\log(en_1)\Big] \wedge \Big(n_1 n_2 \log\frac{e}{\epsilon}\Big).$$

This bounds $\log N(\epsilon; \mathbb{C}_{\mathsf{BISO}}, \|\cdot\|_\infty)$ and the same argument as before yields the bound for $\mathbb{C}^{\mathsf{r,c}}_{\mathsf{Perm}}$. □

A.4 Technical lemmas used in Chapter 3

We now proceed to the technical lemmas used to prove Theorems 3.4.1 and 3.4.2.

Proof of Lemma 3.5.3

We split the proof of Lemma 3.5.3 into two cases.

Case 1 First, suppose that $|B| \geq \frac{n_1 n_2}{N}\log(n_1 n_2)$. In view of the condition

$$\sum_{\ell \in B} M^*_{v,\ell} - \sum_{\ell \in B} M^*_{u,\ell} > 2\eta_B$$

and the definition of a topological sort, Lemma A.6.5 (which is stated and proved shortly) with $a_i = \sum_{\ell \in B} M^*_{i,\ell}$, $b = a$, $\pi = \widehat{\pi}$ and $\tau = 2\eta_B$ yields

$$\Big|\sum_{\ell \in B}(M^*_{\widehat{\pi}(i),\ell} - M^*_{i,\ell})\Big| \leq 2\eta_B \leq 96(\zeta+1)\sqrt{\frac{n_1 n_2}{N}|B|\log(n_1 n_2)},$$

for all $i \in [n_1]$.

Case 2 Otherwise, we have $|B| \leq \frac{n_1 n_2}{N}\log(n_1 n_2)$. It then follows that

$$\Big|\sum_{\ell \in B}(M^*_{\widehat{\pi}(i),\ell} - M^*_{i,\ell})\Big| \leq 2|B| \leq 2\sqrt{\frac{n_1 n_2}{N}|B|\log(n_1 n_2)},$$

where we have used the fact that $M \in [0,1]^{n_1 \times n_2}$.

Since the columns of M^* are all non-decreasing, we have

$$\sum_{j \in B}\big|M^*_{\widehat{\pi}(i),j} - M^*_{i,j}\big| = \Big|\sum_{j \in B}(M^*_{\widehat{\pi}(i),j} - M^*_{i,j})\Big|,$$

so the proof is complete. □

Proof of Lemma 3.5.4

Let $a = \min_{i\in[n]} v_i$ and $b = \max_{i\in[n]} v_i = a + \mathrm{var}(v)$. Since the quantities in the inequality remain the same if we replace v by $-v$, we assume without loss of generality that $b \geq 0$. If $a \leq 0$, then $\|v\|_\infty \leq b - a = \mathrm{var}(v)$. If $a > 0$, then $a \leq \|v\|_1/n$ and $\|v\|_\infty = b \leq \|v\|_1/n + \mathrm{var}(v)$. Hence, in any case we have $\|v\|_2^2 \leq \|v\|_\infty \|v\|_1 \leq [\|v\|_1/n + \mathrm{var}(v)]\|v\|_1$. □

Proof of Lemma 3.5.5

Since A has increasing rows, for any $i, i_2 \in [n]$ with $i \leq i_2$ and any $j, j_2 \in J_k$, we have

$$\begin{aligned} A_{i_2,j} - A_{i,j} &= (A_{i_2,j} - A_{i_2,a_k}) + (A_{i_2,a_k} - A_{i,b_k}) + (A_{i,b_k} - A_{i,j}) \\ &\leq (A_{i_2,b_k} - A_{i_2,a_k}) + (A_{i_2,j_2} - A_{i,j_2}) + (A_{i,b_k} - A_{i,a_k}). \end{aligned}$$

Choosing $j_2 = \arg\min_{r\in J_k}(A_{i_2,r} - A_{i,r})$, we obtain

$$A_{i_2,j} - A_{i,j} \leq (A_{i_2,b_k} - A_{i_2,a_k}) + (A_{i,b_k} - A_{i,a_k}) + \frac{1}{m_k}\sum_{r\in J_k}(A_{i_2,r} - A_{i,r}).$$

Together with the assumption on π, this implies that

$$|A_{\pi(i),j} - A_{i,j}| \leq \underbrace{A_{\pi(i),b_k} - A_{\pi(i),a_k}}_{x_{i,k}} + \underbrace{A_{i,b_k} - A_{i,a_k}}_{y_{i,k}} + \frac{1}{m_k}\underbrace{\sum_{r\in J_k}|A_{\pi(i),r} - A_{i,r}|}_{z_{i,k}}.$$

Hence, it follows that

$$\begin{aligned} \sum_{i=1}^n\sum_{j=1}^m (A_{i,j} - A_{\pi(i),j})^2 &= \sum_{i=1}^n\sum_{k=1}^\ell\sum_{j\in J_k}(A_{i,j} - A_{\pi(i),j})^2 \\ &\leq \sum_{i=1}^n\sum_{k=1}^\ell\sum_{j\in J_k}|A_{i,j} - A_{\pi(i),j}|(x_{i,k} + y_{i,k} + z_{i,k}/m_k) \\ &= \sum_{i=1}^n\sum_{k=1}^\ell z_{i,k}(x_{i,k} + y_{i,k} + z_{i,k}/m_k). \end{aligned}$$

According to the assumptions, we have

1. $\sum_{k=1}^\ell x_{i,k} \leq 1$ and $\sum_{i=1}^n x_{i,k} \leq \chi$ for any $i \in [n], k \in [\ell]$;

2. $\sum_{k=1}^\ell y_{i,k} \leq 1$ and $\sum_{i=1}^n y_{i,k} \leq \chi$ for any $i \in [n], k \in [\ell]$;

3. $z_{i,k} \leq \rho_k$ and $\sum_{k=1}^\ell z_{i,k} \leq \rho$ for any $i \in [n], k \in [\ell]$.

Consequently, the following bounds hold:

1. $\sum_{i=1}^n \sum_{k=1}^\ell z_{i,k} x_{i,k} \leq \sum_{i=1}^n \sum_{k=1}^\ell \rho_k x_{i,k} \leq \chi \sum_{k=1}^\ell \rho_k$;
2. $\sum_{i=1}^n \sum_{k=1}^\ell z_{i,k} y_{i,k} \leq \sum_{i=1}^n \sum_{k=1}^\ell \rho_k y_{i,k} \leq \chi \sum_{k=1}^\ell \rho_k$;
3. $\sum_{i=1}^n \sum_{k=1}^\ell z_{i,k}^2 / m_k \leq \sum_{i=1}^n \sum_{k=1}^\ell z_{i,k} \cdot \max_{k \in [\ell]} (\rho_k / m_k) \leq n\rho \max_{k \in [\ell]} (\rho_k / m_k)$.

Combining these inequalities yields the claim. □

A.5 Technical results on the isotonic projection used in Chapter 4

In this section, we collect some technical results on the isotonic projection onto piecewise constant hyper-rectangular partitions. This is the operator given by $\mathfrak{B}(\,\cdot\,; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)$, which was defined in equation (4.41). Let us begin by defining some other helpful notation. Let $\mathcal{C}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)$ denote the set of all tensors in $\mathbb{R}_{d,n}$ that are piecewise constant on the d-dimensional blocks specified by the Cartesian products of one-dimensional partitions $\mathsf{bl}_1, \ldots, \mathsf{bl}_d$. Define the operators $\mathcal{P} : \mathbb{R}_{d,n} \to \mathbb{R}_{d,n}$ and $\mathcal{A} : \mathbb{R}_{d,n} \to \mathbb{R}_{d,n}$ as projection operators onto the sets $\mathcal{M}(\mathbb{L}_{d,n}; \pi_1, \ldots, \pi_d)$ and $\mathcal{C}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)$, respectively, i.e., for each $\theta \in \mathbb{R}_{d,n}$, we have

$$\mathcal{P}(\theta; \pi_1, \ldots, \pi_d) \in \underset{\theta' \in \mathcal{M}(\mathbb{L}_{d,n}; \pi_1, \ldots, \pi_d)}{\operatorname{argmin}} \|\theta - \theta'\|_2^2, \text{ and}$$
$$\mathcal{A}(\theta; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) \in \underset{\theta' \in \mathcal{C}(\mathbb{L}_{d,n}; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)}{\operatorname{argmin}} \|\theta - \theta'\|_2^2.$$

Recall our notion of a permutation that is faithful to a one-dimensional ordered partition from the proof of Lemma 4.6.4(b). Finally, let $\overline{\theta}_S$ denote the average of the entries of $\theta \in \mathbb{R}_{d,n}$ on the set $S \subseteq \mathbb{L}_{d,n}$.

Our first technical lemma demonstrates that the operator $\mathfrak{B}$ can be written as a composition of the operators $\mathcal{P}$ and $\mathcal{A}$, i.e., in order to project onto the class of isotonic tensors that are piecewise constant on hyper-rectangular blocks given by a d-dimensional ordered partition, it suffices to first average all entries within each block, and then project the result onto the class of isotonic tensors whose partial orderings are consistent with the corresponding one-dimensional ordered partitions.

Lemma A.5.1 (Composition). *For each $j \in [d]$, let $\pi_j \in \mathfrak{S}_{n_1}$ be any permutation that is faithful to the ordered partition bl_j. Then, for each $\theta \in \mathbb{R}_{d,n}$, we have*

$$\mathfrak{B}(\theta; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) = \mathcal{P}(\ \mathcal{A}(\theta; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)\ ; \pi_1, \ldots, \pi_d).$$

Proof. To fix notation, suppose that bl_j is a partition of $[n_j]$ into s_j blocks, and that $\prod_{j=1}^d s_j = s$. Note that the ordered partitions $\mathsf{bl}_1, \ldots, \mathsf{bl}_d$ induce a hyper-rectangular partition of the lattice $\mathbb{L}_{d,n}$ into s pieces. Index each of these hyper-rectangles by the corresponding member of the smaller lattice $\mathbb{L}_{d,s_1,\ldots,s_d}$, and for each $x \in \mathbb{L}_{d,s_1\ldots,s_d}$, let $B_x \subseteq \mathbb{L}_{d,n}$ denote the indices of hyper-rectangle x. With this notation, the projection operator for any $\theta \in \mathbb{R}_{d,n}$ takes the form

$$\mathfrak{B}(\theta; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) \in \underset{\mu \in \mathcal{M}(\mathbb{L}_{d,s_1,\ldots,s_d})}{\operatorname{argmin}} \sum_{x \in \mathbb{L}_{d,s_1,\ldots,s_d}} \sum_{w \in B_x} (\theta_w - \mu_x)^2 .$$

The inner sum in the objective can be written as

$$\begin{aligned}\sum_{w \in B_x} (\theta_w - \mu_x)^2 &= \sum_{w \in B_x} \left(\theta_w - \overline{\theta}_{B_x}\right)^2 + \sum_{w \in B_x} \left(\overline{\theta}_{B_x} - \mu_x\right)^2 + 2\left(\overline{\theta}_{B_x} - \mu_x\right) \sum_{w \in B_x} \left(\theta_w - \overline{\theta}_{B_x}\right) \\ &= \sum_{w \in B_x} \left(\theta_w - \overline{\theta}_{B_x}\right)^2 + |B_x| \cdot \left(\overline{\theta}_{B_x} - \mu_x\right)^2, \qquad \text{(A.35)}\end{aligned}$$

where equation (A.35) follows since $\sum_{w \in B_x} \left(\theta_w - \overline{\theta}_{B_x}\right) = 0$. Putting together the pieces and noting that the first term of inequality (A.35) does not depend on μ, we have

$$\mathfrak{B}(\theta; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) \in \underset{\mu \in \mathcal{M}(\mathbb{L}_{d,s_1,\ldots,s_d})}{\operatorname{argmin}} \sum_{x \in \mathbb{L}_{d,s_1,\ldots,s_d}} |B_x| \cdot \left(\overline{\theta}_{B_x} - \mu_x\right)^2 .$$

The proof is completed by noting that $\mathcal{A}(\theta; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)$ is equal to $\overline{\theta}_{B_x}$ on each block B_x, and so the optimization problem above can be viewed as the projection of $\mathcal{A}(\theta; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)$ onto any set $\mathcal{M}(\mathbb{L}_{d,n}; \pi_1, \ldots, \pi_d)$ such that the permutations $\pi_1, \ldots, \pi_d$ are faithful to the ordered partitions $\mathsf{bl}_1, \ldots, \mathsf{bl}_d$, respectively. □

Lemma A.5.2 (ℓ_∞-contraction)**.** *For each $\theta, \theta' \in \mathbb{R}_{d,n}$, ordered partitions $\mathsf{bl}_1, \ldots, \mathsf{bl}_d$, and permutations $\pi_1, \ldots, \pi_d$, we have*

$$\|\mathcal{A}(\theta; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) - \mathcal{A}(\theta'; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)\|_\infty \leq \|\theta - \theta'\|_\infty \textit{ and} \qquad \text{(A.36a)}$$

$$\|\mathcal{P}(\theta; \pi_1, \ldots, \pi_d) - \mathcal{P}(\theta'; \pi_1, \ldots, \pi_d)\|_\infty \leq \|\theta - \theta'\|_\infty. \qquad \text{(A.36b)}$$

Consequently,

$$\|\mathfrak{B}(\theta; \mathsf{bl}_1, \ldots, \mathsf{bl}_d) - \mathfrak{B}(\theta'; \mathsf{bl}_1, \ldots, \mathsf{bl}_d)\|_\infty \leq \|\theta - \theta'\|_\infty. \qquad \text{(A.36c)}$$

Proof. Owing to Lemma A.5.1, equation (A.36c) follows directly from equations (A.36a) and (A.36b). Equation (A.36a) is also immediate, since the operator $\mathcal{A}$ simply averages entries within each partition, and the averaging operation is trivially ℓ_∞-contractive.

The proof of equation (A.36b) is slightly more involved. First, since the ℓ_∞ norm is invariant to the labeling of the entries of the tensor, it suffices to establish the result when $\pi_j = \mathsf{id}$ for all $j \in [d]$. We use the notation $\mathcal{P}(\cdot) := \mathcal{P}(\,\cdot\,; \mathsf{id}, \ldots, \mathsf{id})$, for convenience. For each $x \in \mathbb{L}_{d,n}$, let $\mathcal{L}(x)$ and $\mathcal{U}(x)$ denote the collections of lower and upper sets containing x, respectively. Recall the min-max characterization of the isotonic projection [264, Chapter 1]: For each tensor $a \in \mathbb{R}_{d,n}$ and $x \in \mathbb{L}_{d,n}$, we have

$$\mathcal{P}(a)(x) = \min_{L \in \mathcal{L}(x)} \max_{U \in \mathcal{U}(x)} \overline{a}_{L \cap U}.$$

Consequently, for each pair of tensors $a, b \in \mathbb{R}_{d,n}$, we obtain the sequence of bounds

$$\begin{aligned}|\mathcal{P}(a)(x) - \mathcal{P}(b)(x)| &= \left| \min_{L \in \mathcal{L}(x)} \max_{U \in \mathcal{U}(x)} \overline{a}_{L \cap U} - \min_{L \in \mathcal{L}(x)} \max_{U \in \mathcal{U}(x)} \overline{b}_{L \cap U} \right| \\ &\leq \max_{L \in \mathcal{L}(x)} \left| \max_{U \in \mathcal{U}(x)} \overline{a}_{L \cap U} - \max_{U \in \mathcal{U}(x)} \overline{b}_{L \cap U} \right| \\ &\leq \max_{L \in \mathcal{L}(x)} \max_{U \in \mathcal{U}(x)} |\overline{a}_{L \cap U} - \overline{b}_{L \cap U}| \\ &\leq \|a - b\|_\infty.\end{aligned}$$

Since this holds for all $x \in \mathbb{L}_{d,n}$, we have proved the claimed result. □

As an immediate corollary of equation (A.36b), we obtain the following result that may be of independent interest.

Corollary A.5.1. *The isotonic projection is ℓ_∞ contractive, i.e., for any $\theta, \theta' \in \mathbb{R}_{d,n}$, we have*

$$\|\mathcal{P}(\theta; \mathsf{id}, \ldots, \mathsf{id}) - \mathcal{P}(\theta'; \mathsf{id}, \ldots, \mathsf{id})\|_\infty \leq \|\theta - \theta'\|_\infty.$$

To the best of our knowledge, similar results are only known when $d = 1$ [338].

A.6 Auxiliary results for Chapter 4

In this section, we collect several results that are used in multiple proofs.

A.6.1 Basic lemmas for least squares estimators

Our first lemma allows us to bound the expected supremum of a Gaussian process over a union of sets in terms of the individual expected suprema. Similar results have appeared in the literature [57, 130]. We state a version that can be readily deduced from [130, Lemma D.1].

Lemma A.6.1. *Let $K \geq 1$, and let ϵ denote a standard Gaussian tensor in $\mathbb{R}_{d,n}$. Suppose that for some positive scalar t, we have $\Theta_1, \ldots, \Theta_K \subseteq \mathbb{B}_2(t)$. There is a universal positive constant C such that*
(a) The supremum of the empirical process satisfies

$$\Pr\left\{ \max_{k \in [K]} \sup_{\theta \in \Theta_k} \langle \epsilon, \theta \rangle \geq \max_{k \in [K]} \mathbb{E}\left[\sup_{\theta \in \Theta_k} \langle \epsilon, \theta \rangle \right] + Ct(\sqrt{\log K} + \sqrt{u}) \right\} \leq e^{-u} \quad \textit{for each } u \geq 0.$$

(b) If, in addition, the all-zero tensor is contained in each individual set $\{\Theta_k\}_{k=1}^K$, then

$$\mathbb{E}\left[\max_{k \in [K]} \sup_{\theta \in \Theta_k} \langle \epsilon, \theta \rangle \right] \leq \max_{k \in [K]} \mathbb{E}\left[\sup_{\theta \in \Theta_k} \langle \epsilon, \theta \rangle \right] + Ct\sqrt{\log K}.$$

Our second lemma bounds the supremum of a Gaussian process over a set that is piecewise constant over known blocks. Recall that for $\theta \in \mathbb{R}_{d,n}$ and $S \subseteq \mathbb{L}_{d,n}$, we let θ_S denote the sub-tensor formed by restricting θ to indices in S. In the statement of the lemma, we also use the notation of stochastic dominance: for a pair of scalar random variables (X_1, X_2), the relation $X_1 \overset{d}{\leq} X_2$ means that $\Pr\{X_1 \geq t\} \leq \Pr\{X_2 \geq t\}$ for each $t \in \mathbb{R}$.

Lemma A.6.2. *Let $B_1, \ldots, B_s$ denote a (known) partition of the lattice $\mathbb{L}_{d,n}$. Let $\Theta \subseteq \mathbb{R}_{d,n}$ denote a collection of tensors such that for each $\theta \in \Theta$ and $\ell \in [s]$, the sub-tensor θ_{B_ℓ} is constant. Let $\epsilon \in \mathbb{R}_{d,n}$ represent a standard Gaussian tensor. Then, for each $t \geq 0$, we have*

$$\sup_{\theta \in \Theta \cap \mathbb{B}_2(t)} \langle \epsilon, \theta \rangle \overset{d}{\leq} t \cdot Y_s,$$

where $Y_s^2 \sim \chi_s^2$. Consequently, we have

$$\mathbb{E} \sup_{\theta \in \Theta \cap \mathbb{B}_2(t)} \langle \epsilon, \theta \rangle \leq t\sqrt{s}.$$

Proof. We focus on proving the first claim, since the second claim follows immediately from it by Jensen's inequality. For each $S \subseteq \mathbb{L}_{d,n}$, we write $\overline{\theta}_S := \frac{1}{|S|} \sum_{x \in S} \theta_x$. For each $\theta \in \Theta$, we have the decomposition

$$\langle \epsilon, \theta \rangle = \sum_{\ell \in [s]} \sum_{x \in B_\ell} \epsilon_x \cdot \theta_x = \sum_{\ell \in [s]} \sqrt{|B_\ell|} \cdot \overline{\theta}_{B_\ell} \cdot \frac{\sum_{x \in B_\ell} \epsilon_x}{\sqrt{|B_\ell|}}.$$

Now define the s-dimensional vectors $\widetilde{\epsilon}$ and $v(\theta)$ via

$$\widetilde{\epsilon}_\ell := \frac{\sum_{x \in B_\ell} \epsilon_x}{\sqrt{|B_\ell|}} \quad \text{and} \quad [v(\theta)]_\ell := \sqrt{|B_\ell|} \cdot \overline{\theta}_{B_\ell}, \quad \text{for each } \ell \in [s].$$

By construction, the vector $\widetilde{\epsilon}$ consists of standard Gaussian entries, and we also have $\|v(\theta)\|_2 = \|\theta\|_2$ for each $\theta \in \Theta$. Combining the pieces with Cauchy–Schwarz inequality yields

$$\sup_{\theta \in \Theta \cap \mathbb{B}_2(t)} \langle \epsilon, \theta \rangle \leq \sup_{\substack{v \in \mathbb{R}^s \\ \|v\|_2 \leq t}} \langle \widetilde{\epsilon}, v \rangle \leq t \cdot \|\widetilde{\epsilon}\|_2,$$

as desired. □

Our third lemma follows almost directly from [323, Theorem 13.5], after a little bit of algebraic manipulation. In order to state the lemma, we require a few preliminaries.

Definition A.6.1. *A set $\mathcal{C}$ is star-shaped if for all $\theta \in \mathcal{C}$ and $\alpha \in [0, 1]$, the inclusion $\alpha\theta \in \mathcal{C}$ holds. We say that $\mathcal{C}$ is additionally non-degenerate if it does not consist solely of the zero element.*

Let ϵ denote a standard Gaussian in $\mathbb{R}_{d,n}$, and suppose that the set $\Theta \subseteq \mathbb{R}_{d,n}$ is star-shaped and non-degenerate. Let $\widehat{\Delta} \in \mathbb{R}_{d,n}$ denote a (random) tensor satisfying the pointwise inequality

$$\|\widehat{\Delta}\|_2^2 \leq \sup_{\substack{\Delta \in \Theta \\ \|\Delta\|_2 \leq \|\widehat{\Delta}\|_2}} \langle \epsilon, \Delta \rangle,$$

and for each $t \geq 0$, define the random variable

$$\xi(t) = \sup_{\substack{\Delta \in \Theta \\ \|\Delta\|_2 \leq t}} \langle \epsilon, \Delta \rangle.$$

Let t_n denote the smallest positive solution to the critical inequality

$$\mathbb{E}[\xi(t)] \leq \frac{t^2}{2}.$$

Such a solution always exists provided Θ is star-shaped and non-degenerate; this can be shown via a standard rescaling argument (see [323, Lemma 13.6]).

We are now ready to state a high probability bound on the error $\|\widehat{\Delta}\|_2^2$.

Lemma A.6.3. *Under the setup above, there is a pair of universal positive constants (c, C) such that*

$$\Pr\left\{\|\widehat{\Delta}\|_2^2 \geq Ct_n^2 + u\right\} \leq \exp\{-cu\} \quad \textit{for all} \quad u \geq 0.$$

Consequently,

$$\mathbb{E}[\|\widehat{\Delta}\|_2^2] \leq C(t_n^2 + 1).$$

Proof. Applying [323, Theorem 13.5] and rescaling appropriately yields the bound

$$\Pr\left\{\|\widehat{\Delta}\|_2^2 \geq 16t_n \cdot \delta\right\} \leq \exp\left\{-\frac{\delta t_n}{2}\right\} \quad \text{for all} \quad \delta \geq t_n.$$

Now note that $t_n > 0$, and that for any $u \geq 0$, we may set $\delta = t_n + \frac{u}{16t_n}$. This yields the bound

$$\Pr\left\{\|\widehat{\Delta}\|_2^2 \geq 16t_n^2 + u\right\} \leq \exp\left\{-\frac{t_n^2}{2}\right\} \cdot \exp\left\{-\frac{u}{32}\right\} \leq \exp\left\{-\frac{u}{32}\right\}.$$

The bound on the expectation follows straightforwardly by integrating the tail bound. □

Our fourth lemma is an immediate corollary of [311, Theorem 2.1], and shows that the error of a least squares estimator—recall our notation from equation (4.5)—over a convex set concentrates around its expected value.

Lemma A.6.4. *Let ϵ denote a standard Gaussian tensor in $\mathbb{R}_{d,n}$, and let $K \subseteq \mathbb{R}_{d,n}$ denote a closed convex set. For a fixed tensor $\theta^* \in K$, let $\widehat{\theta} = \widehat{\theta}_{\mathsf{LSE}}(K, \theta^* + \epsilon)$. Then for each $u \geq 0$:*
(a) The ℓ_2 norm of the error satisfies the two-sided tail bound

$$\Pr\left\{\left|\|\widehat{\theta} - \theta^*\|_2 - \mathbb{E}[\|\widehat{\theta} - \theta^*\|_2]\right| \geq \sqrt{2u}\right\} \leq e^{-u}.$$

(b) The squared ℓ_2 norm of the error satisfies the one-sided tail bound

$$\Pr\left\{\|\widehat{\theta} - \theta^*\|_2^2 \geq 2\mathbb{E}[\|\widehat{\theta} - \theta^*\|_2^2] + 4u\right\} \leq e^{-u}.$$

Proof. Part (a) of the lemma follows directly by rescaling the terms in [311, Theorem 2.1]. Part (b) of the lemma follows from part (a) by noting that if $\|\widehat{\theta} - \theta^*\|_2 - \mathbb{E}[\|\widehat{\theta} - \theta^*\|_2]| \leq \sqrt{2u}$, then

$$\|\widehat{\theta} - \theta^*\|_2^2 \leq 2\left(\mathbb{E}[\|\widehat{\theta} - \theta^*\|_2]\right)^2 + 2 \cdot 2u \leq 2\mathbb{E}[\|\widehat{\theta} - \theta^*\|_2^2] + 4u.$$

□

A.6.2 Some other useful lemmas

We first state a useful (deterministic) lemma regarding permutations, which generalizes [211, Lemma A.10].

Lemma A.6.5. *Let $\{a_i\}_{i=1}^n$ be a non-decreasing sequence of real numbers, let $\{b_i\}_{i=1}^n$ be a sequence of real numbers, and let τ be a positive scalar. If π is a permutation in $\mathfrak{S}_n$ such that $\pi(i) < \pi(j)$ whenever $b_j - b_i > \tau$, then $|a_{\pi(i)} - a_i| \leq \tau + 2\|b - a\|_\infty$ for all $i \in [n]$. Here, we have defined the vectors $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_n)$.*

Proof. The proof is by contradiction. Letting $\Delta = b - a$, assume that $a_j - a_{\pi(j)} > \tau + 2\|\Delta\|_\infty$ for some index $j \in [n]$. Since π is a bijection, there must exist—by the pigeonhole principle—an index $i \leq \pi(j)$ such that $\pi(i) \geq \pi(j)$. Hence, we have

$$b_j - b_i = a_j - a_i + \Delta_j - \Delta_i \geq a_j - a_{\pi(j)} + \Delta_j - \Delta_i > \tau + 2\|\Delta\|_\infty - 2\|\Delta\|_\infty,$$

which contradicts the assumption that $\pi(i) < \pi(j)$ whenever $b_j - b_i > \tau$.

On the other hand, suppose that $a_{\pi(j)} - a_j > \tau + 2\|\Delta\|_\infty$ for some $j \in [n]$. Since π is a bijection, there must exist an index $i \geq \pi(j)$ such that $\pi(i) \leq \pi(j)$. In this case, we have

$$b_i - b_j = a_i - a_j + \Delta_i - \Delta_j \geq a_{\pi(j)} - a_j + \Delta_i - \Delta_j > \tau,$$

which also leads to a contradiction. □

Our next technical lemma is a basic result about random variables.

Lemma A.6.6. *Let (X, Y, Z) denote a triple of real-valued random variables defined on a common probability space, with $X^2 \leq Z^2$ almost surely. Let $\mathcal{E}$ be a measurable event such that on the event $\mathcal{E}$, we have $X^2 \leq Y^2$. Then,*

$$\mathbb{E}[X^2] \leq \mathbb{E}[Y^2] + \sqrt{\mathbb{E}[Z^4]} \cdot \sqrt{\Pr\{\mathcal{E}^c\}}.$$

Proof. Since $X^2 \leq Y^2$ on $\mathcal{E}$, we have $X^2 \leq Y^2 \mathbf{1}\{\mathcal{E}\} + X^2 \mathbf{1}\{\mathcal{E}^c\}$. Consequently,

$$\begin{aligned}\mathbb{E}[X^2] &\leq \mathbb{E}[Y^2 \mathbf{1}\{\mathcal{E}\}] + \mathbb{E}[X^2 \mathbf{1}\{\mathcal{E}^c\}] \\ &\leq \mathbb{E}[Y^2] + \mathbb{E}[Z^2 \mathbf{1}\{\mathcal{E}^c\}] \\ &\leq \mathbb{E}[Y^2] + \sqrt{\mathbb{E}[Z^4]} \cdot \sqrt{\mathbb{E}[\mathbf{1}\{\mathcal{E}^c\}]},\end{aligned}$$

where the final inequality is an application of Cauchy–Schwarz inequality. □

Our third lemma is an elementary type of rearrangement inequality. For a permutation $\pi \in \mathfrak{S}_n$ and vector $v \in \mathbb{R}^n$, we use the notation $v\{\pi\}$ to denote the vector formed by permuting the entries of v according to π, so that $v\{\pi\} = (v_{\pi(1)}, \ldots, v_{\pi(n)})$. Let $\mathbf{1}_n$ denote the n-dimensional all-ones vector.

Lemma A.6.7. *Let $v \in \mathbb{R}^n$ with $\overline{v} = \left(\frac{1}{n}\sum_{i=1}^n v_i\right) \cdot \mathbf{1}_n$. Then, we have*

$$\|v - \overline{v}\|_2^2 \leq \max_{\pi \in \mathfrak{S}_n} \|v - v\{\pi\}\|_2^2.$$

Proof. First note that $\overline{v} = \frac{1}{|\mathfrak{S}_n|}\sum_{\pi \in \mathfrak{S}_n} v\{\pi\}$, so that we have

$$\|v - \overline{v}\|_2^2 = \left\| v - \frac{1}{|\mathfrak{S}_n|}\sum_{\pi \in \mathfrak{S}_n} v\{\pi\} \right\|_2^2 \overset{\text{(i)}}{\leq} \frac{1}{|\mathfrak{S}_n|}\sum_{\pi \in \mathfrak{S}_n} \|v - v\{\pi\}\|_2^2 \leq \max_{\pi \in \mathfrak{S}_n} \|v - v\{\pi\}\|_2^2,$$

where step (i) follows from Jensen's inequality. □

Finally, we state an elementary lemma that bounds the number of distinct one-dimensional ordered partitions satisfying certain conditions. Recall that $\mathfrak{P}_L$ denotes the set of all one-dimensional ordered partitions of the set $[n_1]$ consisting of exactly L blocks. Also recall that $\mathfrak{P}_k^{\max}$ denotes all one-dimensional ordered partitions of $[n_1]$ in which the largest block has size at least k.

Lemma A.6.8. *For each $n_1 \geq 2$, the following statements hold:*
(a) For any $L \in [n_1]$, we have

$$\left|\bigcup_{\ell=1}^{L} \mathfrak{P}_\ell\right| = L^{n_1}, \quad \textit{and so} \quad |\mathfrak{P}| = \left|\bigcup_{\ell=1}^{n_1} \mathfrak{P}_\ell\right| = (n_1)^{n_1}.$$

(b) For any $k^* \in [n_1]$, we have

$$|\mathfrak{P}_{k^*}^{\max}| \leq (n_1)^{3(n_1 - k^*)}.$$

Proof. The second claim of part (a) follows from the first. In order to prove the first claim, note that each $i \in [n_1]$ can be placed into any one of the L blocks, and each different choice yields a different element of the set $\cup_{\ell=1}^{L} \mathfrak{P}_\ell$.

We now proceed to prove part (b). First, note that the claim is immediately true whenever $k^* \geq n_1 - 1$, since

$$|\mathfrak{P}_{n_1}^{\max}| = 1 \quad \text{and} \quad |\mathfrak{P}_{n_1-1}^{\max}| = 2n_1.$$

Consequently, we focus our proof on the case $k^* \leq n_1 - 2$, in which case $n_1 - k^* + 1 \leq \frac{3}{2}(n_1 - k^*)$. Suppose we are interested in bounding the number of one-dimensional ordered partitions in which the largest block has size at least k and the number of blocks is at most s_1. Then there are $\binom{n_1}{k}$ distinct ways of choosing the first k elements of the largest block and s_1 ways of choosing the position of the largest block. After having done this, the remaining $n_1 - k$ elements of $[n_1]$ can be placed in any of the s_1 blocks. Finally, note that $s_1 \leq n_1 - k + 1$, so that the number of such one-dimensional ordered partitions is bounded above by

$$\binom{n_1}{k^*} \cdot s_1 \cdot s_1^{n_1-k} \leq \binom{n_1}{k^*} \cdot (n_1 - k + 1)^{n_1-k+1}.$$

Choosing $k = k^*$ yields the bound

$$\begin{aligned} |\mathfrak{P}_{k^*}^{\max}| &\leq \binom{n_1}{k^*} \cdot (n_1 - k^* + 1)^{n_1-k^*+1} \\ &= \frac{n_1!}{(k^*)!} \cdot (n_1 - k^* + 1) \cdot \frac{(n_1 - k^* + 1)^{n_1-k^*+1}}{(n_1 - k^* + 1)!} \\ &\leq \frac{n_1!}{(k^*)!} \cdot \sqrt{n_1 - k^* + 1} \cdot e^{n_1-k^*+1} \cdot (2\pi)^{-1/2} \\ &\leq (n_1)^{n_1-k^*} \cdot \sqrt{n_1 - k^* + 1} \cdot e^{n_1-k^*+1} \cdot (2\pi)^{-1/2} \end{aligned}$$

where the second inequality uses the bound $n! \geq \left(\frac{n}{e}\right)^n \sqrt{2\pi n}$ given by Stirling's approximation. Now note that for each $n_1 \geq 2$, we have $e/\sqrt{2\pi} \leq \sqrt{n_1}$. Combining this with the bound $n_1 - k^* + 1 \leq n_1$ and putting together the pieces, we have

$$|\mathfrak{P}_{k^*}^{\max}| \leq (n_1)^{n_1-k^*+1} \cdot e^{n_1-k^*} \leq (n_1)^{3(n_1-k^*)},$$

where the final inequality is a consequence of the bounds $n_1 - k^* + 1 \leq \frac{3}{2}(n_1 - k^*)$ and $e \leq (n_1)^{3/2}$ for each $n_1 \geq 2$. □

Appendix B

Technical material for part II

B.1 Technical lemmas used in Chapter 5

In this section, we collect statements and proofs of some technical lemmas used in the proofs of our results concerning the AM algorithm.

B.1.1 Bounds on the "volumes" of wedges in $\mathbb{R}^d$

For a pair of scalars (w, w') and d-dimensional vectors (u, u'), recall that we define the *wedge* formed by the $d+1$-dimensional vectors $v = (u,\ w)$ and $v' = (u',\ w')$ as the region

$$W(v, v') = \{x \in \mathbb{R}^d : (\langle x,\ u\rangle + w) \cdot (\langle x,\ u'\rangle + w') \leq 0\}.$$

Note that the wedge is a purely geometric object.

For any set $\mathcal{C} \subseteq \mathbb{R}^d$, let

$$\mathrm{vol}(\mathcal{C}) = \Pr_{X \sim \mathcal{N}(0, I_d)} \{X \in \mathcal{C}\}$$

denote the volume of the set under the measure corresponding to the covariate distribution.

We now bound the volume of a wedge for the Gaussian distribution.

Lemma B.1.1. *Suppose that for a pair of scalars (w, w'), d-dimensional vectors (u, u'), and $v = (u,\ w)$ and $v' = (u',\ w')$, we have $\frac{\|v-v'\|}{\|u\|} < 1/2$. Then, there is a positive constant C such that*

$$\mathrm{vol}(W(v, v')) \leq C \frac{\|v - v'\|}{\|u\|} \log^{1/2}\left(\frac{2\|u\|}{\|v - v'\|}\right).$$

Proof of Lemma B.1.1

Using the notation $\xi = (x,\ 1) \in \mathbb{R}^{d+1}$ to denote the appended covariate, we have

$$\mathrm{vol}(W(v, v')) = \Pr\left\{\langle \xi,\ v\rangle \cdot \langle \xi,\ v'\rangle \leq 0\right\},$$

where the probability is computed with respect to Gaussian measure.

In order to prove a bound on this probability, we begin by bounding the associated indicator random variable as

$$\begin{aligned} \mathbf{1}\left\{\langle \xi,\, v\rangle \cdot \langle \xi,\, v'\rangle \leq 0\right\} &= \mathbf{1}\left\{\langle \xi,\, v'-v\rangle^2 \geq \langle \xi,\, v\rangle^2\right\} \\ &\leq \mathbf{1}\left\{\langle \xi,\, v'-v\rangle^2 \geq t\right\} + \mathbf{1}\left\{\langle \xi,\, v\rangle^2 \leq t\right\}, \end{aligned} \tag{B.1}$$

where inequality (B.1) holds for all $t \geq 0$. In order to bound the expectation of the second term, we write

$$\begin{aligned} \Pr\left\{\langle \xi,\, v\rangle^2 \leq t\right\} &= \Pr\left\{\|u\|^2 \chi^2_{nc} \leq t\right\} \\ &\overset{\text{(i)}}{\leq} \left(\frac{et}{\|u\|^2}\right)^{1/2} \end{aligned}$$

where χ^2_{nc} is a non-central chi-square random variable centered at $\frac{w}{\|u\|}$, and step (i) follows from standard χ^2 tail bounds (see Lemma B.1.6).

It remains to control the expectation of the first term on the RHS of inequality (B.1). We have

$$\begin{aligned} \Pr\left\{\langle \xi,\, v'-v\rangle^2 \geq t\right\} &\leq \Pr\left\{2\langle x,\, u'-u\rangle^2 + 2(w'-w)^2 \geq t\right\} \\ &\leq \Pr\left\{\|u-u'\|^2 \chi^2 \geq \frac{t}{2} - \|v-v'\|^2\right\}. \end{aligned}$$

Now, invoking a standard sub-exponential tail bound on the upper tail of a χ^2 random variable yields

$$\begin{aligned} \Pr\left\{\langle \xi,\, v'-v\rangle^2 \geq t\right\} &\leq c_1 \exp\left(-\frac{c_2}{\|u-u'\|^2}\left\{\frac{t}{2} - \|v-v'\|^2\right\}\right) \\ &\leq c_1 \exp\left(-\frac{c_2}{\|v-v'\|^2}\left\{\frac{t}{2} - \|v-v'\|^2\right\}\right). \end{aligned}$$

Putting all the pieces together, we obtain

$$\mathrm{vol}(W(v,v')) \leq c_1 \exp\left(-\frac{c_2}{\|v-v'\|^2}\left\{\frac{t}{2} - \|v-v'\|^2\right\}\right) + \left(\frac{et}{\|u\|^2}\right)^{1/2}.$$

Substituting $t = 2c\,\|v-v'\|^2 \log(2\|u\|/\|v-v'\|)$, which is a valid choice provided $\frac{\|v-v'\|}{\|u\|} < 1/2$, yields the desired result. □

B.1.2 Growth Functions and Uniform Empirical Concentration

We now briefly introduce growth functions and uniform laws derived from them, and refer the interested reader to Mohri et al. [225] for a more in-depth exposition on these topics.

We define growth functions in the general multi-class setting [75]. Let $\mathcal{X}$ denote a set, and let $\mathcal{F}$ denote a family of functions mapping $\mathcal{X} \mapsto \{0, 1, \ldots, k-1\}$. The *growth function* $\Pi_{\mathcal{F}} : \mathbb{N} \to \mathbb{R}$ of $\mathcal{F}$ is defined via

$$\Pi_{\mathcal{F}}(n) := \max_{x_1, \ldots, x_n \in \mathcal{X}} |\{\{f(x_1), f(x_2), \ldots, f(x_n)\} \ : \ f \in \mathcal{F}\}| .$$

In words, it is the cardinality of all possible labelings of n points in the set $\mathcal{X}$ by functions in the family $\mathcal{F}$.

A widely studied special case arises in the case $k = 2$, with the class of binary functions. In this case, a natural function class $\mathcal{F}$ is formed by defining $\mathcal{C}$ to be a family of subsets of $\mathcal{X}$, and identifying each set $C \in \mathcal{C}$ with its indicator function $f_C : = 1_C : \mathcal{X} \to \{0, 1\}$. In this case, define $\mathcal{F}_{\mathcal{C}} = \{f_C : C \in \mathcal{C}\}$. A bound on the growth function for such binary function provides following guarantee for the uniform convergence for the empirical measures of sets belonging to $\mathcal{C}$.

Lemma B.1.2 (Theorem 2 in [315]). *Let $\mathcal{C}$ be a family of subsets of a set $\mathcal{X}$. Let μ be a probability measure on $\mathcal{X}$, and let $\hat{\mu}_m := \frac{1}{m}\sum_{i=1}^m \delta_{X_i}$ be the empirical measure obtained from m independent copies of a random variable X with distribution μ. For every u such that $m \geq 2/u^2$, we have*

$$\Pr\left\{\sup_{C \in \mathcal{C}} |\hat{\mu}_m(C) - \mu(C)| \geq u\right\} \leq 4\Pi_{\mathcal{F}_{\mathcal{C}}}(2m)\exp(-mu^2/16). \tag{B.2}$$

We conclude this section by collecting some results on the growth functions of various function classes. For our development, it will be specialize to the case $\mathcal{X} = \mathbb{R}^d$.

Define the class of binary functions $\mathcal{F}_{\mathcal{H}}$ as the set of all functions of the form

$$f_{\theta,b}(x) := \frac{\mathsf{sgn}(\langle x, \, \theta\rangle + b) + 1}{2};$$

specifically, let $\mathcal{F}_{\mathcal{H}} := \left\{f_{\theta,b} : \theta \in \mathbb{R}^d, b \in \mathbb{R}\right\}$. In particular, these are all functions that can be formed by a d-dimensional hyperplane.

Using the shorthand $B_1^k = \{B_1, \ldots, B_k\}$, define the binary function

$$g_{\theta_1^k, b_1^k}(x) := \prod_{i=1}^k f_{\theta_i, b_i}(x),$$

and the binary function class corresponding to the intersection of k hyperplanes

$$\mathcal{G}_{\mathcal{H}^k} := \left\{g_{\theta_1^k, b_1^k} : \theta_1, \ldots, \theta_k \in \mathbb{R}^d \ , \ b_1, \ldots, b_k \in \mathbb{R}\right\}.$$

Finally, we are interested in the argmax function over hyperplanes. Here, define the function

$$m_{\theta_1^k, b_1^k}(x) := \operatorname*{argmax}_{j \in [k]} \left(\langle \theta_j, \, x\rangle + b_j\right) - 1,$$

mapping $\mathbb{R}^d \mapsto \{0, \ldots, k-1\}$. The function class that collects all such functions is given by

$$\mathcal{M}_k := \left\{ m_{\theta_1^k, b_1^k} : \theta_1, \ldots, \theta_k \in \mathbb{R}^d \ , \ b_1, \ldots, b_k \in \mathbb{R} \right\}.$$

The following results bound the growth functions of each of these function classes. We first consider the function classes $\mathcal{F}_{\mathcal{H}}$ and $\mathcal{G}_{\mathcal{H}^k}$, for which bounds on the VC dimension directly yield bounds on the growth function.

Lemma B.1.3 (Sauer-Shelah (e.g. Section 3 of Mohri et al. [225]))**.** *We have*

$$\Pi_{\mathcal{F}_{\mathcal{H}}}(n) \le \left(\frac{en}{d+1} \right)^{d+1}, \text{ and} \tag{B.3}$$

$$\Pi_{\mathcal{G}_{\mathcal{H}^k}}(n) \le \left(\frac{en}{d+1} \right)^{k(d+1)}. \tag{B.4}$$

The following bound on the growth function of the class $\mathcal{M}_k$ is also known.

Lemma B.1.4 (Theorem 3.1 of Daniely et al. [75])**.** *For an absolute constant C, we have*

$$\Pi_{\mathcal{M}_k}(n) \le \left(\frac{en}{Ck(d+1)\log(kd)} \right)^{Ck(d+1)\log(kd)}.$$

B.1.3 Singular value bound

We now state and prove a technical lemma that bound the maximum singular value of a matrix whose rows are drawn from a sub-Gaussian distribution.

Lemma B.1.5. *Suppose that the covariates are drawn i.i.d. from a η-sub-Gaussian distribution. Then for a fixed subset $S \in [n]$ of size ℓ and each $t \ge 0$, we have*

$$\Pr\left\{ \lambda_{\max}\left(\Xi_S^\top \Xi_S \right) \ge \ell + \widetilde{\eta}^2 (\sqrt{\ell d} + d + \ell t) \right\} \le 2e^{-\ell \min\{t, t^2\}},$$

where $\widetilde{\eta} = \max\{\eta, 1\}$.

Proof of Lemma B.1.5

Let $\{z_i\}_{i=1}^{\ell}$ denote i.i.d. Rademacher variables, and collect these in an ℓ-dimensional vector z. Let $D = \mathrm{diag}(z)$ denote a diagonal matrix, and note that by unitary invariance of the singular values, the singular values of the matrix $\widetilde{\Xi}_S = D\Xi_S$ are the same as those of Ξ_S.

By construction, the matrix $\widetilde{\Xi}_S$ has i.i.d. rows, and the i-th row is given by $z_i(x_i, \ 1)$. For a $d+1$ dimensional vector $\widetilde{\lambda} = (\lambda, \ w)$ with $\lambda \in \mathbb{R}^d$ and $w \in \mathbb{R}$, we have

$$\begin{aligned}
\mathbb{E}\left[\exp(\langle \widetilde{\lambda}, \ z_i(x_i, \ 1) \rangle) \right] &= \frac{e^w}{2} \cdot \mathbb{E}\left[\exp(\langle \lambda, \ x_i \rangle) \right] + \frac{e^{-w}}{2} \cdot \mathbb{E}\left[\exp(-\langle \lambda, \ x_i \rangle) \right] \\
&= \exp(\|\lambda\|^2 \eta^2 / 2) \cdot \frac{1}{2} \left(e^w + e^{-w} \right) \\
&\le \exp(\|\lambda\|^2 \eta^2 / 2) \cdot \exp(w^2/2) \le \exp(\|\widetilde{\lambda}\|^2 \widetilde{\eta}^2 / 2).
\end{aligned}$$

where we have used the fact that x_i is zero-mean and η sub-Gaussian.

Since the rows of $\widetilde{\Xi}_S$ are i.i.d., zero-mean, and $\widetilde{\eta}$-sub-Gaussian, applying [323, Theorem 6.2] immediately yields the lemma. □

B.1.4 Anti-concentration of χ^2 random variable

The following lemma shows the anti-concentration of the central and non-central χ^2 random variable.

Lemma B.1.6. *Let Z_ℓ and Z'_ℓ denote central and non-central χ^2 random variables with ℓ degrees of freedom, respectively. Then for all $p \in [0, \ell]$, we have*

$$\Pr\{Z'_\ell \le p\} \le \Pr\{Z_\ell \le p\} \le \left(\frac{p}{\ell} \exp\left(1 - \frac{p}{\ell}\right)\right)^{\ell/2} = \exp\left(-\frac{\ell}{2}\left[\log\frac{\ell}{p} + \frac{p}{\ell} - 1\right]\right) \tag{B.5}$$

Proof of Lemma B.1.6

The fact that $Z'_\ell \overset{st.}{\le} Z_\ell$ follows from standard results that guarantee that central χ^2 random variables stochastically dominate their non-central counterparts.

The tail bound is a simple consequence of the Chernoff bound. In particular, we have for all $\lambda > 0$ that

$$\begin{aligned}\Pr\{Z_\ell \le p\} &= \Pr\{\exp(-\lambda Z_\ell) \ge \exp(-\lambda p)\}\\ &\le \exp(\lambda p)\mathbb{E}\left[\exp(-\lambda Z_\ell)\right]\\ &= \exp(\lambda p)(1 + 2\lambda)^{-\frac{\ell}{2}}.\end{aligned} \tag{B.6}$$

where in the last step, we have used $\mathbb{E}\left[\exp(-\lambda Z_\ell)\right] = (1 + 2\lambda)^{-\frac{\ell}{2}}$, which is valid for all $\lambda > -1/2$. Minimizing the last expression over $\lambda > 0$ then yields the choice $\lambda^* = \frac{1}{2}\left(\frac{\ell}{p} - 1\right)$, which is greater than 0 for all $0 \le p \le \ell$. Substituting this choice back into equation (B.6) proves the lemma. □

B.2 Technical lemmas used in Chapter 6

We now collect some technical lemmas that were used in the proofs of our main results.

B.2.1 A recursion formula

We present a general recursion formula that is used to bound the error in multiple proofs.

Lemma B.2.1. *Consider any sequence of positive reals $\{a_i\}_{i \ge 0}$ satisfying the sequence of inequalities*

$$a_{t+1} \le C_1 + C_2 \left(\frac{a_t + C_3}{n}\right)^\gamma \text{ for each integer } t \ge 0,$$

where the tuple (C_1, C_2, C_3) represent some arbitrary positive scalars, n represents a positive integer, and we have the inclusion $\gamma \in (0,1)$. Define the shorthand $\rho := \left(\frac{1}{2C_2}\right)^{(1-\gamma)^{-1}} a_0$. Then there is an absolute constant c such that for all $T \geq \log_{\gamma^{-1}} \max\{\log n^{\gamma(1-\gamma)^{-1} \vee 1}, \log \rho\}$, we have

$$a_T \leq c \left\{ C_1 + C_2 \left(\frac{C_3}{n}\right)^{\gamma} + (2C_2)^{(1-\gamma)^{-1}} \cdot n^{-\gamma(1-\gamma)^{-1}} \right\}.$$

Proof. First, note the fact that two positive scalars a and b and $\gamma \in (0,1]$, we have $(a+b)^\gamma \leq a^\gamma + b^\gamma$. Thus, a consequence of the recursive inequality above is the relation

$$\begin{aligned} a_{t+1} &\leq C_1 + C_2 \left(\frac{C_3}{n}\right)^{\gamma} + C_2 \left(\frac{a_t}{n}\right)^{\gamma} \\ &\leq 2 \max \left\{ C_1 + C_2 \left(\frac{C_3}{n}\right)^{\gamma}, C_2 \left(\frac{a_t}{n}\right)^{\gamma} \right\}. \end{aligned}$$

Since the first term above is a constant, it suffices to provide upper bounds on the recursion

$$b_{t+1} \leq 2C_2 \left(\frac{b_t}{n}\right)^{\gamma}$$

with the initial condition $b_0 = a_0$.

We now claim that for all $t \geq 1$, the following upper bound holds:

$$b_t \leq (2C_2)^{(1-\gamma)^{-1}(1-\gamma^t)} \cdot n^{-\gamma(1-\gamma)^{-1} \cdot (1-\gamma^t)} \cdot b_0^{\gamma^t} \tag{B.7}$$

$$= (2C_2)^{(1-\gamma)^{-1}} \cdot n^{-\gamma(1-\gamma)^{-1}} \cdot \left(n^{\gamma(1-\gamma)^{-1}}\right)^{\gamma^t} \cdot \left(\left(\frac{1}{2C_2}\right)^{(1-\gamma)^{-1}} b_0\right)^{\gamma^t}, \tag{B.8}$$

where equality (B.8) follows by computation.

Taking this claim as given for the moment, note that if $t \geq \log_{\gamma^{-1}} \log x$ for a scalar $x \geq 1$, then we have $\gamma^t \leq (\log x)^{-1}$. Also note that for each $x \in \mathbb{R}$, we have $x^{(\log x)^{-1}} = e$ by definition. We now split the proof into two cases, using the shorthand $\rho := \left(\frac{1}{2C_2}\right)^{(1-\gamma)^{-1}} b_0$.

Case 1; $\rho \leq 1$: In this case, it suffices to take $t \geq t_0 := \log_{\gamma^{-1}} \log n^{\gamma(1-\gamma)^{-1} \vee 1}$, in which case we have

$$b_t \leq e(2C_2)^{(1-\gamma)^{-1}} \cdot n^{-\gamma(1-\gamma)^{-1}}.$$

Case 2; $\rho > 1$: Now take $t \geq t_0 \vee \log_{\gamma^{-1}} \log \rho$ where t_0 was defined in case 1 above. Then, we again have

$$b_t \leq e^2(2C_2)^{(1-\gamma)^{-1}} \cdot n^{-\gamma(1-\gamma)^{-1}}.$$

Combining the two cases with the setup above completes the proof of the lemma.

It remains to establish claim (B.7), for which we use an inductive argument. The base case follows from the one-step definition of the recursion. Assuming the induction hypothesis—that the claim is true for some positive integer t—evaluating the recursion yields

$$\begin{aligned} b_{t+1} &\leq (2C_2)\left(\frac{b_t}{n}\right)^{\gamma} \\ &\leq (2C_2)\cdot(2C_2)^{\gamma\cdot(1-\gamma)^{-1}\left(1-\gamma^t\right)} \times \left(\frac{1}{n}\right)^{\gamma} \cdot (1/n)^{\gamma\cdot\gamma(1-\gamma)^{-1}\cdot\left(1-\gamma^t\right)} \times b_0^{\gamma^{t+1}} \\ &= (2C_2)^{(1-\gamma)^{-1}\left(1-\gamma^{t+1}\right)} \cdot (1/n)^{\gamma(1-\gamma)^{-1}\left(1-\gamma^{t+1}\right)} \cdot b_0^{\gamma^{t+1}}, \end{aligned}$$

thereby establishing the induction. □

B.2.2 Properties of truncated Gaussians

Let $\Phi(\cdot)$ denote the d-dimensional standard Gaussian PDF. For $a < b$, let $m_2(a,b)$ denote the second moment of a univariate standard Gaussian truncated to lie in the interval $[a,b]$, and let $\gamma = \min\{1, m_2(a,b)\}$. Finally, let κ denote the Gaussian volume of the interval $[a,b]$.

Lemma B.2.2. *Let $w_1, w_2, \ldots, w_n$ denote i.i.d. draws from a Gaussian truncated to the interval $[a,b]$. There is a pair of absolute constants (c_1, c_2) such that if $\kappa^2 n \geq c_1 \log^2(1/\kappa)$, then*

$$\Pr\left\{\frac{1}{n}\sum_{i=1}^{n} w_i^2 \leq \frac{1}{2}\kappa^2\right\} \leq c_1 \exp\left(-c_2 \frac{n\kappa^3}{\log^2(1/\kappa)}\right).$$

Proof. The proof follows immediately from Lemma 4 of Ghosh et al. [119]. In particular, a slight modification of their lemma, specialized (in their notation) to $d = 1$ and with $n\kappa$ samples, yields the following claim. There is a pair of universal constants (c_1, c_2) such that if $\kappa^2 n \geq c_1 \log^2(1/\kappa)$, then

$$\Pr\left\{\frac{1}{n}\sum_{i=1}^{n} w_i^2 \leq \frac{1}{2}\kappa^2\right\} \leq \exp\left(-\frac{n\kappa^3}{\log^2(1/\kappa)}\right).$$

Adjusting the constant factors completes the proof. □

Lemma B.2.3. *Consider a matrix X consisting of $n \geq p$ i.i.d. rows drawn from the distribution*

$$g(x) = \frac{\mathbf{1}\left\{x_1 \in [\ell, r]\right\}}{\Pr\{x_1 \in [\ell, r]\}} \cdot \Phi(x)$$

for each $x \in \mathbb{R}^p$. Then for all $t \geq 0$, we have

$$\Pr\left\{\sigma_{\min}(X^\top X/n) \leq \gamma - c\sqrt{p/n} - \frac{t}{\sqrt{n}}\right\} \leq e^{-t^2/2}.$$

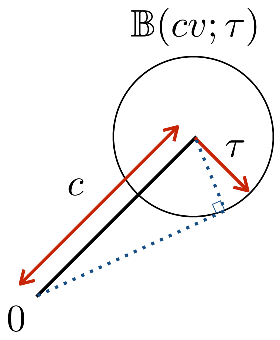


Figure B.1: Center of the circle denotes the point cv; circle denotes valid set of u. Clearly, the u that makes the largest angle with v is given by the tangent to the circle (in blue).

Proof. Let $Y \sim g$ denote the truncated random variable. We claim that

$$\mathbb{E}[YY^\top] \succeq \gamma I, \text{ and that } Y \text{ is sub-Gaussian with parameter at most } 2.$$

Given this claim, the proof of the theorem follows immediately by applying Remark 5.40 of [319].

Proving the claim is also straightforward. Indeed, for any $v \in \mathbb{R}^p$, we have

$$\begin{aligned}\mathbb{E}\langle Y,\, v\rangle^2 &= v_1^2\mathbb{E}[Y_1^2] + \sum_{i\neq 1} v_i^2\\ &= v_1^2 m_2(a,b) + \sum_{i\neq 1} v_i^2.\end{aligned}$$

Minimizing the above expression over unit norm v yields

$$\inf_{v:\|v\|=1} \mathbb{E}\langle Y,\, v\rangle^2 = \min\{1, m_2(a,b)\} = \gamma.$$

In order to show that the random vector Y is sub-Gaussian, it suffices to show that $\langle Y,\, v\rangle$ is 2-sub-Gaussian for each unit vector v. Since the truncation operation only influences the one-dimensional RV Y_1, it suffices to show that Y_1 is 2-sub-Gaussian. Once again, we invoke a standard truncation lemma by symmetrization (e.g. Ledoux [190]), which yields the result. □

B.2.3 Angles and norms

The following lemma collects an elementary fact about angles between vectors and distances between their scaled counterparts.

Lemma B.2.4. *Given a unit norm vector v and a pair of positive scalars (c,τ) obeying the relation $\tau \le c$, suppose that a vector u satisfies*

$$\|u - cv\|^2 \le \tau^2. \tag{B.9}$$

Then, we have

$$\sin \angle(u, v) \leq \frac{\tau}{c}.$$

Proof. We provide a simple proof by picture in Figure B.1. In particular, denoting the ball of radius r centered at x by $\mathbb{B}(x; r)$, condition (B.9) is equivalent to the inclusion $u \in \mathbb{B}(cv; \tau)$. Clearly, the vector u that maximizes the angle between u and v is given by the tangent to this ball from the origin. In this particular case, we have

$$\sin \angle(u, v) = \frac{\tau}{c},$$

and this establishes the proof. □

Appendix C

Technical material for part III

C.1 Calculations for the "hard" sub-class in Section 7.3

Recall from equation (8.30) our previous calculation of the value function and standard deviation, from which we have

$$\|\sigma(\theta^*)\|_\infty = \nu(1-\tau)\frac{\sqrt{p(1-p)}}{1-\gamma p}, \qquad \|(\mathbf{I}-\gamma\mathbf{P})^{-1}\sigma(\theta^*)\|_\infty = \nu(1-\tau)\frac{\sqrt{p(1-p)}}{(1-\gamma p)^2},$$

and $\|\theta^*\|_{\text{span}} = \nu(1-\tau)\frac{1}{1-\gamma p}$. Substituting in our choices $\nu = 1$, $p = \frac{4\gamma-1}{3\gamma}$, and $\tau = 1-(1-\gamma)^\alpha$ and simplifying by employing inequality (8.32), we have

$$\|\sigma(\theta^*)\|_\infty \sim \left(\frac{1}{1-\gamma}\right)^{0.5-\alpha}, \ \|(\mathbf{I}-\gamma\mathbf{P})^{-1}\sigma(\theta^*)\|_\infty \sim \left(\frac{1}{1-\gamma}\right)^{1.5-\alpha}, \ \text{ and } \ \|\theta^*\|_{\text{span}} \sim \left(\frac{1}{1-\gamma}\right)^{1-\alpha},$$

for each discount factor $\gamma \geq \frac{1}{2}$. Here, the $\sim$ notation indicates that the LHS can be sandwiched between two terms that are proportional to the RHS such that the factors of proportionality are strictly positive and γ-independent.

For the plug-in estimator, its performance will be determined by the maximum of the two terms

$$\frac{\|(\mathbf{I}-\gamma\mathbf{P})^{-1}\sigma(\theta^*)\|_\infty}{\sqrt{N}} \sim \frac{1}{\sqrt{N}}\left(\frac{1}{1-\gamma}\right)^{1.5-\alpha} \quad \text{and} \quad \frac{\|\theta^*\|_{\text{span}}}{(1-\gamma)N} \sim \frac{1}{N}\left(\frac{1}{1-\gamma}\right)^{2-\alpha}.$$

In the regime $N \gtrsim \frac{1}{1-\gamma}$, the first term will be dominant.

C.2 Dependence of plug-in error on span semi-norm

In this section, we state and prove a proposition that provides a family of MRPs in which the ℓ_∞-error of the plug-in estimator can be completely characterized by the span semi-norm of the optimal value function.

Proposition C.2.1. *Suppose that the rewards are observed noiselessly, with $\rho(r) = 0$. There is a pair of universal positive constants (c_1, c_2) such that for any triple of positive scalars (ζ, N, D), there is a D-state MRP for which*

$$\|\theta^*\|_\infty = \zeta \quad \textit{and} \quad \frac{\|\sigma(\theta^*)\|_\infty}{N} \leq \frac{3}{\sqrt{D}} \cdot \frac{\zeta}{N}, \tag{C.1}$$

and for which the error of the plug-in estimator satisfies

$$c_1 \gamma \frac{\zeta}{N} \overset{(a)}{\leq} \mathbb{E}\left[\|\widehat{\theta}_{\mathsf{plug}} - \theta^*\|_\infty\right] \overset{(b)}{\leq} c_2 \gamma \frac{\zeta \log(1 + D/3)}{N} \left\{(\log\log(1 + D/3))^{-1} \wedge 1\right\}. \tag{C.2}$$

A few comments are in order. First, note that equation (C.1) guarantees that we have

$$\frac{1}{\sqrt{N}} \cdot \|(\mathbf{I} - \gamma \mathbf{P})^{-1} \sigma(\theta^*)\|_\infty \lesssim \frac{1}{\sqrt{D}} \frac{1}{(1-\gamma)N} \|\theta^*\|_{\text{span}}$$

for large values of the dimension D, so that the first term in the guarantee (8.1b) is dominated by the second. In particular, suppose that $D \gg \frac{1}{(1-\gamma)^2}$; then we have

$$\frac{1}{\sqrt{N}} \cdot \|(\mathbf{I} - \gamma \mathbf{P})^{-1} \sigma(\theta^*)\|_\infty \ll \frac{1}{N} \|\theta^*\|_{\text{span}}.$$

In other words, if our analysis was loose in that the error of the plug-in estimator depended only on the functional $\frac{1}{\sqrt{N}} \cdot \|(\mathbf{I} - \gamma \mathbf{P})^{-1} \sigma(\theta^*)\|_\infty$, then it would be impossible to prove a lower bound that involves the quantity $\frac{\|\theta^*\|_{\text{span}}}{N}$. On the other hand, equation (C.2) shows that this such a lower bound can indeed be proved: the plug-in error is characterized precisely by the quantity $\gamma \frac{\|\theta^*\|_{\text{span}}}{N}$ up to a logarithmic factor in the dimension D.

Second, note that while equation (C.2) shows that the plug-in error must have some span semi-norm dependence, it falls short of showing the stronger lower bound

$$c_1 \frac{\zeta}{(1-\gamma)N} \leq \mathbb{E}\left[\|\widehat{\theta}_{\mathsf{plug}} - \theta^*\|_\infty\right], \tag{C.3}$$

which would show, for instance, that Corollary 8.2.1(a) is sharp up to a logarithmic factor. We conjecture that there is an MRP for which the bound (C.3) holds.

Finally, it is worth commenting on the logarithmic factor that appears in the upper bound of equation (C.2). Note that for sufficiently large D, the logarithmic factor is proportional to $\log D / \log\log D$. This is consequence of applying Bennett's inequality instead of Bernstein's inequality, and we conjecture that the same factor ought to replace the factor $\log D$ factor multiplying the span semi-norm in the upper bound (8.1b).

C.2.1 Proof of Proposition C.2.1

In order to prove Proposition C.2.1, it suffices to construct an MRP satisfying condition (C.1) and compute its plug-in estimator in closed form. With this goal in mind, suppose that for simplicity that D is divisible by three, and consider $D/3$ copies of the 3-state MRP from Figure 8.2(a). By construction, we have $\|\sigma(\theta^*)\|_\infty = \mu\sqrt{q(1-q)}$, and $\|\theta^*\|_{\text{span}} = \frac{\mu}{(1-\gamma)}$. Setting $q = \frac{10}{ND}$, we see that condition (C.1) is immediately satisfied with $\zeta = \frac{\mu}{(1-\gamma)}$.

It remains to verify the claim (C.2). Note that the plug-in estimator for this MRP can be computed in closed form. In particular, it is straightforward to verify that for each state i having reward $\mu/2$, we have

$$\frac{N(1-\gamma)}{\gamma\mu}\left(\widehat{\theta}_{\mathsf{plug}}(i) - \theta^*_i\right) \overset{d}{=} \mathsf{Bin}(N, q) - Nq, \tag{C.4}$$

where we have used the notation $\widehat{\theta}_{\mathsf{plug}}(i)$ to denote the i-th entry of the vector $\widehat{\theta}_{\mathsf{plug}}$. Furthermore, these $D/3$ random variables are independent. Thus, the (scaled) ℓ_∞-error of the plug-in estimator is equal to the maximum absolute deviation in a collection of independent binomial random variables.

Proof of inequality (C.2)**, part (a):** The following technical lemma provides a lower bound on the deviation of binomials, and its proof is postponed to the end of this section.

Lemma C.2.1. *Let $X_1, \ldots, X_k$ denote independent random variables with distribution $\mathsf{Bin}\left(n, \frac{1}{3kn}\right)$. Let $Y_j = X_j - \mathbb{E}[X_j]$ for each $1 \leq j \leq k$. Then, we have*

$$\mathbb{E}\left[\max_{1\leq j\leq k} |Y_j|\right] \geq \frac{4}{9}.$$

Applying Lemma C.2.1 with $k = D/3$ in conjunction with the characterization (C.4), and substituting our choices of the pair (μ, q) yields

$$\mathbb{E}\left[\|\widehat{\theta}_{\mathsf{plug}} - \theta^*\|_\infty\right] \geq \frac{4}{9} \cdot \frac{\zeta\gamma}{N}.$$

Proof of inequality (C.2)**, part (b):** Corollary 3.1(ii) and Lemma 3.3 of Wellner [331] yield, to the best of our knowledge, the sharpest available upper bound on the maximum absolute deviation of $\mathsf{Bin}(n, q)$ random variables in the regime $nq(1-q) \ll 1$:

$$\mathbb{E}\left[\max_{1\leq j\leq k} |Y_j|\right] \leq \sqrt{12} \cdot \frac{\log(1+k)}{\log\log(1+k)} \quad \text{if} \quad \log(1+k) \geq 5. \tag{C.5}$$

Combining this bound with the Bernstein bound when k is small, and substituting the various quantities completes the proof.

Proof of Lemma C.2.1: Employing the shorthand $q = \frac{1}{3kn}$, we have

$$\begin{aligned}
\mathbb{E}\left[\max_{1\leq j\leq k} |Y_j|\right] &\geq (1-nq)\cdot \Pr\left\{\max_{1\leq j\leq k} X_j \geq 1\right\} \\
&= (1-nq)\cdot(1-(1-q)^{nk}) \\
&\geq \frac{2}{3}\cdot\left(1-\sqrt[3]{\left(1-\frac{1}{3nk}\right)^{3nk}}\right) \\
&\geq \frac{2}{3}\cdot\left(1-e^{-1/3}\right) \geq \frac{4}{9}.
\end{aligned}$$

C.3 Technical lemmas used in Chapter 9

Let us now collect technical lemmas used in Chapter 9 along with their proofs.

C.3.1 Proofs of auxiliary lemmas for Proposition 9.1.1

In this subsection, we provide proofs of the auxiliary lemmas that underlie the proof of Proposition 9.1.1.

Proof of Lemma 9.4.1

The proof is basically a lengthy computation. For clarity, let us decompose the procedure into three steps. In the first step, we compute an explicit form for the inverse information matrix $J_\vartheta^\dagger$. In the second step, we evaluate the gradient $\nabla\psi(\vartheta)$. In the third and final step, we use the result in the previous two steps to prove the claim (9.15) of the lemma.

Step 1: In the first step, we evaluate $J_\vartheta^\dagger$. Recall that our data $(\mathbf{Z}, R)$ is generated as follows. We generate the matrix $\mathbf{Z}$ and the vector R independently. Each row of $\mathbf{Z}$ is generated independently. Its j-th row, denoted by z_j, is sampled from a multinomial distribution with parameter p_j, where p_j denotes the j-th row of $\mathbf{P}$. The vector R is sampled from $\mathcal{N}(r, \sigma_r^2\mathbf{I})$. Because of this independence structure, the Fisher information J_ϑ is a block diagonal matrix of the form

$$J_\vartheta = \begin{bmatrix} J_{p_1} & 0 & 0\ldots & 0 & 0 \\ 0 & J_{p_2} & 0\ldots & 0 & 0 \\ 0 & 0 & \ldots & 0 & 0 \\ 0 & 0 & 0\ldots & J_{p_D} & 0 \\ 0 & 0 & 0\ldots & 0 & J_r \end{bmatrix}.$$

Here each sub-block matrix J_{p_j} is the Fisher information corresponding to the model where a single data Z_j is sampled from the multinomial distribution with parameter p_j, and J_r is the Fisher information corresponding to the model in which a single data point R is sampled from $\mathcal{N}(r, \sigma_r^2\mathbf{I})$. Thus, the inverse Fisher information $J_\vartheta^\dagger$ is also a block diagonal matrix of the form

$$J_\vartheta^\dagger = \begin{bmatrix} J_{p_1}^\dagger & 0 & 0\dots & 0 & 0 \\ 0 & J_{p_2}^\dagger & 0\dots & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 \\ 0 & 0 & 0\dots & J_{p_D}^\dagger & 0 \\ 0 & 0 & 0\dots & 0 & J_r^\dagger \end{bmatrix}. \tag{C.6}$$

It is easy to compute $J_{p_j}^\dagger$ and $J_r^\dagger$:

$$J_{p_j}^\dagger = \mathrm{diag}(p_j) - p_j p_j^T = \mathrm{cov}(Z_j - p_j) \qquad \text{for } j \in [D], \text{ and} \tag{C.7a}$$

$$J_r^\dagger = J_r^{-1} = \sigma_r^2 I. \tag{C.7b}$$

For a vector $q \in \mathbb{R}^D$, we use $\mathrm{diag}(q) \in \mathbb{R}^{D\times D}$ to denote the diagonal matrix with diagonal entries q_j.

Step 2: In the second step, we evaluate $\nabla\psi(\vartheta)$. Recall that $\psi(\vartheta) = (\mathbf{I} - \gamma\mathbf{P})^{-1}r$. It is straightforward to see that

$$\nabla_r\psi(\vartheta) = (\mathbf{I} - \gamma\mathbf{P})^{-1}. \tag{C.8}$$

Below we evaluate $\nabla_{p_j}\psi(\theta)$ for $j \in [D]$, where p_j is the j-th row of the matrix $\mathbf{P}$. We show that

$$\nabla_{p_j}\psi(\vartheta) = \gamma(\mathbf{I} - \gamma\mathbf{P})^{-1}e_j\theta^T. \tag{C.9}$$

Here we recall $\theta = \psi(\vartheta) = (\mathbf{I} - \gamma\mathbf{P})^{-1}r$.

To prove Eq. (C.9), we start with the following elementary fact: for the matrix inverse mapping $A \to A^{-1}$, we have $\frac{\partial A^{-1}}{\partial A_{jk}} = -A^{-1}e_j e_k^T A^{-1}$ for all $j, k \in [D]$. Combining this fact with chain rule, we find that

$$\frac{\partial\psi(\vartheta)}{\partial\mathbf{P}_{jk}} = \gamma(\mathbf{I} - \gamma\mathbf{P})^{-1}e_j e_k^T(\mathbf{I} - \gamma\mathbf{P})^{-1}r = \gamma(\mathbf{I} - \gamma\mathbf{P})^{-1}e_j\theta^T e_k,$$

valid for all $j, k \in [D]$. This immediately implies Eq. (C.9) since p_j is the vector with coordinates $\mathbf{P}_{jk}$.

Step 3: In the third step, we evaluate $\nabla\psi(\vartheta)^T J_\vartheta^\dagger\nabla\psi(\vartheta)$. From Eq. (C.6), we observe that the inverse Fisher information $J_\vartheta^\dagger$ has a block structure. Consequently, we can write

$$\nabla\psi(\vartheta)^T J_\vartheta^\dagger\nabla\psi(\vartheta) = \sum_{j\in[D]} \nabla_{p_j}\psi(\vartheta)^T J_{p_j}^\dagger \nabla_{p_j}\psi(\vartheta) + \nabla_R\psi(\vartheta)^T J_R^\dagger \nabla_R\psi(\vartheta). \tag{C.10}$$

Combining Eqs. (C.7b) and (C.8) yields

$$\nabla_R \psi(\vartheta)^T J_R^\dagger \nabla_R \psi(\vartheta) = \sigma_r^2 (\mathbf{I} - \gamma \mathbf{P})^{-1} (\mathbf{I} - \gamma \mathbf{P})^{-T}. \tag{C.11}$$

Combining Eqs. (C.7a) and (C.9) yields

$$\begin{aligned} \nabla_{p_j} \psi(\vartheta)^T J_{p_j}^\dagger \nabla_{p_j} \psi(\vartheta) &= \gamma^2 (\mathbf{I} - \gamma \mathbf{P})^{-1} e_j \theta^T \operatorname{cov}(Z_j - p_j) \theta e_j^T (\mathbf{I} - \gamma \mathbf{P})^{-T} \\ &= \gamma^2 (\mathbf{I} - \gamma \mathbf{P})^{-1} e_j \operatorname{cov}((Z_j - p_j)^T \theta) e_j^T (\mathbf{I} - \gamma \mathbf{P})^{-T}, \end{aligned}$$

valid for each $j \in [D]$. Summing over $j \in [D]$ then leads to

$$\begin{aligned} \sum_{j \in [D]} \nabla_{p_j} \psi(\vartheta)^T J_{p_j}^\dagger \nabla_{p_j} \psi(\vartheta) &= \gamma^2 (\mathbf{I} - \gamma \mathbf{P})^{-1} \Big(\sum_{j \in [D]} e_j \operatorname{cov}((Z_j - p_j)^T \theta) e_j^T \Big) (\mathbf{I} - \gamma \mathbf{P})^{-T} \\ &= \gamma^2 (\mathbf{I} - \gamma \mathbf{P})^{-1} \Sigma_{\mathbf{P}}(\theta) (\mathbf{I} - \gamma \mathbf{P})^{-T}, \end{aligned} \tag{C.12}$$

where the last line uses the definition of $\Sigma_{\mathbf{P}}(\theta)$ in Eq. (7.10). Finally, substituting Eq. (C.11) and Eq. (C.12) into Eq. (C.10) yields the claim (9.15), which completes the proof of Lemma 9.4.1.

C.3.2 Proofs of auxiliary lemmas for Theorem 9.1.1

In this appendix, we detailed proofs of the auxiliary lemmas that underlie the proof of the non-asymptotic local minimax lower bound stated in Theorem 9.1.1.

Proof of Lemma 9.4.2

The proof uses the standard device of reducing estimation to testing (see, e.g., [34, 302, 323]). The first step is to lower bound the minimax risk over $\mathcal{P}$ and $\mathcal{P}'$ by its averaged risk:

$$\inf_{\hat{\theta}_N} \max_{\mathcal{P} \in \{\mathcal{P}, \mathcal{P}'\}} \mathbb{E}_{\mathcal{P}} \left[\|\theta - \theta(\mathcal{P})\|_\infty \right] \geq \frac{1}{2} \left(\mathbb{E}_{P^N} \left[\|\hat{\theta}_N - \theta\|_\infty \right] + \mathbb{E}_{P'^N} \left[\|\hat{\theta}_N - \theta'\|_\infty \right] \right). \tag{C.13}$$

By Markov's inequality, for any $\delta \geq 0$, we have

$$\mathbb{E}_{P^N} \left[\|\hat{\theta}_N - \theta\|_\infty \right] + \mathbb{E}_{P'^N} \left[\|\hat{\theta}_N - \theta'\|_\infty \right] \geq \delta \left[P^N \left(\|\hat{\theta}_N - \theta\|_\infty \geq \delta \right) + P'^N \left(\|\hat{\theta}_N - \theta'\|_\infty \geq \delta \right) \right].$$

If we define $\delta_{01} \overset{\text{def}}{=} \frac{1}{2} \|\theta - \theta'\|_\infty$, then we have the implication

$$\|\theta - \theta\|_\infty < \delta_{01} \quad \Longrightarrow \quad \|\theta - \theta'\|_\infty > \delta_{01}, \tag{C.14}$$

from which it follows that

$$\begin{aligned} \mathbb{E}_{P^n} \left[\|\hat{\theta}_n - \theta\|_\infty \right] + \mathbb{E}_{P'^n} \left[\|\hat{\theta}_n - \theta'\|_\infty \right] &\geq \delta_{01} \left[1 - P^n(\|\hat{\theta}_n - \theta\|_\infty < \delta_{01}) + P'^n(\|\hat{\theta}_n - \theta'\|_\infty \geq \delta_{01}) \right] \\ &\geq \delta_{01} \left[1 - P^n(\|\hat{\theta}_N - \theta'\|_\infty \geq \delta_{01}) + P'^n(\|\hat{\theta}_N - \theta'\|_\infty \geq \delta_{01}) \right] \\ &\geq \delta_{01} \left[1 - \|P^n - P'^n\|_{\mathrm{TV}} \right] \geq \delta_{01} \left[1 - \sqrt{2} d_{\mathrm{hel}}(P^n, P'^n)^2 \right]. \end{aligned}$$

The tensorization property of Hellinger distance (cf. Section 15.1 in [323]) guarantees that

$$d_{\mathrm{hel}}(P^N, P'^N)^2 = 1 - \big(1 - d_{\mathrm{hel}}(P, P')^2\big)^N \le N\ d_{\mathrm{hel}}(P, P')^2.$$

Thus, we have proved that

$$\inf_{\hat{\theta}_N} \max_{\mathcal{Q} \in \{\mathcal{P}, \mathcal{P}'\}} \mathbb{E}_{\mathcal{Q}} \left[\big\| \theta - \theta(\mathcal{Q}) \big\|_\infty \right] \ge \frac{1}{4} \big\| \theta(\mathcal{P}) - \theta(\mathcal{P}') \big\|_\infty \cdot \Big(1 - \sqrt{2} N \cdot d_{\mathrm{hel}}(P, P')^2\Big)_+ .$$

Taking the supremum over all the possible alternatives $\mathcal{P}' \in \mathcal{S}$ yields

$$\mathfrak{M}_N(\mathcal{P}; \mathcal{S}) \ge \sup_{\mathcal{P}' \in \mathcal{S}} \frac{1}{4} \cdot \sqrt{N} \big\| \theta(\mathcal{P}) - \theta(\mathcal{P}') \big\|_\infty \cdot \Big(1 - \sqrt{2} N \cdot d_{\mathrm{hel}}(P, P')^2\Big)_+ . \tag{C.15}$$

A calculation shows that this bound implies the claim in Lemma 9.4.2.

Proof of Lemma 9.4.3

Recall the shorthand $\Delta_{\mathbf{P}} = \mathbf{P} - \mathbf{P}'$ and $\Delta_r = r - r'$, and let $\theta^* \equiv \theta(\mathcal{P})$. We prove that $\|\theta(\mathcal{P}) - \theta(\mathcal{P}')\|_\infty$ is lower bounded by

$$\big\| \gamma (\mathbf{I} - \gamma \mathbf{P})^{-1} \Delta_{\mathbf{P}} \theta^* + (\mathbf{I} - \gamma \mathbf{P})^{-1} \Delta_r \big\|_\infty - \left(\frac{\gamma \left\| \Delta_{\mathbf{P}} \right\|_\infty}{(1-\gamma)} \big\| \gamma (\mathbf{I} - \gamma \mathbf{P})^{-1} \Delta_{\mathbf{P}} \theta^* \big\|_\infty + \frac{\gamma \left\| \Delta_{\mathbf{P}} \right\|_\infty \left\| \Delta_r \right\|_\infty}{(1-\gamma)^2} \right). \tag{C.16}$$

Since $\theta(\mathcal{P}) = (\mathbf{I} - \gamma \mathbf{P})^{-1} r$ and $\theta(\mathcal{P}') = (\mathbf{I} - \gamma \mathbf{P}')^{-1} r'$ by definition, if we introduce the shorthand $M_{\mathbf{P}} = (\mathbf{I} - \gamma \mathbf{P})^{-1} - (\mathbf{I} - \gamma \mathbf{P}')^{-1}$, some elementary calculation gives the identity

$$\theta(\mathcal{P}) - \theta(\mathcal{P}') = M_{\mathbf{P}} r + (\mathbf{I} - \gamma \mathbf{P})^{-1} \Delta_r + M_{\mathbf{P}} \Delta_r. \tag{C.17}$$

Now we find a new expression for $M_{\mathbf{P}} = (\mathbf{I} - \gamma \mathbf{P})^{-1} - (\mathbf{I} - \gamma \mathbf{P}')^{-1}$ that is easy to control. Recall the elementary identity $A_1^{-1} = A_0^{-1} + A_1^{-1}(A_0 - A_1) A_0^{-1}$ for any matrices A_0, A_1. Thus,

$$\begin{aligned} M_{\mathbf{P}} &= (\mathbf{I} - \gamma \mathbf{P})^{-1} - (\mathbf{I} - \gamma \mathbf{P}')^{-1} \\ &= \gamma (\mathbf{I} - \gamma \mathbf{P}')^{-1} (\mathbf{P} - \mathbf{P}') (\mathbf{I} - \gamma \mathbf{P})^{-1} \\ &= \gamma (\mathbf{I} - \gamma \mathbf{P})^{-1} (\mathbf{P} - \mathbf{P}') (\mathbf{I} - \gamma \mathbf{P})^{-1} + \gamma^2 (\mathbf{I} - \gamma \mathbf{P}')^{-1} (\mathbf{P} - \mathbf{P}') (\mathbf{I} - \gamma \mathbf{P})^{-1} (\mathbf{P} - \mathbf{P}') (\mathbf{I} - \gamma \mathbf{P})^{-1} \\ &= \gamma (\mathbf{I} - \gamma \mathbf{P})^{-1} \Delta_{\mathbf{P}} (\mathbf{I} - \gamma \mathbf{P})^{-1} + \gamma^2 (\mathbf{I} - \gamma \mathbf{P}')^{-1} \Delta_{\mathbf{P}} (\mathbf{I} - \gamma \mathbf{P})^{-1} \Delta_{\mathbf{P}} (\mathbf{I} - \gamma \mathbf{P})^{-1}. \end{aligned}$$

Substituting this identity into Eq. (C.17), we obtain

$$\theta(\mathcal{P}) - \theta(\mathcal{P}') = \gamma (\mathbf{I} - \gamma \mathbf{P})^{-1} \Delta_{\mathbf{P}} \theta^* + (\mathbf{I} - \gamma \mathbf{P})^{-1} \Delta_r + \mathcal{R}_{01}, \tag{C.18}$$

where the remainder term $\mathcal{R}_{01}$ takes the form

$$\mathcal{R}_{01} = \gamma^2 (\mathbf{I} - \gamma \mathbf{P}')^{-1} \Delta_{\mathbf{P}} (\mathbf{I} - \gamma \mathbf{P})^{-1} \Delta_{\mathbf{P}} \theta^* + M_{\mathbf{P}} \Delta_r.$$

Since $(1-\gamma)(\mathbf{I} - \gamma \mathbf{P}')^{-1}$ is a probability transition matrix, it follows that $\|(1-\gamma)(\mathbf{I} - \gamma \mathbf{P}')^{-1}\|_\infty \le 1$. Thus, the remainder term $\mathcal{R}_{01}$ satisfies the bound

$$\left\| \mathcal{R}_{01} \right\|_\infty \le \frac{\gamma}{(1-\gamma)} \left\| \Delta_{\mathbf{P}} \right\|_\infty \big\| \gamma (\mathbf{I} - \gamma \mathbf{P})^{-1} \Delta_{\mathbf{P}} \theta^* \big\|_\infty + \frac{\gamma}{(1-\gamma)^2} \left\| \Delta_{\mathbf{P}} \right\|_\infty \left\| \Delta_r \right\|_\infty .$$

The claimed lower bound (C.16) now follows from Eq. (C.18) and the triangle inequality. It is clear that Eq. (C.16) implies the claim in the lemma statement once we restrict $\mathcal{P}' \in \mathcal{S}_1$ and $\mathcal{P}' \in \mathcal{S}_2$.

Proof of Lemma 9.4.4

Throughout the proof, we use $(\mathbf{Z}, R)$ (respectively $(\mathbf{Z}', R')$) to denote a sample drawn from the distribution P (respectively from the distribution P'). We use $P_{\mathbf{Z}}, P_R$ (respectively $P'_{\mathbf{Z}}, P'_R$) to denote the marginal distribution of $\mathbf{Z}, R$ (respectively $\mathbf{Z}'$, R'). By the independence of $\mathbf{Z}$ and R (and similarly for $(\mathbf{Z}', R')$, the joint distributions have the product form

$$P = P_{\mathbf{Z}} \otimes P_R, \quad \text{and} \quad P' = P'_{\mathbf{Z}} \otimes P'_R. \tag{C.19}$$

Proof of part (a): Let $\mathcal{P}' = (\mathbf{P}', R') \in \mathcal{S}_1$ (so $r' = r$). Because of the independence between $\mathbf{Z}$ and R (see Eq. (C.19)) and $r = r'$, we have that

$$d_{\text{hel}}(P, P')^2 = d_{\text{hel}}(P_{\mathbf{Z}}, P_{\mathbf{Z}'})^2.$$

Note that the rows of $\mathbf{Z}$ and $\mathbf{Z}'$ are independent. Thus, if we let $\mathbf{Z}_i, \mathbf{Z}'_i$ denote the i-th rows of $\mathbf{Z}$ and $\mathbf{Z}'$, we have

$$d_{\text{hel}}(P_{\mathbf{Z}}, P_{\mathbf{Z}'})^2 = 1 - \prod_i \left(1 - d_{\text{hel}}(P_{\mathbf{Z}_i}, P_{\mathbf{Z}'_i})^2\right) \leq \sum_i d_{\text{hel}}(P_{\mathbf{Z}_i}, P_{\mathbf{Z}'_i})^2.$$

Now, note that $\mathbf{Z}_i$ and $\mathbf{Z}'_i$ have multinomial distribution with parameters $\mathbf{P}_i$ and $\mathbf{P}'_i$, where $P_{0,i}, P_{1,i}$ are the i-th row of P_0 and P_1. Thus, we have

$$d_{\text{hel}}(P_{\mathbf{Z}_i}, P_{\mathbf{Z}'_i})^2 \leq \frac{1}{2} D_{\chi^2}\left(P_{\mathbf{Z}'_i} \| P_{\mathbf{Z}_i}\right) = \frac{1}{2} \sum_j \frac{\left(\mathbf{P}_{i,j} - \mathbf{P}'_{i,j}\right)^2}{\mathbf{P}_{i,j}}.$$

Putting together the pieces yields the desired upper bound (9.21a).

Proof of part (b): Let $\mathcal{P}' = (\mathbf{P}', R') \in \mathcal{S}_2$ (so $\mathbf{P}' = \mathbf{P}$). Given the independence between $\mathbf{Z}$ and R (see Eq. (C.19)) and $\mathbf{P} = \mathbf{P}'$, we have the relation $d_{\text{hel}}(P, P')^2 = d_{\text{hel}}(P_R, P_{R'})^2$. Note that $R \sim \mathcal{N}(r, \mathbf{I})$ and $R' \sim \mathcal{N}(r', \mathbf{I})$. Thus, we have

$$d_{\text{hel}}(P_R, P_{R'})^2 \leq D_{\text{kl}}\left(P_R \| P_{R'}\right) = \frac{1}{2\sigma_r^2} \|r - r'\|_2^2,$$

as claimed.

Proof of Lemma 9.4.5

We now specify how to construct the probability matrix $\bar{\mathbf{P}}$ that satisfies the desired properties stated in Lemma 9.4.5. We introduce the shorthand notation $\bar{\theta} = \mathbf{P}\theta^*$, and $\mathbf{U} = (\mathbf{I} - \gamma\mathbf{P})^{-1}$. Let $\bar{\ell} \in [D]$ be an index such that

$$\bar{\ell} \in \underset{\ell \in [D]}{\operatorname{argmax}} \left(e_\ell^\top (\mathbf{I} - \gamma\mathbf{P})^{-1} \Sigma(\theta) (\mathbf{I} - \gamma\mathbf{P})^{-\top} e_\ell\right)^{1/2} = \underset{\ell \in [D]}{\operatorname{argmax}} \left(\sum_i \mathbf{U}_{\ell,i}^2 \sigma_i^2(\theta^*)\right)^{1/2}$$

We construct the matrix $\bar{\mathbf{P}}$ entrywise as follows:

$$\bar{\mathbf{P}}_{i,j} = \mathbf{P}_{i,j} + \frac{1}{\nu\sqrt{2N}} \cdot \mathbf{P}_{i,j}\mathbf{U}_{\bar{\ell},i}(\theta^*_j - \bar{\theta}_i)$$

for $\nu \equiv \nu(\mathbf{P}, \theta^*) = \Big(\sum_i \mathbf{U}^2_{\bar{\ell},i}\sigma^2(\theta_i)\Big)^{1/2}$. Now we show that $\bar{\mathbf{P}}$ satisfy the following properties:

(P1) The matrix $\bar{\mathbf{P}}$ is a probability transition matrix.

(P2) It satisfies the constraint $\sum_{i,j} \frac{\left((\mathbf{P}-\bar{\mathbf{P}})_{i,j}\right)^2}{\mathbf{P}_{i,j}} \leq \frac{1}{2N}$.

(P3) It satisfies the inequalities

$$\big\|\mathbf{P} - \bar{\mathbf{P}}\big\|_\infty \leq \frac{1}{\sqrt{2N}}, \quad \text{and} \quad \big\|\gamma(\mathbf{I} - \gamma\mathbf{P})^{-1}(\mathbf{P} - \bar{\mathbf{P}})\theta^*\big\|_\infty \geq \frac{\gamma}{\sqrt{2N}} \cdot \nu(\mathbf{P}, \theta^*). \tag{C.20}$$

We prove each of these properties in turn.

Proof of (P1): For each row $i \in [D]$, we have

$$\sum_j \bar{\mathbf{P}}_{i,j} = \sum_j \mathbf{P}_{i,j} + \frac{1}{\nu\sqrt{2N}}\mathbf{U}_{\bar{\ell},i}\sum_j \mathbf{P}_{i,j}(\theta^*_j - \bar{\theta}_i) = \sum_j \mathbf{P}_{i,j} = 1, \tag{C.21}$$

thus showing that $\bar{\mathbf{P}}\mathbf{1} = \mathbf{1}$ as desired. Moreover, since $(1-\gamma)\mathbf{U} = (1-\gamma)(\mathbf{I} - \gamma\mathbf{P})^{-1}$ is a probability transition matrix, we have the bound $|\mathbf{U}_{\bar{\ell},i}| \leq \frac{1}{1-\gamma}$. By the triangle inequality, we have

$$2\|\theta^*\|_{\mathrm{span}} \geq |\theta^*_j - \bar{\theta}_i|.$$

Thus, our assumption on the sample size N implies that $\nu\sqrt{N} \geq \frac{2}{1-\gamma}\|\theta^*\|_{\mathrm{span}} \geq |\mathbf{U}_{\bar{l},i}(\theta^*_j - \bar{\theta}_j)|$, which further implies that

$$\bar{\mathbf{P}}_{i,j} = \mathbf{P}_{i,j}\left(1 + \frac{1}{\nu\sqrt{2N}} \cdot \mathbf{U}_{\bar{\ell},i}(\theta^*_j - \bar{\theta}_i)\right) \geq 0.$$

In conjunction with the property $\bar{\mathbf{P}}\mathbf{1} = \mathbf{1}$, we conclude that $\bar{\mathbf{P}}$ is a probability transition matrix, as claimed.

Proof of (P2): We begin by observing that $(\Delta_{\mathbf{P}})_{i,j} = \frac{1}{\nu\sqrt{2N}} \cdot \mathbf{P}_{i,j}\mathbf{U}_{\bar{\ell},i}(\theta^*_j - \bar{\theta}_i)$. Now it is simple to check that

$$\sum_{i,j} \frac{\left((\Delta_{\mathbf{P}})_{i,j}\right)^2}{\mathbf{P}_{i,j}} = \frac{1}{2N\nu^2}\sum_{i,j}\mathbf{P}_{i,j}\mathbf{U}^2_{\bar{\ell},i}(\theta^*_j - \bar{\theta}_i)^2 \overset{(i)}{=} \frac{1}{2N\nu^2}\sum_i \mathbf{U}^2_{\bar{\ell},i}\sigma^2_i(\theta^*) = \frac{1}{2N}, \tag{C.22}$$

where in step (i), we use $\sigma^2_i(\theta^*) = \sum_j \mathbf{P}_{i,j}(\theta^*_j - \bar{\theta}_i)^2$ for each i, as the i-th row of our observation $\mathbf{Z}$ is a multinomial distribution with mean specified by the i-th row of $\mathbf{P}$. This proves that $\bar{\mathbf{P}}$ satisfies the constraint, as desired.

Proof of (P3): In order to verify the first inequality, we note that for any row i,

$$\sum_j |(\Delta_{\mathbf{P}})_{i,j}| \overset{(i)}{\leq} \Big(\sum_j \frac{(\Delta_{\mathbf{P}})^2_{i,j}}{\mathbf{P}_{i,j}} \Big)^{1/2} \leq \Big(\sum_{i,j} \frac{(\Delta_{\mathbf{P}})^2_{i,j}}{\mathbf{P}_{i,j}} \Big)^{1/2} \overset{(ii)}{=} \frac{1}{\sqrt{2N}},$$

where step (i) follows from the Cauchy-Schwartz inequality, and step (ii) follows by the previously established Property 2. Taking the maximum over row i yields

$$\|\Delta_{\mathbf{P}}\|_\infty = \max_i \Big\{ \sum_j |(\Delta_{\mathbf{P}})_{i,j}| \Big\} \leq \frac{1}{\sqrt{2N}},$$

thus establishing the first claimed inequality in Eq. (C.20).

In order to establish the second inequality in Eq. (C.20), our starting point is the lower bound

$$\big\|\gamma(\mathbf{I} - \gamma\mathbf{P})^{-1}\Delta_{\mathbf{P}}\theta^*\big\|_\infty \geq \big|e_{\bar{\ell}}^T \gamma(\mathbf{I} - \gamma\mathbf{P})^{-1}\Delta_{\mathbf{P}}\theta^*\big| = \gamma \cdot \Big| \sum_{i,j} \mathbf{U}_{\bar{\ell},i}(\Delta_{\mathbf{P}})_{i,j}\theta^*_j \Big|.$$

It is straightforward to check that

$$\sum_{i,j} \mathbf{U}_{\bar{\ell},i}(\Delta_{\mathbf{P}})_{i,j}\theta^*_j \overset{(i)}{=} \sum_{i,j} \mathbf{U}_{\bar{\ell},i}(\Delta_{\mathbf{P}})_{i,j}(\theta^*_j - \bar{\theta}_i) = \frac{1}{\nu\sqrt{2N}} \sum_{i,j} \mathbf{P}_{i,j}\mathbf{U}^2_{\bar{\ell},i}(\theta^*_j - \bar{\theta}_i)^2 \overset{(ii)}{=} \frac{\nu}{\sqrt{2N}}.$$

Here step (i) follows from the fact that $\sum_j (\Delta_{\mathbf{P}})_{i,j} = 0$ for all i (as $\Delta_{\mathbf{P}}\mathbf{1} = \bar{\mathbf{P}}\mathbf{1} - \mathbf{P}\mathbf{1} = 0$); whereas step (ii) follows from our previous calculation (see Eq. (C.22)) showing that

$$\sum_{i,j} \mathbf{P}_{i,j}\mathbf{U}^2_{\bar{\ell},i}(\theta^*_j - \bar{\theta}_i)^2 = \nu^2.$$

Thus, we have verified the second inequality in Eq. (C.20).

C.3.3 Proofs of auxiliary lemmas for Theorem 9.3.1

This appendix is devoted to the proofs of auxiliary lemmas involved in the proof of Theorem 9.3.1.

Proof of Lemma 9.4.7:

In this section, we prove all three parts of Lemma 9.4.7, which provides high-probability upper bounds on the suboptimality gap at the end of each epoch. Parts (a), (b) and (c), respectively, of Lemma 9.4.7 provides guarantees for the recentered linear stepsize, polynomially-decaying stepsizes and constant stepsize. In order to de-clutter the notation, we omit the dependence on the epoch m in the operators and epoch initialization $\overline{\theta}_m$. In order to distinguish between the total sample size N and the recentering sample size at epoch m, we retain the notation N_m for the recentering sample size.

Proof of part (a): We begin by rewriting the update Eq, (9.26b) in a form suitable for application of general results from [324]. Subtracting off the fixed point $\widehat{\theta}$ of the operator $\mathcal{J}$, we find that

$$\theta_{k+1} - \widehat{\theta} = (1 - \alpha_k)\left(\theta_k - \widehat{\theta}\right) + \alpha_k \left\{\mathcal{J}_k(\theta_k) - \widehat{\theta}\right\}.$$

Note that the operator $\theta \mapsto \widehat{\mathcal{J}}_k(\theta)$ is γ-contractive in the ℓ_∞-norm and monotonic with respect to the orthant ordering; consequently, Corollary 1 from [324] can be applied. In applying this corollary, the effective noise term is given by

$$W_k := \mathcal{J}_k(\widehat{\theta}) - \mathcal{J}(\widehat{\theta}) \; = \; \left\{\widehat{\mathcal{T}}_k(\widehat{\theta}) - \widehat{\mathcal{T}}_k(\overline{\theta})\right\} - \left\{\mathcal{T}(\widehat{\theta}) - \mathcal{T}(\overline{\theta})\right\}.$$

With this setup, by adapting Corollary 1 from [324] we have

$$\left\|\theta_{K+1} - \widehat{\theta}\right\|_\infty \leq \frac{2}{1 + (1-\gamma)K} \left\{\|\overline{\theta} - \widehat{\theta}\|_\infty + \sum_{k=1}^{K} \|V_\ell\|_\infty\right\} + \|V_{K+1}\|_\infty, \tag{C.23a}$$

where the auxiliary stochastic process $\{V_k\}_{k \geq 1}$ evolves according to the recursion

$$V_{k+1} = \big(1 - \alpha_k\big) V_k + \alpha_k W_k. \tag{C.23b}$$

We claim that the ℓ_∞-norm of this process can be bounded with high probability as follows:

Lemma C.3.1. *Consider any sequence of stepsizes $\{\alpha_k\}_{k \geq 1}$ in $(0, 1)$ such that*

$$(1 - \alpha_{k+1})\alpha_k \leq \alpha_{k+1}. \tag{C.24}$$

Then for any tolerance level $\delta > 0$, we have

$$\mathbb{P}\left[\|V_{\ell+1}\|_\infty \geq 4\|\widehat{\theta} - \overline{\theta}\|_\infty \sqrt{\alpha_\ell}\sqrt{\log(8KMD/\delta)}\right] \leq \frac{\delta}{2KM}. \tag{C.25}$$

See Appendix C.3.3 for a proof of this claim. For future reference, note that all three stepsize choices (7.8a)–(7.8c) satisfy the condition (C.24).

Substituting the bound (C.25) into the relation (C.23a) yields

$$\begin{aligned}\left\|\theta_{K+1} - \widehat{\theta}\right\|_\infty &\leq c\left\{\frac{\|\overline{\theta} - \widehat{\theta}\|_\infty}{1 + (1-\gamma)K} + \frac{\|\overline{\theta} - \widehat{\theta}\|_\infty}{(1-\gamma)^{3/2}\sqrt{K}}\right\}\sqrt{\log(8KMD/\delta)} \\ &\leq c\|\overline{\theta} - \widehat{\theta}\|_\infty \left\{\frac{\sqrt{\log(8KMD/\delta)}}{1 + (1-\gamma)K} + \frac{\sqrt{\log(8KMD/\delta)}}{(1-\gamma)^{3/2}\sqrt{K}}\right\},\end{aligned}$$

with probability at least $1 - \frac{\delta}{2M}$. Combining the last bound with the fact that $KM = \frac{N}{2}$ we find that for all $K \geq c_1 \frac{\log(8ND/\delta)}{(1-\gamma)^3}$, we have

$$\|\theta_{K+1} - \widehat{\theta}\|_\infty \leq \tfrac{1}{8}\|\overline{\theta} - \widehat{\theta}\|_\infty \leq \tfrac{1}{8}\|\overline{\theta} - \theta^*\|_\infty + \tfrac{1}{8}\|\widehat{\theta} - \theta^*\|_\infty,$$

which completes the proof of part (a).

Proof of part (b): The proof of part (b) is similar to that of part (a). In particular, adapting Corollary 2 from the paper [324] for polynomial steps, we have

$$\|\theta_{k+1} - \widehat{\theta}\|_\infty \le e^{-\frac{1-\gamma}{1-\omega}\left(k^{1-\omega}-1\right)}\|\overline{\theta} - \widehat{\theta}\|_\infty + e^{-\frac{1-\gamma}{1-\omega}k^{1-\omega}}\sum_{\ell=1}^{k}\frac{e^{\frac{1-\gamma}{1-\omega}\ell^{1-\omega}}}{\ell^{\omega}}\|V_\ell\|_\infty + \|V_{k+1}\|_\infty. \tag{C.26}$$

Recall that polynomial stepsize (7.8b) satisfies the conditions of Lemma C.3.1. Consequently, applying the bound from Lemma C.3.1 we find that

$$\|\theta_{k+1} - \widehat{\theta}\|_\infty \le \|\overline{\theta} - \widehat{\theta}\|_\infty \left\{ e^{-\frac{1-\gamma}{1-\omega}\left(k^{1-\omega}-1\right)} + 4\sqrt{\log(8KMD/\delta)}\left(e^{-\frac{1-\gamma}{1-\omega}k^{1-\omega}}\sum_{\ell=1}^{k}\frac{e^{\frac{1-\gamma}{1-\omega}\ell^{1-\omega}}}{\ell^{3\omega/2}} + \frac{1}{k^{\omega/2}}\right)\right\}. \tag{C.27}$$

It remains to bound the coefficient of $\|\overline{\theta} - \widehat{\theta}\|_\infty$ in the last equation, and we do so by using the following lemma from [324]:

Lemma C.3.2 (Bounds on exponential-weighted sums)**.** *There is a universal constant c such that for all $\omega \in (0, 1)$ and for all $k \ge \left(\frac{3\omega}{2(1-\gamma)}\right)^{\frac{1}{1-\omega}}$, we have*

$$e^{-\frac{1-\gamma}{1-\omega}k^{1-\omega}}\sum_{\ell=1}^{k}\frac{e^{\frac{1-\gamma}{1-\omega}\ell^{1-\omega}}}{\ell^{3\omega/2}} \le c\left\{\frac{e^{-\frac{1-\gamma}{1-\omega}\left(k^{1-\omega}\right)}}{(1-\gamma)^{\frac{1}{1-\omega}}} + \frac{1}{(1-\gamma)}\frac{1}{k^{\omega/2}}\right\}.$$

Substituting the last bound in Eq. (C.26) yields

$$\begin{aligned}
&\left\|\theta_{k+1} - \widehat{\theta}\right\|_\infty \\
&\le c\|\overline{\theta} - \widehat{\theta}\|_\infty \left\{ e^{-\frac{1-\gamma}{1-\omega}\left(k^{1-\omega}-1\right)} + 4\sqrt{\log(8KMD/\delta)}\left(\frac{e^{-\frac{1-\gamma}{1-\omega}\left(k^{1-\omega}-1\right)}}{(1-\gamma)^{\frac{1}{1-\omega}}} + \frac{1}{(1-\gamma)}\frac{1}{k^{\omega/2}} + \frac{1}{k^{\omega/2}}\right)\right\} \\
&\le c\|\overline{\theta} - \widehat{\theta}\|_\infty \cdot \sqrt{\log(8KMD/\delta)}\left\{5\cdot\frac{e^{-\frac{1-\gamma}{1-\omega}\left(k^{1-\omega}-1\right)}}{(1-\gamma)^{\frac{1}{1-\omega}}} + \frac{2}{(1-\gamma)k^{\omega/2}}\right\}.
\end{aligned}$$

Finally, doing some algebra and using the fact that $KM = \frac{N}{2}$ we find that there is an absolute constant c such that for all K lower bounded as $K \ge c\log(4ND/\delta)\cdot\left(\frac{1}{1-\gamma}\right)^{\frac{1}{1-\omega}\vee\frac{2}{\omega}}$, we have

$$\|\theta_{K+1} - \widehat{\theta}\|_\infty \le \frac{\|\overline{\theta} - \widehat{\theta}\|_\infty}{8} \le \frac{1}{8}\|\overline{\theta} - \theta^*\|_\infty + \frac{1}{8}\|\widehat{\theta} - \theta^*\|_\infty.$$

The completes the proof of part (b).

Proof of part (c): Invoking Theorem 1 from [324], we have $\|\theta_K - \widehat{\theta}\|_\infty \leq a_K + b_K + \|V_K\|_\infty$. For a constant stepsize $\alpha_k = \alpha$, the pair (a_K, b_K) is given by

$$\begin{aligned} b_K &= \left\|\bar{\theta} - \widehat{\theta}\right\|_\infty \cdot (1 - \alpha(1-\gamma))^{K-1}, \\ a_K &= \gamma\alpha \left\|V_k\right\|_\infty + \gamma\alpha \left\|V_\ell\right\|_\infty \sum_{k=1}^{K-1} \left\{(1-(1-\gamma)\alpha)^{K-k}\right\} \\ &\overset{(i)}{\leq} \|\bar{\theta} - \widehat{\theta}\|_\infty \cdot \left(2\gamma\alpha^{\frac{3}{2}}\sqrt{\log(8KMD/\delta)} + \frac{2\gamma\alpha^{\frac{1}{2}}}{1-\gamma}\sqrt{\log(8KMD/\delta)}\right), \end{aligned}$$

where inequality (i) follows by substituting $\alpha_k = \alpha$, and using the bound on $\|V_\ell\|_\infty$ from Lemma C.3.1.

It remains to choose the pair (α, K) such that $\|\theta_{K+1} - \widehat{\theta}\|_\infty \leq \frac{1}{8}\|\bar{\theta} - \widehat{\theta}\|_\infty$. Doing some simple algebra and using the fact that $KM = \frac{N}{2}$ we find that it is sufficient to choose the pair (α, K) satisfying the conditions

$$0 < \alpha \leq \frac{(1-\gamma)^2}{\log(4ND/\delta)} \cdot \frac{1}{5^2 \cdot 32^2}, \quad \text{and} \quad K \geq 1 + \frac{2\log 16}{\log\left(\frac{1}{1-\alpha(1-\gamma)}\right)}.$$

With this choice, we have

$$\|\theta_{K+1} - \widehat{\theta}\|_\infty \leq \frac{\|\bar{\theta} - \widehat{\theta}\|_\infty}{8} \leq \frac{1}{8}\left\|\bar{\theta} - \theta^*\right\|_\infty + \frac{1}{8}\left\|\widehat{\theta} - \theta^*\right\|_\infty,$$

which completes the proof of part (c).

Proof of Lemma 9.4.8

Recall our shorthand notation for the local complexities (7.11). The following lemma characterizes the behavior of various random variables as a function of these complexities. In stating the lemma, we let $\widehat{\mathbf{P}}_n$ be a sample transition matrix constructed as the average of n i.i.d. samples, and let $\widehat{r}_n$ denote the reward vector constructed as the average of n i.i.d. samples.

Lemma C.3.3. *Each of the following statements holds with probability exceeding* $1 - \frac{\delta}{M}$*:*

$$\|(\mathbf{I} - \gamma\mathbf{P})^{-1}(\widehat{\mathbf{P}}_n - \mathbf{P})\theta^*\|_\infty \leq 2\nu(\mathbf{P}, \theta^*) \cdot \sqrt{\frac{\log(4DM/\delta)}{n}} + 4 \cdot b(\theta^*) \cdot \frac{\log(4DM/\delta)}{n}, \textit{ and}$$

$$\|(\mathbf{I} - \gamma\mathbf{P})^{-1}(\widehat{r}_n - r)\|_\infty \leq 2\rho(\mathbf{P}, r) \cdot \sqrt{\frac{\log(4DM/\delta)}{n}}.$$

Proof. Entry ℓ of the vector $(\mathbf{I} - \gamma\mathbf{P})^{-1}(\widehat{\mathbf{P}}_n - \mathbf{P})\theta^*$ is zero mean with variance given by the ℓ^{th} diagonal entry of the matrix $(\mathbf{I} - \gamma\mathbf{P})^{-1}\Sigma(\theta^*)(\mathbf{I} - \gamma\mathbf{P})^{-T}$, and is bounded by $b(\theta^*)$ almost surely. Consequently, applying the Bernstein bound in conjunction with the union bound completes the proof of the first claim. In order to establish the second claim, note that the vector $(\mathbf{I} - \gamma\mathbf{P})^{-1}(\widehat{r}_n - r)$ has sub-Gaussian entries, and apply the Hoeffding bound in conjunction with the union bound. □

In light of Lemma C.3.3, note that it suffices to establish the inequality

$$\Pr\left\{\|\widehat{\theta}_m - \theta^*\|_\infty \geq \frac{\|\overline{\theta} - \theta^*\|_\infty}{9} + \|(\mathbf{I} - \gamma \mathbf{P})^{-1}(\widehat{\mathbf{P}}_{N_m} - \mathbf{P})\theta^*\|_\infty + \|(\mathbf{I} - \gamma \mathbf{P})^{-1}(\widehat{r}_{N_m} - r)\|_\infty\right\}$$
$$\leq \frac{\delta}{2M}, \tag{C.28}$$

where we have let $\widehat{\mathbf{P}}_{N_m}$ and $\widehat{r}_{N_m}$ denote the empirical mean of the observed transitions and rewards in epoch m, respectively. The proof of Lemma 9.4.8 follows from Eq. (C.28) by a union bound.

Establishing the bound (C.28)**:** Since the epoch number m should be clear from context, let us adopt the shorthand $\widehat{\theta} \equiv \widehat{\theta}_m$, along with the shorthand $\widehat{r} \equiv \widehat{r}_{N_m}$ and $\widehat{\mathbf{P}} \equiv \widehat{\mathbf{P}}_{N_m}$. Note that $\widehat{\theta}$ is the fixed point of the following operator:

$$\mathscr{J}(\theta) := \mathcal{T}(\theta) - \mathcal{T}(\overline{\theta}) + \widetilde{\mathcal{T}}_{N_m}(\overline{\theta}) = \underbrace{\widehat{r} + \gamma \left(\widehat{\mathbf{P}} - \mathbf{P}\right) \overline{\theta}}_{\widetilde{r}} + \gamma \mathbf{P}\theta,$$

where we have used the fact that $\widetilde{T}_{N_m}(\theta) = \widehat{r} + \gamma \widehat{\mathbf{P}}\theta$.

Thus, we have $\widehat{\theta} = (\mathbf{I} - \gamma \mathbf{P})^{-1}\widetilde{r}$, so that $\widehat{\theta} - \theta^* = (\mathbf{I} - \gamma \mathbf{P})^{-1}\left(\widetilde{r} - r\right)$. Also note that we have

$$\widetilde{r} - r = \widehat{r} + \gamma\left(\widehat{\mathbf{P}} - \mathbf{P}\right)\overline{\theta} - r = \widehat{r} - r + \gamma\left(\widehat{\mathbf{P}} - \mathbf{P}\right)\theta^* + \gamma\left(\widehat{\mathbf{P}} - \mathbf{P}\right)\left(\overline{\theta} - \theta^*\right),$$

so that putting together the pieces and using the triangle inequality yields the bound

$$\begin{aligned}
&\|\widehat{\theta} - \theta^*\|_\infty \\
&\leq \|(\mathbf{I} - \gamma \mathbf{P})^{-1}(\widehat{r} - r)\|_\infty + \gamma\|(\mathbf{I} - \gamma \mathbf{P})^{-1}\left(\widehat{\mathbf{P}} - \mathbf{P}\right)\theta^*\|_\infty + \gamma\|(\mathbf{I} - \gamma \mathbf{P})^{-1}\left(\widehat{\mathbf{P}} - \mathbf{P}\right)\left(\overline{\theta} - \theta^*\right)\|_\infty \\
&\leq \|(\mathbf{I} - \gamma \mathbf{P})^{-1}(\widehat{r} - r)\|_\infty + \gamma\|(\mathbf{I} - \gamma \mathbf{P})^{-1}\left(\widehat{\mathbf{P}} - \mathbf{P}\right)\theta^*\|_\infty + \frac{\gamma}{1-\gamma}\|\left(\widehat{\mathbf{P}} - \mathbf{P}\right)\left(\overline{\theta} - \theta^*\right)\|_\infty.
\end{aligned}$$

Note that the random vector $\left(\widehat{\mathbf{P}} - \mathbf{P}\right)\left(\overline{\theta} - \theta^*\right)$ is the empirical average of N_m i.i.d. random vectors, each of which is bounded entrywise by $2\|\overline{\theta} - \theta^*\|_\infty$. Consequently, by a combination of Hoeffding's inequality and the union bound, we find that

$$\left\|\left(\widehat{\mathbf{P}} - \mathbf{P}\right)\left(\overline{\theta} - \theta^*\right)\right\|_\infty \leq 4\|\overline{\theta} - \theta^*\|_\infty\sqrt{\frac{\log(8DM/\delta)}{N_m}},$$

with probability at least $1 - \frac{\delta}{4M}$. Thus, provided $N_m \geq 4^2 \cdot 9^2 \cdot \frac{\gamma^2}{(1-\gamma)^2}\log(8DM/\delta)$ for a large enough constant c_1, we have

$$\frac{\gamma}{1-\gamma}\left\|\left(\widehat{\mathbf{P}} - \mathbf{P}\right)\left(\overline{\theta} - \theta^*\right)\right\|_\infty \leq \frac{\|\overline{\theta} - \theta^*\|_\infty}{9}.$$

This completes the proof.

Proof of Lemma C.3.1

Recall that by definition, the stochastic process $\{V_k\}_{k\geq 1}$ evolves according to the linear recursion $V_k = (1-\alpha_k)V_{k-1} + \alpha_k W_{k-1}$, where the effective noise sequence $\{W_k\}_{k\geq 0}$ satisfies the uniform bound

$$\|W_k\|_\infty \leq \left\|\widehat{\mathcal{T}}_k(\widehat{\theta}) - \widehat{\mathcal{T}}_k(\overline{\theta})\right\|_\infty + \|\mathcal{T}(\widehat{\theta}) - \mathcal{T}(\overline{\theta})\|_\infty \leq \underbrace{2\|\widehat{\theta} - \overline{\theta}\|_\infty}_{:=b} \quad \text{for all } k \geq 0.$$

Moreover, we have $\mathbb{E}[W_k] = 0$ by construction so that each entry of the random vector W_k is a zero-mean sub-Gaussian random variable with sub-Gaussian parameter at most $2\|\widehat{\theta} - \overline{\theta}\|_\infty$. Consequently, by known properties of sub-Gaussian random variables (cf. Chapter 2 in [323]), we have

$$\log \mathbb{E}\left[e^{sW_k(x)}\right] \leq \frac{s^2 b^2}{8} \qquad \text{for all scalars } s \in \mathbb{R}, \text{ and states } x. \tag{C.29}$$

We complete the proof by using an inductive argument to upper bound the moment generating function of the random variable V_ℓ; given this inequality, we can then apply the Chernoff bound to obtain the stated tail bounds. Beginning with the bound on the moment generating function, we claim that

$$\log \mathbb{E}\left[e^{sV_k(x)}\right] \leq \tfrac{s^2 \alpha_k b^2}{8} \qquad \text{for all scalars } s \in \mathbb{R} \text{ and states } x. \tag{C.30}$$

We prove this claim via induction on k.

Base case: For $k = 1$, we have

$$\log \mathbb{E}\left[e^{sV_1(x)}\right] = \log \mathbb{E}\left[e^{s\alpha_1 W_0(x)}\right] \leq \tfrac{s^2 \alpha_1^2 b^2}{8},$$

where the first equality follows from the definition of V_1, and the second inequality follows by applying the bound (C.29).

Inductive step: We now assume that the bound (C.30) holds for some iteration $k \geq 1$ and prove that it holds for iteration $k+1$. Recalling the definition of V_k, and the independence of the random variables V_k and W_k, we have

$$\begin{aligned}
\mathbb{E}\left[e^{sV_{k+1}(x)}\right] &= \log \mathbb{E}\left[e^{s(1-\alpha_k)V_k(x)}\right] + \log \mathbb{E}\left[e^{s\alpha_k W_k(x)}\right] \\
&\leq \tfrac{s^2(1-\alpha_k)^2 \alpha_{k-1} b^2}{8} + \tfrac{s^2 \alpha_k^2 b^2}{8} \\
&\overset{(i)}{\leq} \tfrac{s^2(1-\alpha_k)\alpha_k b^2}{8} + \tfrac{s^2 \alpha_k^2 b^2}{8} \\
&= \tfrac{s^2 \alpha_k b^2}{8},
\end{aligned}$$

where inequality (i) follows from the assumed condition (C.24) on the stepsizes.

Simple algebra yields that all the stepsize choices (7.8a)–(7.8c) satisfy the condition (C.24). Finally, combining the bound (C.30) with the Chernoff bounding technique along with a union bound over iterations $k = 1, \ldots K$ yields

$$\mathbb{P}\left[\|V_\ell\|_\infty \geq 2b\sqrt{\alpha_{\ell-1}}\sqrt{\log(8KMD/\delta)}\right] \leq \frac{\delta}{8KM},$$

as claimed.

Bibliography

[1] A. Abid, A. Poon, and J. Zou. “Linear Regression with Shuffled Labels”. In: *ArXiv e-prints* (May 2017). arXiv: `1705.01342 [stat.ML]`.

[2] A. Agarwal, N. Jiang, and S. M. Kakade. “Reinforcement Learning: Theory and Algorithms”. In: *Technical Report, CS Department, UW Seattle* (2019).

[3] A. Agarwal, S. Kakade, and L. F. Yang. “Model-Based Reinforcement Learning with a Generative Model is Minimax Optimal”. In: *Proceedings of the 33rd Conference On Learning Theory*. Proceedings of Machine Learning Research. 2020.

[4] C. M. Andersen and R. Bro. “Practical aspects of PARAFAC modeling of fluorescence excitation-emission data”. In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 17.4 (2003), pp. 200–215.

[5] E. Aras, K.-Y. Lee, A. Pananjady, and T. A. Courtade. “A Family of Bayesian Cramér-Rao Bounds, and Consequences for Log-Concave Priors”. In: *2019 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2019, pp. 2699–2703.

[6] M. G. Azar, R. Munos, M. Ghavamzadeh, and H. J. Kappen. “Speedy Q-learning”. In: *Advances in Neural Information Processing Systems*. 2011, pp. 2411–2419.

[7] M. G. Azar, R. Munos, and H. J. Kappen. “Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model”. In: *Machine Learning* 91 (2013), pp. 325–349.

[8] D. Babichev and F. Bach. “Slice inverse regression with score functions”. In: *Electronic Journal of Statistics* 12.1 (2018), pp. 1507–1543.

[9] F. Bach and E. Moulines. “Non-asymptotic analysis of stochastic optimization algorithms for machine learning”. In: *Advances in neural information processing systems*. Dec. 2011.

[10] L. Baird. “Residual algorithms: Reinforcement learning with function approximation”. In: *Machine Learning Proceedings 1995*. Elsevier, 1995, pp. 30–37.

[11] F. Balabdaoui, C. R. Doss, and C. Durot. “Unlinked monotone regression”. In: *arXiv preprint arXiv:2007.00830* (2020).

[12] F. Balabdaoui, C. Durot, and H. Jankowski. “Least squares estimation in the monotone single index model”. In: *Bernoulli* 25.4B (Nov. 2019), pp. 3276–3310.

[13] A. V. Balakrishnan. “On the problem of time jitter in sampling”. In: *IRE Transactions on Information Theory* 8.3 (1962), pp. 226–236.

[14] S. Balakrishnan, M. J. Wainwright, and B. Yu. “Statistical guarantees for the EM algorithm: From population to sample-based analysis”. In: *The Annals of Statistics* 45.1 (2017), pp. 77–120.

[15] G. Balázs. “Convex regression: theory, practice, and applications”. PhD thesis. University of Alberta, 2016.

[16] T. P. Ballinger and N. T. Wilcox. “Decisions, error and heterogeneity”. In: *The Economic Journal* 107.443 (1997), pp. 1090–1105.

[17] L. Baltrunas, T. Makcinskas, and F. Ricci. “Group Recommendations with Rank Aggregation and Collaborative Filtering”. In: *Proceedings of the Fourth ACM Conference on Recommender Systems*. RecSys ’10. Barcelona, Spain: ACM, 2010, pp. 119–126.

[18] B. Barak, S. Hopkins, J. Kelner, P. K. Kothari, A. Moitra, and A. Potechin. “A nearly tight sum-of-squares lower bound for the planted clique problem”. In: *SIAM Journal on Computing* 48.2 (2019), pp. 687–735.

[19] R. E. Barlow, D. J. Bartholomew, J. M. Bremner, and H. D. Brunk. *Statistical inference under order restrictions. The theory and application of isotonic regression*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, London-New York-Sydney, 1972, pp. xii+388.

[20] W. Barnett. “The modern theory of consumer behavior: Ordinal or cardinal?” In: *Quarterly Journal of Austrian Economics* 6.1 (2003), pp. 41–65.

[21] P. L. Bartlett and S. Mendelson. “Rademacher and Gaussian complexities: Risk bounds and structural results”. In: *Journal of Machine Learning Research* 3.Nov (2002), pp. 463–482.

[22] E. M. Beale and R. J. Little. “Missing values in multivariate analysis”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 37.1 (1975), pp. 129–145.

[23] P. C. Bellec and A. B. Tsybakov. “Sharp oracle bounds for monotone and convex regression through aggregation.” In: *The Journal of Machine Learning Research* 16 (2015), pp. 1879–1892.

[24] M. E. Ben-Akiva, S. R. Lerman, and S. R. Lerman. *Discrete choice analysis: theory and application to travel demand*. Vol. 9. MIT press, 1985.

[25] J. O. Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[26] Q. Berthet and P. Rigollet. “Optimal detection of sparse principal components in high dimension”. In: *The Annals of Statistics* 41.4 (2013), pp. 1780–1815.

[27] D. P. Bertsekas. *Dynamic programming and stochastic control*. Vol. 1. Belmont, MA: Athena Scientific, 1995.

[28] D. Bertsekas. *Dynamic programming and stochastic control*. Vol. 2. Belmont, MA: Athena Scientific, 1995.

[29] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

[30] J. Bhandari, D. Russo, and R. Singal. “A Finite Time Analysis of Temporal Difference Learning With Linear Function Approximation”. In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, June 2018, pp. 1691–1692.

[31] K. Bhatia, A. Pananjady, P. L. Bartlett, A. Dragan, and M. J. Wainwright. “Preference learning along multiple criteria: A game theoretic perspective”. In: *Preprint* (2020).

[32] P. J. Bickel, A. Chen, and E. Levina. “The method of moments and degree distributions for network models”. In: *The Annals of Statistics* 39.5 (Oct. 2011), pp. 2280–2301.

[33] P. J. Bickel, C. A. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore, 1993.

[34] L. Birgé. “Approximation dans les espaces métriques et théorie de l’estimation”. In: *Zeitschrift für Wahrscheinlichkeitstheorie und verwebte Gebiet* 65 (1983), pp. 181–238.

[35] L. Birgé and P. Massart. “Rates of convergence for minimum contrast estimators”. In: *Probability Theory and Related Fields* 97.1-2 (1993), pp. 113–150.

[36] V. S. Borkar. “Asynchronous stochastic approximations”. In: *SIAM Journal on Control and Optimization* 36.3 (1998), pp. 840–851.

[37] V. S. Borkar and S. P. Meyn. “The ODE method for convergence of stochastic approximation and reinforcement learning”. In: *SIAM Journal on Control and Optimization* 38.2 (2000), pp. 447–469.

[38] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[39] J. A. Boyan. “Technical update: Least-squares temporal difference learning”. In: *Machine learning* 49.2-3 (2002), pp. 233–246.

[40] R. A. Bradley and M. E. Terry. “Rank analysis of incomplete block designs. I. The method of paired comparisons”. In: *Biometrika* 39 (1952), pp. 324–345.

[41] S. J. Bradtke and A. G. Barto. “Linear least-squares algorithms for temporal difference learning”. In: *Machine learning* 22.1-3 (1996), pp. 33–57.

[42] M. Braverman and E. Mossel. “Noisy sorting without resampling”. In: *Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms*. ACM, New York, 2008, pp. 268–276.

[43] M. Brennan and G. Bresler. “Reducibility and Statistical-Computational Gaps from Secret Leakage”. In: *Proceedings of the 33rd Conference On Learning Theory*. Proceedings of Machine Learning Research. 2020.

[44] M. Brennan, G. Bresler, and W. Huleihel. “Reducibility and Computational Lower Bounds for Problems with Planted Sparse Structure”. In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, June 2018, pp. 48–166.

[45] D. R. Brillinger. “A generalized linear model with “Gaussian” regressor variables”. In: *A Festschrift for Erich L. Lehmann*. Wadsworth Statist./Probab. Ser. Wadsworth, Belmont, CA, 1983, pp. 97–114.

[46] E. Bronshtein. “ε-entropy of convex sets and functions”. In: *Siberian Mathematical Journal* 17.3 (1976), pp. 393–398.

[47] H. D. Brunk. “Maximum likelihood estimates of monotone parameters”. In: *The Annals of Mathematical Statistics* (1955), pp. 607–616.

[48] T. T. Cai and M. G. Low. “An adaptation theory for nonparametric confidence intervals”. In: *The Annals of Statistics* 32.5 (2004), pp. 1805–1840.

[49] T. Cai and M. Low. “A framework for estimating convex functions”. In: *Statistica Sinica* 25 (2015), pp. 423–456.

[50] E. J. Candes and T. Tao. “Near-optimal signal recovery from random projections: Universal encoding strategies?” In: *Information Theory, IEEE Transactions on* 52.12 (2006), pp. 5406–5425.

[51] A. Carpentier and T. Schlueter. “Learning Relationships between Data Obtained Independently”. In: *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*. 2016, pp. 658–666.

[52] A. T. Chaganty and P. Liang. “Spectral experts for estimating mixtures of linear regressions”. In: *International Conference on Machine Learning*. 2013, pp. 1040–1048.

[53] S. Chan and E. Airoldi. “A consistent histogram estimator for exchangeable graph models”. In: *International Conference on Machine Learning*. 2014, pp. 208–216.

[54] S. R. Chandukala, J. Kim, T. Otter, and G. M. Allenby. *Choice models in marketing: Economic assumptions, challenges and trends*. Now Publishers Inc, 2008.

[55] S. Chatterjee and S. Mukherjee. “Estimation in Tournaments and Graphs Under Monotonicity Constraints”. In: *IEEE Transactions on Information Theory* 65.6 (June 2019), pp. 3525–3539.

[56] S. Chatterjee, A. Guntuboyina, and B. Sen. “On matrix estimation under monotonicity constraints”. In: *Bernoulli* 2 (May 2018), pp. 1072–1100.

[57] S. Chatterjee and J. Lafferty. “Adaptive risk bounds in unimodal regression”. In: *Bernoulli* 25.1 (2019), pp. 1–25.

[58] S. Chatterjee. “A New Perspective on Least Squares under Convex Constraint”. In: *The Annals of Statistics* 42.6 (Dec. 2014), pp. 2340–2381. arXiv: `1402.0830`.

[59] S. Chatterjee. “Matrix estimation by universal singular value thresholding”. In: *The Annals of Statistics* 43.1 (2015), pp. 177–214.

[60] X. Chen, P. N. Bennett, K. Collins-Thompson, and E. Horvitz. “Pairwise ranking aggregation in a crowdsourced setting”. In: *Proceedings of the sixth ACM international conference on Web search and data mining*. ACM. 2013, pp. 193–202.

[61] Y. Chen and R. J. Samworth. “Generalized additive and index models with shape constraints”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 78.4 (2016), pp. 729–754.

[62] Y. Chen, Y. Chi, J. Fan, C. Ma, and Y. Yan. “Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization”. In: *arXiv preprint arXiv:1902.07698* (2019).

[63] Y. Chen, J. Fan, C. Ma, and K. Wang. “Spectral method and regularized MLE are both optimal for top-K ranking”. In: *The Annals of Statistics* 47.4 (2019), pp. 2204–2235.

[64] Y. Chen and C. Suh. “Spectral MLE: Top-k rank aggregation from pairwise comparisons”. In: *International Conference on Machine Learning*. 2015, pp. 371–380.

[65] Z. Chen, S. T. Maguluri, S. Shakkottai, and K. Shanmugam. “Finite-Sample Analysis of Stochastic Approximation Using Smooth Convex Envelopes”. In: *arXiv preprint arXiv:2002.00874* (2020).

[66] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. *Double/debiased machine learning for treatment and structural parameters*. 2018.

[67] V. Chernozhukov, M. Goldman, V. Semenova, and M. Taddy. “Orthogonal machine learning for demand estimation: High dimensional causal inference in dynamic panels”. In: *arXiv preprint arXiv:1712.09988* (2017).

[68] S. Chrétien and A. O. Hero. “On EM algorithms and their proximal generalizations”. In: *ESAIM: Probability and Statistics* 12 (2008), pp. 308–326.

[69] O. Collier and A. S. Dalalyan. “Minimax Rates in Permutation Estimation for Feature Matching”. In: *Journal of Machine Learning Research* 17.6 (2016), pp. 1–31.

[70] T. A. Courtade, M. Fathi, and A. Pananjady. “Existence of Stein kernels under a spectral gap, and discrepancy bounds”. In: *Ann. Inst. H. Poincaré Probab. Statist.* 55.2 (May 2019), pp. 777–790.

[71] T. A. Courtade, M. Fathi, and A. Pananjady. “Quantitative stability of the entropy power inequality”. In: *IEEE Transactions on Information Theory* 64.8 (2018), pp. 5691–5703.

[72] G. Dalal, B. Szörényi, G. Thoppe, and S. Mannor. “Finite sample analyses for TD(0) with function approximation”. In: *Thirty-Second AAAI Conference on Artificial Intelligence*. 2018.

[73] A. S. Dalalyan, A. Juditsky, and V. Spokoiny. “A new algorithm for estimating the effective dimension-reduction subspace”. In: *Journal of Machine Learning Research* 9.Aug (2008), pp. 1647–1678.

[74] N. Dalvi, A. Dasgupta, R. Kumar, and V. Rastogi. “Aggregating crowdsourced binary ratings”. In: *Proceedings of the 22nd international conference on World Wide Web*. ACM. 2013, pp. 285–294.

[75] A. Daniely, S. Sabato, and S. S. Shwartz. “Multiclass learning approaches: A theoretical comparison with implications”. In: *Advances in Neural Information Processing Systems*. 2012, pp. 485–493.

[76] C. Dann and E. Brunskill. “Sample complexity of episodic fixed-horizon reinforcement learning”. In: *Advances in Neural Information Processing Systems*. 2015, pp. 2818–2826.

[77] C. Dann, G. Neumann, and J. Peters. “Policy evaluation with temporal differences: A survey and comparison”. In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 809–883.

[78] S. Dasgupta and A. Gupta. “An elementary proof of a theorem of Johnson and Lindenstrauss”. In: *Random Structures & Algorithms* 22.1 (2003), pp. 60–65.

[79] C. Daskalakis, C. Tzamos, and M. Zampetakis. “Ten Steps of EM Suffice for Mixtures of Two Gaussians”. In: *30th Annual Conference on Learning Theory*. 2017.

[80] A. P. Dawid and A. M. Skene. “Maximum likelihood estimation of observer error-rates using the EM algorithm”. In: *Applied Statistics* (1979), pp. 20–28.

[81] V. de la Pena and E. Giné. *Decoupling: From dependence to independence*. Springer Science & Business Media, 2012.

[82] M. H. DeGroot and P. K. Goel. “Estimation of the correlation coefficient from a broken random sample”. In: *The Annals of Statistics* 8.2 (1980), pp. 264–278.

[83] H. Deng and C.-H. Zhang. “Isotonic regression in multidimensional spaces and graphs”. In: *arXiv preprint arXiv:1812.08944* (2018).

[84] A. M. Devraj and S. P. Meyn. “Fastest convergence for Q-learning”. In: *arXiv preprint arXiv:1707.03770* (2017).

[85] S. Dirksen. “Tail bounds via generic chaining”. In: *Electronic Journal of Probability* 20 (2015).

[86] T. T. Doan, S. T. Maguluri, and J. Romberg. “Finite-time performance of distributed temporal difference learning with linear function approximation”. In: *arXiv preprint arXiv:1907.12530* (2019).

[87] D. L. Donoho and R. C. Liu. *Geometrizing Rates of Convergence I*. Tech. rep. 137. University of California, Berkeley, Department of Statistics, 1987.

[88] D. L. Donoho and R. C. Liu. “Geometrizing Rates of Convergence II”. In: *Annals of Statistics* 19.2 (1991), pp. 633–667.

[89] R. Dudeja and D. Hsu. “Learning Single-Index Models in Gaussian Space”. In: *Conference On Learning Theory*. 2018, pp. 1887–1930.

[90] R. Durrett. *Essentials of stochastic processes*. Vol. 1. Springer, 1999.

[91] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. “Rank aggregation methods for the web”. In: *Proceedings of the 10th International Conference on World Wide Web*. ACM. 2001, pp. 613–622.

[92] R. L. Dykstra and T. Robertson. “An algorithm for isotonic regression for two or more independent variables”. In: *The Annals of Statistics* (1982), pp. 708–716.

[93] G. Elhami, A. Scholefield, B. B. Haro, and M. Vetterli. “Unlabeled sensing: Reconstruction algorithm and theoretical guarantees”. In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. Mar. 2017, pp. 4566–4570.

[94] V. Emiya, A. Bonnefoy, L. Daudet, and R. Gribonval. “Compressed sensing with unknown sensor permutation”. In: *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE. 2014, pp. 1040–1044.

[95] E. Even-Dar and Y. Mansour. “Learning rates for Q-learning”. In: *Journal of machine learning research* 5 (2003), pp. 1–25.

[96] M. Falahatgar, A. Orlitsky, V. Pichapati, and A. T. Suresh. “Maximum Selection and Ranking under Noisy Comparisons”. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by D. Precup and Y. W. Teh. Vol. 70. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia: PMLR, June 2017, pp. 1088–1096.

[97] V. F. Farias, S. Jagabathula, and D. Shah. “A nonparametric approach to modeling choice with limited data”. In: *Management science* 59.2 (2013), pp. 305–322.

[98] U. Feige and R. Krauthgamer. “The probable value of the Lovász–Schrijver relaxations for maximum independent set”. In: *SIAM Journal on Computing* 32.2 (2003), pp. 345–370.

[99] W. Feller. *An Introduction to Probability Theory and its Applications: Volume II*. New York: John Wiley and Sons, 1966.

[100] M. Fickus, D. G. Mixon, A. A. Nelson, and Y. Wang. “Phase retrieval from very few measurements”. In: *Linear Algebra and its Applications* 449 (2014), pp. 475–499.

[101] C. Fienup and J. Dainty. “Phase retrieval and image reconstruction for astronomy”. In: *Image Recovery: Theory and Application* 231 (1987), p. 275.

[102] J. R. Fienup. “Phase retrieval algorithms: a comparison”. In: *Applied optics* 21.15 (1982), pp. 2758–2769.

[103] P. C. Fishburn. “Binary choice probabilities: on the varieties of stochastic transitivity”. In: *Journal of Mathematical psychology* 10.4 (1973), pp. 327–352.

[104] N. Flammarion, C. Mao, and P. Rigollet. “Optimal rates of statistical seriation”. In: *Bernoulli* 25.1 (2019), pp. 623–653.

[105] F. Fogel, R. Jenatton, F. Bach, and A. d’Aspremont. “Convex relaxations for permutation problems”. In: *Advances in Neural Information Processing Systems*. 2013, pp. 1016–1024.

[106] F. Fogel, I. Waldspurger, and A. d'Aspremont. "Phase retrieval for imaging problems". In: *Mathematical programming computation* 8.3 (2016), pp. 311–335.

[107] K. Fokianos, A. Leucht, and M. H. Neumann. "On Integrated L_1 Convergence Rate of an Isotonic Regression Estimator for Multivariate Observations". In: *arXiv preprint arXiv:1710.04813* (2017).

[108] D. J. Foster and V. Syrgkanis. "Orthogonal statistical learning". In: *arXiv preprint arXiv:1901.09036* (2019).

[109] J. Frank, S. Mannor, and D. Precup. "Reinforcement learning in the presence of rare events". In: *Proceedings of the International conference on Machine learning*. ACM. 2008, pp. 336–343.

[110] R. Ganti, N. Rao, R. M. Willett, and R. Nowak. "Learning single index models in high dimensions". In: *arXiv preprint arXiv:1506.08910* (2015).

[111] C. Gao. "Phase Transitions in Approximate Ranking". In: *arXiv preprint arXiv:1711.11189* (2017).

[112] C. Gao, F. Han, and C.-H. Zhang. "On estimation of isotonic piecewise constant signals". In: *The Annals of Statistics* 48.2 (2020), pp. 629–654.

[113] C. Gao, Y. Lu, and H. H. Zhou. "Rate-optimal graphon estimation". In: *The Annals of Statistics* 43.6 (Dec. 2015), pp. 2624–2652.

[114] F. Gao and J. A. Wellner. "Entropy estimate for high-dimensional monotonic functions". In: *Journal of Multivariate Analysis* 98.9 (2007), pp. 1751–1764.

[115] F. Gao and J. A. Wellner. "Entropy of Convex Functions on $\mathbb{R}^d$". In: *Constructive approximation* 46.3 (2017), pp. 565–592.

[116] R. J. Gardner. *Geometric Tomography*. 2nd ed. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 2006.

[117] R. Gerchberg and W. Saxton. "A practical algorithm for the determination of phase from image and diffraction plane pictures". In: *Optik* 35 (1972), pp. 237–246.

[118] A. Ghosh, A. Pananjady, A. Guntuboyina, and K. Ramchandran. "Max-affine regression with universal parameter estimation for small-ball designs". In: *2020 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2020.

[119] A. Ghosh, A. Pananjady, A. Guntuboyina, and K. Ramchandran. "Max-Affine Regression: Provable, Tractable, and Near-Optimal Statistical Estimation". In: *arXiv preprint arXiv:1906.09255* (2019).

[120] D. Ghoshdastidar and A. Dukkipati. "Consistency of spectral hypergraph partitioning under planted partition model". In: *The Annals of Statistics* 45.1 (2017), pp. 289–315.

[121] E. N. Gilbert. "A comparison of signalling alphabets". In: *Bell Labs Technical Journal* 31.3 (1952), pp. 504–522.

[122] E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Vol. 40. Cambridge University Press, 2016.

[123] A. Gosavi. "Reinforcement learning: A tutorial survey and recent advances". In: *INFORMS Journal on Computing* 21.2 (2009), pp. 178–192.

[124] H. Grad. "Note on N-dimensional hermite polynomials". In: *Communications on Pure and Applied Mathematics* 2.4 (1949), pp. 325–330.

[125] J. Gregor and F. R. Rannou. "Three-dimensional support function estimation and application for projection magnetic resonance imaging". In: *International journal of imaging systems and technology* 12.1 (2002), pp. 43–50.

[126] P. Groeneboom and K. Hendrickx. "Estimation in monotone single-index models". In: *Statistica Neerlandica* 73.1 (2019), pp. 78–99.

[127] P. Groeneboom, G. Jongbloed, and J. A. Wellner. "Estimation of a convex function: characterizations and asymptotic theory". In: *The Annals of Statistics* 29.6 (2001), pp. 1653–1698.

[128] S. J. Grotzinger and C. Witzgall. "Projections onto order simplexes". In: *Appl. Math. Optim.* 12.3 (1984), pp. 247–270.

[129] A. Guntuboyina. "Optimal rates of convergence for convex set estimation from support functions". In: *The Annals of Statistics* 40.1 (2012), pp. 385–411.

[130] A. Guntuboyina, D. Lieu, S. Chatterjee, and B. Sen. "Adaptive risk bounds in univariate total variation denoising and trend filtering". In: *The Annals of Statistics* 48.1 (2020), pp. 205–229.

[131] A. Guntuboyina and B. Sen. "Covering numbers for convex functions". In: *IEEE Transactions on Information Theory* 59.4 (2013), pp. 1957–1965.

[132] A. Guntuboyina and B. Sen. "Global risk bounds and adaptation in univariate convex regression". In: *Probability Theory and Related Fields* 163.1-2 (2015), pp. 379–411.

[133] S. Haghighatshoar and G. Caire. "Signal recovery from unlabeled samples". In: *arXiv preprint arXiv:1701.08701* (2017).

[134] B. Hajek, S. Oh, and J. Xu. "Minimax-optimal inference from partial rankings". In: *Advances in Neural Information Processing Systems*. 2014, pp. 1475–1483.

[135] J. Hájek. "Local asymptotic minimax and admissibility in estimation". In: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*. 1972, pp. 175–194.

[136] Q. Han. "Global empirical risk minimizers with "shape constraints" are rate optimal in general dimensions". In: *arXiv preprint arXiv:1905.12823* (2019).

[137] Q. Han, T. Wang, S. Chatterjee, and R. J. Samworth. "Isotonic regression in general dimensions". In: *The Annals of Statistics* 47.5 (Oct. 2019), pp. 2440–2471.

[138] Q. Han and J. A. Wellner. “Multivariate convex regression: global risk bounds and adaptation”. In: *arXiv preprint arXiv:1601.06844* (2016).

[139] L. A. Hannah and D. B. Dunson. “Multivariate convex regression with adaptive partitioning”. In: *The Journal of Machine Learning Research* 14.1 (2013), pp. 3261–3294.

[140] J.-B. Hardouin and M. Mesbah. “Clustering binary variables in subscales using an extended Rasch model and Akaike information criterion”. In: *Communications in Statistics-Theory and Methods* 33.6 (2004), pp. 1277–1294.

[141] R. W. Harrison. “Phase problem in crystallography”. In: *JOSA a* 10.5 (1993), pp. 1046–1055.

[142] H. O. Hartley. “Maximum likelihood estimation from incomplete data”. In: *Biometrics* 14.2 (1958), pp. 174–194.

[143] S. B. K. Hopkins. “Statistical inference and the sum of squares method”. PhD thesis. 2018.

[144] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge: Cambridge University Press, 1985.

[145] J. L. Horowitz. *Semiparametric and nonparametric methods in econometrics*. Vol. 12. Springer, 2009.

[146] M. Hristache, A. Juditsky, J. Polzehl, and V. Spokoiny. “Structure adaptive approach for dimension reduction”. In: *The Annals of Statistics* 29.6 (2001), pp. 1537–1566.

[147] D. J. Hsu, K. Shi, and X. Sun. “Linear regression without correspondence”. In: *Advances in Neural Information Processing Systems*. 2017, pp. 1531–1540.

[148] D. Hsu and S. M. Kakade. “Learning mixtures of spherical Gaussians: moment methods and spectral decompositions”. In: *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*. ACM. 2013, pp. 11–20.

[149] X. Huang and A. Madan. “CAP3: A DNA sequence assembly program”. In: *Genome Research* 9.9 (1999), pp. 868–877.

[150] T. Jaakkola, M. I. Jordan, and S. P. Singh. “Convergence of stochastic iterative dynamic programming algorithms”. In: *Advances in neural information processing systems*. 1994, pp. 703–710.

[151] R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. “Adaptive mixtures of local experts.” In: *Neural computation* 3.1 (1991), pp. 79–87.

[152] M. Jerrum. “Large cliques elude the Metropolis process”. In: *Random Structures & Algorithms* 3.4 (1992), pp. 347–359.

[153] M. R. Jerrum, L. G. Valiant, and V. V. Vazirani. “Random generation of combinatorial structures from a uniform distribution”. In: *Theoretical Computer Science* 43 (1986), pp. 169–188.

[154] N. Jiang and A. Agarwal. “Open problem: The dependence of sample complexity lower bounds on planning horizon”. In: *Proceedings of the Conference On Learning Theory*. 2018, pp. 3395–3398.

[155] C. Jin, Y. Zhang, S. Balakrishnan, M. J. Wainwright, and M. I. Jordan. “Local maxima in the likelihood of Gaussian mixture models: Structural results and algorithmic consequences”. In: *Advances in neural information processing systems*. 2016, pp. 4116–4124.

[156] R. Johnson and T. Zhang. “Accelerating stochastic gradient descent using predictive variance reduction”. In: *Advances in Neural Information Processing Systems*. 2013, pp. 315–323.

[157] I. M. Johnstone. “Function estimation and Gaussian sequence models”. In: (2002).

[158] M. I. Jordan and R. A. Jacobs. “Hierarchical mixtures of experts and the EM algorithm”. In: *Neural computation* 6.2 (1994), pp. 181–214.

[159] A. B. Kahn. “Topological sorting of large networks”. In: *Communications of the ACM* 5.11 (1962), pp. 558–562.

[160] S. Kakade. “On the sample complexity of reinforcement learning”. PhD thesis. 2003.

[161] S. M. Kakade, V. Kanade, O. Shamir, and A. Kalai. “Efficient learning of generalized linear and single index models with isotonic regression”. In: *Advances in Neural Information Processing Systems*. 2011, pp. 927–935.

[162] A. T. Kalai and R. Sastry. “The Isotron Algorithm: High-Dimensional Isotonic Regression.” In: *COLT*. Citeseer. 2009.

[163] M. Kanter and H. Proppe. “Reduction of variance for Gaussian densities via restriction to convex sets”. In: *Journal of Multivariate Analysis* 7.1 (1977), pp. 74–81.

[164] D. R. Karger, S. Oh, and D. Shah. “Budget-optimal crowdsourcing using low-rank matrix approximations”. In: *Communication, Control, and Computing (Allerton), 2011 49th Annual Allerton Conference on*. IEEE. 2011, pp. 284–291.

[165] D. R. Karger, S. Oh, and D. Shah. “Iterative learning for reliable crowdsourcing systems”. In: *Advances in neural information processing systems*. 2011, pp. 1953–1961.

[166] M. Kearns and S. Singh. “Finite-sample convergence rates for Q-learning and indirect algorithms”. In: *Advances in neural information processing systems*. 1999.

[167] R. W. Keener. *Theoretical statistics: Topics for a core course*. Springer, 2011.

[168] L. Keller, M. J. Siavoshani, C. Fragouli, K. Argyraki, and S. Diggavi. “Identity aware sensor networks”. In: *INFOCOM 2009, IEEE*. IEEE. 2009, pp. 2177–2185.

[169] K. Khamaru, A. Pananjady, F. Ruan, M. J. Wainwright, and M. I. Jordan. “Is temporal difference learning optimal? An instance-dependent analysis”. In: *arXiv preprint arXiv:2003.07337* (2020).

[170] A. G. Kök, M. L. Fisher, and R. Vaidyanathan. “Assortment planning: Review of literature and industry practice”. In: *Retail supply chain management*. Springer, 2008, pp. 99–153.

[171] T. G. Kolda, B. W. Bader, and J. P. Kenny. “Higher-order web link analysis using multilinear algebra”. In: *Fifth IEEE International Conference on Data Mining (ICDM’05)*. IEEE. 2005, 8–pp.

[172] A. N. Kolmogorov and V. M. Tikhomirov. “ε-entropy and ε-capacity of sets in function spaces”. In: *Uspekhi Matematicheskikh Nauk* 14.2 (1959), pp. 3–86.

[173] V. Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Vol. 2033. Springer Science & Business Media, 2011.

[174] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. “Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion”. In: *The Annals of Statistics* 39.5 (2011), pp. 2302–2329.

[175] N. Korda and P. La. “On TD(0) with function approximation: Concentration bounds and a centered variant with exponential convergence”. In: *International Conference on Machine Learning*. 2015, pp. 626–634.

[176] M. R. Kosorok. *Introduction to empirical processes and semiparametric inference.* Springer, 2008.

[177] D. H. Krantz. “The scaling of small and large color differences.” PhD thesis. 1965.

[178] A. K. Kuchibhotla, R. K. Patra, and B. Sen. “Efficient estimation in convex single index models”. In: *arXiv preprint arXiv:1708.00145* (2017).

[179] D. Kunisky, A. S. Wein, and A. S. Bandeira. “Notes on computational hardness of hypothesis testing: Predictions using the low-degree likelihood ratio”. In: *arXiv preprint arXiv:1907.11636* (2019).

[180] G. Kur, F. Gao, A. Guntuboyina, and B. Sen. “Convex Regression in Multidimensions: Suboptimality of Least Squares Estimators”. In: *arXiv preprint arXiv:2006.02044* (2020).

[181] J. Kwon, W. Qian, C. Caramanis, Y. Chen, and D. Davis. “Global Convergence of the EM Algorithm for Mixtures of Two Component Linear Regression”. In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Ed. by A. Beygelzimer and D. Hsu. Vol. 99. Proceedings of Machine Learning Research. Phoenix, USA: PMLR, June 2019, pp. 2055–2110.

[182] R. Kyng, A. Rao, and S. Sachdeva. “Fast, provable algorithms for isotonic regression in all ℓ_p-norms”. In: *Advances in neural information processing systems*. 2015, pp. 2719–2727.

[183] C. Lakshminarayanan and C. Szepesvari. “Linear Stochastic Approximation: How Far Does Constant Step-Size and Iterate Averaging Go?” In: *Proceedings of the International Conference on Artificial Intelligence and Statistics*. 2018, pp. 1347–1355.

[184] T. Lattimore and M. Hutter. “Near-optimal PAC bounds for discounted MDPs”. In: *Theoretical Computer Science* 558 (2014), pp. 125–143.

[185] B. Laurent and P. Massart. “Adaptive estimation of a quadratic functional by model selection”. In: *The Annals of Statistics* (2000), pp. 1302–1338.

[186] L. Le Cam. "Limits of experiments". In: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*. 1972, pp. 245–261.

[187] L. Le Cam and G. L. Yang. *Asymptotics in Statistics: Some Basic Concepts*. Springer, 2000.

[188] G. Lecué and M. Lerasle. "Robust machine learning by median-of-means: Theory and practice". In: *Ann. Statist.* 48.2 (Apr. 2020), pp. 906–931.

[189] G. Lecué and S. Mendelson. "Regularization and the small-ball method II: Complexity dependent error rates". In: *The Journal of Machine Learning Research* 18.1 (2017), pp. 5356–5403.

[190] M. Ledoux. *The concentration of measure phenomenon*. 89. American Mathematical Soc., 2001.

[191] C. Lee and D. Shah. "Reducing Crowdsourcing to Graphon Estimation, Statistically". In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. Vol. 84. Proceedings of Machine Learning Research. PMLR, 2018, pp. 1741–1750.

[192] E. L. Lehmann and G. Casella. *Theory of point estimation*. Vol. 31. Springer Science & Business Media, 1998.

[193] O. V. Lepski. "On a problem of adaptive estimation in Gaussian white noise". In: *Theory of Probability & Its Applications* 35.3 (1991), pp. 454–466.

[194] O. V. Lepski and V. G. Spokoiny. "Optimal pointwise adaptive methods in nonparametric estimation". In: *The Annals of Statistics* (1997), pp. 2512–2546.

[195] K.-C. Li. "On principal Hessian directions for data visualization and dimension reduction: Another application of Stein's lemma". In: *Journal of the American Statistical Association* 87.420 (1992), pp. 1025–1039.

[196] K.-C. Li. "Sliced inverse regression for dimension reduction". In: *Journal of the American Statistical Association* 86.414 (1991), pp. 316–327.

[197] Q. Li and J. S. Racine. *Nonparametric econometrics: theory and practice*. Princeton University Press, 2007.

[198] E. Lim and P. W. Glynn. "Consistency of multidimensional convex regression". In: *Operations Research* 60.1 (2012), pp. 196–208.

[199] D. V. Lindley. "On a measure of the information provided by an experiment". In: *The Annals of Mathematical Statistics* (1956), pp. 986–1005.

[200] R. J. A. Little and D. B. Rubin. *Statistical analysis with missing data*. John Wiley & Sons, 2014.

[201] A. Liu and A. Moitra. "Better Algorithms for Estimating Non-Parametric Models in Crowd-Sourcing and Rank Aggregation". In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by J. Abernethy and S. Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 2780–2829.

[202] Q. Liu, J. Peng, and A. T. Ihler. “Variational inference for crowdsourcing”. In: *Advances in neural information processing systems*. 2012, pp. 692–700.

[203] P. Loh and M. J. Wainwright. “Corrupted and missing predictors: Minimax bounds for high-dimensional linear regression”. In: *Information Theory Proceedings (ISIT), 2012 IEEE International Symposium on*. IEEE. 2012, pp. 2601–2605.

[204] R. D. Luce. *Individual choice behavior: A theoretical analysis*. Wiley New York, 1959.

[205] Y. Luo and A. R. Zhang. “Tensor Clustering with Planted Structures: Statistical Optimality and Computational Limits”. In: *arXiv preprint arXiv:2005.10743* (2020).

[206] R. Ma, T. T. Cai, and H. Li. “Optimal and Adaptive Estimation of Extreme Values in the Permuted Monotone Matrix Model”. In: *arXiv preprint arXiv:1911.12516* (2019).

[207] Z. Ma and Y. Wu. “Computational barriers in minimax submatrix detection”. In: *The Annals of Statistics* 43.3 (2015), pp. 1089–1116.

[208] A. Magnani and S. P. Boyd. “Convex piecewise-linear fitting”. In: *Optimization and Engineering* 10.1 (2009), pp. 1–17.

[209] O.-A. Maillard, T. A. Mann, and S. Mannor. “How hard is my MDP? The distribution-norm to the rescue”. In: *Advances in Neural Information Processing Systems*. 2014, pp. 1835–1843.

[210] D. Malik, A. Pananjady, K. Bhatia, K. Khamaru, P. L. Bartlett, and M. J. Wainwright. “Derivative-Free Methods for Policy Optimization: Guarantees for Linear Quadratic Systems”. In: *Journal of Machine Learning Research* 21.21 (2020), pp. 1–51.

[211] C. Mao, A. Pananjady, and M. J. Wainwright. “Towards optimal estimation of bivariate isotonic matrices with unknown permutations”. In: *The Annals of Statistics, to appear* (2019+).

[212] C. Mao, J. Weed, and P. Rigollet. “Minimax Rates and Efficient Algorithms for Noisy Sorting”. In: *Proceedings of Algorithmic Learning Theory*. Ed. by F. Janoos, M. Mohri, and K. Sridharan. Vol. 83. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 821–847.

[213] M. Marques, M. Stovsić, and J. Costeira. “Subspace matching: Unique solution to point matching with geometric constraints”. In: *Computer Vision, IEEE 12th International Conference on*. IEEE. 2009, pp. 1288–1294.

[214] J. Marschak and D. Davidson. *Experimental tests of stochastic decision theory*. Tech. rep. Cowles Foundation for Research in Economics, Yale University, 1957.

[215] P. Massart. *Concentration inequalities and model selection*. Vol. 6. Springer.

[216] R. Mazumder, A. Choudhury, G. Iyengar, and B. Sen. “A computational framework for multivariate convex regression and its variants”. In: *Journal of the American Statistical Association* 114.525 (2019), pp. 318–331.

[217] P. McCullagh and J. A. Nelder. *Generalized linear models*. Monographs on Statistics and Applied Probability. Second edition [of MR0727836]. Chapman & Hall, London, 1989, pp. xix+511.

[218] D. McFadden. "Conditional logit analysis of qualitative choice behavior". In: *Frontiers in Econometrics* (1973), pp. 105–142.

[219] D. McFadden. "Econometric models of probabilistic choice". In: *Structural analysis of discrete data with econometric applications* (1981), pp. 198–272.

[220] D. H. McLaughlin and R. D. Luce. "Stochastic transitivity and cancellation of preferences between bitter-sweet solutions". In: *Psychonomic Science* 2.1-12 (1965), pp. 89–90.

[221] A. M. Medina and M. Mohri. "Learning theory and algorithms for revenue optimization in second price auctions with reserve". In: *Proceedings of the 31st International Conference on Machine Learning (ICML-14)*. 2014, pp. 262–270.

[222] S. Mendelson. "Learning without concentration". In: *Conference on Learning Theory*. 2014, pp. 25–39.

[223] S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann. "Reconstruction and subgaussian operators in asymptotic geometric analysis". In: *Geometric and Functional Analysis* 17.4 (2007), pp. 1248–1282.

[224] L. Mirsky. "A dual of Dilworth's decomposition theorem". In: *The American Mathematical Monthly* 78.8 (1971), pp. 876–877.

[225] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[226] A. Montanari. "Computational implications of reducing data to sufficient statistics". In: *Electronic Journal of Statistics* 9.2 (2015), pp. 2370–2390.

[227] J. Morgenstern and T. Roughgarden. "Learning simple auctions". In: *Conference on Learning Theory*. 2016, pp. 1298–1318.

[228] A. Narayanan and V. Shmatikov. "Robust de-anonymization of large sparse datasets". In: *Security and Privacy, 2008. SP 2008. IEEE Symposium on*. IEEE. 2008, pp. 111–125.

[229] S. Negahban, S. Oh, and D. Shah. "Rank centrality: Ranking from pairwise comparisons". In: *Operations Research* 65.1 (2016), pp. 266–287.

[230] S. Negahban, S. Oh, K. K. Thekumparampil, and J. Xu. "Learning from comparisons and choices". In: *The Journal of Machine Learning Research* 19.1 (2018), pp. 1478–1572.

[231] S. Negahban and M. J. Wainwright. "Restricted strong convexity and (weighted) matrix completion: Optimal bounds with noise". In: *Journal of Machine Learning Research* 13 (May 2012), pp. 1665–1697.

[232] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. "Robust stochastic approximation approach to stochastic programming". In: *SIAM Jour. Opt.* 19.4 (2009), pp. 1574–1609.

[233] A. S. Nemirovski, B. T. Polyak, and A. B. Tsybakov. "Convergence rate of nonparametric estimates of maximum-likelihood type". In: *Problemy Peredachi Informatsii* 21.4 (1985), pp. 17–33.

[234] A. S. Nemirovsky and D. B. Yudin. *Problem Complexity and Method Efficiency in Optimization*. New York: John Wiley and Sons, 1983.

[235] J. Neter, M. H. Kutner, C. J. Nachtsheim, and W. Wasserman. *Applied linear statistical models*. Vol. 4. Irwin Chicago, 1996.

[236] P. Netrapalli, P. Jain, and S. Sanghavi. "Phase retrieval using alternating minimization". In: *Advances in Neural Information Processing Systems*. 2013, pp. 2796–2804.

[237] M. Neykov, Z. Wang, and H. Liu. "Agnostic estimation for misspecified phase retrieval models". In: *Advances in Neural Information Processing Systems*. 2016, pp. 4089–4097.

[238] J. Neyman. "Optimal asymptotic tests of composite hypotheses". In: *Probability and statistics* (1959), pp. 213–234.

[239] A. Nguyen, C. Piech, J. Huang, and L. Guibas. "Codewebs: scalable homework search for massive open online programming courses". In: *Proceedings of the 23rd international conference on World wide web*. 2014, pp. 491–502.

[240] D. Oleson, A. Sorokin, G. Laughlin, V. Hester, J. Le, and L. Biewald. "Programmatic gold: Targeted and scalable quality assurance in crowdsourcing". In: *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*. 2011.

[241] S. Oyama, Y. Baba, Y. Sakurai, and H. Kashima. "Accurate integration of crowdsourced labels using workers' self-reported confidence scores". In: *Twenty-Third International Joint Conference on Artificial Intelligence*. 2013.

[242] A. Pananjady and T. A. Courtade. "The effect of local decodability constraints on variable-length compression". In: *IEEE Transactions on Information* 64.4 (2018), pp. 2593–2608.

[243] A. Pananjady and D. P. Foster. "Single-index models in the high signal regime". In: *Preprint* (2020).

[244] A. Pananjady, C. Mao, V. Muthukumar, M. J. Wainwright, and T. A. Courtade. "Worst-case versus average-case design for estimation from partial pairwise comparisons". In: *The Annals of Statistics* 48.2 (2020), pp. 1072–1097.

[245] A. Pananjady and R. J. Samworth. "Isotonic regression with unknown permutations: Statistics, computation, and adaptation". In: *Preprint* (2020).

[246] A. Pananjady and M. J. Wainwright. "Value function estimation in Markov reward processes: Instance-dependent ℓ_∞ -bounds for policy evaluation". In: *arXiv preprint arXiv:1909.08749* (2019).

[247] A. Pananjady, M. J. Wainwright, and T. A. Courtade. "Denoising linear models with permuted data". In: *2017 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2017, pp. 446–450.

[248] A. Pananjady, M. J. Wainwright, and T. A. Courtade. “Linear regression with shuffled data: Statistical and computational limits of permutation recovery”. In: *IEEE Transactions on Information Theory* 64.5 (2017), pp. 3286–3300.

[249] D. Park, J. Neeman, J. Zhang, S. Sanghavi, and I. Dhillon. “Preference completion: Large-scale collaborative ranking from pairwise comparisons”. In: *International Conference on Machine Learning*. 2015, pp. 1907–1916.

[250] J. Pazis, R. E. Parr, and J. P. How. “Improving PAC exploration using the median of means”. In: *Advances in Neural Information Processing Systems*. 2016, pp. 3898–3906.

[251] L. Pitsoulis and P. M. Pardalos. “Quadratic assignment problem”. In: *Encyclopedia of Optimization*. Ed. by C. A. Floudas and P. M. Pardalos. Boston, MA: Springer US, 2001, pp. 2075–2107.

[252] R. L. Plackett. “The analysis of permutations”. In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 24.2 (1975), pp. 193–202.

[253] Y. Plan and R. Vershynin. “The generalized lasso with non-linear observations”. In: *IEEE Transactions on information theory* 62.3 (2016), pp. 1528–1537.

[254] Y. Plan, R. Vershynin, and E. Yudovina. “High-dimensional estimation with geometric constraints”. In: *Information and Inference: A Journal of the IMA* 6.1 (2017), pp. 1–40.

[255] B. T. Polyak and A. B. Juditsky. “Acceleration of stochastic approximation by averaging”. In: *SIAM J. Control and Optimization* 30.4 (1992), pp. 838–855.

[256] A. B. Poore and S. Gadaleta. “Some assignment problems arising from multiple target tracking”. In: *Mathematical and Computer Modelling* 43.9 (2006), pp. 1074–1091.

[257] J. L. Prince and A. S. Willsky. “Reconstructing convex sets from support line measurements”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12.4 (1990), pp. 377–389.

[258] M. L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. Wiley, 2005.

[259] A. Rajkumar and S. Agarwal. “A statistical convergence perspective of algorithms for rank aggregation from pairwise data”. In: *International Conference on Machine Learning*. 2014, pp. 118–126.

[260] A. Rajkumar and S. Agarwal. “When can we rank well from comparisons of O $(n\log(n))$ non-actively chosen pairs?” In: *Conf. on Learning Theory*. 2016, pp. 1376–1401.

[261] A. Rakhlin, K. Sridharan, and A. B. Tsybakov. “Empirical entropy, minimax regret and minimax risk”. In: *Bernoulli* 23.2 (2017), pp. 789–824.

[262] P. Rigollet and J. Weed. “Uncoupled isotonic regression via minimum Wasserstein deconvolution”. In: *Information and Inference: A Journal of the IMA* 8.4 (2019), pp. 691–717.

[263] H. Robbins and S. Monro. “A stochastic approximation method”. In: *The Annals of Mathematical Statistics* (1951), pp. 400–407.

[264] T. Robertson, F. T. Wright, and R. L. Dykstra. *Order restricted statistical inference*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Ltd., Chichester, 1988.

[265] T. Robertson and F. Wright. “Consistency in generalized isotonic regression”. In: *The Annals of Statistics* (1975), pp. 350–362.

[266] P. M. Robinson. “Root-N-consistent semiparametric regression”. In: *Econometrica: Journal of the Econometric Society* (1988), pp. 931–954.

[267] W. S. Robinson. “A method for chronologically ordering archaeological deposits”. In: *American Antiquity* (1951), pp. 293–301.

[268] C. Rose, I. S. Mian, and R. Song. “Timing Channels with Multiple Identical Quanta”. In: *arXiv preprint arXiv:1208.1070* (2012).

[269] M. Rudelson and R. Vershynin. “Hanson-Wright inequality and sub-Gaussian concentration”. In: *Electronic Communications in Probability* 18 (2013).

[270] D. Ruppert. *Efficient estimators from a slowly convergent Robbins-Monro process*. Tech. rep. 781. Cornell University, 1988.

[271] H. Sedghi, M. Janzamin, and A. Anandkumar. “Provable tensor methods for learning mixtures of generalized linear models”. In: *Artificial Intelligence and Statistics*. 2016, pp. 1223–1231.

[272] E. Seijo and B. Sen. “Nonparametric least squares estimation of a multivariate convex regression function”. In: *The Annals of Statistics* 39.3 (2011), pp. 1633–1657.

[273] N. B. Shah, S. Balakrishnan, J. Bradley, A. Parekh, K. Ramchandran, and M. J. Wainwright. “Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence”. In: *Journal of Machine Learning Research* 17 (2016), Paper No. 58, 47.

[274] N. B. Shah, S. Balakrishnan, A. Guntuboyina, and M. J. Wainwright. “Stochastically transitive models for pairwise comparisons: Statistical and computational issues”. In: *IEEE Trans. Inform. Theory* 63.2 (2017), pp. 934–959.

[275] N. B. Shah, S. Balakrishnan, and M. J. Wainwright. “A permutation-based model for crowd labeling: Optimal estimation and robustness”. In: *arXiv preprint arXiv:1606.09632* (2016).

[276] N. B. Shah, S. Balakrishnan, and M. J. Wainwright. “Low Permutation-rank Matrices: Structural Properties and Noisy Completion”. In: *Journal of Machine Learning Research* 20.101 (2019), pp. 1–43.

[277] N. B. Shah, S. Balakrishnan, and M. J. Wainwright. “Feeling the Bern: Adaptive estimators for Bernoulli probabilities of pairwise comparisons”. In: *IEEE Transactions on Information Theory* 65.8 (2019), pp. 4854–4874.

[278] C. E. Shannon. “Probability of error for optimal codes in a Gaussian channel”. In: *Bell System Technical Journal* 38.3 (1959), pp. 611–656.

[279] Y. Shuo Tan and R. Vershynin. “Phase Retrieval via Randomized Kaczmarz: Theoretical Guarantees”. In: *arXiv e-prints*, arXiv:1706.09993 (June 2017), arXiv:1706.09993. arXiv: `1706.09993 [math.NA]`.

[280] A. Sidford, M. Wang, X. Wu, and Y. Ye. “Variance Reduced Value Iteration and Faster Algorithms for Solving Markov Decision Processes”. In: *Proceedings of the Symposium on Discrete Algorithms (SODA)*. 2018.

[281] A. Sidford, M. Wang, X. Wu, L. Yang, and Y. Ye. “Near-optimal time and sample complexities for solving Markov decision processes with a generative model”. In: *Advances in Neural Information Processing Systems*. 2018, pp. 5186–5196.

[282] D. Silver, R. S. Sutton, and M. Müller. “Reinforcement Learning of Local Shape in the Game of Go.” In: *Proceedings of the International Joint Conferences on Artificial Intelligence*. Vol. 7. 2007, pp. 1053–1058.

[283] M. Simchowitz and K. G. Jamieson. “Non-asymptotic gap-dependent regret bounds for tabular MDPs”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 1151–1160.

[284] M. Simchowitz, H. Mania, S. Tu, M. I. Jordan, and B. Recht. “Learning without mixing: Towards a sharp analysis of linear system identification”. In: *arXiv preprint arXiv:1802.08334* (2018).

[285] M. Slawski and E. Ben-David. “Linear regression with sparsely permuted data”. In: *Electronic Journal of Statistics* 13.1 (2019), pp. 1–36.

[286] Y. S. Soh. “Fitting Convex Sets to Data: Algorithms and Applications”. PhD thesis. California Institute of Technology, 2019.

[287] Y. S. Soh and V. Chandrasekaran. “Fitting Tractable Convex Sets to Support Function Evaluations”. In: *arXiv preprint arXiv:1903.04194* (2019).

[288] D. Song and D. Shah. “Learning mixture model with missing values and its application to rankings”. In: *arXiv preprint arXiv:1812.11917* (2018).

[289] R. Srikant and L. Ying. “Finite-Time Error Bounds For Linear Stochastic Approximation and TD Learning”. In: *Proceedings of the Thirty-Second Conference on Learning Theory*. Ed. by A. Beygelzimer and D. Hsu. Vol. 99. Proceedings of Machine Learning Research. Phoenix, USA: PMLR, June 2019, pp. 2803–2830.

[290] C. Stein. “Efficient nonparametric testing and estimation”. In: *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*. 1956, pp. 187–195.

[291] Q. F. Stout. “Isotonic regression for multiple independent variables”. In: *Algorithmica* 71.2 (2015), pp. 450–470.

[292] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. 2nd. Cambridge, MA: MIT Press, 2018.

[293] R. S. Sutton. “Learning to predict by the methods of temporal differences”. In: *Machine learning* 3.1 (1988), pp. 9–44.

[294] C. Szepesvári. *Algorithms for reinforcement learning*. Morgan-Claypool, 2009.

[295] V. B. Tadic. “On the almost sure rate of convergence of linear stochastic approximation algorithms”. In: *IEEE Transactions on Information Theory* 50.2 (2004), pp. 401–409.

[296] C. Thrampoulidis, E. Abbasi, and B. Hassibi. “Lasso with non-linear measurements is equivalent to one with linear measurements”. In: *Advances in Neural Information Processing Systems*. 2015, pp. 3420–3428.

[297] C. Thrampoulidis and A. S. Rawat. “The PhaseLift for Non-quadratic Gaussian Measurements”. In: *arXiv preprint arXiv:1712.03638* (2017).

[298] S. Thrun and J. J. Leonard. “Simultaneous localization and mapping”. In: *Springer Handbook of Robotics*. Springer Berlin Heidelberg, 2008, pp. 871–889.

[299] L. L. Thurstone. “A law of comparative judgment.” In: *Psychological review* 34.4 (1927), p. 273.

[300] P. Tseng. “An analysis of the EM algorithm and entropy-like proximal point methods”. In: *Mathematics of Operations Research* 29.1 (2004), pp. 27–44.

[301] J. N. Tsitsiklis and B. Van Roy. “Analysis of temporal-diffference learning with function approximation”. In: *Advances in neural information processing systems*. 1997, pp. 1075–1081.

[302] A. B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer, 2009.

[303] Y. Z. Tsypkin and B. T. Polyak. “Attainable accuracy of adaptation algorithms”. In: *Doklady Akademii Nauk*. Vol. 218. Russian Academy of Sciences. 1974, pp. 532–535.

[304] M. B. A. Turlach. “Package ‘Sleuth3’”. In: (2019).

[305] A. Tversky. “Elimination by aspects: A theory of choice.” In: *Psychological review* 79.4 (1972), p. 281.

[306] T. Ueno, M. Kawanabe, T. Mori, S.-i. Maeda, and S. Ishii. “A semiparametric statistical approach to model-free policy evaluation”. In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 1072–1079.

[307] J. Unnikrishnan, S. Haghighatshoar, and M. Vetterli. “Unlabeled sensing with random linear measurements”. In: *IEEE Transactions on Information Theory* 64.5 (2018), pp. 3237–3253.

[308] S. van de Geer. “Estimating a regression function”. In: *The Annals of Statistics* (1990), pp. 907–924.

[309] S. A. van de Geer. *Applications of empirical process theory*. Vol. 91. 2000.

[310] S. A. van de Geer. *Regression analysis and empirical processes*. Vol. 45. CWI Tract. Stichting Mathematisch Centrum, Centrum voor Wiskunde en Informatica, Amsterdam, 1988, pp. vi+161.

[311] S. van de Geer and M. J. Wainwright. “On concentration for (regularized) empirical risk minimization”. In: *Sankhya A* 79.2 (2017), pp. 159–200.

[312] A. W. van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge University press, 2000.

[313] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. With applications to statistics. Springer-Verlag, New York, 1996, pp. xvi+508.

[314] W. H. van Schuur. “Mokken scale analysis: Between the Guttman scale and parametric item response theory”. In: *Political Analysis* 11.2 (2003), pp. 139–163.

[315] V. N. Vapnik and A. Y. Chervonenkis. “The uniform convergence of frequencies of the appearance of events to their probabilities”. In: *Doklady Akademii Nauk*. Vol. 181. 4. Russian Academy of Sciences. 1968, pp. 781–783.

[316] R. R. Varshamov. “Estimate of the number of signals in error correcting codes”. In: *Dokl. Akad. Nauk SSSR*. Vol. 117. 1957, pp. 739–741.

[317] S. S. Vempala. “Learning convex concepts from Gaussian distributions with PCA”. In: *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. IEEE. 2010, pp. 124–130.

[318] R. Vershynin. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge University Press, 2018.

[319] R. Vershynin. “Introduction to the non-asymptotic analysis of random matrices”. In: *arXiv preprint arXiv:1011.3027* (2010).

[320] A. Vince. “A rearrangement inequality and the permutahedron”. In: *Amer. Math. Monthly* 97.4 (1990), pp. 319–323.

[321] H.-T. Wai, M. Hong, Z. Yang, Z. Wang, and K. Tang. “Variance Reduced Policy Evaluation with Smooth Function Approximation”. In: *Advances in Neural Information Processing Systems*. 2019, pp. 5776–5787.

[322] M. J. Wainwright. “Information-Theoretic Limits on Sparsity Recovery in the High-Dimensional and Noisy Setting”. In: *IEEE Transactions on Information Theory* 55.12 (Dec. 2009), pp. 5728–5741.

[323] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge University Press, 2019.

[324] M. J. Wainwright. “Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for Q-learning”. In: *arXiv preprint arXiv:1905.06265* (2019).

[325] M. J. Wainwright. “Variance-reduced Q-learning is minimax optimal”. In: *arXiv preprint arXiv:1906.04697* (2019).

[326] A. Wald. “Contributions to the theory of statistical estimation and testing hypotheses”. In: *The Annals of Mathematical Statistics* 10.4 (1939), pp. 299–326.

[327] I. Waldspurger. “Phase Retrieval With Random Gaussian Sensing Vectors by Alternating Projections”. In: *IEEE Transactions on Information Theory* 64.5 (May 2018), pp. 3301–3312.

[328] T. Wang, Q. Berthet, and R. J. Samworth. “Statistical and computational trade-offs in estimation of sparse principal components”. In: *The Annals of Statistics* 44.5 (2016), pp. 1896–1930.

[329] L. Wasserman. *All of nonparametric statistics*. Springer Science & Business Media, 2006.

[330] T. Weissman, E. Ordentlich, G. Seroussi, S. Verdu, and M. J. Weinberger. “Inequalities for the L1 deviation of the empirical distribution”. In: *Hewlett-Packard Labs, Tech. Rep* (2003).

[331] J. A. Wellner. “The Bennett-Orlicz norm”. In: *Sankhya A* 79.2 (2017), pp. 355–383.

[332] J. Whitehill, T.-f. Wu, J. Bergsma, J. R. Movellan, and P. L. Ruvolo. “Whose vote should count more: Optimal integration of labels from labelers of unknown expertise”. In: *Advances in neural information processing systems*. 2009, pp. 2035–2043.

[333] C. J. Wu. “On the convergence properties of the EM algorithm”. In: *The Annals of Statistics* 11.1 (1983), pp. 95–103.

[334] C. C. Xinyang Yi and S. Sanghavi. “Solving a Mixture of Many Random Linear Equations by Tensor Decomposition and Alternating Minimization”. In: (2016).

[335] J. Xu, D. J. Hsu, and A. Maleki. “Global analysis of expectation maximization for mixtures of two Gaussians”. In: *Advances in Neural Information Processing Systems*. 2016, pp. 2676–2684.

[336] L. Xu and M. I. Jordan. “On convergence properties of the EM algorithm for Gaussian mixtures”. In: *Neural computation* 8.1 (1996), pp. 129–151.

[337] T. Xu, Z. Wang, Y. Zhou, and Y. Liang. “Reanalysis of Variance Reduced Temporal Difference Learning”. In: *arXiv preprint arXiv:2001.01898* (2020).

[338] F. Yang and R. F. Barber. “Contraction and uniform convergence of isotonic regression”. In: *Electronic Journal of Statistics* 13.1 (2019), pp. 646–677.

[339] Z. Yang, K. Balasubramanian, and H. Liu. “High-dimensional non-Gaussian single index models via thresholded score function estimation”. In: *International Conference on Machine Learning*. 2017, pp. 3851–3860.

[340] Z. Yang, Z. Wang, H. Liu, Y. Eldar, and T. Zhang. “Sparse nonlinear regression: Parameter estimation under nonconvexity”. In: *International Conference on Machine Learning*. 2016, pp. 2472–2481.

[341] Z. Yang, L. F. Yang, E. X. Fang, T. Zhao, Z. Wang, and M. Neykov. “Misspecified Nonconvex Statistical Optimization for Phase Retrieval”. In: *arXiv preprint arXiv:1712.06245* (2017).

[342] X. Yi, Z. Wang, C. Caramanis, and H. Liu. “Optimal linear estimation under unknown nonlinear transform”. In: *Advances in neural information processing systems*. 2015, pp. 1549–1557.

[343] K. Yoshihara. “Simple proofs for the strong converse theorems in some channels”. In: *Kodai Mathematical Seminar Reports*. Vol. 16. Dept. of Mathematics, Tokyo Institute of Technology. 1964, pp. 213–222.

[344] B. Yu. “Rates of convergence for empirical processes of stationary mixing sequences”. In: *The Annals of Probability* (1994), pp. 94–116.

[345] A. Zanette and E. Brunskill. “Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds”. In: *arXiv preprint arXiv:1901.00210* (2019).

[346] A. Zhang and D. Xia. “Tensor SVD: Statistical and computational limits”. In: *IEEE Transactions on Information Theory* 64.11 (2018), pp. 7311–7338.

[347] C.-H. Zhang. “Risk bounds in isotonic regression”. In: *The Annals of Statistics* 30.2 (2002), pp. 528–555.

[348] T. Zhang. “Phase Retrieval by Alternating Minimization With Random Initialization”. In: *IEEE Transactions on Information Theory* 66.7 (2020), pp. 4563–4573.

[349] T. Zhang. “Phase retrieval using alternating minimization in a batch setting”. In: *Applied and Computational Harmonic Analysis* (2019).

[350] Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan. “Spectral methods meet EM: A provably optimal algorithm for crowdsourcing”. In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 3537–3580.

[351] K. Zhong, P. Jain, and I. S. Dhillon. “Mixed linear regression with multiple components”. In: *Advances in neural information processing systems*. 2016, pp. 2190–2198.

[352] D. Zhou, J. Huang, and B. Schölkopf. “Learning with hypergraphs: Clustering, classification, and embedding”. In: *Advances in neural information processing systems*. 2007, pp. 1601–1608.

[353] J. Zhou, A. Bhattacharya, A. H. Herring, and D. B. Dunson. “Bayesian factorizations of big sparse tensors”. In: *Journal of the American Statistical Association* 110.512 (2015), pp. 1562–1576.

[354] Y. Zhu, S. Chatterjee, J. Duchi, and J. Lafferty. “Local minimax complexity of stochastic convex optimization”. In: *Advances in Neural Information Processing Systems*. 2016, pp. 3431–3439.