Improved Algorithms and Upper Bounds in Differential Privacy

Arun Ganesh

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2022-53
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2022/EECS-2022-53.html

May 10, 2022

Improved Algorithms and Upper Bounds in Differential Privacy

by

Arun Ganesh

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Satish Rao, Chair
Professor Jelani Nelson
Associate Professor Nikhil Srivastava

Spring 2022

Improved Algorithms and Upper Bounds in Differential Privacy

Abstract

Improved Algorithms and Upper Bounds in Differential Privacy

by

Arun Ganesh

Doctor of Philosophy in Computer Science

University of California, Berkeley

Professor Satish Rao, Chair

Differential privacy is the now de facto industry standard for ensuring privacy while publicly releasing statistics about a sensitive database. At a high level, differentially private algorithms add noise to the statistics they compute, so an adversary cannot with high confidence guess if any individual is in the database as any individual's effect on the statistics can be replicated by the noise.

The fundamental paradigm in differential privacy is the *privacy-accuracy trade-off*: A differentially private algorithm's output can be made more accurate by reducing the amount of noise added, but in doing so the privacy guarantee decays. Current state-of-the-art algorithms often require practitioners to choose between a large drop in accuracy compared to non-private algorithms, or very weak privacy guarantees. Improving the trade-off between privacy, accuracy, would ideally allow practitioners to get the "best of both worlds." Sometimes, *efficiency* is also traded off with privacy and accuracy. That is, despite differential privacy being an information-theoretic guarantee, an inefficient (and thus impractical to implement) algorithm may still obtain a better privacy-accuracy trade-off than the best known efficient algorithm. Designing efficient algorithms that match the privacy-accuracy trade-off of known inefficient algorithms thus is also of practical importance.

In this thesis, we consider counting queries, private log-strongly concave sampling, and private convex optimization, all fundamental problems in sampling and optimization, and give algorithms for each problem improving the privacy-accuracy trade-off or efficiency when compared to the previous state of the art algorithms. For counting queries, we show that adding noise from a "Generalized Gaussian" gives better worst-case accuracy compared to Gaussian noise. For private log-strongly concave sampling, we show that the discrete Langevin dynamics allows one to efficiently approximately sample from a target distribution while preserving privacy, a commonly needed primitive in private optimization. For private convex optimization, we show that in some settings (including e.g. private linear regression), if

we are given a sufficient amount of public data, we can obtain a better dependence on the dimensionality of the problem than differentially private gradient descent.

To Amma and Appa, who crossed an ocean for their children.

Contents

List of Figures

List of Tables

Acknowledgments

The writing of this thesis, much less my entire PhD program, would not have been nearly as enjoyable or fruitful had it not been for the guidance of my advisor, Satish Rao. Satish went above and beyond to make sure I could focus on working on problems that interested me with minimal distractions, facilitated opportunities to do so, and was always a supportive and friendly force to help combat the feeling of inadequacy that is sometimes inevitable during grad school. Apart from just helping me be successful in publishing papers, Satish also taught me to have a more holistic approach to being a researcher and teacher. I will be happy if in the future I am able to demonstrate even a small fraction of the wisdom and kindness Satish showed as an advisor.

In addition to Satish, I am thankful to Prasad Raghavendra, Jelani Nelson, and Nikhil Srivastava for kindly agreeing to be my committee members. I am especially thankful to Jelani, who generously funded me for a semester, allowing me to continue to focus on my research.

Pursuing a PhD is one of the best decisions I've made in my life, and one I likely would not have made had it not been for Debmalya Panigrahi, my undergraduate research mentor. Debmalya reached out to me while I was still unsure if research was something I wanted to pursue, and gave my undergraduate self far more time and energy than I could have asked for in helping me pursue my first results and get over the hurdles of being a novice researcher, for which I am eternally thankful.

My passion for research in differential privacy was fostered largely through two internships at Google, during which I was lucky to have Kunal Talwar, Abhradeep Guha Thakurta, and Shuang Song as my hosts. I am thankful to them for helping me quickly learn about new aspects of the field, and their patience and support as I was doing so.

I am thankful to all my co-authors, for humoring my often half-baked ideas on how to make progress on research problems, and for their optimism even when our projects seemed to have hit dead ends. In addition to Debmalya, Kunal, Abhradeep, and Shuang, this includes Ehsan Amid, Yossi Azar, Rong Ge, Tomasz Kociumaka, Andrea Lincoln, Bruce Maggs, Rajiv Mathews, Swaroop Ramaswamy, Barna Saha, Thomas Steinke, Vinith Suriyakumar, Aaron Sy, Om Thakkar, Jalaj Upadhyay, Qiuyi Zhang, and Jiazheng Zhao.

A PhD program can easily become a lonely-feeling experience, so I am thankful to the many fellow grad students who both helped me create fond memories and who supported me through the less fond ones. I am especially thankful to my former fellow first-years Efe Aras, Tarun Kathuria, Rachel Lawrence, Chinmay Nirkhe, Nick Spooner, Elizabeth Yang, and Morris Yau, who made what could have been a very scary transition to grad school a fun one instead. I am also thankful to the friends I made through Smash At Berkeley for giving me an escape from the PhD life when needed.

Lastly but most importantly, I am thankful to my parents Ganesan and Geetha, who made endless sacrifices so that I would have the opportunities that I did, and to my brother Ashwin and sister-in-law Sarah, who helped me become the nerd I am today.

Chapter 1

Introduction and Preliminaries

1.1 Introduction

Most questions in data analysis can be formulated as instantiations of the following question: Given a database D, output some statistic $f(D)$ about the database. For example, in an election, D may be a database of votes, and $f(D)$ might be the number of votes for a given individual. This framework also captures much more complicated problems. For example, D may be a database of labelled examples for a classification problem, and $f(D)$ might be the neural network achieving the highest accuracy or lowest loss on D.

Unfortunately, even if f is a very complicated function of D, publicly releasing $f(D)$ may leak sensitive information contained in the dataset D. A recent infamous example of [21] demonstrates an attack that can extract sensitive information such as phone numbers from GPT-2, one of OpenAI's generative text models. Another classic example is the attack of [68] on the Netflix Prize dataset, which uses public IMDB ratings to uniquely identify individuals' appearances (and thus their private Netflix ratings) in the Netflix Prize dataset, *despite the fact that this dataset was anonymized.*

In light of these attacks, differential privacy, originally proposed by [33], has become the industry standard for performing data analysis while preserving privacy of individuals whose data appears in the database D. Let D be a database, and let D' be an "adjacent" database, i.e. the same database but with one individual's data changed arbitrarily (say, we flip every bit in some representation of that individual's data). Informally, differential privacy adds noise to the statistics computed, such that an adversary trying to learn sensitive information about this individual in D cannot be sure if we used D or D' to compute our statistics, since the noise could replicate the effects on the statistic of changing from D to D' or vice-versa. In particular, the simplest notion of differential privacy, ϵ-differential privacy, roughly says that someone with no prior information on whether we used D or D' cannot guess which we used with probability higher than $\frac{e^\epsilon}{e^\epsilon+1}$. As we add more noise, ϵ gets closer to 0, and we approach "perfect privacy."

Note that as the noise we add increases, the noise dominates the signal of the true statistic and in turn the privacy guarantee strengthens but the accuracy of our noisy statistic decreases. This is known as the *privacy-accuracy trade-off.* This trade-off is a central paradigm in differential privacy, and many algorithmic results in the differential privacy literature can be viewed as attempting to optimize this trade-off for a given problem, i.e. give an algorithm with the best possible accuracy that achieves a specific level of privacy or vice-versa. This is an important practical problem, as practitioners may be hesitant to utilize differential privacy if it requires a large sacrifice in accuracy. For example, for the CIFAR-10 benchmark for image classification, there is an abundance of non-private models that achieve 99% accuracy (see e.g. [28]). However, the best models trained using ϵ-differential privacy for reasonable values of ϵ obtain accuracy closer to 70% (see e.g. [72]). For this reason, practitioners often use "unreasonable" values of ϵ in training machine learning models. For example [80] states that Apple performs data analysis with $\epsilon = 16t$, where t is the number of days since data collection on a user began. For this setting of ϵ, even for $t = 1$ we have $\frac{e^\epsilon}{e^\epsilon+1} \approx .99999988$, i.e. the privacy guarantee is somewhat vacuous. By finding ways to improve the privacy-accuracy tradeoff for fundamental problems, we can incentivize practitioners to use more reasonable values of ϵ in practice and give more meaningful privacy guarantees to their user base.

Of course, efficiency of these algorithms is also an important concern; there are some settings where the algorithm with the best privacy-accuracy trade-off is much less efficient than another algorithm with a worse privacy-accuracy trade-off. For example, a fundamental tool in differential privacy is the exponential mechanism of [62]. The exponential mechanism says that if the loss out of outputting the statistic x is $\ell(x)$, we output x with probability/density proportional to $e^{-c\ell(x)}$, where c is a constant depending on ϵ and the "sensitivity" of the loss function to the database. The exponential mechanism is used in a wide variety of theoretical results in the differential privacy literature, due to its general nature and ease of analysis. However, in the worst case, implementing the exponential mechanism requires computing $\ell(x)$ for *every possible statistic we could output*, or if the statistics are real-valued, computing a complex integral over all statistics that could be output. In turn, many of the results using the exponential mechanism as a building block are impractical unless they show the exponential mechanism can be implemented efficiently. So for practical purposes, it is not just important to improve privacy-accuracy trade-offs, but also pursue efficient algorithms obtaining the best possible privacy-accuracy trade-off, or close to it.

In this thesis, we improve the privacy-accuracy trade-off and/or efficiency of algorithms for consider several fundamental problems in the differential privacy literature.

1.2 Differential Privacy and Basic Mechanisms

To formally define differential privacy, we will view a database $D \in \mathcal{D}^*$ as a (multi-)set of data points in $\mathcal{D}$, one per individual, and we will say that two databases D, D' are *adjacent* if they differ by at most one individual's data point (that is, $|D \setminus D'|, |D' \setminus D| \leq 1$). We will denote adjacency by $D \sim D'$. We let $\mathcal{M} : \mathcal{D}^* \to \mathcal{X}$ be a randomized algorithm that takes a dataset D and outputs a point in the space of statistics $\mathcal{X}$. We will use $\mathcal{M}(D)$ to denote the random variable of $\mathcal{M}$'s output when given database D as input.

The simplest notion of differential privacy is pure differential privacy.

Definition 1 (Pure Differential Privacy [33])**.** *A randomized algorithm $\mathcal{M}$ is ϵ-differentially private if for any two adjacent databases D, D', and any (measurable) subset $S \subseteq \mathcal{X}$, of the range of $\mathcal{M}$, we have:*

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(D') \in S].$$

Arguably the simplest example of a pure differentially private algorithm is the Laplace Mechanism:

Example 2 (Laplace Mechanism [33])**.** *Suppose we have a vector function $\phi : \mathcal{D} \to \mathbb{R}^k$, and we want to compute $\sum_{d \in D} \phi(d)$, which we will denote $\phi(D)$ for brevity. Suppose we additionally have the guarantee that*

$$\max_{D \sim D'} \left|\left|\phi(D) - \phi(D')\right|\right|_1 \leq \Delta_1.$$

The multivariate Laplace distribution (with mean zero) on $\mathbb{R}^k$ with scale λ, denoted Lap(λ), has density function $p(\mathbf{x}) \propto \exp(-\left|\left|\mathbf{x}\right|\right|_1 / \lambda)$. The Laplace mechanism for privately computing $\phi(D)$ samples $\mathbf{x} \sim$ Lap(Δ_1/ϵ), and outputs $\phi(D) + \mathbf{x}$.

The following is now well-known:

Fact 3 ([33])**.** *The Laplace Mechanism is ϵ-differentially private. Furthermore, if the output of the Laplace Mechanism is $\tilde{\phi}$, we have $\mathbb{E}\left[\left|\left|\tilde{\phi} - \phi(D)\right|\right|_1 / k\right] = \Delta_1/\epsilon$, i.e. the average amount of noise added to each coordinate of $\phi(D)$, is Δ_1/ϵ.*

The Laplace mechanism cleanly illustrates the *privacy-accuracy trade-off*, a core paradigm in differential privacy. As ϵ increases, our privacy guarantee worsens, but our accuracy guarantee improves, and the opposite is true as ϵ decreases. The privacy-accuracy trade-off is well-demonstrated by the settings of $\epsilon = 0, \infty$. When $\epsilon = 0$, we have a perfect privacy guarantee, i.e. an adversary can learn nothing about our database. However, as ϵ approaches 0, the Laplace mechanism approaches a "uniform distribution over the reals," i.e. the statistic we output is independent of the database, and thus not useful. When $\epsilon = \infty$, we have no privacy guarantee, i.e. we might as well just publish the database. However, the statistic we publish has perfect accuracy. In turn, the goal in designing differentially private algorithms is to optimize this trade-off. In other words, for a fixed ϵ we want to provide an algorithm with as good as possible an accuracy guarantee, or equivalently for a fixed accuracy guarantee we want to provide that guarantee using the smallest ϵ possible. Improving this trade-off is an important practical question, as often in practice we need to set ϵ to be quite large to see the desired level of accuracy.

For many problems, ϵ-differential privacy is a strong constraint, and slightly relaxing this constraint allows us to obtain much better results in terms of accuracy. A common relaxation that we will use in all results in this thesis is approximate differential privacy:

Definition 4 (Approximate Differential Privacy [33]). *A randomized algorithm $\mathcal{M}$ is (ϵ, δ)-differentially private if for any two adjacent databases D, D', and any (measurable) subset $S \subseteq \mathcal{X}$, of the range of $\mathcal{M}$, we have:*

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(D') \in S] + \delta.$$

When δ is sufficiently small, $(\epsilon, 0)$- and (ϵ, δ)-differential privacy are nearly identical in terms of their privacy guarantee. However, even with small δ theoretical upper bounds under approximate differential privacy often improve substantially over upper bounds under pure differential privacy.

The simplest approximate differentially private mechanism is the Gaussian mechanism.

Example 5 (Gaussian Mechanism [33]). *Consider the problem of privately computing $\phi(D)$ as defined in Example 2, but instead we have the guarantee that*

$$\max_{D \sim D'} ||\phi(D) - \phi(D')||_2 \leq \Delta_2.$$

The Gaussian mechanism outputs $\tilde{\phi}$ sampled from $N(\phi(D), \sigma^2 \cdot \mathbb{I})$ for $\sigma = \frac{\Delta_2 \sqrt{2 \ln(1.25/\delta)}}{\epsilon}$.

Fact 6 ([33]). *The Gaussian mechanism is (ϵ, δ)-differentially private and*

$$\mathbb{E}\left[\frac{\left|\left|\tilde{\phi} - \phi(D)\right|\right|_2}{\sqrt{k}}\right] = O(\Delta_2 \sqrt{\ln(1/\delta)}/\epsilon).$$

The Laplace and Gaussian mechanisms can both be viewed as instances of the much more general exponential mechanism.

Example 7 (Exponential Mechanism [62]). *Suppose we want to privately output the element in $\mathcal{X}$ that minimizes some loss function $\ell(x; D)$ depending on the database D. Suppose that for all $x \in \mathcal{X}$:*

$$\max_{D \sim D'} |\ell(x; D) - \ell(x; D')| \leq \Delta.$$

The exponential mechanism outputs $\tilde{x}$ from $\mathcal{X}$ with probability (or density) proportional to $\exp(-\frac{2\epsilon\ell(x;D)}{\Delta})$.

Theorem 8 ([62]). *The exponential mechanism is ϵ-differentially private. Furthermore, let $\mathcal{G}$ be any subset of $\mathcal{X}$. Then if $\mathcal{X}$ is finite and x is the (random) output of the exponential mechanism:*

$$\mathbb{E}\left[\ell(x;D) - \max_{x\in\mathcal{G}} \ell(x;D)\right] = O\left(\frac{\Delta \log\left(\frac{|\mathcal{X}|}{|\mathcal{G}|}\right)}{\epsilon}\right).$$

If instead $\mathcal{X}, \mathcal{G}$ are subsets of $\mathbb{R}^k$, then the above claim holds with $\log(|\mathcal{X}|/|\mathcal{G}|)$ replaced with $\log(\mathit{Vol}(\mathcal{X})/\mathit{Vol}(\mathcal{G}))$.

Many mechanisms in differential privacy can be viewed as instantiations of the exponential mechanism, and/or use the exponential mechanism as a black box. For example, in Example 2, we can set $\ell(x; D) = ||x - \phi(D)||_1$, retrieving the Laplace mechanism (up to a factor of 2).

1.3 Properties of Differential Privacy

Differential privacy has several properties that are convenient for designing and analyzing differentially private algorithms. A widely used property is the post-processing property; it follows from the data processing inequality in information theory.

Theorem 9 (Post-Processing). *Let $\mathcal{M}(D)$ be an (ϵ, δ)-differentially private algorithm outputting a (random) element in $\mathcal{X}$. Let f be an arbitrary randomized function from $\mathcal{X}$ to $\mathcal{X}'$, that is independent of the database D. Let $\mathcal{M}'$ be the mechanism that takes $x = \mathcal{M}(D)$ and outputs $f(x)$. Then $\mathcal{M}'$ is (ϵ, δ)-differentially private.*

For a proof reference, see [32, Proposition 2.1]. As an example, if we are using the Laplace mechanism to compute the number of users who satisfy a certain predicate, the only possible answers to this question are non-negative integers, whereas the Laplace mechanism can output any real. The post-processing property says that if after running the Laplace mechanism, we round all negative numbers to 0, and round all positive reals to the nearest integer, we do not violate differential privacy. Intuitively, an adversary trying to learn our dataset could have performed the rounding on their own since it does not require access to the dataset, so by performing this step for them, they gain no information about the dataset.

Another important property is composition, which allows us to analyze the privacy of a chain of mechanisms, each of which is private.

Theorem 10 (Adaptive Composition [33]). *Let $\mathcal{M}_1, \mathcal{M}_2$ be two mechanisms, where $\mathcal{M}_1$ takes elements of $\mathcal{D}$ as input and outputs elements in $\mathcal{X}_1$, and $\mathcal{M}_2$ takes elements of $\mathcal{X}_1 \times \mathcal{D}$ and outputs elements in $\mathcal{X}_2$. Suppose $\mathcal{M}_1$ is (ϵ_1, δ_1)-differentially private, and $\mathcal{M}_2$ is (ϵ_2, δ_2)-differentially private. Let $\mathcal{M}$ be the mechanism that samples x_1 from $\mathcal{M}_1(D)$, then samples x_2 from $\mathcal{M}_2(x_1, D)$, and outputs (x_1, x_2). Then $\mathcal{M}$ is $(\epsilon_1 + \epsilon_2, \delta_1 + \delta_2)$-differentially private.*

The above theorem can be composed multiple times to analyze a chain of mechanisms of arbitrary length rather than length 2. As an example, we can view Example 2 when every coordinate has the same sensitivity to the database Δ, we have $\Delta_1 = k\Delta$. Then, we can view the Laplace mechanism as the composition of k mechanisms each outputting a single real $\tilde{\phi}_i$ with ϵ/k-differential privacy, rather than a single mechanism outputting a k-dimensional vector $\tilde{\phi}$. Theorem 10 shows that k ϵ/k-differentially private mechanisms are collectively ϵ-differentially private, giving an alternate proof for the privacy of the Laplace mechanism. Of course, here the mechanisms are non-adaptive, i.e. can be run in parallel, whereas Theorem 10 allows for adaptive/sequential mechanisms.

1.4 Differentially Private Stochastic Convex Optimization

Definition 11 (Stochastic Optimization)**.** *Stochastic optimization is the following problem: Given a set of models $\mathcal{C} \subseteq \mathbb{R}^k$ and a data universe $\mathcal{D}$, we have a loss function $\ell : \mathcal{C} \times \mathcal{D} \to \mathbb{R}$. There is an unknown distribution τ over the data universe $\mathcal{D}$, and we are given n samples D, each i.i.d sampled from τ. Using this samples, our goal is to find a point θ (approximately) minimizing the population loss $\mathcal{L}^*(\theta) := \mathbb{E}_{d\sim\tau}\left[\ell(\theta; d)\right]$.*

Here, the stochasticity lies in the distribution τ over the samples we receive. Let $\theta^* = \arg\min_{\theta\in\mathcal{C}} \mathcal{L}^*(\theta)$ be the true minimizer of $\mathcal{L}^*$. θ^* is often referred to as the population minimizer. Since we only have access to samples from τ, we cannot hope to exactly compute θ^*, so the strategy is to instead (approximately) compute the empirical minimizer, i.e. the minimizer θ_{emp} of the empirical loss $\ell(\theta; D) := \frac{1}{|D|}\sum_{d\in D} \ell(\theta; d)$, and then argue that the generalization error $\mathcal{L}^*(\theta_{emp}) - \mathcal{L}^*(\theta^*)$ is not too large.

Stochastic optimization as a general framework captures the problem of training deep learning models, as the weights of a deep learning model can be viewed as a high-dimensional vector, and deep learning models are trained by minimizing the average per-example loss on a training set. As the aforementioned attack of [21] demonstrates, in practice, many deep learning models are trained on data sets containing sensitive data, and thus publishing these models may violate the privacy of users. In turn, we can instead solve differentially private stochastic optimization.

Definition 12 (Differentially Private Stochastic Optimization)**.** *Differentially private stochastic optimization is the same as stochastic optimization as defined in Definition 11, but instead we output a random θ, and our algorithm's distribution over θ must satisfy differential privacy with respect to the dataset D.*

Most algorithms for stochastic optimization only access D via a gradient oracle for the empirical loss, i.e. an oracle that computes $\frac{1}{|D|}\nabla \sum_{d\in D} \ell(\theta; d)$ for a given θ[1]. In turn, in order for a stochastic optimization algorithm to be differentially private, by the post-processing

[1] In theory and practice, for various reasons we may not actually want our oracle to compute the "full gradient" on D, but instead on a (random) subset of examples in D. For brevity, we will only consider algorithms that only compute full gradients in this thesis.

property it suffices if it uses a differentially private gradient oracle. Since differential privacy requires some sort of bounded sensitivity, we will usually assume $\ell(\theta; d)$ is L-Lipschitz with respect to the ℓ_2-norm, i.e. $||\nabla \ell(\theta; d)||_2 \leq L$ for all θ, d[2]. Given this assumption, we can simply use the Gaussian mechanism to compute a single gradient with differential privacy, giving the desired private gradient oracle.

The simplest and most ubiquitous example of a differentially private stochastic optimization algorithm is differentially private gradient descent (DP-GD). DP-GD is nearly the same as gradient descent, which repeatedly performs the update $\theta_{t+1} = \Pi_{\mathcal{C}}(\theta_t - \eta_t \mathbf{g}_t)$, where η_t is a learning rate function, $\mathbf{g}_t$ is the gradient at θ_t, and $\Pi_{\mathcal{C}}$ is the Euclidean projection into $\mathcal{C}$. The only difference is that, as per the previous discussion of the Gaussian mechanism, DP-GD uses the noisy gradient $\mathbf{g}_t + \mathbf{b}_t$ in place of $\mathbf{g}_t$, where $\mathbf{b}_t$ is sampled from a Gaussian distribution. By post-processing, the privacy guarantee of each iteration of DP-GD is the same as the privacy guarantee of the Gaussian mechanism used to compute the noisy gradient. We could use Theorem 10 to get a privacy guarantee for the entirety of DP-GD from the privacy guarantee for a single iteration. However, it turns out this gives a sub-optimal analysis. In the next section, we discuss Rènyi-divergences, which can be used to give a better privacy analysis for DP-GD. Note that convexity is not necessarily for the privacy guarantee, so DP-GD is still applicable for non-convex model training problems. When the loss function is convex as well as L-Lipschitz, it's known that DP-GD with appropriate parameter settings has empirical loss within $O(\frac{L||\mathcal{C}||\sqrt{k}}{\epsilon n})$ of the optimal solution for the empirical loss minimization problem. The generalization error can be shown to be at most $O(L\,||\mathcal{C}||\,/\sqrt{n})$, giving an overall population loss within $O(L\,||\mathcal{C}||\,(\frac{\sqrt{k}}{\epsilon n} + \frac{1}{\sqrt{n}}))$ of θ^*.

1.5 Rènyi-Divergences

Rènyi-divergences, defined by [74], are a generalization of other divergences such as the KL-divergence and max-divergence.

Definition 13 (Rènyi-Divergence [74]). *Let P and Q be two distributions. For simplicity, we will assume $supp(P) = supp(Q) \subseteq \mathbb{R}^k$. The α-Rènyi-divergence between P and Q for $\alpha > 0, \alpha \neq 1$ is defined as:*

$$R_\alpha(P||Q) = \frac{1}{\alpha - 1} \int_{supp(Q)} \frac{P(x)^\alpha}{Q(x)^{\alpha-1}} dx = \frac{1}{\alpha - 1} \mathbb{E}_{x \sim P} \left[\frac{P(x)^{\alpha-1}}{Q(x)^{\alpha-1}} \right] = \frac{1}{\alpha - 1} \mathbb{E}_{x \sim Q} \left[\frac{P(x)^\alpha}{Q(x)^\alpha} \right]$$

For $\alpha = 1, \infty$, the α-Rènyi-divergence is defined by taking the limit of $R_\alpha(P||Q)$ as α approaches $1, \infty$ respectively, and are equal to the KL-divergence and max-divergence respectively.

[2] In practice, one can enforce this assumption by "clipping" the gradients, i.e. rescaling any per-sample gradients that have ℓ_2-norm larger than L so they have ℓ_2-norm exactly L; since this is more of a practical rather than theoretical concern, we will largely ignore it here.

Rènyi-divergences can be used to give privacy guarantees due to the following theorem:

Theorem 14 (Proposition 3 of [66]). *Let $\mathcal{M}$ be a mechanism taking inputs in $\mathcal{D}$ and outputting random elements of $\mathcal{X}$ such that for some α, ϵ:*

$$\max_{D \sim D'} R_\alpha(\mathcal{M}(D)||\mathcal{M}(D')) \leq \epsilon.$$

Then for any $0 < \delta \leq 1$, $\mathcal{M}$ is $(\epsilon + \frac{\ln(1/\delta)}{\alpha-1}, \delta)$-differentially private.

In other words, to show (ϵ, δ)-differential privacy of a mechanism $\mathcal{M}$ it suffices to bound the α-Rènyi-divergence of $\mathcal{M}(D)$ and $\mathcal{M}(D')$ by $\epsilon/2$ for $\alpha = 1 + \frac{2\ln(1/\delta)}{\epsilon}$. Rènyi-divergences in combination with the above theorem provide another way to analyze the Gaussian mechanism due to the following well-known fact:

Fact 15 (Example 3 of [36]).

$$R_\alpha(N(\mathbf{0}, \sigma^2 \mathbb{I}_k)||N(\mathbf{x}, \sigma^2 \mathbb{I}_k)) \leq \frac{\alpha \left|\left|\mathbf{x}\right|\right|_2^2}{2\sigma^2}.$$

Furthermore, Rènyi-divergences satisfy an adaptive composition theorem similar to Theorem 10, which e.g., [1] used to provide a tighter analysis for the privacy of DP-GD.

Theorem 16 (Proposition 1 of [66]). *Let $\mathcal{M}_1, \mathcal{M}_2$ be two mechanisms, where $\mathcal{M}_1$ takes elements of $\mathcal{D}$ as input and outputs elements in $\mathcal{X}_1$, and $\mathcal{M}_2$ takes elements of $\mathcal{X}_1 \times \mathcal{D}$ and outputs elements in $\mathcal{X}_2$. Suppose $\mathcal{M}_1$ satisfies $R_\alpha(\mathcal{M}_1(D)||\mathcal{M}_1(D')) \leq \epsilon_1$ for all $D \sim D'$, and $\mathcal{M}_2$ satisfies $R_\alpha(\mathcal{M}_2(x, D)||\mathcal{M}_2(x, D')) \leq \epsilon_2$ for all $x \in \mathcal{X}_1, D \sim D'$. Let $\mathcal{M}$ be the mechanism that samples x_1 from $\mathcal{M}_1(D)$, then samples x_2 from $\mathcal{M}_2(x_1, D)$, and outputs (x_1, x_2). Then $\mathcal{M}$ satisfies $R_\alpha(\mathcal{M}(D)||\mathcal{M}(D')) \leq \epsilon_1 + \epsilon_2$ for all $D \sim D'$.*

In particular, since in DP-GD the full gradients on D and D' differ by at most L/n in ℓ_2-norm by the Lipschitz assumption, if we sample $\mathbf{g}_t$ from $N(\sigma^2 \mathbb{I}_k)$ in DP-GD and run DP-GD for T iterations, we get that DP-GD satisfies

$$R_\alpha(\mathcal{M}(D)||\mathcal{M}(D')) \leq \frac{T\alpha L^2}{2\sigma^2 n^2}.$$

Then, by setting $\sigma^2 = O\left(\frac{TL^2 \ln(1/\delta)}{\epsilon^2 n^2}\right)$, for $\alpha = 1 + \frac{2\ln(1/\delta)}{\epsilon}$ we get that the α-Rènyi-divergence is at most $\epsilon/2$ as desired. This saves logarithmic factors over analyses based on other standard composition theorems. Rènyi-divergences satisfy a number of other properties that will be useful for the results in Chapter 3; we defer discussing those properties to that chapter.

1.6 Our Results and Organization

Chapter 2 [Generalized Gaussians and The Counting Queries Problem]: We consider the counting queries problem, one of the simplest yet most fundamental problems in differential privacy. In this problem, we wish to answer k questions about the database D, each of the form "how many users in D satisfy a predicate ϕ_i?" Since each user can affect the answers to these questions by at most 1, this is equivalent to outputting a vector $\phi(D)$ under differential privacy with the guarantee that $||\phi(D) - \phi(D')||_\infty \leq 1$ for any adjacent D, D'. Our goal is to minimize the maximum absolute error over the answers to the questions. For approximate differential privacy, the simplest mechanism for this problem is the Gaussian mechanism. The average error of the Gaussian mechanism is proportional to $\sqrt{k}$, but the maximum error is proportional to $\sqrt{k \log k}$, since the largest coordinate of a k-dimensional Gaussian is $\sqrt{\log k}$ times larger than the average coordinate. We show that using Generalized Gaussians, whose density is proportional to $e^{-||x||_p^p}$ rather than $e^{-||x||_2^2}$, we can improve the dependence on k to $\sqrt{k \log \log k}$. Furthermore, by composing with the sparse vector mechanism, which roughly speaking trims the largest coordinates of a vector, we can improve the dependence to $\sqrt{k \log \log \log k}$. This chapter is based on results previously appearing in [41].

Chapter 3 [Efficient Private Log-Strongly Concave Sampling]: We next consider finding settings in which one can efficiently implement the exponential mechanism. In particular, we consider when the loss function ℓ has support over $\mathbb{R}^k$ and is strongly convex and smooth, i.e. we want to sample from the distribution with density p proportional to $e^{-\ell(x)}$. While exactly sampling from this distribution may not be efficient, we can approximately sample by using a finite-time solution for a stochastic differential equation called the Langevin dynamics whose stationary distribution is p. It is still non-obvious how to efficiently exactly solve the stochastic differential equation even in finite-time, but we can consider a discretization whose solution can be found efficiently given a gradient oracle for ℓ. However, the notion of approximate sampling needs to be chosen carefully; if, e.g. we have a sampler for a distribution that is close to p in a metric such as total variation distance or K-L divergence, using the approximate sampler in place of an exact sampler for our exponential mechanism may violate privacy. We show that using a number of gradient oracle calls near-linear in k, the discrete finite-time solution converges to within a small Rènyi-divergence of p, which implies an efficient sampler for log-strongly concave densities that preserves privacy when used in place of an exact sampler. This chapter is based on results previously appearing in [40].

Chapter 4 [Public Data-Augmented Stochastic Optimization]: In the last chapter, we consider how we can use a small amount of public data, i.e. data whose privacy we do not need to preserve, to improve private learning, using private stochastic optimization as a proxy for the learning problem. The go-to algorithm for private optimization is DP-GD. We propose an alternative algorithm based on mirror descent. Mirror descent takes a function Ψ that is strictly convex (such that $\nabla\Psi$ is a bijective map), and uses the update equation $\nabla\Psi(\mathbf{x}_{t+1}) = \nabla\Psi(\mathbf{x}_t) - \eta_t \mathbf{g}_t$. At a high level, mirror descent adapts gradient descent to the geometry of Ψ by reshaping the gradient to be smaller in directions where Ψ has large convexity, and larger in directions where Ψ has small convexity. Our algorithm, PDA-DPMD, chooses Ψ to be the loss function on the public data. We show that given enough public data, PDA-DPMD has excess loss with a better dependence on the dimension than DP-GD. As an example, for private linear regression, the excess loss of DP-GD depends on the *minimum* eigenvalue of the Hessian of the loss function, whereas we show that with enough public data PDA-DPMD's excess loss depends on the *average* eigenvalue of the Hessian of the loss function, i.e. PDA-DPMD outperforms DP-GD in non-isotropic settings. This chapter is based on results previous appearing in [4].

Chapter 2

Generalized Gaussians and The Counting Queries Problem

2.1 Introduction and Problem Definition

The counting queries problem is a special case of the problem of outputting $\phi(D)$ as defined in Example 2. In the counting queries problem, the vector function ϕ can only take on values in $[0,1]^k$. It is called the counting queries problem as in the special case where ϕ can only take on values in $\{0,1\}^k$ we can view each coordinate of $\phi(d)$ as being equal to 1 if d satisfies some predicate, and 0 if d does not satisfy this predicate. In turn, the entries of $\phi(D)$ are answers to counting queries, i.e. queries of the form "how many individuals in D satisfy the predicate ϕ_i?" The counting queries problem is a fundamental problem in data analysis. For example, the counting queries problem captures the powerful statistical query model of [54].

The goal of the counting queries problem is to find a mechanism $\mathcal{M}(D)$ that outputs $\tilde{\phi}$, a noisy version $\phi(D)$ while minimizing some function of the error $\tilde{\phi} - \phi(D)$. A well-studied notion of error is the expected average error, i.e. $\mathbb{E}_{\tilde{\phi} \sim \mathcal{M}(D)}\left[\left|\left|\tilde{\phi} - \phi(D)\right|\right|_1 / k\right]$. For ϵ-differential privacy, the Laplace mechanism of [33] achieves expected average error $O(k/\epsilon)$, and there is a matching lower bound due to [46]. For (ϵ, δ)-differential privacy, the Gaussian mechanism of [33] achieves expected average error $O(\sqrt{k \log(1/\delta)}/\epsilon)$, and again there is a matching lower bound due to [78].

A more challenging notion of error is the expected maximum error, i.e.

$$\mathbb{E}_{\tilde{\phi} \sim \mathcal{M}(D)}\left[\left|\left|\tilde{\phi} - \phi(D)\right|\right|_\infty\right].$$

This notion of error is more challenging in the sense that for any vector $\mathbf{x}$, we have $||\mathbf{x}||_1 / k \leq ||\mathbf{x}||_\infty$, since the average absolute coordinate is clearly smaller than the largest absolute coordinate. So of course, the expected maximum error of any mechanism is at least its expected average error. For this notion of error, the Laplace and Gaussian mechanisms are not optimal. For example, the expected maximum error of the Laplace mechanism is $O(k \log k/\epsilon)$, as the largest coordinate of a random vector sampled from the Laplace distribution has absolute value that is in expectation $\log k$ times larger than the average coordinate. On the other hand, if one uses the exponential mechanism with loss function $\ell(\tilde{\phi}) = \left|\left|\tilde{\phi} - \phi(D)\right|\right|_\infty$, the expected maximum error is $O(k/\epsilon)$, matching the best possible expected average error [78].

For approximate differential privacy, on the other hand, until recently it was not known whether one could get expected maximum error of $O(\sqrt{k \log(1/\delta)}/\epsilon)$, also matching the best possible expected average error. The Gaussian mechanism here achieves expected maximum error $O(\sqrt{k \log k \log(1/\delta)}/\epsilon)$, as a multivariate Gaussian's largest coordinate has absolute value $\sqrt{\log k}$ times larger than its average coordinate. Prior to the results in this section, the best known expected maximum error was $O(\sqrt{k \log \log k \log(1/\delta)}/\epsilon)$, due to [78]. The mechanism achieving this result starts with the Gaussian mechanism, and then uses the sparse vector mechanism to correct coordinates with large error.

2.2 Our Results and Technical Overview

Our first result is as follows:

Theorem 17. *For all $1 \leq p \leq \log k$, $\epsilon \leq O(1)$, $\delta \in [2^{-O(k/p)}, 1/k]$, there exists a (ϵ, δ)-differentially private mechanism $\mathcal{M}$ that takes in a database D and counting queries ϕ and outputs a random $\tilde{\phi} \in \mathbb{R}^k$ such that for some sufficiently large constant c, and all $t \geq 0$:*

$$\Pr_{\tilde{\phi} \sim \mathcal{M}(D)} \left[\left|\left|\tilde{\phi} - \phi(D)\right|\right|_\infty \geq \frac{ct\sqrt{kp}\log^{1/p} k \sqrt{\log(1/\delta)}}{\epsilon} \right] \leq e^{-t^p \log k}$$

In particular, this implies:

$$\mathbb{E}_{\tilde{\phi} \sim \mathcal{M}(D)} \left[\left|\left|\tilde{\phi} - \phi(D)\right|\right|_\infty \right] = O\left(\frac{\sqrt{kp}\log^{1/p} k \sqrt{\log(1/\delta)}}{\epsilon} \right).$$

We also have for all $1 \leq q \leq p$:

$$\mathbb{E}_{\tilde{\phi} \sim \mathcal{M}(D)} \left[\frac{\left|\left|\tilde{\phi} - \phi(D)\right|\right|_q}{k^{1/q}} \right] = O\left(\frac{\sqrt{kp \log(1/\delta)}}{\epsilon} \right).$$

We note that the lower bound on δ in Theorem 17 can easily be removed: if δ is smaller than $2^{-O(k/p)}$, we can instead use the exponential mechanism achieving average maximum error $O(k/\epsilon)$, which matches the error guarantees of Theorem 17 in this range of δ. The mechanism is simply to add noise from a Generalized Gaussian with shape p and an appropriate scale parameter σ, i.e. with density proportional to $\exp(-(\left|\left|\mathbf{x}\right|\right|_p/\sigma)^p)$. In our analysis, we arrive at the bounds $c \leq 2094$ and $\sigma \leq 262 \cdot \frac{\sqrt{kp\log(1/\delta)}}{\epsilon}$, although we did not attempt to optimize the constants in favor of a simpler analysis and presentation. We believe the multiplicative constants in both bounds can be substantially improved with a more careful analysis.

Setting $p = \Theta(\log\log k)$, this result matches the asymptotic error bound of [78]. However, this result improves on [78] qualitatively. Although the mechanism of [78] is already not too complex, the Generalized Gaussian mechanism we use is even simpler, just adding noise from a well-known distribution. Notably, Generalized Gaussian mechanisms retain the property of the Gaussian mechanism that the noise added to each entry of $\phi(D)$ is independent (unlike the mechanism of [78], which uses dependent noise), and that the noise has a known closed-form distribution that is easy to sample from[1]. To the best of our knowledge, this is the first analysis giving privacy guarantees for Generalized Gaussian mechanisms besides that in [57]. Even then, [57] does not give any closed-form bounds on the value of σ needed for privacy. This analysis may be of independent interest for other applications where one would normally use the Gaussian mechanism, but may want to use a Generalized Gaussian mechanism with $p > 2$ to trade average-case error guarantees for better worst-case error guarantees.

We give a summary of our analysis here. We first need to determine what value of σ causes the Generalized Gaussian mechanism to be private. Viewing the Generalized Gaussian mechanism as an instance of the exponential mechanism of [62], this reduces to deriving a tail bound on $||\mathbf{x}+1||_p^p - ||\mathbf{x}||_p^p$ for $\mathbf{x}$ sampled from the noise distribution. If p is even this is roughly equal to $p\sum_{j=1}^{k} x_j^{p-1}$. By a Chernoff bound on the signs of each random variable in the sum, this is roughly tail bounded by the sum of $\sqrt{k\log(1/\delta)}$ of the x_j^{p-1} random variables. These variables are distributed according to a *Generalized Gamma* distribution, which we prove is sub-gamma. This gives us the desired tail bound, and thus an upper bound on the σ needed to ensure (ϵ,δ)-differential privacy. To prove the error guarantees, we derive tail bounds on the ℓ_p-norm of $\mathbf{x}$ sampled from Generalized Gaussian distributions, as well as on the coordinates of points sampled from unit-radius ℓ_p-spheres, the latter of which is done by upper bounding the volume of "sphere caps" of these spheres.

Building on this result, we improve the previous best-known ℓ_∞ error for answering counting queries with (ϵ,δ)-differential privacy:

Theorem 18. *For all $\epsilon \leq O(1)$, $\delta \in [2^{-O(k/\log\log\log k)}, 1/k]$, $t \in [0, O(\frac{\log k}{\log\log k})]$, there exists a (ϵ,δ)-differentially private mechanism $\mathcal{M}$ that takes in a database D and counting queries ϕ and outputs a random $\tilde{\phi} \in \mathbb{R}^k$ such that for a sufficiently large constant c:*

$$\Pr_{\tilde{\phi}\sim\mathcal{M}(D)}\left[\left|\left|\tilde{\phi}-\phi(D)\right|\right|_\infty \geq \frac{ct\sqrt{k\log\log\log k\log(1/\delta)}}{\epsilon}\right] \leq e^{-\log^t k}.$$

In particular, if we choose e.g. $t = 2$ we get:

[1] see e.g. https://sccn.ucsd.edu/wiki/Generalized_Gaussian_Probability_Density_Function.

$$\mathbb{E}_{\tilde{\phi}\sim\mathcal{M}(D)}\left[\left|\left|\tilde{\phi}-\phi(D)\right|\right|_{\infty}\right]=O\left(\frac{\sqrt{k\log\log\log k\log(1/\delta)}}{\epsilon}\right).$$

Again, the lower bound on δ can easily be removed. We arrive at this result by improving upon Generalized Gaussian mechanisms in the same manner [78] improves upon the Gaussian mechanism: After sampling $\mathbf{x}$ from a Generalized Gaussian, we apply the sparse vector mechanism to $\mathbf{x}$ to get $\tilde{\mathbf{x}}$ which satisfies $||\mathbf{x}-\tilde{\mathbf{x}}||_{\infty} \ll ||\mathbf{x}||_{\infty}$. We then just output $\tilde{\phi} = \phi(D)+\mathbf{x}-\tilde{\mathbf{x}}$. Similarly to [78], the major technical component is showing that at least $k/\log^{\Omega(1)} k$ entries of x are small with high probability, which we do by deriving tail bounds on ℓ_p-spheres. This is necessary for the sparse vector mechanism to satisfy that $||\mathbf{x}-\tilde{\mathbf{x}}||_{\infty}$ is, roughly speaking, the $(k/\log^{\Omega(1)} k)$-th largest entry of $\mathbf{x}$ rather than the largest entry with high probability.

2.3 Other Related Work

Following our work, [25] and [42] independently gave mechanisms with optimal expected ℓ_∞-error $O(\sqrt{k\log(1/\delta)}/\epsilon)$, quantitatively improving our results. Since in practice $\sqrt{\log\log k}$ is unlikely to be much larger than the constants hidden by the asymptotic notation (e.g., using the natural log, $\sqrt{\log\log k} = 2$ for $k \approx 5\cdot 10^{23}$), the qualitative differences between our results and these two results make our results still of interest to e.g. practitioners; we highlight those differences here.

The result of [25] remarkably uses a bounded noise distribution, and in turn the maximum error is unconditionally bounded, rather than the average maximum error being bounded, in contrast with Generalized Gaussian mechanisms whose maximum error can be arbitrarily large. However, a bounded noise distribution cannot e.g. satisfy group differential privacy for all group sizes simultaneously, whereas Generalized Gaussian mechanisms can. Also, while both results simply add noise, Generalized Gaussians are more well-studied than the noise distribution of [25] and can be sampled by simplying powering and rescaling samples from Gamma random variables, which should make them easier to implement in practice.

The result of [42] at a high level adds noise and then repeatedly applies the sparse vector mechanism to correct entries with large noise, in contrast to just adding noise. In addition, their result uses arguably even simpler sampling primitives than ours (it only needs to sample Laplace distributions and permutations of lists), but their overall mechanism needs a somewhat more involved iterative approach rather than a one-shot sample. Finally, as presented the resulting noise distribution from their overall mechanism does not have e.g. a closed-form or independent entries which may be desirable.

2.4 Preliminaries

Our main idea is to add noise drawn from a Generalized Gaussian distribution, defined as follows:

Definition 19 (Generalized Gaussian)**.** *The (multivariate) Generalized Gaussian distribution with shape p and scale σ, denoted $GGauss(p, \sigma)$, is the distribution over $\mathbf{x} \in \mathbb{R}^k$ with density function proportional to $\exp(-(||\mathbf{x}||_p / \sigma)^p)$.*

Note that when $p = 1$, this is just the Laplace distribution, and when $p = 2$, this is just the Gaussian distribution.

Sub-Gamma Random Variables

The following facts about sub-gamma random variables will be useful in our analysis:

Definition 20 (Sub-Gamma Random Variable)**.** *A random variable X is sub-gamma to the right with variance v and scale c if:*

$$\forall \lambda \in (0, 1/c) : \mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] \leq \exp\left(\frac{\lambda^2 v}{2(1 - c\lambda)}\right).$$

Here, we use the convention $1/c = \infty$ if $c = 0$. We denote the class of such random variables $\Gamma^+(v, c)$. Similarly, a random variable X is sub-gamma to the left with variance v and scale c, if $-X \in \Gamma^+(v, c)$, i.e.:

$$\forall \lambda \in (0, 1/c) : \mathbb{E}[\exp(\lambda(\mathbb{E}[X] - X))] \leq \exp\left(\frac{\lambda^2 v}{2(1 - c\lambda)}\right).$$

We denote the class of such random variables $\Gamma^-(v, c)$.

We refer the reader to [20] for a textbook reference for this definition and proofs of the following facts.

Fact 21. *If for $i \in [n]$ we have a random variable $X_i \in \Gamma^+(v_i, c_i)$, then $X = \sum_{i \in [n]} X_i$ satisfies $X \in \Gamma^+(\sum_{i \in [n]} v_i, \max_{i \in [n]} c_i)$ (and the same relation holds for $\Gamma^-(v, c)$).*

Lemma 22. *If $X \in \Gamma^+(v, c)$ then for all $t > 0$:*

$$\Pr[X > \mathbb{E}[X] + \sqrt{2vt} + ct] \leq e^{-t}.$$

Similarly, if $X \in \Gamma^-(v, c)$ then for all $t > 0$:

$$\Pr[X < \mathbb{E}[X] - \sqrt{2vt} - ct] \leq e^{-t}.$$

Fact 23. *Let $X \sim Gamma(a)$, i.e. X has pdf satisfying:*

$$p(x) \propto x^{a-1} e^{-x}.$$

Then X satisfies $X \in \Gamma^+(a, 1)$ and $X \in \Gamma^-(a, 0)$.

Other Probability Facts

We will use the following standard fact to relate distributions of variables to the distributions of their powers:

Fact 24 (Change of Variables for Powers)**.** *Let X be distributed over $(0, \infty)$ with pdf proportional to $f(x)$. Let Y be the random variable X^c for $c > 0$. Then Y has pdf proportional to $y^{\frac{1}{c}-1} f(y^{\frac{1}{c}})$.*

Finally, we'll use the following standard tail bounds:

Lemma 25 (Laplace Tail Bound)**.** *Let X be a Laplace random variable with scale b, $Lap(b)$. That is, X has pdf proportional to $\exp(-|x|/b)$. Then we have $\Pr[|x| \geq tb] \leq e^{-t}$.*

Lemma 26 (Chernoff Bound)**.** *Let $X_1, X_2, \ldots X_k$ be independent Bernoulli random variables. Let $\mu = \mathbb{E}\left[\sum_{i \in [k]} X_i\right]$. Then for $t \in [0, 1]$, we have:*

$$\Pr\left[\sum_{i \in [k]} X_i \geq (1+t)\mu\right] \leq \exp\left(-\frac{t^2 \mu}{3}\right).$$

2.5 The Generalized Gaussian Mechanism

In this section, we analyze the Generalized Gaussian mechanism that given database D, samples $\mathbf{x} \sim GGauss(p, \sigma)$ and outputs $\tilde{\phi} = \phi(D) + \mathbf{x}$. We denote this mechanism $\mathcal{M}_\sigma^p$. When $p = 1$ this is the Laplace mechanism, and when $p = 2$ this is the Gaussian mechanism.

Privacy Guarantees

We first determine what σ is needed to make this mechanism private. We start with the following lemma, which gives a tail bound on the change in the "utility" function $\left|\left|\tilde{\phi} - \phi(D)\right|\right|_p^p$ if $\phi(D)$ changes by $\Delta \in [-1, 1]^k$:

Lemma 27. *Let $\mathbf{x} \in \mathbb{R}^k$ be sampled from $GGauss(p, \sigma)$. Then for $4 \leq p \leq \log k$ that is an even integer, $\delta \in [2^{-O(k/p)}, 1/k]$, and any $\Delta \in [-1, 1]^k$ we have with probability $1 - \delta$:*

$$||\mathbf{x} - \Delta||_p^p - ||\mathbf{x}||_p^p \leq 32pk^{1/p-1/2}\sqrt{p \log(1/\delta)}\, ||\mathbf{x}||_p^{p-1} + 2k^{\frac{p}{2}}p^2$$

We remark that the requirement that p be an even integer can be dropped by generalizing the proofs in this section appropriately. However, we can reduce proving Theorem 17 for all p to proving it for only even p by rounding p up to the nearest even integer (at the loss of a multiplicative constant of at most $\sqrt{2}$), and only considering even p simplifies the presentation. So, we stick to considering only even p.

Proof. By symmetry of $GGauss(p, \sigma)$ we can assume Δ has all negative entries. Then we have:

$$||\mathbf{x} - \Delta||_p^p - ||\mathbf{x}||_p^p = \sum_{i=1}^{k} ((x_i - \Delta_i)^p - x_i^p)$$

$$= \sum_{i=1}^{k} \int_{x_i}^{x_i - \Delta} py^{p-1} \mathrm{d}y \leq \sum_{i=1}^{k} \int_{x_i}^{x_i - \Delta} p(x_i - \Delta_i)^{p-1} \mathrm{d}y \leq p \sum_{i=1}^{k} (x_i - \Delta_i)^{p-1} \leq p \sum_{i=1}^{k} (x_i + 1)^{p-1}.$$

We want to replace the terms $(x_i + 1)^{p-1}$ with terms x_i^{p-1} since the latter's distribution is more easily analyzed. To do so, we use the following observation:

Fact 28. *If $p \leq \sqrt{k}/2$:*

- *If $x_i > \sqrt{k}$, then we have $(x_i + 1)^{p-1} \leq \left(1 + \frac{1}{\sqrt{k}}\right)^{p-1} x_j^{p-1} \leq \left(1 + \frac{2p}{\sqrt{k}}\right) x_j^{p-1}$.*
- *If $|x_i| \leq \sqrt{k}$, then we have $(x_i + 1)^{p-1} - x_i^{p-1} \leq (\sqrt{k} + 1)^{p-1} - \sqrt{k}^{p-1} \leq 2k^{\frac{p}{2}-1} p$.*
- *If $x_i < -\sqrt{k}$, then we have $(x_i + 1)^{p-1} \leq \left(1 - \frac{1}{\sqrt{k}}\right)^{p-1} x_j^{p-1} \leq \left(1 - \frac{2p}{\sqrt{k}}\right) x_j^{p-1}$.*

Fact 28 gives:

$$\sum_{i=1}^{k} (x_i + 1)^{p-1} \leq \left(1 - \frac{2p}{\sqrt{k}}\right) \sum_{i: x_i < 0} x_i^{p-1} + \left(1 + \frac{2p}{\sqrt{k}}\right) \sum_{i: x_i \geq 0} x_i^{p-1} + 2k^{\frac{p}{2}} p.$$

It now suffices to show that:

$$-\left(1 - \frac{2p}{\sqrt{k}}\right) \sum_{i: x_i < 0} |x_i|^{p-1} + \left(1 + \frac{2p}{\sqrt{k}}\right) \sum_{i: x_i \geq 0} |x_i|^{p-1} \leq 32k^{1/p - 1/2} \sqrt{p \log(1/\delta)} \, ||\mathbf{x}||_p^{p-1}, \quad (2.1)$$

with probability at least $1 - \delta$. Note that each x_i is sampled independently with probability proportional to $\exp(-(|x_i|/\sigma)^p)$. Since multiplying x by a constant rescales both sides of (2.1) by the same multiplicative factor, it suffices to show (2.1) when each x_i is independently sampled with probability proportional to $\exp(-|x_i|^p)$, i.e. when $\sigma = 1$. By change of variables, $y_i = |x_i|^{p-1}$ is sampled from the distribution with pdf proportional to $y_i^{\frac{1}{p-1}-1} \exp(-y_i^{\frac{p}{p-1}})$. This is the Generalized Gamma random variable with parameters $(\frac{1}{p-1}, \frac{p}{p-1})$, which we denote $GGamma(\frac{1}{p-1}, \frac{p}{p-1})$. We show the following property of this random variable in Section 2.7:

Lemma 29. *For any $p \geq 4$, let Y be the random variable $GGamma(\frac{1}{p-1}, \frac{p}{p-1})$, let $\mu = \mathbb{E}[Y]$. Then $\mu \in [1/p, 1.2/p), Y \in \Gamma^+(\mu, 1)$, and $Y \in \Gamma^-(\mu, 3/2)$.*

Let k' be the number of positive coordinates in x. A Chernoff bound gives that $k' \leq \frac{k}{2} + 3\sqrt{k \log \frac{1}{\delta}}$ with probability $1 - \delta/3$. By Lemma 29 and Fact 21 $\sum_{i:x_i<0} |x_i|^{p-1}$ is in $\Gamma^-((k-k')\mu, 3/2)$ and $\sum_{i:x_i \geq 0} |x_i|^{p-1}$ is in $\Gamma^+(k'\mu, 1)$ for μ as defined in Lemma 29. We now apply Lemma 22 with $t = \log(6/\delta)$ to each sum. Since $\delta \geq 2^{-O(k/\sqrt{p})}$, $\log(6/\delta) = O(\sqrt{k\log(1/\delta)/p})$, i.e. we are still in the range of δ for which the square-root term in the tail bound of Lemma 22 is the linear term ct. So Lemma 22 gives that:

$$\Pr\left[\sum_{i:x_i<0} |x_i|^{p-1} < (k-k')\mu - 2\sqrt{2k\mu\log(1/\delta)}\right] \leq \delta/6,$$

$$\Pr\left[\sum_{i:x_i\geq 0} |x_i|^{p-1} > k'\mu + 2\sqrt{2k\mu\log(1/\delta)}\right] \leq \delta/6.$$

Combined with the Chernoff bound, this gives that with probability $1 - 2\delta/3$:

$$\begin{aligned}
&-\left(1 - \frac{2p}{\sqrt{k}}\right)\sum_{i:x_i<0}|x_i|^{p-1} + \left(1 + \frac{2p}{\sqrt{k}}\right)\sum_{i:x_i\geq 0}|x_i|^{p-1} \\
\leq& -\left(1 - \frac{2p}{\sqrt{k}}\right)\left((k-k')\mu - (2\sqrt{2})\sqrt{k\mu\log(1/\delta)}\right) \qquad (2.2)\\
&+\left(1 + \frac{2p}{\sqrt{k}}\right)\left(k'\mu + (2\sqrt{2})\sqrt{k\mu\log(1/\delta)}\right) \\
\leq& (2k'-k)\mu + (2\sqrt{k}p)\mu + (4\sqrt{2})\sqrt{k\mu\log(1/\delta)} \\
\leq& 6\mu\sqrt{k\log(1/\delta)} + (2\sqrt{k}p)\mu + (5\sqrt{2})\mu\sqrt{kp\log(1/\delta)} \\
\leq& 16k\mu \cdot \frac{\sqrt{p\log(1/\delta)}}{\sqrt{k}}. \qquad (2.3)
\end{aligned}$$

In the last step, we use that $p \leq \log k \leq \log(1/\delta)$ for the range of p, δ we consider. On the other hand, by Fact 21 $\sum_{i\in[k]} |x_i|^{p-1} = ||\mathbf{x}||_{p-1}^{p-1}$ is sampled from a random variable in $\Gamma^-(k\mu, 3/2)$ and thus by Lemma 29 and Lemma 22 is at least $k\mu/2$ with probability at least $1 - \delta/3$, i.e. $k\mu \leq 2\,||\mathbf{x}||_{p-1}^{p-1}$ with probability at least $1-\delta/3$. Combined with (2.3) by a union bound we get with probability $1 - \delta$:

$$-\left(1 - \frac{2p}{\sqrt{k}}\right)\sum_{i:x_i<0}|x_i|^{p-1} + \left(1 + \frac{2p}{\sqrt{k}}\right)\sum_{i:x_i\geq 0}|x_i|^{p-1} \leq 32\frac{\sqrt{p\log(1/\delta)}}{\sqrt{k}} \cdot ||\mathbf{x}||_{p-1}^{p-1}$$

Finally, by the Cauchy-Schwarz inequality for any $a \leq b$ and k-dimensional x we have $||\mathbf{x}||_a \leq k^{1/a-1/b}\,||\mathbf{x}||_b$. So, $||\mathbf{x}||_{p-1}^{p-1} \leq k^{1/p}\,||\mathbf{x}||_p^{p-1}$, giving (2.1) with probability $1-\delta$ as desired. □

Given Lemma 27, determining the value of σ that makes $\mathcal{M}_\sigma^p$ private is fairly straightforward:

Lemma 30. *Let $\mathcal{M}_\sigma^p$ be the mechanism such that $\mathcal{M}_\sigma^p(D)$ samples $\mathbf{x} \in \mathbb{R}^k$ from $\mathbf{x} \sim$ GGauss(p, σ) and outputs $\tilde{\phi} = \phi(D) + \mathbf{x}$. For $4 \leq p \leq \log k$ that is an even integer, $\epsilon \leq O(1)$, $\delta \in [2^{-O(k/p)}, 1/k]$, and*

$$\sigma = \Theta\left(\frac{\sqrt{kp\log(1/\delta)}}{\epsilon}\right),$$

$\mathcal{M}_\sigma^p$ is (ϵ, δ)-differentially private.

Proof. It suffices to show that for any D, D', letting $\Delta = \phi(D') - \phi(D)$:

$$\Pr_{\tilde{\phi}\sim\mathcal{M}_\sigma^p(D)}\left[\log\left(\frac{\Pr[\mathcal{M}_\sigma^p(D) = \tilde{\phi}]}{\Pr[\mathcal{M}_\sigma^p(D') = \tilde{\phi}]}\right) \leq \epsilon\right] = \Pr_{\tilde{\phi}\sim\mathcal{M}_\sigma^p(D)}\left[\frac{||\mathbf{x} - \Delta||_p^p - ||\mathbf{x}||_p^p}{\sigma^p} \leq \epsilon\right] \geq 1 - \delta.$$

Here, we abuse notation by letting Pr also denote a likelihood function. By Lemma 27 we now have with probability $1 - \delta/2$:

$$||\mathbf{x} - \Delta||_p^p - ||\mathbf{x}||_p^p \leq 64pk^{1/p-1/2}\sqrt{p\log(1/\delta)}\,||\mathbf{x}||_p^{p-1} + 2p^2k^{\frac{p}{2}}.$$

The pdf of the rescaled norm $r = ||\mathbf{x}||_p/\sigma$ is proportional to $r^{k-1}\exp(-r^p)$ over $(0, \infty)$ (the r^{k-1} appears because the $(k-1)$-dimensional surface area of the ℓ_p-sphere of radius r is proportional to r^{k-1}). Letting R denote r^p, the pdf of R is proportional to $R^{\frac{k}{p}-1}\exp(-R)$ by change of variables, i.e. R is the random variable $Gamma(\frac{k}{p})$. Then by the Gamma tail bound, with probability at least $1 - e^{-.001k/p} > 1 - \delta/2$, R is contained in $[\frac{k}{2p}, \frac{2k}{p}]$, so $||\mathbf{x}||_p$ is contained in $[\sigma\left(\frac{k}{2p}\right)^{1/p}, \sigma\left(\frac{2k}{p}\right)^{1/p}]$. Then by a union bound, with probability $1 - \delta$:

$$\frac{||\mathbf{x} - \Delta||_p^p - ||\mathbf{x}||_p^p}{\sigma^p} \leq \frac{128p^{1/p}\sqrt{kp\log(1/\delta)}}{\sigma} + \frac{4p^2k^{\frac{p}{2}}}{\sigma^p}.$$

Noting that $n^{1/n}$ is contained within $[1, e^{1/e}]$ for all $n \geq 1$, letting

$$\sigma = 185 \cdot \frac{\sqrt{kp\log(1/\delta)}}{\epsilon},$$

we get that $\frac{||\mathbf{x}-\Delta||_p^p - ||\mathbf{x}||_p^p}{\sigma^p} \leq \epsilon$ with probability $1 - \delta$ as desired. □

Error Guarantees

In this section, we analyze the ℓ_∞ error of $\mathcal{M}_\sigma^p$, for a given choice of δ in the range specified in Lemma 30. We give an expected error bound, and also a tail bound on the error. The error analysis follows almost immediately from the following lemma, which bounds the fraction of a sphere cap's volume with a large first coordinate:

Lemma 31. *Let* $\mathbf{x}$ *be chosen uniformly at random from a* k*-dimensional* ℓ_p*-sphere with arbitrary radius, i.e. the set of points with* $||\mathbf{x}||_p = R$ *for some* R*, for* $p \geq 1$*. Then we have:*

$$\Pr[|x_1| \geq r\,||\mathbf{x}||_p] \leq (1 - r^p)^{(k-1)/p} \leq \exp\left(-\frac{(k-1)r^p}{p}\right)$$

This lemma or one providing a similar bound likely already exists in the literature, but we are unaware of a reference for it. So, for completeness we give a full proof here. To prove this lemma we'll need the following lemma about convex bodies.

Lemma 32. *Let* $A \subseteq B \subset \mathbb{R}^k$ *be two compact convex bodies with* A *contained in* B*, and* A', B' *be their respective boundaries. Then* $Vol_{k-1}(A') \leq Vol_{k-1}(B')$*, where* Vol_{k-1} *denotes the* $(k-1)$*-dimensional volume.*

Proof. For any compact convex body S and its boundary S', the $(k-1)$-dimensional volume of S' satisfies:

$$\mathrm{Vol}_{k-1}(S') \propto \int_{\mathbb{S}^k} \mathrm{Vol}_{k-1}(\pi_{\theta^\top} S)\mathrm{d}\theta,$$

Where $\mathbb{S}^k$ is the k-dimensional unit sphere and $\pi_{\theta^\top} S$ is the orthogonal projection of S onto the subspace of $\mathbb{R}^k$ orthogonal to θ (see e.g. Section 5.5 of [55] for a proof of this fact). Since $A \subseteq B$ it follows that for all θ we have $\mathrm{Vol}_{k-1}(\pi_{\theta^\top} A) \leq \mathrm{Vol}_{k-1}(\pi_{\theta^\top} B)$ and so $\mathrm{Vol}_{k-1}(A') \leq \mathrm{Vol}_{k-1}(B')$. □

The idea behind the proof of Lemma 31 is to show that the region of the ℓ_p-ball with large positive first coordinate is contained within a smaller ℓ_p-ball, and then apply Lemma 32:

Proof of Lemma 31. By rescaling, we can assume $||\mathbf{x}||_p = 1$ and instead show:

$$\Pr[|x_1| \geq r] \leq (1 - r^p)^{(k-1)/p}$$

$$\Pr[|x_1| \geq r] = \frac{\mathrm{Vol}_{k-1}\left(\{\mathbf{x} : |x_1| \geq r, ||\mathbf{x}||_p = 1\}\right)}{\mathrm{Vol}_{k-1}\left(\mathbf{x} : ||\mathbf{x}||_p = 1\right)} = \frac{\mathrm{Vol}_{k-1}\left(\{\mathbf{x} : x_1 \geq r, ||\mathbf{x}||_p = 1\}\right)}{\mathrm{Vol}_{k-1}\left(\{\mathbf{x} : x_1 \geq 0, ||\mathbf{x}||_p = 1\}\right)},$$

Where Vol_{k-1} denotes the $(k-1)$-dimensional volume. To bound this ratio, let $\mathbf{v}$ be the vector $(r, 0, 0, \ldots, 0)$, and consider the (compact, convex) body $B_1 = \{\mathbf{x} : x_1 \geq r, ||\mathbf{x} - \mathbf{v}||_p \leq$

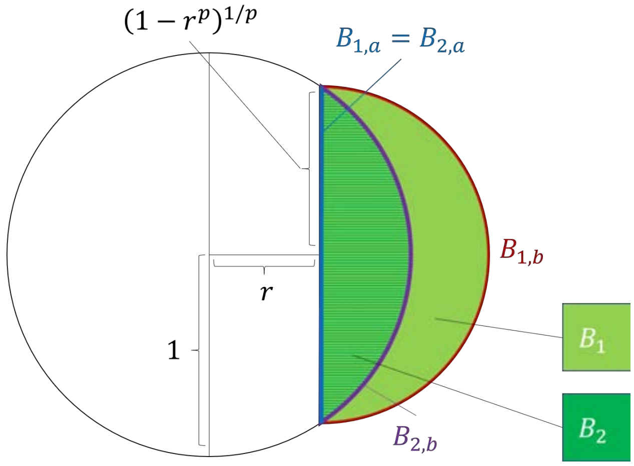


Figure 2.1: A picture of the bodies in the proof of Lemma 31 for $p = 2, k = 2$. B_2 has stripes that are the same color as $B_1 \setminus B_2$ to emphasize that B_1 contains B_2.

$(1-r^p)^{1/p}\}$. We have $r^p + (1-r)^p \leq 1$ for $0 \leq r \leq 1$, so B_1 contains the (also compact, convex) body $B_2 = \{\mathbf{x} : x_1 \geq r, ||\mathbf{x}||_p \leq 1\}$. Then by Lemma 32 the $(k-1)$-dimensional surface area of B_1 is larger than that of B_2. The boundary of B_1 is the union of the bodies $B_{1,a} := \{\mathbf{x} : x_1 = r, ||\mathbf{x}-\mathbf{v}||_p \leq (1-r^p)^{1/p}\}$ and $B_{1,b} := \{\mathbf{x} : x_1 \geq r, ||\mathbf{x}-\mathbf{v}||_p = (1-r^p)^{1/p}\}$, whose intersection has $(k-1)$-dimensional volume 0. Similarly, the boundary of B_2 is the union of the bodies $B_{2,a} := \{\mathbf{x} : x_1 = r, ||\mathbf{x}||_p \leq 1\}$ and $B_{2,b} := \{\mathbf{x} : x_1 \geq r, ||\mathbf{x}||_p = 1\}$, whose intersection has $(k-1)$-dimensional volume 0. See Figure 2.1 for an example of a picture of all of these bodies.

Nothing that $B_{1,a} = B_{2,a}$, we conclude that $\mathrm{Vol}_{k-1}(B_{1,b}) \geq \mathrm{Vol}_{k-1}(B_{2,b})$. Now we have:

$$\frac{\mathrm{Vol}_{k-1}\left(\{\mathbf{x} : x_1 \geq r, ||\mathbf{x}||_p = 1\}\right)}{\mathrm{Vol}_{k-1}\left(\{\mathbf{x} : x_1 \geq 0, ||\mathbf{x}||_p = 1\}\right)} \leq \frac{\mathrm{Vol}_{k-1}(\{\mathbf{x} : x_1 \geq r, ||\mathbf{x}-\mathbf{v}||_p = (1-r^p)^{1/p}\})}{\mathrm{Vol}_{k-1}\left(\{\mathbf{x} : x_1 \geq 0, ||\mathbf{x}||_p = 1\}\right)}.$$

The body in the numerator of the final expression is the body in the denominator, but shifted by $\mathbf{v}$ and rescaled by $(1-r^p)^{1/p}$ in every dimension. So, the final ratio is at most $(1-r^p)^{(k-1)/p}$. □

Corollary 33. *Let $\mathbf{x}$ be chosen uniformly at random from a k-dimensional ℓ_p-sphere with arbitrary radius for $p \geq 1$. Then we have:*

$$\Pr[||\mathbf{x}||_\infty \geq r\, ||\mathbf{x}||_p] \leq k \cdot \exp\left(-\frac{(k-1)r^p}{p}\right)$$

Proof. This follows from Lemma 31 and a union bound over all k coordinates (which have identical marginal distributions). □

Combining this corollary with Lemma 30, it is fairly straightforward to prove our first main result:

Theorem 34. *Let $\mathcal{M}_\sigma^p$ be the mechanism that takes in database D and counting queries ϕ, samples $\mathbf{x} \in \mathbb{R}^k$ from GGauss(p, σ), and outputs $\tilde{\phi} = \phi(D) + \mathbf{x}$. For $4 \leq p \leq \log k$ that is an even integer, For $\epsilon \leq O(1)$, $\delta \in [2^{-O(k/p)}, 1/k]$, and*

$$\sigma = 185 \cdot \frac{\sqrt{kp \log(1/\delta)}}{\epsilon},$$

$\mathcal{M}_\sigma^p$ is (ϵ, δ)-differentially private and for some sufficiently large constant c, and all $t \geq 0$:

$$\Pr_{\tilde{\phi} \sim \mathcal{M}_\sigma^p(D)} \left[\left|\left| \tilde{\phi} - \phi(D) \right|\right|_\infty \geq 1480t \cdot \frac{\sqrt{kp} \log^{1/p} k \sqrt{\log(1/\delta)}}{\epsilon} \right] \leq e^{-t^p \log k} + e^{-.001k/p}$$

Proof. The privacy guarantee follows from Lemma 30.

For the tail bound, if $\left|\left| \tilde{\phi} - \phi(D) \right|\right|_\infty > 1480t \cdot \frac{\sqrt{k} \log^{1/p} k \sqrt{p \log(1/\delta)}}{\epsilon}$ we have either $||\mathbf{x}||_p \geq 370 \cdot \frac{k^{1/2+1/p}\sqrt{p \log(1/\delta)}}{\epsilon}$ or $||\mathbf{x}||_\infty > \frac{4t \log^{1/p} k}{k^{1/p}} ||\mathbf{x}||_p$. Recall that $(||\mathbf{x}||_p / \sigma)^p$ is distributed according to a *Gamma*$(\frac{k}{p})$ random variable, and thus by a Gamma tail bound exceeds $2k/p$ with probability at most $e^{-.001k/p}$. In turn, $||\mathbf{x}||_p \geq 370 \cdot \frac{k^{1/2+1/p}\sqrt{p \log(1/\delta)}}{\epsilon} \geq \left(\frac{2k}{p}\right)^{1/p} \sigma$ with at most this probability. Then it follows by setting $r = \frac{4t \log^{1/p} k}{k^{1/p}}$ in Corollary 33 and a union bound that:

$$\Pr\left[\left|\left| \tilde{\phi} - \phi(D) \right|\right|_\infty \geq 1480t \cdot \frac{\sqrt{k} \log^{1/p} k \sqrt{p \log(1/\delta)}}{\epsilon} \right] \leq \Pr\left[||\mathbf{x}||_\infty \geq \frac{4t \log^{1/p} k}{k^{1/p}} ||\mathbf{x}||_p \right]$$

$$+ e^{-.001k/p} \leq \exp\left(-\frac{(k-1) 4^p t^p \log k}{kp} \right) + e^{-.001k/p} \leq e^{-t^p \log k} + e^{-.001k/p}.$$

□

This proves Theorem 17, up to some details. We first need the following corollary of Lemma 31:

Corollary 35. *Let $\mathbf{x}$ be chosen uniformly at random from a k-dimensional ℓ_p-sphere with arbitrary radius for $p \geq 1$. Then we have:*

$$\mathbb{E}[||\mathbf{x}||_\infty] \leq \frac{5 \log^{1/p} k}{k^{1/p}} ||\mathbf{x}||_p$$

Proof. Since $||\mathbf{x}||_\infty / ||\mathbf{x}||_p$ takes values in $[0,1]$, by Lemma 31 we have:

$$\begin{aligned}
\mathbb{E}[||\mathbf{x}||_\infty / ||\mathbf{x}||_p] &= \int_0^1 \Pr[||\mathbf{x}||_\infty / ||\mathbf{x}||_p \geq r]\mathrm{d}r \\
&\leq \int_0^{\frac{2^{1+1/p}\log^{1/p}k}{k^{1/p}}} 1\mathrm{d}r + \int_{\frac{2^{1+1/p}\log^{1/p}k}{k^{1/p}}}^1 k\cdot\exp\left(-\frac{(k-1)r^p}{p}\right)\mathrm{d}r \\
&\leq \frac{2^{1+1/p}\log^{1/p}k}{k^{1/p}} + \int_{\frac{2^{1+1/p}\log^{1/p}k}{k^{1/p}}}^1 k\cdot\exp\left(-\frac{(k-1)2^{p+1}\log k}{kp}\right)\mathrm{d}r \\
&\leq \frac{2^{1+1/p}\log^{1/p}k}{k^{1/p}} + \int_{\frac{2^{1+1/p}\log^{1/p}k}{k^{1/p}}}^1 k\cdot\exp\left(-2\log k\right)\mathrm{d}r \\
&\leq \frac{2^{1+1/p}\log^{1/p}k}{k^{1/p}} + \frac{1}{k} \\
&\leq \frac{5\log^{1/p}k}{k^{1/p}}.
\end{aligned}$$

Here we use that $2^p \geq p$ for all $p \geq 1$ and that $(1-\frac{c}{n})^n \leq e^{-c}$ for all $c \geq 0$. □

Proof of Theorem 17. We use Theorem 34 after rounding p up to the nearest even integer (this loses at most a multiplicative constant in the resulting error bounds). If the constant hidden in $\Theta(\log\log k)$ is a sufficiently large function of c_1, this gives the desired tail bound, up to the additive $e^{-.001k/p}$ in the probability bound (which may be larger than the $e^{-t^p\log k}$ term for large values of p). To remove the additive $e^{-.001k/p}$: if the less than $e^{-.001k/p} \leq \delta$ probability event that $(||\mathbf{x}||_p/\sigma)^p$ exceeds $2k/p$ occurs, we can instead just output $\tilde{d} = d$, i.e. instead set $x = 0$. This gives an $(\epsilon, 2\delta)$-private mechanism that always satisfies $(||\mathbf{x}||_p/\sigma)^p \leq 2k/p$, and then we can rescale our choice of δ appropriately. The tail bound can now be derived as in the proof of Theorem 34. Similarly, since we always have $(||\mathbf{x}||_p/\sigma)^p \leq 2k/p$, the expectation of $||\mathbf{x}||_\infty$ follows from Corollary 35. Finally, the expectation of $||\mathbf{x}||_q$ for $1 \leq q \leq p$ follows by using Jensen's inequality twice and the unconditional upper bound on $||\mathbf{x}||_p^p$:

$$\begin{aligned}
\mathbb{E}[||\mathbf{x}||_q] \leq \mathbb{E}[||\mathbf{x}||_q^q]^{1/q} = k^{1/q}\mathbb{E}[|x_1|^q]^{1/q} \leq k^{1/q}\mathbb{E}[|x_1|^p]^{1/p} = k^{1/q-1/p}\mathbb{E}[||\mathbf{x}||_p^p] \\
\leq k^{1/q-1/p}\cdot(2k/p)^{1/p}\sigma = O(k^{1/q}\sigma).
\end{aligned}$$

□

2.6 Composition with Sparse Vector

In this section, we generalize the mechanism of [78], which is a composition of the Gaussian mechanism and sparse vector mechanism of [34], by analyzing a composition of $\mathcal{M}_\sigma^p$ and the sparse vector mechanism instead[2]. The guarantees given by sparse vector can be given in the following form that we will use:

Theorem 36 (Sparse Vector)**.** *For every $k \geq 1, c_{SV} \leq k, \epsilon_{SV}, \delta_{SV}, \beta_{SV} > 0$, and*

$$\alpha_{SV} \geq O\left(\frac{\sqrt{c_{SV}\log(1/\delta_{SV})}\log(k/\beta_{SV})}{\epsilon_{SV}}\right),$$

there exists a mechanism SV that takes as input a dataset D and counting queries ϕ and outputs $\tilde{\phi} \in \mathbb{R}^k$ such that:

- *Letting $\mathbf{x} \sim \mathbf{x}'$ if $||\mathbf{x} - \mathbf{x}'||_\infty \leq 1$, SV is $(\epsilon_{SV}, \delta_{SV})$-differentially private.*
- *If at most c_{SV} entries of $\phi(D)$ have absolute value strictly greater than $\alpha_{SV}/2$, then:*

$$\Pr_{\tilde{\phi}\sim SV(D)}\left[\left|\left|\tilde{\phi} - \phi(D)\right|\right|_\infty \geq \alpha_{SV}\right] \leq \beta_{SV}.$$

- *Regardless of the value of $\tilde{\phi}$ we have for all $t \geq 0$:*

$$\Pr_{\tilde{\phi}\sim SV(\mathbf{x})}\left[\left|\left|\tilde{\phi} - \phi(D)\right|\right| \geq \max\left\{||\phi(D)||_\infty, t\sqrt{k\log(1/\delta_{SV})}/\epsilon_{SV}\right\}\right] \leq ke^{-\Omega(t)}.$$

Proof. The mechanism is given by modifying the NumericSparse algorithm given as Algorithm 3 in [32] by outputting 0 instead of $\perp$ or 0 for all remaining queries instead of halting prematurely. The first two properties follow from the associated proofs in that text.

The third property follows because for all entries of $\tilde{\phi}$ that SV does not output as 0 (for which the error, i.e. corresponding entry of $\tilde{\phi} - \phi(D)$, is of course bounded by $||\phi(D)||_\infty$), the error is drawn from $Lap(b)$ where $b = O(\sqrt{k\log(1/\delta_{SV})}/\epsilon_{SV})$. So the maximum error for these (at most $c_{SV} \leq k$) entries is stochastically dominated by the maximum of the absolute value of k of these Laplace random variables, which is at most tb with probability ke^{-t}. □

We now prove Theorem 37, from which Theorem 18 follows up to some minor details:

Theorem 37. *For any $4 \leq p \leq \log k$ that is an even integer, $\epsilon \leq O(1)$, $\delta \in [2^{-O(k/p)}, 1/k]$, and $t \in [0, O(\frac{\log k}{\log\log k})]$, there exists a (ϵ, δ)-differentially private mechanism $\mathcal{M}$ that takes in a database D and counting queries ϕ and outputs a random $\tilde{\phi} \in \mathbb{R}^k$ such that for a sufficiently large constant c:*

$$\Pr_{\tilde{\phi}\sim\mathcal{M}(D)}\left[\left|\left|\tilde{\phi} - \phi(D)\right|\right|_\infty \geq \frac{ct\sqrt{kp\log(1/\delta)}(\log\log k)^{1/p}}{\epsilon}\right] \leq e^{-\log^t k}.$$

[2] Unlike its preprint, the journal version of [78] uses a slightly different mechanism based on the exponential mechanism in place of the sparse vector mechanism. A similar change can likely be made to the mechanism given in this section; we stick to using the sparse vector mechanism for a slightly simpler proof.

Proof. The mechanism is as follows: We sample $\mathbf{x} \sim GGauss(p, \sigma)$ for

$$\sigma = \Theta\left(\frac{\sqrt{kp\log(1/\delta)}}{\epsilon}\right),$$

If $||\mathbf{x}||_p^p > 2k\sigma^p/p$, we output $\phi(D)$. Otherwise, we instantiate SV from Theorem 36 with parameters:

$$\alpha_{SV} = 12t(\log\log k)^{1/p}\sigma \leq \frac{ct\sqrt{kp\log(1/\delta)}(\log\log k)^{1/p}}{\epsilon}, \qquad c_{SV} = 4k/\log^{2+2t} k,$$

$$\epsilon_{SV} = \epsilon/2, \qquad \delta_{SV} = \delta/3, \qquad \beta_{SV} = \exp(-\log^t k)/2.$$

We create an arbitrary database D_{synth} and set of counting queries ϕ_{synth} such that $\phi_{synth}(D_{synth}) = \mathbf{x}$, and let $\tilde{\mathbf{x}}$ be the output of SV on D_{synth}, ϕ_{synth}. We then output $\tilde{\phi} = d + x - \hat{x}$.

First, note that:

$$\frac{\sqrt{c_{SV}\log(1/\delta_{SV})}\log(k/\beta_{SV})}{\epsilon_{SV}} \leq \frac{\sqrt{\frac{16k}{\log^{2+2t}k}\log(1/\delta)}(\log k + \log^t k)}{\epsilon} \leq \frac{4\sqrt{k\log(1/\delta)}}{\epsilon},$$

i.e. α satisfies the requirements of Theorem 36 as long as the constant hidden in the $\Theta(\cdot)$ notation in the choice of σ is sufficiently large.

To analyze the privacy guarantee, this is the composition of:

- The mechanism of Theorem 34, which if the constant hidden in the $\Theta(\cdot)$ in the expression for σ is sufficiently large, is $(\epsilon/2, \delta/3)$-differentially private.
- The SV mechanism of Theorem 36, with parameters set so it is $(\epsilon/2, \delta/3)$-differentially private.
- The event that $||\mathbf{x}||_p^p > 2k\sigma^p/p$, causing us to release the database, which we recall from the Proof of Theorem 34 happens with probability at most $2^{-\Omega(k/p)} \leq \delta/3$.

By composition, we get that the mechanism is (ϵ, δ)-differentially private as desired.

To show the tail bound on ℓ_∞-error: If $||\mathbf{x}||_p^p > 2k\sigma^p/p$, then we have $\tilde{\phi} = \phi(D)$, so trivially the tail bound is satisfied. So, it suffices to show that conditional on $||\mathbf{x}||_p^p \leq 2k\sigma^p/p$ occurring, we have the tail bound. By a union bound, the guarantees of Theorem 36 give that $\left|\left|\tilde{\phi} - \phi(D)\right|\right|_\infty = ||\mathbf{x} - \hat{\mathbf{x}}||_\infty \leq \alpha_{SV}$ (i.e the tail bound is satisfied) if at most $4k/\log^{2+2t} k$ entries of $\mathbf{x}$ have absolute value greater than $\alpha_{SV}/2$ with probability less than, say, $e^{-2\log^t k}$. Using $r = 3t\frac{(\log\log k)^{1/p}}{k^{1/p}}$ in Lemma 31 and a union bound with the $1 - \delta/3$ probability event that $||\mathbf{x}||_p \leq (2k/p)^{1/p}\sigma$, for each coordinate x_i of $\mathbf{x}$ we have:

$$|x_i| \geq \alpha_{SV}/2 = 6t(\log\log k)^{1/p}\sigma = 2rk^{1/p}\sigma \geq r\,||\mathbf{x}||_p\,,$$

with probability at most $\frac{1}{\log^{2+2t} k} + 2^{-\Omega(k/p)} \leq \frac{2}{\log^{2+2t} k}$. Since we sample $\mathbf{x}$ with probability proportional to $\exp(-\sum_{i\in[k]}|x_i|^p/\sigma^p)$, each coordinate's distribution is independent, so using a Chernoff bound we conclude that with probability $e^{-\Omega(k/\log^{2+2t} k)} \leq e^{-2\log^t k}$ at most $4k/\log^{2+2t} k$ coordinates have absolute value greater than α_{SV} as desired. □

Proof of Theorem 18. The tail bound in Theorem 18 follows immediately from Theorem 37 by choosing p to be an even integer satisfying $p = \Theta(\log\log\log k)$.

For the expectation, we use the tail bound of Theorem 18. We have:

$$\mathbb{E}_{\tilde{d}\sim\mathcal{M}(d)}\left[\left|\left|\tilde{\phi} - \phi(D)\right|\right|_\infty\right] = \int_0^\infty \Pr\left[\left|\left|\tilde{\phi} - \phi(D)\right|\right|_\infty \geq s\right] \mathrm{d}s$$

$$= \int_0^a \Pr\left[\left|\left|\tilde{\phi} - \phi(D)\right|\right|_\infty \geq s\right] \mathrm{d}s + \int_a^b \Pr\left[\left|\left|\tilde{\phi} - \phi(D)\right|\right|_\infty \geq s\right] \mathrm{d}s + \int_b^\infty \Pr\left[\left|\left|\tilde{\phi} - \phi(D)\right|\right|_\infty \geq s\right] \mathrm{d}s.$$

We choose $a = \frac{2c\sqrt{k\log\log\log k\log(1/\delta)}}{\epsilon}$, $b = \frac{k\sqrt{\log(1/\delta)}}{\epsilon}$. The integral over $[0,a]$ is of course bounded by a. By Theorem 37, the integral over $[a,b]$ is bounded by $b\cdot e^{-\log^2 k} \leq \frac{\sqrt{\log(1/\delta)}}{\epsilon} \leq a$. Finally, to bound the third term, recall that the mechanism of Theorem 37 outputs $\phi(D)$ (i.e. effectively chooses $\mathbf{x}, \hat{\mathbf{x}} = 0$ instead) if $||\mathbf{x}||_p$ is too large. So, unconditionally we have:

$$||\mathbf{x}||_\infty \leq ||\mathbf{x}||_p \leq (2k/p)^{1/p}\sigma \leq \frac{2c\sqrt{k\log\log\log k\log(1/\delta)}}{\epsilon} \leq b.$$

So by the third property in Theorem 36 we have for $s \in [b, \infty)$:

$$\Pr_{\tilde{d}\sim\mathcal{M}(d)}\left[\left|\left|\tilde{\phi} - \phi(D)\right|\right|_\infty \geq s\right] = \Pr_{x,\hat{x}}\left[||x - \hat{x}||_\infty \geq s\right] \leq ke^{-\Omega(s/(\sqrt{k\log(1/\delta)}/\epsilon))}.$$

And so by change of variables, with $s' = s/(\sqrt{k\log(1/\delta)}/\epsilon)$:

$$\int_b^\infty \Pr\left[\left|\left|\tilde{\phi} - \phi(D)\right|\right|_\infty \geq s\right] \mathrm{d}s \leq \frac{\sqrt{k\log(1/\delta)}}{\epsilon}\int_{\sqrt{k}}^\infty ke^{-\Omega(s')}\mathrm{d}s' \leq \frac{k^{1.5}\sqrt{\log(1/\delta)}}{\epsilon}\cdot e^{-\Omega(\sqrt{k})} \leq a.$$

So we conclude

$$\mathbb{E}_{\tilde{d}\sim\mathcal{M}(d)}\left[\left|\left|\tilde{\phi} - \phi(D)\right|\right|_\infty\right] \leq 3a = O\left(\frac{\sqrt{k\log\log\log k\log(1/\delta)}}{\epsilon}\right),$$

as desired. □

2.7 Concentration of Generalized Gammas

In this section we consider the Generalized Gamma random variable $GGamma(a, b)$ parameterized by a, b with pdf:

$$p(x) = \frac{bx^{a-1}e^{-x^b}}{\Gamma(a/b)}, x \in (0, \infty).$$

Where the Gamma function $\Gamma(x)$ is defined over the positive reals as

$$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\mathrm{d}x.$$

We recall that $\Gamma(z)$ is a continuous analog of the factorial in that it satisfies $\Gamma(x+1) = x \cdot \Gamma(x)$. When $b = 1$, $GGamma(a, b)$ is exactly the Gamma random variable $Gamma(a)$ (we will use $Gamma$ to denote the random variable and Γ to denote the function to avoid ambiguous notation).

We want to show that sums of $GGamma(\frac{1}{p-1}, \frac{p}{p-1})$ random variables concentrate nicely. To do this, we will show that they are sub-gamma:

To show that $GGamma(\frac{1}{p-1}, \frac{p}{p-1})$ are sub-gamma, we will relate the moment-generating function of $GGamma(\frac{1}{p-1}, \frac{p}{p-1})$ to that of the Gamma random variable with the same mean using the following facts:

Fact 38. *For a Generalized Gamma random variable $X \sim GGamma(a, b)$ the moments are $\mathbb{E}[X^r] = \frac{\Gamma((a+r)/b)}{\Gamma(a/b)}$. In particular, for a Gamma random variable $X \sim Gamma(a)$ the moments are $\mathbb{E}[X^r] = \frac{\Gamma(a+r)}{\Gamma(a)}$.*

See e.g. Section 17.8.7 of [49] for a derivation of this fact. Note here that $GGamma(\frac{1}{p-1}, \frac{p}{p-1})$ has mean $\mu = 1/\Gamma(1/p)$. To relate the moments of Generalized Gamma random variables to Gamma random variables' we note the following about μ:

Fact 39. *For all $p \geq 2$, we have $\frac{1}{p} \leq \frac{1}{\Gamma(1/p)} \leq \frac{1.2}{p}$.*

Putting it all together, we get the following lemmas, which combined with Fact 39 give us Lemma 29:

Lemma 40. *Let $Y = GGamma(\frac{1}{p-1}, \frac{p}{p-1})$ for $p \geq 2$. Then, for $\mu = \mathbb{E}[Y] = \frac{1}{\Gamma(1/p)}$, we have $Y \in \Gamma^+(\mu, 1)$.*

Proof. We compare the moment-generating function of (the centered version of) Y to that of $X = Gamma(\mu)$ where $\mu = \mathbb{E}[Y]$. X is in $\Gamma(\mu, 1)$ so it suffices to show Y's moment generating function is smaller than X's. First, looking at the moment generating function of Y, we have:

$$\begin{aligned}
\mathbb{E}[e^{\lambda Y}] &= 1+\lambda\mu+\sum_{r=2}^{\infty}\left[\frac{\lambda^r}{r!}\mathbb{E}[Y^r]\right]\\
&= 1+\lambda\mu+\sum_{r=2}^{\infty}\left[\frac{\lambda^r}{r!}\frac{\Gamma(\frac{1}{p}+\frac{r(p-1)}{p})}{\Gamma(\frac{1}{p})}\right]\\
&\overset{(a)}{\leq} 1+\lambda\mu+\sum_{r=2}^{\infty}\left[\frac{\lambda^r}{r!}\frac{\Gamma(\frac{1}{p}+r)}{\Gamma(\frac{1}{p})}\right]\\
&\overset{(b)}{\leq} 1+\lambda\mu+\sum_{r=2}^{\infty}\left[\frac{\lambda^r}{r!}\frac{\Gamma(\mu+r)}{\Gamma(\mu)}\right] = \mathbb{E}[e^{\lambda X}].
\end{aligned}$$

(a) follows because the Gamma function is monotonically increasing in the range $[1.5,\infty)$. (b) follows because $\mu=\frac{1}{\Gamma(1/p)}\geq 1/p$ for $p\geq 1$, and because for positive integers r, $\frac{\Gamma(x+r)}{\Gamma(x)}=\prod_{i=0}^{r-1}(x+i)$ is monotonically increasing in x. Since $X\in\Gamma^+(\mu,1)$ and X,Y have the same mean, we have that $Y\in\Gamma^+(\mu,1)$ as well. □

Lemma 41. *Let $Y = GGamma(\frac{1}{p-1},\frac{p}{p-1})$ for $p\geq 3$. Then, for $\mu=\mathbb{E}[Y]=\frac{1}{\Gamma(1/p)}$, we have $Y\in\Gamma^-(\mu,3/2)$.*

Proof. Similarly to the previous lemma, we have for all $0\leq\lambda\leq 2/3$:

$$\begin{aligned}
&\mathbb{E}[e^{-\lambda Y}]\\
&= 1-\lambda\mu+\sum_{r=2}^{\infty}\left[\frac{(-\lambda)^r}{r!}\frac{\Gamma(\frac{1}{p}+\frac{r(p-1)}{p})}{\Gamma(\frac{1}{p})}\right]\\
&= 1-\lambda\mu+\sum_{r=1}^{\infty}\left[\frac{\lambda^{2r}}{(2r)!}\cdot\frac{\Gamma(\frac{1}{p}+2r\frac{p-1}{p})}{\Gamma(\frac{1}{p})}\left(1-\frac{\lambda}{2r+1}\cdot\frac{\Gamma(\frac{1}{p}+(2r+1)\frac{p-1}{p})}{\Gamma(\frac{1}{p}+2r\frac{p-1}{p})}\right)\right]\\
&= 1-\lambda\mu+\sum_{r=1}^{\infty}\left[\frac{\lambda^{2r}}{(2r)!}\cdot\frac{\Gamma(\frac{1}{p}+2r)}{\Gamma(\frac{1}{p})}\left(\frac{\Gamma(\frac{1}{p}+2r\frac{p-1}{p})}{\Gamma(\frac{1}{p}+2r)}-\frac{\lambda}{2r+1}\cdot\frac{\Gamma(\frac{1}{p}+(2r+1)\frac{p-1}{p})}{\Gamma(\frac{1}{p}+2r)}\right)\right]\\
&\overset{(c)}{\leq} 1-\lambda\mu+\sum_{r=1}^{\infty}\left[\frac{\lambda^{2r}}{(2r)!}\cdot\frac{\Gamma(\frac{1}{p}+2r)}{\Gamma(\frac{1}{p})}\left(1-\frac{\lambda}{2r+1}\cdot\frac{\Gamma(\frac{1}{p}+2r+1)}{\Gamma(\frac{1}{p}+2r)}\right)\right]\\
&\overset{(d)}{\leq} 1-\lambda\mu+\sum_{r=1}^{\infty}\left[\frac{\lambda^{2r}}{(2r)!}\cdot\frac{\Gamma(\mu+2r)}{\Gamma(\mu)}\left(1-\frac{\lambda}{2r+1}\cdot\frac{\Gamma(\mu+2r+1)}{\Gamma(\mu+2r)}\right)\right]\\
&= 1-\lambda\mu+\sum_{r=2}^{\infty}\left[\frac{(-\lambda)^r}{r!}\cdot\frac{\Gamma(\mu+r)}{\Gamma(\mu)}\right] = \mathbb{E}[e^{-\lambda X}].
\end{aligned}$$

Which, up to proving $(c), (d)$ hold, shows that $Y \in \Gamma^-(\mu, 3/2)$ since X and Y have the same mean and $X \in \Gamma^-(\mu, 0) \subset \Gamma^-(\mu, 3/2)$. (c) follows because the change in each term in the sum is

$$\frac{\lambda^{2r}}{(2r)!}\frac{1}{\Gamma\left(\frac{1}{p}\right)}\cdot$$

$$\left[\Gamma\left(\frac{1}{p}+2r\right)-\Gamma\left(\frac{1}{p}+2r\frac{p-1}{p}\right)-\frac{\lambda}{2r+1}\left(\Gamma\left(\frac{1}{p}+2r+1\right)-\Gamma\left(\frac{1}{p}+(2r+1)\frac{p-1}{p}\right)\right)\right].$$

To show this expression is non-negative, it suffices to show that just the term in the brackets is positive, or equivalently, for all $r \geq 2, p \geq 3$:

$$\Gamma\left(\frac{1}{p}+2r\right)\left(1-\frac{\Gamma\left(\frac{1}{p}+2r\frac{(p-1)}{p}\right)}{\Gamma\left(\frac{1}{p}+2r\right)}\right) \geq$$

$$\frac{\lambda}{2r+1}\Gamma\left(\frac{1}{p}+2r+1\right)\left(1-\frac{\Gamma\left(\frac{1}{p}+(2r+1)\frac{p-1}{p}\right)}{\Gamma\left(\frac{1}{p}+2r+1\right)}\right).$$

Since we have $\Gamma\left(\frac{1}{p}+2r+1\right) = (\frac{1}{p}+2r)\Gamma\left(\frac{1}{p}+2r\right) \leq (2r+1)(\frac{1}{p}+2r)$, it further suffices to just show:

$$f(r,p) := \frac{\left(1-\frac{\Gamma(\frac{1}{p}+2r\frac{(p-1)}{p})}{\Gamma(\frac{1}{p}+2r)}\right)}{\left(1-\frac{\Gamma(\frac{1}{p}+(2r+1)\frac{p-1}{p})}{\Gamma(\frac{1}{p}+2r+1)}\right)} \geq \lambda.$$

For any fixed $r \geq 2$, one can verify analytically that $f(r,p)$ is monotonically decreasing in p over $p \in [1, \infty)$ and the limit as p goes to infinity is $g(r) := \frac{2r\psi(2r)}{(2r+1)\psi(2r+1)}$ where ψ is the digamma function $\psi(x) = \frac{\frac{d}{dx}\Gamma(x)}{\Gamma(x)}$. One can also verify analytically that $g(r)$ is monotonically increasing, and $g(2) \approx .6672$. So, for all $r \geq 2, p \geq 3$ we have $f(r,p) > 2/3$ and thus for $\lambda \in [0, 2/3]$, the inequality (c) is satisfied.

(d) follows by looking at the function

$$z(x) = \frac{\Gamma(x+r)}{\Gamma(x)}\left(1-\frac{\lambda}{r+1}\cdot\frac{\Gamma(x+r+1)}{\Gamma(x+r)}\right) = \left(1-\frac{\lambda(x+r)}{r+1}\right)\prod_{i=0}^{r-1}(x+i).$$

For $r \geq 2, \lambda \leq 1$, one can verify analytically that $z(x)$ is monotonically increasing in the interval $(0, 1/2] \supseteq (0, \frac{1.2}{p}] \supseteq (0, \mu]$. Since $\mu \geq \frac{1}{p}$, this gives that each term in the right-hand-side of (d) is larger than the corresponding term on the left-hand-side. □

Chapter 3

Efficient Private Log-Strongly Concave Sampling

3.1 Introduction and Problem Definition

There is a large class of mechanisms in the differential privacy literature that use the exponential mechanism (Example 7) with appropriate score functions, use it as a subroutine, or sample from $\exp(-f)$ for some function f. This includes differentially private mechanisms for several important problems, such as PCA [23, 53], functional PCA [9], answering counting queries [47], robust regression [6], some combinatorial optimization problems [43], k-means clustering [37], optimization of dispersed functions [11], convex optimization [13, 64], Bayesian data analysis [65, 27, 86, 87, 39], linear and quantile regression [73], etc.

When the range of outputs $\mathcal{X}$ is finite and small, this sampling is straightforward. Several differentially private mechanisms instantiate the exponential mechanism where $\mathcal{X} = \mathbb{R}^k$, in which case this sampling is not straightforward. Such sampling problems are not new and often occur in statistics and machine learning settings. The common practical approach is to use heuristic MCMC samplers such as Gibbs sampling, which often works well in problems arising in practice. However, given that convergence is not guaranteed, the resulting algorithms may not be differentially private. Indeed one can construct simple score functions on the hypercube for which the natural Metropolis chain run for any polynomial time leads to a non-private algorithm. There are also well-known complexity-theoretic barriers in exactly sampling from $\exp(-f)$ if f is not required to be convex.

Several applications however involve convex functions f, which is the focus of this chapter. This is the problem of sampling from a log-concave distribution, which has attracted a lot of interest. Here, there are two broad lines of work. The classical results in this line of work (e.g. [5, 60]) show that given an oracle for computing the function, one can sample from a distribution that is δ-close to the target distribution in time polynomial in k and $\log \frac{1}{\delta}$. Here the closeness is measured in statistical distance. By itself, this does not suffice to give a differentially private algorithm, as differential privacy requires closeness in more stringent notions of distance. The fact that the time complexity is logarithmic in $\frac{1}{\delta}$ however allows for an exponentially small statistical distance in polynomial time. This immediately yields (ϵ, δ)-differentially private algorithms, and with some additional work can also yield ϵ-differentially private algorithms [47]. Techniques from this line of work can also sometimes apply to non-convex f of interest. Indeed [53] designed a polynomial time algorithm for the case of f being a Rayleigh quotient to allow for efficient private PCA.

The runtime of these log-concave sampling algorithms however involves large polynomials. A beautiful line of work has reduced the dependence (of the number of function oracle calls) on the dimension from roughly k^{10} in [5] to k^3 in [59, 60]. Nevertheless, the algorithms still fall short of being efficient enough to be implementable in practice for large k. A second, more recent, line of work [26, 30] have shown that "first order" Markov Chain Monte Carlo (MCMC) algorithms such as Langevin MCMC and Hamiltonian MCMC enjoy fast convergence, and have better dependence on the dimension. These algorithms are typically simpler and more practical but have polynomial dependence on the closeness parameter ϵ. This polynomial dependence on ϵ makes the choice of distance more important. Indeed these algorithms have been analyzed for various measures of distance between distributions such as statistical distance, KL-divergence and Wasserstein distance.

These notions of distance however do not lead to efficient differentially private algorithms. This motivates the question of establishing rapid mixing in Rènyi-divergence for these algorithms. This is the question we address in this work, and show that when f is smooth and strongly convex, discretized Langevin dynamics converge in iteration complexity near-linear in the dimension. This gives more efficient differentially private algorithms for sampling for such f.

[82] recently studied this question, partly for similar reasons. They considered the Unadjusted (i.e., overdamped) Langevin Algorithm and showed that when the (discretized) Markov chain satisfies suitable mixing properties (e.g. Log Sobolev inequality), then the discrete process converges in Rènyi-divergence to *a* stationary distribution. However this stationary distribution of the discretized chain is different from the target distribution. The Rènyi-divergence between the stationary distribution and $\exp(-f)$ is not very well-understood [75, 88], and it is conceivable that the stationary distribution of the discrete process is *not* close in Rènyi-divergence to the target distribution and thus may not be differentially private. Thus the question of designing fast algorithms that sample from a distribution close to the distribution $\exp(-f)$ in Rènyi-divergence was left open.

In this work we use a novel approach to address these questions of fast sampling from $\exp(-f)$ using the discretized Langevin Algorithm. Interestingly, we borrow tools commonly used in differential privacy, though applied in a way that is not very intuitive from a privacy point of view. We upper bound the Rènyi-divergence between the output of the discrete Langevin Algorithm run for T steps, and the output of the continuous process run for time $T\eta$. The continuous process is known [82] to converge very quickly in Rènyi-divergence to the target distribution. This allows us to assert closeness (in Rènyi-divergence) of the output of the discrete algorithm to the target distribution. This bypasses the question of the bias of the stationary distribution of the discrete process. Moreover, this gives us a differentially private algorithm with iteration complexity near-linear in the dimension. Our result applies to log-smooth and strongly log-concave distributions. While results of this form may also be provable using methods from optimal transport, we believe that our techniques are simpler and more approachable to the differential privacy community, and may be more easily adaptable to other functions f of interest.

Our approach is general and simple. We show that it can be extended to the *underdamped* Langevin dynamics which have a better dependence on dimension, modulo proving fast mixing for the continuous process. As a specific application, we show how our results lead to faster algorithms for implementing the mechanisms in [64].

3.2 Our Results and Technical Overview

Properties of f	Process	η
1-strongly convex, M-smooth	Overdamped	$\tilde{O}\left(\frac{1}{\tau M^4 \ln^2 \alpha} \cdot \frac{\epsilon^2}{k}\right)$ (Theorem 52)
L-Lipschitz, M-smooth	Overdamped	$\tilde{O}\left(\frac{1}{\tau M^4 \ln^2 \alpha} \cdot \frac{\epsilon^2}{L^2+k}\right)$ (Theorem 56)
1-strongly convex, M-smooth	Underdamped	$\tilde{O}\left(\frac{1}{\tau M \ln \alpha} \cdot \frac{\epsilon}{\sqrt{k}}\right)$ (Theorem 65)

Table 3.1: Summary of results. For each family of functions and process (either overdamped or underdamped Langevin dynamics), an upper bound is listed on the timestep length η needed to ensure the α-Rényi-divergence between the discrete and continuous processes is at most ϵ after time τ.

Our results are summarized in Table 3.1. Combined with results from [82] on the convergence of the continuous process, the first result gives the following algorithmic guarantee, our main result:

Theorem 42. *Fix any $\alpha \geq 1$. Let S be a distribution satisfying $S(x) \propto e^{-f(x)}$ for 1-strongly convex and M-smooth f with global minimum at 0. Let P be the distribution arrived at by running discretized overdamped Langevin dynamics using f with step size $\eta = \tilde{O}(\frac{1}{\tau M^4 \ln^2 \alpha} \cdot \frac{\epsilon^2}{k})$ for continuous time $\tau = O(\alpha \ln \frac{k \ln M}{\epsilon})$ (i.e. for $\tilde{O}(\frac{\alpha^2 M^4 k}{\epsilon^2})$ steps) from initial distribution $N(0, \mathbb{I}_k)$. Then we have $R_\alpha(P||S), R_\alpha(S||P) \leq \epsilon$.*

This is the first algorithmic result for sampling from log-smooth and strongly log-concave distributions with low error in Rényi-divergence without additional assumptions. In particular, if for $\alpha = 1 + 2\log(1/\delta)/\epsilon$ we have $R_\alpha(P||S), R_\alpha(S||P) \leq \epsilon/2$, then by Theorem 14 we have that P, S satisfy divergence bounds corresponding to (ϵ, δ)-differential privacy. In turn, given any mechanism that outputs S, S' on adjacent databases satisfying (ϵ, δ)-differential privacy and the strong convexity and smoothness conditions, Theorem 42 and standard composition theorems gives a mechanism that outputs P, P' for these databases such that the mechanism satisfies $(3\epsilon, 3\delta)$-differential privacy, P, P' are possible to efficiently sample from, and P, P' obtain utility guarantees comparable to those of S, S'.

All results in Figure 3.1 are achieved using a similar analysis, which we describe here. Instead of directly bounding the divergence between the discrete and continuous processes, we instead bound the divergence between the discrete processes using step sizes $\eta, \eta/c$. Our resulting bound does not depend on c, so we can take the limit as c goes to infinity and the latter approaches the continuous process. Suppose within each step of size η, neither process moves more than r away from the position at the start of this step. Then by smoothness, in each interval of length η/c the distance between the gradient steps between the two processes is upper bounded by $Lr\frac{\eta}{c}$. Our divergence bound thus worsens by at most $R_\alpha(N(0, \frac{2\eta}{c})||N(\mathbf{x}, \frac{2\eta}{c}))$ where $\mathbf{x}$ is a vector with $||\mathbf{x}||_2 \leq Mr\frac{\eta}{c}$. The divergence between shifted Gaussians is given by Fact 15, giving us a divergence bound.

Of course, since the movement due to Brownian motion can be arbitrarily large, there is no unconditional bound on r. Instead, we derive tail bounds for r, giving a divergence bound (depending on δ) between the two processes conditioned on a probability $1-\delta$ event for every δ. We then show a simple lemma which says that conditional upper bounds on the larger moments of a random variable give an unconditional upper bound on the expectation of that random variable. By the definition of Rényi-divergence, $\exp((\alpha'-1)R_{\alpha'}(P||Q))$ is a moment of $\exp((\alpha-1)R_\alpha(P||Q))$ for $\alpha' > \alpha$, so we can apply this lemma to our conditional bound on α'-Rényi-divergence to get an unconditional bound on α-Rényi-divergence via Jensen's inequality.

Finally, since our analysis only needs smoothness, the radius tail bound, and the fact that the process is a composition of gradient steps with Gaussian noise, our analysis easily extends to sampling from Lipschitz rather than strongly convex functions and analyzing the underdamped Langevin dynamics.

As an immediate application, we recall the work of [64], who give a (ϵ, δ)-differentially private mechanism that (approximately) samples from a Gibbs posterior with a strongly log-concave prior, for applications such as mean estimation and logistic regression. Their iteration complexity of $\tilde{O}(k^3/\delta^2)$ proved in Proposition 13 of [64] gets improved to $\tilde{O}(k/\epsilon^4)$ using our main result. We note that the privacy parameters in (ϵ, δ)-differential privacy that one typically aims for are ϵ being constant, and δ being negligible.

3.3 Other Related Work

Following our work, [35] improved the dependence of the iteration complexity needed to converge to within Rènyi-divergence ϵ of the stationary distribution from $1/\epsilon^2$ to $1/\epsilon$, using a weaker assumption called strong dissipativity, which can be viewed as a strong convexity condition holding only for pairs of points that are sufficiently far away. Also following our work, [24] showed that for two instances of the discrete Langevin dynamics on different loss functions, as long as the two loss functions' gradients differ by a bounded amount, they can derive a bound on the Rènyi-divergence between the two instances' final values that does not go to infinity as time goes to infinity. This is in contrast with our divergence bounds, which effectively bound the divergence between two entire chains rather than a single value in the chains, and which do go to infinity as the number of iterations goes to infinity. While both these results improve upon ours quantitatively in the setting of Theorem 42, we remark that qualitatively our analysis remains relatively simple compared to both these papers, and as previously mentioned our analysis is arguably much more general and applicable to other settings.

3.4 Preliminaries

Langevin Dynamics and Basic Assumptions

For the majority of the chapter we focus on the overdamped Langevin dynamics in $\mathbb{R}^k$, given by the following stochastic differential equation (SDE):

$$\mathrm{d}\mathbf{x}_t = -\nabla f(\mathbf{x}_t)\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,$$

Where B_t is a standard k-dimensional Brownian motion. Under mild assumptions (such as strong convexity of f), it is known that the stationary distribution of the SDE is the distribution p satisfying $p(\mathbf{x}) \propto e^{-f(\mathbf{x})}$. Algorithmically, it is easier to use the following discretization with *steps* of size η:

$$\mathrm{d}\mathbf{x}_t = -\nabla f(\mathbf{x}_{\lfloor \frac{t}{\eta} \rfloor \eta})\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,$$

i.e., we only update the gradient used in the SDE at the beginning of each step. Restricted to the position at times that are multiples of η, equivalently:

$$\mathbf{x}_{(i+1)\eta} = \mathbf{x}_{i\eta} - \eta\nabla f(\mathbf{x}_{i\eta}) + \xi_i.$$

Where $\xi_i \sim N(0, 2\eta\mathbb{I}_d)$ are independent samples. Throughout the chapter, when we refer to the result of running a Langevin dynamics for *continuous time* t, we mean the distribution $\mathbf{x}_t$, *not* the distribution $\mathbf{x}_{t\eta}$. When the iteration complexity (i.e. number of steps) is of interest, we may refer to running a Langevin dynamics for continuous time $T\eta$ equivalently as the result of running it for T steps (of size η).

A similarly defined second order process is the underdamped Langevin dynamics, given by the following SDE (parameterized by $\gamma, \mu > 0$):

$$\mathrm{d}\mathbf{v}_t = -\gamma\mathbf{v}_t\mathrm{d}t - \mu\nabla f(\mathbf{x}_t)\mathrm{d}t + \sqrt{2\gamma\mu}\mathrm{d}B_t, \qquad \mathrm{d}\mathbf{x}_t = \mathbf{v}_t\mathrm{d}t.$$

Again, under mild assumptions it is known that the stationary distribution of this SDE is the distribution p satisfying $p(\mathbf{x}) \propto e^{-(f(\mathbf{x})+||v||_2^2/2\mu)}$, so that the marginal on $\mathbf{x}$ is as desired. Algorithmically, it is easier to use the following discretization:

$$\mathrm{d}\mathbf{v}_t = -\gamma\mathbf{v}_t\mathrm{d}t - \mu\nabla f(\mathbf{x}_{\lfloor \frac{t}{\eta} \rfloor \eta})\mathrm{d}t + \sqrt{2\gamma\mu}\mathrm{d}B_t, \qquad \mathrm{d}\mathbf{x}_t = \mathbf{v}_t\mathrm{d}t. \tag{3.1}$$

In the majority of the chapter we consider sampling from distributions given by m-strongly convex, M-smooth functions f. To simplify the presentation, we also assume f is twice-differentiable, so these conditions on f can be expressed as: for all $\mathbf{x}$, $m\mathbb{I}_k \preccurlyeq \nabla^2 f(\mathbf{x}) \preccurlyeq M\mathbb{I}_k$. We make two additional simplifying assumptions: The first is that the minimum point of f is at $\mathbf{0}$, as if f's true minimum is $\mathbf{x}^* \neq \mathbf{0}$, we can sample from $g(\mathbf{x}) := f(\mathbf{x} - \mathbf{x}^*)$ and then shift our sample by $\mathbf{x}^*$ to get a sample from f instead ($\mathbf{x}^*$ can be found using e.g. gradient descent). The second is that $m = 1$, as if $m \neq 1$, we can sample from $g(\mathbf{x}) = f(\frac{1}{\sqrt{m}}\mathbf{x})$ and rescale our sample by $\sqrt{m}$ instead.

Rényi-divergences

We state here additional facts about Rényi-divergences needed in this chapter that are not covered in Chapter 1.

Definition 43 (Negative Rényi-divergence). *The definition of α-Rényi-divergence can be extended to negative α using the identity $R_{1-\alpha}(P||Q) = \frac{1-\alpha}{\alpha} R_\alpha(Q||P)$.*

Fact 44 (Monotonicity, Theorem 3 of [36]). *For any distributions P, Q and $\alpha_1 \leq \alpha_2$ we have $R_{\alpha_1}(P||Q) \leq R_{\alpha_2}(P||Q)$.*

Fact 45 (Post-Processing, Theorem 9 of [36]). *For any sample spaces $\mathcal{X}, \mathcal{Y}$, distributions X_1, X_2 over $\mathcal{X}$, and any function $f : \mathcal{X} \to \mathcal{Y}$ we have $R_\alpha(f(X_1)||f(X_2)) \leq R_\alpha(X_1||X_2)$.*

Fact 46 (Weak Triangle Inequality, Proposition 11 of [66]). *For any $\alpha > 1$, $p, q > 1$ satisfying $1/p + 1/q = 1$ and distributions P, Q, R with the same support:*

$$R_\alpha(P||R) \leq \frac{\alpha - 1/p}{\alpha - 1} R_{p\alpha}(P||Q) + R_{q(\alpha-1/p)}(Q||R).$$

3.5 Langevin Dynamics with Bounded Movements

As a first step, we analyze the divergence between the discrete and continuous processes conditioned on the event $\mathcal{E}_r$ that throughout each step of size η they stay within a ball of radius r around their location at the start of the step. We will actually analyze the divergence between two discrete processes with steps of size η and η/c respectively, and obtain a bound on their divergence independent of c. The former is exactly the discrete Langevin dynamics with step size η. Then taking the limit of the latter, as c goes to infinity, the former is exactly the discrete Langevin dynamics with step size η and the latter is the continuous Langevin dynamics. Thus, and so the same bound applies to the divergence between the discrete and continuous processes. We set up discretized overdamped Langevin dynamics with step sizes $\eta, \eta/c$ as random processes which record the position at each time that is a multiple of η/c.

Let $\mathbf{x}_t$ denote the position of the chain using step size η at continuous time t, and $\mathbf{x}'_t$ denote the position of the chain using step size η/c at time t. If $\mathcal{E}_r$ does not hold at time t^*

(more formally, if $\max_{t\in[0,t^*]} \left|\left|\mathbf{x}_t - \mathbf{x}_{\lfloor t/\eta \rfloor \eta}\right|\right|_2 > r$), we will instead let $\mathbf{x}_t = \perp$ for all $t \geq t^*$. We want to bound the divergence after T steps of size η, i.e. the divergence between the distributions of $\mathbf{x}_{T\eta}$ and $\mathbf{x}'_{T\eta}$. Let $X_{0:j}$ denote the distribution of $\{\mathbf{x}_{i\eta/c}\}_{0\leq i\leq j}$, and define $X'_{0:j}$ analogously. By post-processing , it suffices to bound the divergence between $X_{0:Tc}$ and $X'_{0:Tc}$. Note that we can sample from $X_{0:Tc}$ (resp $X'_{0:Tc}$) by starting with a sample $\{\mathbf{x}_0\}$ (resp $\{\mathbf{x}'_0\}$) from the distribution X_0 from which we start the Langevin dynamics, and applying the following randomized update Tc times:

- To draw a sample from $X_{0:Tc}$, given a sample $\{\mathbf{x}_{i\eta/c}\}_{0\leq i\leq j}$ from $X_{0:j}$:
  - If $\mathbf{x}_{j\eta/c} = \perp$ append $\mathbf{x}_{(j+1)\eta/c} = \perp$ to $\{\mathbf{x}_{i\eta/c}\}_{0\leq i\leq j}$ to get a sample from $X_{0:j+1}$.
  - Otherwise, append $\mathbf{x}_{(j+1)\eta/c} = \mathbf{x}_{j\eta/c} - \frac{\eta}{c}\nabla f(\mathbf{x}_{\lfloor j/c \rfloor \eta}) + \xi_j$, where $\xi_j \sim N(\mathbf{0}, \frac{2\eta}{c} I_d)$ to get a sample from $X_{0:j+1}$. Then if $\left|\left|\mathbf{x}_{(j+1)\eta/c} - \mathbf{x}_{\lfloor (j+1)/c \rfloor \eta}\right|\right|_2 > r$ (i.e. $\mathcal{E}_r$ no longer holds) replace $\mathbf{x}_{(j+1)\eta/c}$ with $\perp$.

  We will denote this update by ψ. More formally, ψ is the map from distributions over to distributions such that $X_{0:j+1} = \psi(X_{0:j})$.

- To draw a sample from $X'_{0:Tc}$, we instead use the update ψ' that is identical to ψ except ψ' uses the gradient at $\mathbf{x}'_{j\eta/c}$ instead of $\mathbf{x}'_{\lfloor j/c \rfloor \eta}$.

We now have $X_{0:Tc} = \psi^{\circ Tc}(X_0)$ and $X'_{0:Tc} = (\psi')^{\circ Tc}(X_0)$, allowing us to use Theorem 16 to bound the divergence between the two distributions:

Lemma 47. *For any M-smooth f, any initial distribution X_0 over $\mathbf{x}_0, \mathbf{x}'_0$, and the distributions over tuples $X_{0:Tc}, X'_{0:Tc}$ as defined above, we have:*

$$R_\alpha(X_{0:Tc}||X'_{0:Tc}), R_\alpha(X'_{0:Tc}||X_{0:Tc}) \leq \frac{T\alpha M^2 r^2 \eta}{4}.$$

Proof. We prove the bound for $R_\alpha(X_{0:Tc}||X'_{0:Tc})$, the bound for $R_\alpha(X'_{0:Tc}||X_{0:Tc})$ follows similarly. Let a tuple $\{\mathbf{x}_{i\eta/c}\}_{0\leq i\leq j}$ be *good* if for $0 \leq i \leq j$ either (i) $\left|\left|\mathbf{x}_{i\eta/c} - \mathbf{x}_{\lfloor i/c \rfloor \eta}\right|\right|_2 \leq r$ (i.e., $\mathcal{E}_r$) or (ii) $\{\mathbf{x}_{\ell\eta/c}\}_{i\leq \ell\leq j}$ are all $\perp$. We claim that for each j, for any point mass distribution $X_{0:j}$ over good $(j+1)$-tuples:

$$R_\alpha(\psi(X_{0:j}), \psi'(X_{0:j})) \leq \frac{\alpha(\frac{Mr\eta}{c})^2}{2\cdot\frac{2\eta}{c}}. \tag{3.2}$$

By Fact 45, we can instead bound the divergence between $\tilde{\psi}(X_{0:j}), \tilde{\psi}'(X_{0:j})$ which are defined equivalently to ψ, ψ' except without the deterministic step of replacing the last entry with $\perp$ if $\mathcal{E}_r$ is violated. If $X_{0:j}$ is a point mass on a good tuple containing $\perp$, then $R_\alpha(\tilde{\psi}(X_{0:j})||\tilde{\psi}'(X_{0:j})) = 0$. For $X_{0:j}$ that is a point mass on a good tuple not containing $\perp$, $R_\alpha(\tilde{\psi}(X_{0:j})||\tilde{\psi}'(X_{0:j}))$ is just the divergence between the final values of $\tilde{\psi}(X_{0:j}), \tilde{\psi}'(X_{0:j})$. The distance between the final values in $\tilde{\psi}(X_{0:j}), \tilde{\psi}'(X_{0:j})$ prior to the addition of Gaussian noise in $\tilde{\psi}, \tilde{\psi}'$ is the value of $\frac{\eta}{c}\left|\left|\nabla f(\mathbf{x}_{j\eta/c}) - \nabla f(\mathbf{x}_{\lfloor j/c \rfloor \eta})\right|\right|_2$ for the single tuple in the support of $X_{0:j}$, which is at most $\frac{Mr\eta}{c}$ by smoothness and because $\mathcal{E}_r$ holds for all good tuples not containing $\perp$. (3.2) now follows by Fact 15.

Then, $X_{0:Tc}, X'_{0:Tc}$ are arrived at by a composition of Tc applications of ψ, ψ' to the same initial distribution X_0. Note that X_0 and the distributions arrived at by applying ψ or ψ' any number of times to X_0 have support only including good tuples. Then combining Theorem 16 (with the sample spaces being good tuples) and (3.2) we have:

$$R_\alpha(X_{0:Tc}||X'_{0:Tc}) \leq Tc \cdot \frac{\alpha\left(\frac{Mr\eta}{c}\right)^2}{2 \cdot \frac{2\eta}{c}} = \frac{T\alpha M^2 r^2 \eta}{4}.$$

□

By taking the limit as c goes to infinity and applying Fact 45 we get:

Corollary 48. *For any M-smooth f and $\eta > 0$, and any initial distribution X_0 let X_t be the distribution over positions x_t arrived at by running the discretized overdamped Langevin dynamics with step size η on f from X_0 for continuous time t, except that $X_t = \perp$ if $\mathcal{E}_r$ does not hold at time t for this chain. Let X'_t be the same but for the continuous overdamped Langevin dynamics. Then for any integer $T \geq 0$:*

$$R_\alpha(X_{T\eta}||X'_{T\eta}), R_\alpha(X'_{T\eta}||X_{T\eta}) \leq \frac{T\alpha M^2 r^2 \eta}{4}.$$

Note that if we are running the process for continuous time τ, then $T = \tau/\eta$. r will end up being roughly proportional to $\sqrt{\eta}$, so the above bound is then roughly proportional to η.

3.6 Removing the Bounded Movement Restriction

In this section, we will prove the following "one-sided" version of Theorem 42:

Theorem 49. *Fix any $\alpha \geq 1$. Let S be a distribution satisfying $S(\mathbf{x}) \propto e^{-f(\mathbf{x})}$ for 1-strongly convex and M-smooth f with global minimum at 0. Let P be the distribution arrived at by running discretized overdamped Langevin dynamics using f with step size $\eta = \tilde{O}(\frac{1}{\tau M^4 \ln^2 \alpha} \cdot \frac{\epsilon^2}{k})$ for continuous time $\tau = \alpha \ln \frac{k \ln M}{\epsilon}$ (i.e. for $\tilde{O}(\frac{\alpha^2 M^4 k}{\epsilon^2})$ steps) from initial distribution $N(\mathbf{0}, \frac{1}{M}\mathbb{I}_k)$. Then we have $R_\alpha(P||S) \leq \epsilon$.*

To remove the assumption that the process never moves more than r away from its original position within each step of size η, we give a tail bound on the maximum value r that the process moves within one of these steps.

Lemma 50. *Let c be a sufficiently large constant. Let $\eta \leq \frac{2}{M+1}$ and let X_0 be an initial distribution over $\mathbb{R}^k$ satisfying that for all $\delta > 0$,*

$$\Pr_{\mathbf{x} \sim X_0}\left[||\mathbf{x}||_2 \leq \frac{c}{2\sqrt{\eta}} \left(\sqrt{k} + \sqrt{\ln(T/\delta)} \right) \right] \geq 1 - \frac{\delta}{4(T+1)}. \tag{3.3}$$

Let $\mathbf{x}_t$ be the random variable given by running the discretized overdamped Langevin dynamics starting from X_0 for continuous time t. Then with probability at least $1-\delta$ over the path $\{\mathbf{x}_t : t \in [0, T\eta]\}$:

$$\forall t \leq T\eta : \left|\left|\mathbf{x}_t - \mathbf{x}_{\lfloor t/\eta \rfloor \eta}\right|\right|_2 \leq cM \left(\sqrt{k} + \sqrt{\ln(T/\delta)} \right) \sqrt{\eta}.$$

Similarly, let $\mathbf{x}_t'$ be the random variable given by running the continuous overdamped Langevin dynamics starting from X_0 for continuous time t. Then with probability at least $1-\delta$ over the path $\{\mathbf{x}_t' : t \in [0, T\eta]\}$:

$$\forall t \leq T\eta : \left|\left|\mathbf{x}_t' - \mathbf{x}_{\lfloor t/\eta \rfloor \eta}'\right|\right|_2 \leq cM \left(\sqrt{k} + \sqrt{\ln(T/\delta)} \right) \sqrt{\eta}.$$

The proof is deferred to Section 3.9. Intuitively, the $\sqrt{\eta}$ accounts for movement due to Brownian motion, which dominates the movement due to the gradient, and $cM(\sqrt{k} + \sqrt{\ln(T/\delta)})$ is a tail bound on norm of the gradient by smoothness. This gives us a bound on the Rényi-divergence between the continuous and discrete processes conditioned on a probability $1-\delta$ event for all $0 < \delta < 1$. By absorbing the failure probability of this event into the probability of large privacy loss in the definition of (ϵ, δ)-differential privacy we can prove iteration complexity bounds matching those in Figure 3.1 for running discretized overdamped Langevin dynamics with (ϵ, δ)-differential privacy without using the tools we develop in the rest of this section. Since these bounds do not improve on those in the ones derived from our final (unconditional) divergence bounds, we omit the proof here.

To prove a Rényi divergence bound, we need to remove the conditioning. We start with the following lemma, which takes bounds on conditional moments and gives an unconditional bound on expectation:

Lemma 51. *Let Y be a random variable distributed over $\mathbb{R}_{\geq 0}$ that has the following property (parameterized by positive parameters $\beta, \gamma < 1, \theta > 1 + \gamma$): For every $0 < \delta < 1/2$, there is a probability at least $1-\delta$ event $\mathcal{E}_\delta$ such that $\mathbb{E}\left[Y^\theta | \mathcal{E}_\delta\right] \leq \frac{\beta}{\delta^\gamma}$. Then we have:*

$$\mathbb{E}[Y] \leq \beta^{\frac{1}{\theta}} \left(\gamma^{\frac{1}{1+\gamma}} + \gamma^{-\frac{\gamma}{1+\gamma}} \right)^{\frac{1+\gamma}{\theta}} \left(\frac{\theta(1+\gamma)}{\theta(1+\gamma) - 1} \right) \leq \beta^{1/\theta} 2^{2/\theta} \frac{\theta}{\theta - 1}.$$

Proof. Let z be an arbitrary parameter, $\eta : [z, \infty) \to (0, 1/2)$ be an arbitrary map, and $\mathcal{E}_\delta$ be the event specified in the lemma statement for $\delta \in (0,1)$. Using the definition of expectation and the property of Y in the lemma statement, we have:

$$\begin{aligned}
\mathbb{E}[Y] &= \int_0^\infty \Pr[Y \geq y]\mathrm{d}y \\
&\leq \int_0^z 1 \ \mathrm{d}y + \int_z^\infty \Pr[Y \geq y]\mathrm{d}y \\
&\leq z + \int_z^\infty \eta(y) + (1-\eta(y)) \Pr[Y \geq y | \mathcal{E}_{\eta(y)}]\mathrm{d}y \\
&\leq z + \int_z^\infty \eta(y) + \Pr[Y \geq y | \mathcal{E}_{\eta(y)}]\mathrm{d}y \\
&= z + \int_z^\infty \eta(y) + \Pr[Y^\theta \geq y^\theta | \mathcal{E}_{\eta(y)}]\mathrm{d}y \\
&\leq z + \int_z^\infty \eta(y) + \frac{\mathbb{E}[Y^\theta | \mathcal{E}_{\eta(y)}]}{y^\theta}\mathrm{d}y \\
&\leq z + \int_z^\infty \eta(y) + \frac{\beta}{\eta(y)^\gamma y^\theta}\mathrm{d}y.
\end{aligned}$$

We now choose $\eta(y) = \left(\frac{\gamma\beta}{y^\theta}\right)^{\frac{1}{1+\gamma}}$ to minimize the value of the expression in the integral. We will eventually choose z such that $0 < \eta(y) < 1/2$ for all $y \geq z$ as promised. Plugging in this choice of η gives the upper bound:

$$\begin{aligned}
\mathbb{E}[Y] &\leq z + \beta^{\frac{1}{1+\gamma}}(\gamma^{\frac{1}{1+\gamma}} + \gamma^{-\frac{\gamma}{1+\gamma}}) \int_z^\infty y^{-\frac{\theta}{1+\gamma}}\mathrm{d}y \\
&= z + \beta^{\frac{1}{1+\gamma}}(\gamma^{\frac{1}{1+\gamma}} + \gamma^{-\frac{\gamma}{1+\gamma}}) \left(\frac{1}{\frac{\theta}{1+\gamma} - 1}\right) \left[y^{1-\frac{\theta}{1+\gamma}}\right]_\infty^z \\
&= z + \beta^{\frac{1}{1+\gamma}}(\gamma^{\frac{1}{1+\gamma}} + \gamma^{-\frac{\gamma}{1+\gamma}}) \left(\frac{1}{\frac{\theta}{1+\gamma} - 1}\right) z^{1-\frac{\theta}{1+\gamma}}.
\end{aligned}$$

We finish by choosing $z = \beta^{\frac{1}{\theta}}\left(\gamma^{\frac{1}{1+\gamma}} + \gamma^{-\frac{\gamma}{1+\gamma}}\right)^{\frac{1+\gamma}{\theta}}$. This gives the upper bound on $\mathbb{E}[Y]$ in the lemma statement. We also verify that $\eta(y)$ is a map to $(0, 1/2)$: $\eta(y) \propto y^{-\frac{\theta}{1+\gamma}}$, giving that $\eta(y) > 0$. For all $y \geq z$, since $\gamma < 1$ we have $\eta(y) \leq \eta(z) = \frac{\gamma}{\gamma+1} < 1/2$. □

Putting it all together, we get the following theorem:

Theorem 52. *For any* 1*-strongly convex,* M*-smooth* f*, let* P *be the distribution of states for discretized overdamped Langevin dynamics with step size* η *and* Q *be the distribution of states for continuous overdamped Langevin dynamics, both run from any initial distribution* X_0 *satisfying* (3.3) *for continuous time* τ *that is a multiple of* η *(i.e. for* τ/η *steps). Then for* $\alpha > 1$*,* $\epsilon > 0$*, if* $\eta = \tilde{O}(\frac{1}{\tau M^4 \ln^2 \alpha} \cdot \frac{\epsilon^2}{k})$ *we have* $R_\alpha(P||Q), R_\alpha(Q||P) \leq \epsilon$*.*

We provide some high level intuition for the proof here. Plugging Lemma 50 into Lemma 47 gives a bound on roughly the α'-Rényi divergence between P conditioned on some probability $1-\delta_1$ event and Q conditioned on some probability $1-\delta_2$ event for every δ_1, δ_2. We apply Lemma 51 once for P and once for Q to remove the conditioning, giving a bound of $\approx \frac{\ln \alpha'}{\alpha'-1}$ on the actual α'-Rényi divergence between P, Q if η is sufficiently small (as a function of α'). Using Jensen's inequality, we can turn this into a bound of ϵ on the α-Rényi divergence between P, Q for any α if α' is large enough (which in turn requires η to be small enough).

Proof of Theorem 52. We prove the bound on $R_\alpha(P||Q)$. Since Corollary 48 provides a "bi-directional" divergence bound, the same proof can be used to bound $R_\alpha(Q||P)$.

For arbitrary δ_1, δ_2, plugging in $r = cM(\sqrt{k} + \sqrt{\ln(T/\delta_1)} + \sqrt{\ln(T/\delta_2)})\sqrt{\eta}$ into Corollary 48 (where c is the constant specified in Lemma 50) and using the definition $T = \tau/\eta$ we get that

$$R_{\alpha'}(X_{T\eta}||X'_{T\eta}) \leq \frac{3\tau\alpha' M^4 c^2(k + \ln(\frac{\tau}{\eta\delta_1}) + \ln(\frac{\tau}{\eta\delta_2}))\eta}{4}$$

for $X_{T\eta}, X'_{T\eta}$ as defined in Corollary 48. Using the definition of Rényi divergence, this gives:

$$\int_{\mathbb{R}^k} \frac{X_{T\eta}(\mathbf{x})^{\alpha'}}{X'_{T\eta}(\mathbf{x})^{\alpha'-1}}\mathrm{d}x \leq \int_{\mathbb{R}^k} \frac{X_{T\eta}(\mathbf{x})^{\alpha'}}{X'_{T\eta}(\mathbf{x})^{\alpha'-1}}\mathrm{d}x + \frac{\Pr_{x\sim X_{T\eta}}[x=\perp]^{\alpha'}}{\Pr_{x\sim X'_{T\eta}}[x=\perp]^{\alpha'-1}} \leq \frac{c_1(\alpha')}{{\delta_1}^{c_2(\alpha')}\delta_2^{c_3(\alpha')}},$$

where:

$$c_1(\alpha') = \exp\left(\frac{3\tau\alpha'(\alpha'-1)M^4c^2(k+2\ln(\frac{\tau}{\eta}))\eta}{4}\right),$$

$$c_2(\alpha') = c_3(\alpha') = \frac{3\tau\alpha'(\alpha'-1)M^4c^2\eta}{4}.$$

Removing the conditioning on the continuous chain: Let $\mathcal{E}_{\delta_1}$ denote the (at least probability $1-\delta_1$) event that the conditions in Lemma 50 are satisfied for the discrete chain and $\mathcal{E}_{\delta_2}$ denote the (at least probability $1-\delta_2$) event that the conditions in Lemma 50 are satisfied for the continuous chain. By Lemma 50, we have $Q(\mathbf{x}) \geq X'_{T\eta}(\mathbf{x}), Q(x|\mathcal{E}_{\delta_2}) \leq \frac{1}{1-\delta_2}X'_{T\eta}(\mathbf{x})$. Then for $\delta_2 < 1/2$:

$$\begin{aligned}\mathbb{E}_{x\sim Q}\left[\frac{X_{T\eta}(\mathbf{x})^{\alpha'}}{Q(\mathbf{x})^{\alpha'}}\middle|\mathcal{E}_{\delta_2}\right] &= \int_{\mathbb{R}^k} Q(x|\mathcal{E}_{\delta_2})\frac{X_{T\eta}(\mathbf{x})^{\alpha'}}{Q(\mathbf{x})^{\alpha'}}\mathrm{d}x \\ &\leq \frac{1}{1-\delta_2}\int_{\mathbb{R}^k}\frac{X_{T\eta}(\mathbf{x})^{\alpha'}}{X'_{T\eta}(\mathbf{x})^{\alpha'-1}}\mathrm{d}x \\ &\leq \frac{2\cdot c_1(\alpha')}{{\delta_1}^{c_2(\alpha')}\delta_2^{c_3(\alpha')}}.\end{aligned}$$

This statement holds independent of δ_2. We will eventually choose α' such that for the choice of η specified in the lemma statement, $c_1(\alpha') < 2, c_3(\alpha') < 1$. Then applying Lemma 51 with $Y = \frac{X_{T\eta}(\mathbf{x})^{\alpha'/2}}{Q(\mathbf{x})^{\alpha'/2}}$ $\theta = 2$, $\beta = \frac{2c_1(\alpha')}{\delta_1^{c_2(\alpha')}}$, $\gamma = c_3(\alpha')$, we get:

$$\mathbb{E}_{x\sim Q}\left[\frac{X_{T\eta}(\mathbf{x})^{\alpha'/2}}{Q(\mathbf{x})^{\alpha'/2}}\right] \leq \frac{8}{\delta_1^{c_2(\alpha')/2}}.$$

Removing the conditioning on the discrete chain: We now turn to removing the conditioning on $\mathcal{E}_{\delta_1}$. Here we need to be a bit more careful since unlike with $X'_{T\eta}(\mathbf{x})$, $X_{T\eta}(\mathbf{x})$ is in the numerator and so the inequality $X_{T\eta}(\mathbf{x}) \leq P(\mathbf{x})$ is facing the wrong way. Since P, Q have the same support, we note that:

$$\begin{aligned}
\mathbb{E}_{x\sim Q}\left[\frac{X_{T\eta}(\mathbf{x})^{\alpha'/2}}{Q(\mathbf{x})^{\alpha'/2}}\right] &= \mathbb{E}_{x\sim P}\left[\frac{X_{T\eta}(\mathbf{x})^{\alpha'/2-1}}{Q(\mathbf{x})^{\alpha'/2-1}} \cdot \frac{X_{T\eta}(\mathbf{x})}{P(\mathbf{x})}\right] \\
&\overset{(\star)}{=} \frac{\alpha'}{2}\mathbb{E}_{x\sim P, y\sim Unif(0,P(\mathbf{x}))}\left[\frac{y^{\alpha'/2-1}}{Q(\mathbf{x})^{\alpha'/2-1}} \cdot \mathbb{I}\left[y \leq X_{T\eta}(\mathbf{x})\right]\right] \\
&= \frac{\alpha'}{2}\mathbb{E}_{x\sim P, y\sim Unif(0,P(\mathbf{x}))}\left[\frac{y^{\alpha'/2-1}}{Q(\mathbf{x})^{\alpha'/2-1}}\middle| y \leq X_{T\eta}(\mathbf{x})\right] \\
&\qquad \cdot \Pr_{x\sim P, y\sim Unif(0,P(\mathbf{x}))}\left[y \leq X_{T\eta}(\mathbf{x})\right] \\
&= \frac{\alpha'}{2}\mathbb{E}_{x\sim P, y\sim Unif(0,P(\mathbf{x}))}\left[\frac{y^{\alpha'/2-1}}{Q(\mathbf{x})^{\alpha'/2-1}}\middle| \mathcal{E}_{\delta_1}\right] \cdot (1-\delta_1).
\end{aligned}$$

$(\star)$ follows as for any given x, we have:

$$\begin{aligned}
X_{T\eta}(\mathbf{x})^{\alpha'/2-1} &= \frac{1}{X_{T\eta}(\mathbf{x})} X_{T\eta}(\mathbf{x})^{\alpha'/2} \\
&= \int_0^{X_{T\eta}(\mathbf{x})} \frac{1}{X_{T\eta}(\mathbf{x})} \frac{\alpha'}{2} y^{\alpha'/2-1} \mathrm{d}y \\
&= \frac{P(\mathbf{x})}{X_{T\eta}(\mathbf{x})} \int_0^{X_{T\eta}(\mathbf{x})} \frac{1}{P(\mathbf{x})} \frac{\alpha'}{2} y^{\alpha'/2-1} \mathrm{d}y \\
&= \frac{P(\mathbf{x})}{X_{T\eta}(\mathbf{x})} \int_0^{P(\mathbf{x})} \frac{1}{P(\mathbf{x})} \frac{\alpha'}{2} y^{\alpha'/2-1} \cdot \mathbb{I}\left[y \leq X_{T\eta}(\mathbf{x})\right] \mathrm{d}y \\
&= \frac{P(\mathbf{x})}{X_{T\eta}(\mathbf{x})} \frac{\alpha'}{2} \mathbb{E}_{y\sim Unif(0,P(\mathbf{x}))}\left[y^{\alpha'/2-1} \cdot \mathbb{I}\left[y \leq X_{T\eta}(\mathbf{x})\right]\right].
\end{aligned}$$

In turn, for all $\delta_1 < 1/2$, we have

$$\mathbb{E}_{x\sim P, y\sim Unif(0,P(\mathbf{x}))}\left[\frac{y^{\alpha'/2-1}}{Q(\mathbf{x})^{\alpha'/2-1}}\middle| \mathcal{E}_{\delta_1}\right] \leq \frac{32}{\alpha' \delta_1^{c_2(\alpha')/2}}.$$

If $c_2(\alpha')/2 < 1/2$ (which is equivalent to $c_2(\alpha') = c_3(\alpha') < 1$), by applying Lemma 51 for $\theta = 2$ with $X = \frac{y^{\alpha'/4-1/2}}{Q(\mathbf{x})^{\alpha'/4-1/2}}, \beta = \frac{32}{\alpha'}, \gamma = c_2(\alpha')/2$ we get:

$$\mathbb{E}_{x\sim P, y\sim Unif(0,P(\mathbf{x}))}\left[\frac{y^{\alpha'/4-1/2}}{Q(\mathbf{x})^{\alpha'/4-1/2}}\right] \leq \frac{19}{\sqrt{\alpha'}} \implies$$

$$\begin{aligned}\mathbb{E}_{x\sim Q}\left[\frac{P(\mathbf{x})^{\alpha'/4+1/2}}{Q(\mathbf{x})^{\alpha'/4+1/2}}\right] &= \left(\frac{\alpha'}{4}+\frac{1}{2}\right)\mathbb{E}_{x\sim P, y\sim Unif(0,P(\mathbf{x}))}\left[\frac{y^{\alpha'/4-1/2}}{Q(\mathbf{x})^{\alpha'/4-1/2}}\right] \\ &\leq \frac{19(\alpha'/4+1/2)}{\sqrt{\alpha'}} \\ &\leq 15\sqrt{\alpha'}.\end{aligned}$$

From moderate α'-Rényi divergence to small α-Rényi divergence: If $\epsilon \geq \frac{3\ln\alpha}{\alpha-1}$, without loss of generality we can assume e.g. $\alpha \geq 4$ (by monotonocity of Rényi divergences, if $\alpha < 4$ it suffices to bound the 4-Rényi divergence instead of the α-Rényi divergence at the loss of a constant in the bound for η). Then for $\alpha' = 4\alpha - 2$ the preceding inequality lets us conclude the lemma holds. Otherwise, for $1 < \kappa < \alpha'/4 + 1/2$, for $\alpha = \frac{\alpha'/4+1/2}{\kappa}$, by Jensen's inequality we get:

$$\frac{1}{\alpha-1}\ln\mathbb{E}_{x\sim Q}\left[\frac{P(\mathbf{x})^{\alpha}}{Q(\mathbf{x})^{\alpha}}\right] \leq \frac{1}{\alpha-1}\ln\left(\mathbb{E}_{x\sim Q}\left[\frac{P(\mathbf{x})^{\alpha\kappa}}{Q(\mathbf{x})^{\alpha\kappa}}\right]^{1/\kappa}\right) \leq \frac{\ln 15 + \frac{1}{2}\ln\alpha + \frac{1}{2}\ln\kappa}{(\alpha-1)\kappa}.$$

Choosing $\kappa = \frac{3\ln\alpha\cdot\ln 1/\epsilon}{(\alpha-1)\epsilon}$ then gives $R_\alpha(P||Q) \leq \epsilon$ as desired (note that for $\epsilon < \frac{3\ln\alpha}{\alpha-1}$ we have $\kappa > 1$ as is required). Now, we just need to verify that $c_1(\alpha') < 2, c_2(\alpha') = c_3(\alpha') < 1$ holds for $\alpha' = \frac{12\alpha\ln\alpha\cdot\ln 1/\epsilon}{(\alpha-1)\epsilon} - 2$. Since $c_2(\alpha') = c_3(\alpha') < \ln(c_1(\alpha'))/k$, it just suffices to show $c_1(\alpha') < 2$. This holds if:

$$\frac{3\tau\alpha'(\alpha'-1)M^4c^2(k+2\ln(\frac{\tau}{\eta}))\eta}{4} < \ln 2,$$

which is given by choosing $\eta = \tilde{O}(\frac{1}{\tau M^4\ln^2\alpha}\cdot\frac{\epsilon^2}{k})$ with a sufficiently small constant hidden in $\tilde{O}$. □

We now apply results from [82] and the weak triangle inequality for Rényi divergence to get a bound on the number of iterations of discrete overdamped Langevin dynamics needed to achieve α-Rényi divergence ϵ:

Lemma 53. *If $R(\mathbf{x}) = e^{-f(\mathbf{x})}$ is a probability distribution over $\mathbb{R}^k$ with mode $\mathbf{0}$ and f is 1-strongly convex and L-smooth, then for all $\alpha \geq 1$ we have:*

$$R_\alpha\left(N\left(\mathbf{0}, \frac{1}{L}\mathbb{I}_k\right)||S\right) \leq \frac{k}{2}\ln M.$$

Proof. This follows from Lemma 4 in [82], which gives the bound $R_\alpha(N(\mathbf{0}, \frac{1}{M}\mathbb{I}_k)||S) \leq f(\mathbf{0}) + \frac{d}{2}\ln\frac{M}{2\pi}$. We then note that the 1-strongly convex, M-smooth f with the maximum $f(\mathbf{0})$ is given when R is $N(\mathbf{0}, \mathbb{I}_k)$, which has density $R(\mathbf{x}) = e^{-\left(\frac{k}{2}\ln(2\pi)+\frac{1}{2}\mathbf{x}^\top\mathbf{x}\right)}$. □

It is well-known that 1-strong convexity of f implies that $S \propto e^{-f}$ satisfies log-Sobolev inequality with constant 1 (see e.g. [10]). We then get:

Lemma 54 (Theorem 2, [82])**.** *Fix any f that is 1-strongly convex. Let Q_t be the distribution arrived at by running overdamped Langevin dynamics using f for continuous time t from initial distribution Q_0. Then for the distribution S satisfying $S(\mathbf{x}) \propto e^{-f(\mathbf{x})}$ and any $\alpha \geq 1$:*

$$R_\alpha(Q_t||S) \leq e^{-2t/\alpha} R_\alpha(Q_0||S).$$

Proof of Theorem 49. We will prove the bound for $\alpha \geq 3/2$ - the bound for $1 \leq \alpha < 3/2$ follows by just applying monotonicity to the bound for $\alpha = 3/2$, at the loss of a multiplicative constant on τ, η, and the iteration complexity.

Let R be the distribution arrived at by running continuous overdamped Langevin dynamics using f for time τ from initial distribution $N(\mathbf{0}, \frac{1}{M}\mathbb{I}_k)$. $N(\mathbf{0}, \frac{1}{M}\mathbb{I}_k)$ satisfies (3.3), so from Theorem 52 we have $R_{2\alpha}(P||Q) \leq \epsilon/3$. From Lemmas 53 and 54 we have $R_{2\alpha}(Q||S) \leq \epsilon/3$. Then, we use the weak triangle inequality of Rényi divergence (Fact 46) with $p, q = 2$ to conclude that $R_\alpha(P||S) \leq \epsilon$. □

With only a minor modification to the analysis of the strongly convex and smooth case, we can also give a discretization error bound when f is L-Lipschitz instead of strongly convex (while still M-smooth). We have the following radius tail bound analogous to Lemma 50:

Lemma 55. *For all $\eta \leq 1$ and any L-Lipschitz, M-smooth f, let $\mathbf{x}_t$ be the random variable given by running the discretized overdamped Langevin dynamics starting from an arbitrary initial distribution for continuous time t. Then with probability $1-\delta$ over $\{\mathbf{x}_t : t \in [0, T\eta]\}$, for all $t \leq T\eta$ and for a sufficiently large constant c:*

$$\left|\left|\mathbf{x}_t - \mathbf{x}_{\lfloor t/\eta \rfloor \eta}\right|\right|_2 \leq c(L + \sqrt{k} + \sqrt{\ln(T/\delta)})\sqrt{\eta}).$$

Similarly, if $\mathbf{x}'_t$ is the random variable given by running continuous overdamped Langevin dynamics starting from an arbitrary initial distribution for time t, with probability $1-\delta$ over $\mathbf{x}'_t$ for all $t \leq T\eta$:

$$\left|\left|x'_t - x'_{\lfloor t/\eta \rfloor \eta}\right|\right|_2 \leq c(L + \sqrt{k} + \sqrt{\ln(T/\delta)})\sqrt{\eta}).$$

The proof is deferred to Section 3.9. This gives:

Theorem 56. *For any L-Lipschitz, M-smooth function f, let P be the distribution of states for discretized overdamped Langevin dynamics with step size η and Q be the distribution of states for continuous overdamped Langevin dynamics, both run from arbitrary initial distribution for continuous time τ that is a multiple of η. Then for $\alpha > 1$, $\epsilon > 0$, if $\eta = \tilde{O}(\frac{1}{\tau M^4 \ln^2 \alpha} \cdot \frac{\epsilon^2}{L^2+k})$ we have $D_\alpha(P||Q), D_\alpha(Q||P) \leq \epsilon$.*

The proof of Theorem 56 follows identically to Theorem 52, except using Lemma 55 instead of Lemma 50.

3.7 Making the Bound Bi-Directional

In this section, we show that with slight modifications to the proof of Theorem 49, $R_\alpha(P||S)$ and $R_\alpha(S||P)$ can be simultaneously bounded, proving Theorem 42.

Note that Theorem 52 provides bounds on both $R_\alpha(P||Q)$ and $R_\alpha(Q||P)$ for Q that is the finite time distribution of the continuous chain. So, we just need to show that the following claim holds: for an appropriate choice of initial distribution, $R_\alpha(Q||S), R_\alpha(S||Q)$ are both small after sufficiently many iterations. To show this claim, we use the following results, all of which are slight modifications of the results in [82]. For completeness, we provide the proofs of these claims at the end of the section. We first need a lemma analogous to Lemma 54 to show that $R_\alpha(S||Q)$ decays exponentially:

Lemma 57. *Fix any f that is 1-strongly convex. Let Q_t be the distribution arrived at by running overdamped Langevin dynamics using f for continuous time t from initial distribution Q_0 such that $-\log Q_0$ is 1-strongly convex. Then for the distribution R satisfying $S(\mathbf{x}) \propto e^{-f(\mathbf{x})}$, any $\alpha > 1$, and any t:*

$$R_\alpha(S||Q_t) \leq e^{-t/\alpha} R_\alpha(S||Q_0).$$

This proof follows similarly to Lemma 2 in [82]. If $R_\alpha(S||Q_0)$ and $R_\alpha(Q_0||S)$ were both initially not too large, Lemma 57 along with Lemma 54 would be enough to arrive at Theorem 42. However, for any initial distribution Q_0, there is some S satisfying the conditions of Lemma 57 such that for sufficiently large α one of $R_\alpha(S||Q_0)$ and $R_\alpha(Q_0||S)$ is infinite. The following hypercontractivity property of the Langevin dynamics gives that as long as $R_\alpha(Q_0||S)$ is finite for some small α, it will become finite for larger α after a short amount of time:

Lemma 58 (Lemma 14, [82])**.** *Fix any f that is 1-strongly convex. Let Q_t be the distribution arrived at by running overdamped Langevin dynamics using f for continuous time t from initial distribution Q_0. Fix any $\alpha_0 > 1$, and let $\alpha_t = 1 + e^{2t}(\alpha_0 - 1)$. Then for the distribution R satisfying $S(\mathbf{x}) \propto e^{-f(\mathbf{x})}$:*

$$R_{\alpha_t}(Q_t||S) \leq \frac{1 - 1/\alpha_0}{1 - 1/\alpha_t} R_{\alpha_0}(Q_0||S).$$

Given this lemma, we can now settle for an initial distribution where $R_\alpha(S||Q_0)$ is not too large for all α, and $R_\alpha(Q_0||S)$ is not too large for α slightly larger than 1. Lemma 58 then says that $R_\alpha(Q_0||S)$ will be eventually be not too large after time $O(\log \alpha)$, at which point we can apply Lemmas 54 and 57. We now just need to show that our choice of initial distribution $N(\mathbf{0}, \mathbb{I}_k)$ satisfies these conditions:

Lemma 59. *Let $Q_0 = N(\mathbf{0}, \mathbb{I}_k)$. If $S(\mathbf{x}) = e^{-f(\mathbf{x})}$ is a probability distribution over $\mathbb{R}^k$ with mode $\mathbf{0}$ and f is 1-strongly convex and M-smooth, then for all $\alpha \geq 1$ we have:*

$$R_\alpha(S||Q_0) \leq k \log M.$$

In addition:

$$R_{1+1/L}(Q_0||S) \leq \frac{kM \log M}{2}.$$

Putting it all together, we can now prove Theorem 42.

Proof of Theorem 42. Let Q_t be the distribution of the continuous overdamped Langevin dynamics using f run from initial distribution $N(\mathbf{0}, \mathbb{I}_k)$ for time t. Assume without loss of generality that $\alpha \geq 2$, since if $\alpha \leq 2$ we can use monotonicity of Rényi-divergences to bound e.g. $R_\alpha(P||S)$ by $R_2(P||S)$.

If τ is at least a sufficiently large constant times $\alpha \ln \frac{k \ln M}{\epsilon}$, Lemma 59 and Lemma 57 give that $R_{2\alpha}(S||Q_\tau) \leq \epsilon/3$. Theorem 52 gives that $R_{2\alpha}(Q_\tau||P) \leq \epsilon/3$. Fact 46 with $p, q = 2$ gives that $R_\alpha(S||P) \leq \epsilon$.

Lemma 58 and Lemma 59 give that at time $t = \frac{1}{2}\log((2\alpha-1)M)$, $R_{2\alpha}(Q_t||S) \leq k \log M$. Then Lemma 54 gives that, $R_{2\alpha}(Q_\tau||S) \leq \epsilon/3$. Theorem 52 gives that $R_{2\alpha}(P||Q_\tau) \leq \epsilon/3$. Fact 46 with $p, q = 2$ again gives that $R_\alpha(P||S) \leq \epsilon$. □

Proof of Lemma 57

To prove Lemma 57, we modify the proofs of Lemma 4 and 5 of [82]. To describe the modifications, we reintroduce the following definitions from that paper:

Definition 60. *We say that a distribution Q has LSI constant κ if for all smooth functions $g : \mathbb{R}^n \to \mathbb{R}$ for which $\mathbb{E}_{x\sim Q}[g(\mathbf{x})^2] < \infty$:*

$$\mathbb{E}_{x\sim Q}\left[g(\mathbf{x})^2 \log\left(g(\mathbf{x})^2\right)\right] - \mathbb{E}_{x\sim Q}\left[g(\mathbf{x})^2\right] \log\left(\mathbb{E}_{x\sim Q}\left[g(\mathbf{x})^2\right]\right) \leq \frac{2}{\kappa}\mathbb{E}_{x\sim Q}\left[||\nabla g(\mathbf{x})||^2\right].$$

Definition 61. *We define for $\alpha \neq 0, 1$:*

$$F_\alpha(Q||S) = \mathbb{E}_{x\sim R}\left[\frac{Q(\mathbf{x})^\alpha}{S(\mathbf{x})^\alpha}\right],$$

$$G_\alpha(Q||S) = \mathbb{E}_{x\sim R}\left[\frac{Q(\mathbf{x})^\alpha}{S(\mathbf{x})^\alpha}\left|\left|\nabla \log \frac{Q(\mathbf{x})}{S(\mathbf{x})}\right|\right|_2^2\right] = \frac{4}{\alpha^2}\mathbb{E}_{x\sim R}\left[\left|\left|\nabla\left(\frac{Q(\mathbf{x})}{S(\mathbf{x})}\right)^{\alpha/2}\right|\right|_2^2\right].$$

For $\alpha = 0, 1$ these quantities are defined as their limit as α goes to $0, 1$ respectively.

Unlike [82], we extend this definition to negative values of α, which allows us to swap the arguments Q, R:

Fact 62. *$F_{1-\alpha}(Q||S) = F_\alpha(S||Q), G_{1-\alpha}(Q||S) = G_\alpha(S||Q)$. We also recall that $R_{1-\alpha}(Q||S) = \frac{1-\alpha}{\alpha} R_\alpha(S||Q)$.*

Proof of Lemma 57. [10] shows that since the initial distribution satisfies that $-\log Q_0$ is 1-strongly convex, Q_0 has LSI constant 1. Consider instead running the discrete overdamped Langevin dynamics with step size η starting with Q_0. In one step, we apply a gradient descent step that is $(1-\eta/2)$-Lipschitz (see e.g. Lemma 3.7 of [45]), and then add Gaussian noise $N(0, 2\eta\mathbb{I}_k)$. Lemma 16 in [82] shows that applying a $(1-\eta/2)$-Lipschitz map to a distribution with LSI constant c results in a distribution with LSI constant at least $c/(1-\eta/2)^2$. Adding Gaussian noise $N(0, 2\eta\mathbb{I}_k)$ to a distribution with LSI constant c results in a distribution with LSI constant at least $\frac{1}{1/c+2\eta}$ (see e.g. Proposition 1.1 of [84]). Putting it together, we get that after one step of the discrete dynamics, the LSI constant of the distribution goes from c to at least:

$$\frac{1}{\frac{(1-\eta/2)^2}{c} + 2\eta} = \frac{c}{1-(1-2c)\eta + \eta^2/4}.$$

Then, we have that $1-(1-2c)\eta+\eta^2/4 \leq 1$, i.e. the LSI constant does not decrease after one step, as long as $\eta \leq 4(1-2c)$. Taking the limit as η goes to 0, we conclude that Q_t's LSI constant can never decrease past 1/2, i.e. Q_t has LSI constant at least 1/2 for all $t \geq 0$.

Now, since Q_t has LSI constant at least 1/2, we can repeat the proof of Lemma 5 in [82] with the distributions swapped to show that $\frac{G_\alpha(S||Q_t)}{F_\alpha(S||Q_t)} \geq \frac{1}{\alpha^2} R_\alpha(S||Q_t)$. Applying Fact 62 to the proof of Lemma 6 in [82], we can show that $\frac{\mathrm{d}}{\mathrm{dt}} R_\alpha(S||Q_t) = -\alpha \frac{G_\alpha(S||Q_t)}{F_\alpha(S||Q_t)}$. Combining these two inequalities and integrating gives the lemma. □

Proof of Lemma 59

The proof of Lemma 59 follows similarly to that of Lemma 53.

Proof of Lemma 59. Since f is 1-strongly convex and M-smooth, we have:

$$f(\mathbf{0}) + \frac{1}{2} ||\mathbf{x}||_2^2 \leq f(\mathbf{x}) \leq f(\mathbf{0}) + \frac{M}{2} ||\mathbf{x}||_2^2 .$$

Then:

$$\begin{aligned}\exp((\alpha-1)R_\alpha(S||Q_0)) &= \int_{\mathbb{R}^k} \frac{S(\mathbf{x})^\alpha}{Q_0(\mathbf{x})^{\alpha-1}}\mathrm{d}x \\ &= (2\pi)^{k(\alpha-1)/2}\int_{\mathbb{R}^k} \exp\left(-\alpha f(\mathbf{x}) + \frac{\alpha-1}{2}\left|\left|\mathbf{x}\right|\right|_2^2\right)\mathrm{d}x \\ &\leq \frac{(2\pi)^{k(\alpha-1)/2}}{e^{\alpha f(\mathbf{0})}}\int_{\mathbb{R}^k} \exp\left(-\frac{1}{2}\left|\left|\mathbf{x}\right|\right|_2^2\right)\mathrm{d}x \\ &= \frac{(2\pi)^{k\alpha/2}}{e^{\alpha f(\mathbf{0})}}.\end{aligned}$$

Taking logs and using that the M-smooth f that minimizes $f(\mathbf{0})$ is $N(\mathbf{0}, \frac{1}{M}\mathbb{I}_k)$ with density $\exp(-\frac{k}{2}\log(2\pi/M) - M\left|\left|\mathbf{x}\right|\right|_2^2)$:

$$R_\alpha(S||Q_0) \leq \frac{\alpha}{\alpha-1}\cdot\left(\frac{k}{2}\log 2\pi - f(\mathbf{0})\right) \leq \frac{\alpha}{\alpha-1}\cdot\frac{k}{2}\log M.$$

For $\alpha \geq 2$, the above bound is thus at most $k\log M$ as desired, and for $1 \leq \alpha \leq 2$ we can just use monotonicity of Rényi-divergences to bound $R_\alpha(S||Q_0)$ by $R_2(S||Q_0)$.

Similarly:

$$\begin{aligned}\exp((1/M)R_{1+1/M}(Q_0||S)) &= \int_{\mathbb{R}^k} \frac{Q_0(\mathbf{x})^{1+1/M}}{S(\mathbf{x})^{1/M}}\mathrm{d}x \\ &= (2\pi)^{-k(1+1/M)/2}\int_{\mathbb{R}^k} \exp\left(-\frac{1+1/M}{2}\left|\left|\mathbf{x}\right|\right|_2^2 + f(\mathbf{x})/M\right)\mathrm{d}x \\ &\leq \frac{e^{f(\mathbf{0})/M}}{(2\pi)^{k(1+1/M)/2}}\int_{\mathbb{R}^k} \exp\left(-\frac{1}{2M}\left|\left|\mathbf{x}\right|\right|_2^2\right)\mathrm{d}x \\ &= \frac{e^{f(\mathbf{0})/M}M^{k/2}}{(2\pi)^{k/2M}}.\end{aligned}$$

Taking logs, and using that the 1-strongly convex f that maximizes $f(\mathbf{0})$ is $N(0, \mathbb{I}_k)$ with density $\exp(-\frac{k}{2}\log(2\pi) - \left|\left|\mathbf{x}\right|\right|_2^2)$:

$$R_{1+1/M}(Q_0||S) \leq M\left[f(\mathbf{0})/M + \frac{k}{2}\log M - \frac{k}{2M}\log(2\pi)\right] \leq \frac{kM\log M}{2}.$$

□

3.8 Underdamped Langevin Dynamics

Our approach can also be used to show a bound on the discretization error of *underdamped* Langevin dynamics. We again start by bounding the divergence between two discrete processes with step sizes η and η/c, whose limits as c goes to infinity are the discretized and continuous underdamped Langevin dynamics. Again let $\mathbf{x}_t$ denote the position of the chain using step size η at continuous time t, and $\mathbf{x}'_t$ denote the position of the chain using step size η/c. Let $\mathbf{v}_t, \mathbf{v}'_t$ denote the same but for velocity instead of position. If e.g. for the first chain we ever have $\left|\left|\mathbf{x}_{t^*} - \mathbf{x}_{\lfloor t^*/\eta \rfloor \eta}\right|\right|_2 > r$ we will let $(\mathbf{x}_t, \mathbf{v}_t)$ equal $\perp$ for all $t \geq t^*$. We want to bound the divergence between the distributions $X_{0:Tc}$ over $\{(\mathbf{x}_{i\eta/c}, \mathbf{v}_{i\eta/c})\}_{0 \leq i \leq Tc}$ and $X'_{0:Tc}$ over $\{(\mathbf{x}'_{i\eta/c}, \mathbf{v}'_{i\eta/c})\}_{0 \leq i \leq Tc}$. A sample from $X_{0:Tc}$ or $X'_{0:Tc}$ can be constructed by applying the following operations Tc times to $\{(\mathbf{x}_0, \mathbf{v}_0)\}$ sampled from an initial distribution X_0:

- To construct a sample from $X_{0:Tc}$, given a sample $\{(\mathbf{x}_{i\eta/c}, \mathbf{v}_{i\eta/c})\}_{0 \leq i \leq j}$ from $X_{0:j}$:
  - If $(\mathbf{x}_{j\eta/c}, \mathbf{v}_{j\eta/c}) = \perp$ append $(\mathbf{x}_{i\eta/c}, \mathbf{v}_{i\eta/c}) = \perp$ to $\{(\mathbf{x}_{i\eta/c}, \mathbf{v}_{i\eta/c})\}_{0 \leq i \leq j}$.
  - Otherwise, append $(\mathbf{x}_{(j+1)\eta/c}, \mathbf{v}_{(j+1)\eta/c})$ where:
    $$\mathbf{v}_{(j+1)\eta/c} = (1 - \gamma\frac{\eta}{c})\mathbf{v}_{j\eta/c} - \mu\frac{\eta}{c}\nabla f(\mathbf{x}_{\lfloor j/c \rfloor \eta}) + \xi_j,$$
    $$\mathbf{x}_{(j+1)\eta/c} = \mathbf{x}_{j\eta/c} + \frac{\eta}{c}\mathbf{v}_{(j+1)\eta/c},$$
    and $\xi_j \sim N(0, 2\gamma\mu\frac{\eta}{c}\mathbb{I}_k)$. Then if $\left|\left|\mathbf{x}_{(j+1)\eta/c} - \mathbf{x}_{\lfloor (j+1)/c \rfloor \eta}\right|\right|_2 > r$ (i.e. $\mathcal{E}_r$ no longer holds) replace $(\mathbf{x}_{(j+1)\eta/c}, \mathbf{v}_{(j+1)\eta/c})$ with $\perp$.

  Let ψ denote this update, i.e. $X_{0:j+1} = \psi(X_{0:j})$.
- To construct a sample from $X'_{0:Tc}$, the update (which we denote ψ') is identical to ψ except we use the gradient at $\mathbf{x}'_{j\eta/k}$ instead of $\mathbf{x}'_{\lfloor j/k \rfloor \eta}$ to compute $\mathbf{v}_{(j+1)\eta/c}$.

We remark that unlike in our analysis of the overdamped Langevin dynamics, for finite c, $X_{0:Tc}, X'_{0:Tc}$ do *not* actually correspond to the SDE (3.1) with step size $\eta, \eta/c$. However, we still have the property that the limit of $X_{0:Tc}$ (resp. $X'_{0:Tc}$) as c goes to infinity follows a discretized (resp. continuous) underdamped Langevin dynamics, which is all that is needed for our analysis. Similarly to the overdamped Langevin dynamics we have:

Lemma 63. *For any M-smooth f and $X_{0:Tc}, X'_{0:Tc}$ as defined above, we have:*

$$R_\alpha(X_{0:Tc}||X'_{0:Tc}), R_\alpha(X'_{0:Tc}||X_{0:Tc}) \leq \frac{T\alpha M^2 r^2 \eta}{4} \cdot \frac{\mu}{\gamma}.$$

The proof follows almost exactly as did the proof of Lemma 47: we note that the updates to position are deterministic, and so by Fact 45 we just need to control the divergence between velocities, which can be done using the same analysis as in Lemma 47. The multiplicative factor of μ/γ appears because the ratio of the Gaussian's standard deviation in any direction to the gradient step's multiplier is $\sqrt{\gamma/\mu}$ times what it was in the overdamped Langevin dynamics. Next, similar to Lemma 50, we have the following tail bound on r:

Lemma 64. *Fix any $\gamma \geq 2$, and define*

$$\mathbf{v}_{\max} := c\sqrt{\gamma\mu}\left(\sqrt{\tau k} + \sqrt{\ln(1/\delta)}\right).$$

Fix any $\eta \leq \frac{\gamma}{\mu L}$, and any distribution over $\mathbf{x}_0, \mathbf{v}_0$ satisfying that

$$\Pr\left[\mu f(\mathbf{x}_0) + \frac{||\mathbf{v}_0||_2^2}{2} \leq \frac{1}{2}\mathbf{v}_{\max}^2\right] \geq 1 - \delta, \tag{3.4}$$

let $\mathbf{x}_t, \mathbf{v}_t$ be the random variable given by running the discretized underdamped Langevin dynamics starting from $\mathbf{x}_0, \mathbf{v}_0$ drawn from this distribution for time t. Then with probability $1 - \delta$ over $\{(\mathbf{x}_t, \mathbf{v}_t) : t \in [0, \tau]\}$, for all $t \leq \tau$ that are multiples of η and for a sufficiently large constant c:

$$||\mathbf{x}_{t+\eta} - \mathbf{x}_t||_2 \leq \mathbf{v}_{\max}\eta.$$

Similarly, if $\mathbf{x}_t$ is the random variable given by running continuous underdamped Langevin dynamics starting from $\mathbf{x}_0, \mathbf{v}_0$ drawn from this distribution for time t, with probability $1 - \delta$ over $\{(\mathbf{x}'_t, v'_t) : t \in [0, \tau]\}$ for all $t \leq \tau$:

$$\left|\left|\mathbf{x}_t - \mathbf{x}_{\lfloor t/\eta \rfloor \eta}\right|\right|_2 \leq \mathbf{v}_{\max}\eta.$$

The proof is deferred to Section 3.9. We note that the correct tail bound likely has a logarithmic dependence on τ and not a polynomial one. However, based on similar convergence bounds (e.g. [82, 61]), we conjecture that the time τ needed for continuous underdamped Langevin dynamics to converge in Rényi-divergence has a logarithmic dependence on $k, 1/\epsilon$. So, improving the dependence on τ in this tail bound will likely not improve the final iteration complexity's dependence on $k, 1/\epsilon$ by more than logarithmic factors. In addition, settling for a polynomial dependence on τ makes the proof rather straightforward. Putting it all together, we get:

Theorem 65. *For any 1-strongly convex, M-smooth function f, let P be the distribution of states for discretized underdamped Langevin dynamics with step size η and Q be the distribution of states for continuous underdamped Langevin dynamics, both run from any initial distribution on $\mathbf{x}_0, \mathbf{v}_0$ satisfying (3.4), for continuous time τ that is a multiple of η. Then for $\alpha > 1$, $\epsilon > 0$, if $\eta = \tilde{O}(\min\{\frac{1}{M\tau\mu\ln\alpha} \cdot \frac{\epsilon}{\sqrt{k}}, \frac{\gamma}{\mu L}\})$ we have $R_\alpha(P||Q), R_\alpha(Q||P) \leq \epsilon$.*

Proof. The proof follows similarly to that of Theorem 52. From Lemma 63, plugging in the tail bound of Lemma 64 for r (which holds since we assume $\eta \leq \frac{\gamma}{\mu L}$) we get the divergence bound:

$$R_{\alpha'}(X_{T,k}, X'_{T,k}) \leq \frac{3\mu\tau\alpha' M^2 c^2(\tau k + \ln(\frac{1}{\delta_1}) + \ln(\frac{1}{\delta_2}))\eta^2}{4}$$

We can then just follow the proof of Theorem 52 as long as:

$$c_1(\alpha') = \exp\left(\frac{3\mu\tau^2 k\alpha'(\alpha'-1)M^2c^2\eta^2}{4}\right) < 2,$$

For $\alpha' = \frac{12\alpha \ln \alpha \ln 1/\epsilon}{(\alpha-1)\epsilon} - 2$. This follows if $\eta = \tilde{O}(\frac{1}{M\tau\mu\ln\alpha} \cdot \frac{\epsilon}{\sqrt{k}})$ as assumed in the lemma statement. □

We give here some intuition for why the proof achieves an iteration complexity for underdamped Langevin dynamics with a quadratically improved dependence on k, ϵ compared to overdamped Langevin dynamics. The tail bound on the maximum movement within each step of size η (and in turn the norm of the discretization error due to the gradient) has a quadratically stronger dependence on η in the underdamped case than in the overdamped case. In turn, in underdamped Langevin dynamics the "privacy loss" of hiding this error with Brownian motion also improves quadratically as a function of η.

3.9 Proofs of Tail Bounds on Movements

In this section we give the proofs of Lemmas 50, 55, and 64, which provide tail bounds for the maximum movement within each step of the Langevin dynamics in the three settings we consider. We first recall some facts about Gaussians, Brownian motion, and gradient descent:

Fact 66 (Univariate Gaussian Tail Bound). *For $x \sim N(0, \sigma^2)$ and any $c \geq 0$, we have*

$$\Pr[x \geq c] = \Pr[x \leq -c] \leq \exp\left(-\frac{c^2}{2\sigma^2}\right).$$

Fact 67 (Isotropic Multivariate Normal Tail Bound). *For $\mathbf{x} \sim N(0, \mathbb{I}_k)$ and any $c \geq 0$, we have*

$$\Pr[||\mathbf{x}||_2 \geq \sqrt{k} + c] \leq \exp\left(-\frac{c^2}{2}\right).$$

Fact 68 (Univariate Brownian Motion Tail Bound). *Let B_t be a standard (one-dimensional) Brownian motion. For any $0 \leq a \leq b$ and $c \geq 0$ we have:*

$$\Pr\left[\sup_{t\in[a,b]} [B_t - B_a] \geq c\right] = 2 \cdot \Pr[N(0, b-a) \geq c] \leq 2\exp\left(-\frac{c^2}{2(b-a)}\right)$$

The preceding fact is also known as *the reflection principle.*

Fact 69 (Multivariate Brownian Motion Tail Bound). *Let B_t be a standard k-dimensional Brownian motion. For any $0 \leq a \leq b$, $c \geq 0$, we have:*

$$\Pr\left[\sup_{t\in[a,b]} ||B_t - B_a||_2 \geq \sqrt{b-a}\left(\sqrt{k} + c\right)\right] \leq 2\exp(-c^2/4).$$

Fact 70 (Discrete Gradient Descent Contracts)**.** *Let $f : \mathbb{R}^k \to \mathbb{R}$ be a 1-strongly convex, M-smooth function. Then for $\eta \leq \frac{2}{M+1}$, we have $||\mathbf{x} - \eta \nabla f(\mathbf{x}) - \mathbf{x}' + \eta \nabla f(\mathbf{x}')||_2 \leq (1 - \frac{\eta M}{M+1}) ||\mathbf{x} - \mathbf{x}'||_2 \leq (1 - \frac{\eta}{2}) ||\mathbf{x} - \mathbf{x}'||_2$ for any $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^k$.*

See e.g. Lemma 3.7 of [45] for a proof of this fact. Since we assume f's global minimum is at $\mathbf{0}$ (and thus $\nabla f(\mathbf{0}) = 0$), as a corollary we have $||\mathbf{x} - \eta \nabla f(\mathbf{x})||_2 \leq (1 - \eta/2) ||\mathbf{x}||_2$. We also have as a corollary:

Fact 71 (Continuous Gradient Descent Contracts)**.** *Let $f : \mathbb{R}^k \to \mathbb{R}$ be a 1-strongly convex, M-smooth function. Then for any $\mathbf{x}_0, \mathbf{x}_0' \in \mathbb{R}^k$ and $\mathbf{x}_t, \mathbf{x}_t'$ that are solutions to the differential equation $d\mathbf{x}_t = -\nabla f(\mathbf{x}_t) dt$ we have $||\mathbf{x}_t - \mathbf{x}_t'||_2 \leq e^{-t/2} ||\mathbf{x}_0 - \mathbf{x}_0'||_2$.*

Proof. This follows by noting that the $\mathbf{x}_t$ is the limit as integer c goes to ∞ of applying c discrete gradient descent steps to $\mathbf{x}_0$ with $\eta = t/c$. So, the contractivity bound we get for $\mathbf{x}_t$ is $||\mathbf{x}_t - \mathbf{x}_t'||_2 \leq \lim_{c \to \infty} (1 - t/2c)^c ||\mathbf{x}_0 - \mathbf{x}_0'||_2 = e^{-t/2} ||\mathbf{x}_0 - \mathbf{x}_0'||_2$. □

Proof of Lemma 50

Proof. We consider the discrete chain first. For each timestep starting at t that is a multiple of η, using smoothness we have:

$$\begin{aligned}\max_{t' \in [t, t+\eta)} ||\mathbf{x}_{t'} - \mathbf{x}_t||_2 &= \max_{t' \in [t, t+\eta)} \left|\left| -(t'-t)\nabla f(\mathbf{x}_t) + \sqrt{2} \int_t^{t'} \mathrm{d}B_s \right|\right|_2 \\ &\leq \eta ||\nabla f(\mathbf{x}_t)||_2 + \sqrt{2} \max_{t' \in [t, t+\eta)} \left|\left| \int_t^{t'} \mathrm{d}B_s \right|\right|_2 \\ &\leq \eta M ||\mathbf{x}_t||_2 + \sqrt{2} \max_{t' \in [t, t+\eta)} \left|\left| \int_t^{t'} \mathrm{d}B_s \right|\right|_2 .\end{aligned}$$

Using the tail bound for multivariate Brownian motion, $\max_{t' \in [t, t+\eta)} \left|\left| \int_t^{t'} \mathrm{d}B_s \right|\right|_2$ is at most $\frac{c}{2\sqrt{2}} \left(\sqrt{k} + \sqrt{\ln(T/\delta)}\right) \sqrt{\eta}$ with probability at least $1 - \frac{\delta}{2T}$ for each timestep. So it suffices to show that with probability at least $1 - \frac{\delta}{2}$, for all $0 \leq t < T\eta$ that are multiples of η, $||\mathbf{x}_t||_2 \leq \frac{c}{2\sqrt{\eta}} \left(\sqrt{k} + \sqrt{\ln(T/\delta)}\right)$. From (3.3), with probability $1 - \frac{\delta}{T+1}$, $||\mathbf{x}_0||_2 \leq \frac{c}{2\sqrt{\eta}} \left(\sqrt{k} + \sqrt{\ln(T/\delta)}\right)$. We will show that if $||\mathbf{x}_t||_2 \leq \frac{c}{2\sqrt{\eta}} \left(\sqrt{k} + \sqrt{\ln(T/\delta)}\right)$ then with probability $1 - \frac{\delta}{T+1}$ we have $||\mathbf{x}_{t+\eta}||_2 \leq \frac{c}{2\sqrt{\eta}} \left(\sqrt{k} + \sqrt{\ln(T/\delta)}\right)$, completing the proof for the discrete case by a union bound. This follows because by Fact 70 the gradient descent step is $(1 - \eta/2)$-Lipschitz for the range of η we consider. This gives that after the gradient descent step but before adding Gaussian noise, $\mathbf{x}_{t+\eta}$ has norm at most

$(1-\eta/2)\left|\left|\mathbf{x}_t\right|\right|_2 \leq (1-\eta/2)\frac{c}{2\sqrt{\eta}}\left(\sqrt{k}+\sqrt{\ln(T/\delta)}\right)$. Then, $\left|\left|\mathbf{x}_{t+\eta}\right|\right|_2 > \frac{c}{2\sqrt{\eta}}\left(\sqrt{k}+\sqrt{\ln(T/\delta)}\right)$ only if $\sqrt{2}\left|\left|\int_t^{t+\eta}\mathrm{d}B_s\right|\right|_2$ is larger than $c\sqrt{\eta}\left(\sqrt{k}+\sqrt{\ln(T/\delta)}\right)$, which happens with probability at most $\frac{\delta}{T+1}$ by the multivariate Gaussian tail bound.

We now consider the continuous chain. For all t that are multiples of η:

$$\begin{aligned}\max_{u\in[t,t+\eta)}\left|\left|\mathbf{x}'_u-\mathbf{x}'_t\right|\right|_2 &= \max_{u\in[t,t+\eta)}\left|\left|\int_t^u -\nabla f(\mathbf{x}'_s)\mathrm{d}s+\sqrt{2}\mathrm{d}B_s\right|\right|_2\\ &\leq \eta L\max_{u\in[t,t+\eta)}\left|\left|\mathbf{x}'_u\right|\right|_2+\max_{u\in[t,t+\eta)}\left|\left|\sqrt{2}\int_t^u\mathrm{d}B_s\right|\right|_2.\end{aligned}$$

As with the discrete chain, the multivariate Brownian motion tail bound gives that

$$\max_{u\in[t,t+\eta)}\left|\left|\sqrt{2}\int_t^u\mathrm{d}B_s\right|\right|_2 \leq \frac{c}{2}\left(\sqrt{k}+\sqrt{\ln(T/\delta)}\right)\sqrt{\eta},$$

with probability at least $1-\frac{\delta}{2T}$. So it suffices to show that at all times between 0 and $T\eta$, $\left|\left|\mathbf{x}'_u\right|\right|_2 \leq \frac{c}{2\sqrt{\eta}}\left(\sqrt{k}+\sqrt{\ln(T/\delta)}\right)$ with probability at least $1-\frac{\delta}{2}$. We first claim that with probability at least $1-\frac{\delta}{4}$, for all t that are multiples of η, $\left|\left|\mathbf{x}'_t\right|\right|_2 \leq \frac{c}{4\sqrt{\eta}}\left(\sqrt{k}+\sqrt{\ln(T/\delta)}\right)$. This is true for $\mathbf{x}'_0$ with probability at least $1-\frac{\delta}{4(T+1)}$ by (3.3). By contractivity of continuous gradient descent, $\mathbf{x}'_{t+\eta}$ is equal to $A\mathbf{x}'_t+\sqrt{2}\int_t^{t+\eta}A'_s\mathrm{d}B_s$ for some A which has eigenvalues in $[-e^{-\eta/2},e^{-\eta/2}]$ and a set of matrices $\{A'_s|s\in[0,\eta]\}$ with eigenvalues in $[-e^{-(\eta-s)/2},e^{-(\eta-s)/2}]$[1]. Then conditioning on the claim holding for $\mathbf{x}'_t$, $\left|\left|\mathbf{x}'_{t+\eta}\right|\right|_2$ exceeds $\frac{c}{4\sqrt{\eta}}\left(\sqrt{k}+\sqrt{\ln(T/\delta)}\right)$ only if the norm of $\sqrt{2}\int_t^{t+\eta}A'_s\mathrm{d}B_s$ exceeds

$$\frac{c(1-e^{-\eta/2})}{4\sqrt{\eta}}\left(\sqrt{k}+\sqrt{\ln(T/\delta)}\right) \geq \frac{c(1-e^{-.5}))\sqrt{\eta}}{4}\left(\sqrt{k}+\sqrt{\ln(T/\delta)}\right).$$

Since Brownian motion is rotationally symmetric, and all A'_s have eigenvalues in $[-1,1]$, this occurs with probability upper bounded by the probability $\sqrt{2}\int_t^{t+\eta}\mathrm{d}B_s$ exceeds this bound, which is at most $\frac{\delta}{4(T+1)}$ by the Brownian motion tail bound. The claim follows by taking a union bound over all t that are multiples of η.

Then, conditioning on the event in the claim, for each corresponding interval $[t,t+\eta)$ since gradient descent contracts we have

[1] In particular, recalling the proof of Facts 70 and 71, we can write A explicitly as $\lim_{k\to\infty}\prod_{j=0}^{k-1}(I_d-\frac{\eta}{k}\nabla^2 f(z_j))$, where z_j is some point on the path from 0 to $\mathbf{x}'_{t+\frac{j\eta}{k}}$. Each A_s can be written similarly, except only considering the gradient descent process from time $t+s$ to $t+\eta$.

$$\begin{aligned}\max_{u\in[t,t+\eta)} ||\mathbf{x}'_u||_2 &\leq ||\mathbf{x}'_t||_2 + \max_{u\in[t,t+\eta)} \left|\left|\sqrt{2}\int_t^u \mathrm{d}B_s\right|\right|_2 \\ &\leq \frac{c}{4\sqrt{\eta}}\left(\sqrt{k}+\sqrt{\ln(T/\delta)}\right) + \max_{u\in[t,t+\eta)} \left|\left|\sqrt{2}\int_t^u \mathrm{d}B_s\right|\right|_2 .\end{aligned}$$

We conclude by using the multivariate Brownian motion tail bound to observe that

$$\max_{u\in[t,t+\eta)} \left|\left|\sqrt{2}\int_t^u \mathrm{d}B_s\right|\right|_2 \leq \frac{c}{4\sqrt{\eta}}\left(\sqrt{k}+\sqrt{\ln(T/\delta)}\right),$$

with probability at least $1-\frac{\delta}{4T}$, and then taking a union bound over all intervals. □

Proof of Lemma 55

Proof. By L-Lipschitzness of f, the movement in any interval of length η due to the gradient step in both the discrete and continuous case is at most $2L\eta$. By the multivariate Brownian motion tail bound, in both the discrete and continuous cases the maximum movement due to the addition of Gaussian noise is at most $c(\sqrt{k}+\sqrt{\ln(T/\delta)})\sqrt{\eta}$ with probability at least $1-\frac{\delta}{T}$ in each interval of length η, and then the lemma follows by a union bound and triangle inequality. □

Proof of Lemma 64

Proof. We can assume $\delta < 1/2$, at a loss of a multiplicative constant. We first focus on the continuous chain. It suffices to show the maximum norm of the velocity over $[0,\tau)$ is $\mathbf{v}_{\max}$ with the desired probability. We will instead focus on bounding the Hamiltonian, defined as follows:

$$\phi_t = \mu f(\mathbf{x}'_t) + ||\mathbf{v}'_t||_2^2/2.$$

Analyzing the rate of change, by Ito's lemma we get

$$\begin{aligned}\mathrm{d}\phi_t &= \frac{\partial\phi_t}{\partial\mathbf{x}'_t}\cdot\mathrm{d}\mathbf{x}'_t + \frac{\partial\phi_t}{\partial\mathbf{v}'_t}\cdot\mathrm{d}\mathbf{v}'_t + \frac{1}{2}\left[\sum_{i,j\in[k]}\frac{\partial^2\phi_t}{\partial(\mathbf{v}'_t)_i\partial(\mathbf{v}'_t)_j}\frac{\mathrm{d}(\mathbf{v}'_t)_i}{\mathrm{d}B_t}\frac{\mathrm{d}(\mathbf{v}'_t)_j}{\mathrm{d}B_t}\right]\mathrm{d}t \\ &= \mu\nabla f(\mathbf{x}'_t)\cdot\mathbf{v}'_t\mathrm{d}t + \mathbf{v}'_t\cdot(-\mu\nabla f(\mathbf{x}'_t)\mathrm{d}t - \gamma\mathbf{v}'_t\mathrm{d}t + \sqrt{2\gamma\mu}\mathrm{d}B_t) + 2\gamma\mu k\cdot\mathrm{d}t \\ &= \gamma(2\mu k - ||\mathbf{v}'_t||_2^2)\mathrm{d}t + \sqrt{2\gamma\mu}(\mathbf{v}'_t\cdot\mathrm{d}B_t).\end{aligned}$$

So, we can write the Hamiltonian at any time as a function of the initial Hamiltonian ϕ_0 and the random variables B_t and $\mathbf{v}'_t$ as:

$$\phi_t = \phi_0 - \gamma \int_0^t ||\mathbf{v}'_s||_2^2 \, \mathrm{d}s + \sqrt{2\gamma\mu} \int_0^t ||\mathbf{v}'_s||_2 \frac{\mathbf{v}'_s}{||\mathbf{v}'_s||_2} \cdot \mathrm{d}B_s + 2\gamma\mu k t.$$

Let V_t denote $\int_0^t ||\mathbf{v}'_s||_2^2 \, \mathrm{d}s$. By scalability of Brownian motion, we can define a Brownian motion B'_t jointly distributed with B_t such that $\mathrm{d}B_t = \frac{1}{||\mathbf{v}'_t||_2} \frac{\mathrm{d}}{\mathrm{d}t} \int_0^{V_t} \mathrm{d}B'_s$. Then, we have:

$$\phi_t = \phi_0 - \gamma V_t + \sqrt{2\gamma\mu} \int_0^{V_t} \frac{\mathbf{v}'_{g(s)}}{\left|\left|\mathbf{v}'_{g(s)}\right|\right|_2} \cdot \mathrm{d}B'_s + 2\gamma\mu k t,$$

Where $g(r)$ is the value r' such that $\int_0^{r'} ||\mathbf{v}'_s||_2^2 \, \mathrm{d}s = r$. We can then use the rotational symmetry of Brownian motion to define another Brownian motion B''_t jointly distributed with B'_t such that $\mathbf{u} \cdot \mathrm{d}B''_t = \frac{\mathbf{v}'_{g(t)}}{\left|\left|\mathbf{v}'_{g(t)}\right|\right|_2} \cdot \mathrm{d}B'_t$ for a fixed unit vector $\mathbf{u}$, giving:

$$\phi_t = \phi_0 - \gamma V_t + \sqrt{2\gamma\mu} \int_0^{V_t} \mathbf{u} \cdot \mathrm{d}B''_s + 2\gamma\mu k t.$$

We will show that with probability at least $1 - \delta$ over B''_t, the maximum of $\phi'(V) := \phi_0 - \gamma V + \sqrt{2\gamma\mu} \int_0^V \mathbf{u} \cdot \mathrm{d}B''_s$ over $V \in [0, \infty)$ is at most $\frac{1}{4}\mathbf{v}_{\max}^2$. Under this event, if c is sufficiently large then for all $t \in [0, \tau)$ we have $\phi_t \leq \frac{1}{4}\mathbf{v}_{\max}^2 + 2\gamma\mu k\tau \leq \frac{1}{2}\mathbf{v}_{\max}^2$, giving the desired velocity bound.

We first claim that with probability at at least $1 - \frac{\delta}{2}$. for all non-negative integers q, we have $\phi'(q\mathbf{v}_{\max}^2) \leq -\frac{(q-1)\mathbf{v}_{\max}^2}{2}$. For sufficiently large c, this holds for $q = 0$ with probability at least $1 - \frac{\delta}{4}$ by (3.4). Conditioning on this event, for $q > 0$ if $\phi'(q\mathbf{v}_{\max}^2) \geq -\frac{(q-1)\mathbf{v}_{\max}^2}{2}$, then:

$$\sqrt{2\gamma\mu} \int_0^{q\mathbf{v}_{\max}^2} \mathbf{u} \cdot \mathrm{d}B''_s = N(0, 2q\gamma\mu\mathbf{v}_{\max}^2) \geq -\frac{(q-1)\mathbf{v}_{\max}^2}{2} - \phi_0 + q\gamma\mathbf{v}_{\max}^2 \geq (\gamma - 1)q\mathbf{v}_{\max}^2,$$

Which occurs with probability at most $\exp(-\frac{(\gamma-1)^2 q^2 \mathbf{v}_{\max}^4}{4q\gamma\mu\mathbf{v}_{\max}^2}) \leq \exp(-\frac{q\mathbf{v}_{\max}^2}{8\mu})$. If the constant c in $\mathbf{v}_{\max}$ is sufficiently large, then this is less than $\frac{\delta^{q+2}}{2}$. Taking a union bound over all q, we get the claim. Next, we claim that in each interval $[q\mathbf{v}_{\max}^2, (q+1)\mathbf{v}_{\max}^2)$, the maximum increase of $\phi'(V)$ is more than $(\frac{q+1}{2})\mathbf{v}_{\max}^2$ with probability at most $\frac{\delta^{q+2}}{2}$. Taking a union bound over all intervals, this claim along with the previous claim this gives the desired bound on $\phi'(V)$ with probability $1 - \delta$. This claim follows by observing that in the interval $[q\mathbf{v}_{\max}^2, (q+1)\mathbf{v}_{\max}^2)$, $\phi'(V)$ increases more than $\max_{V \in [q\mathbf{v}_{\max}^2, (q+1)\mathbf{v}_{\max}^2)} \left[\int_{q\mathbf{v}_{\max}^2}^V \mathbf{u} \cdot \mathrm{d}B''_s\right]$, which is at most $(\frac{q+1}{2})\mathbf{v}_{\max}^2$ with probability at most $\exp(-\frac{(\frac{q+1}{2})^2\mathbf{v}_{\max}^4}{8\mathbf{v}_{\max}^2}) \leq \frac{\delta^{q+1}}{2}$.

The discrete chain is analyzed similarly. We have:

$$\begin{aligned}
\mathrm{d}\phi_t &= \frac{\partial \phi_t}{\partial \mathbf{x}_t} \cdot \mathrm{d}\mathbf{x}_t + \frac{\partial \phi_t}{\partial \mathbf{v}_t} \cdot \mathrm{d}\mathbf{v}_t + \frac{1}{2}\left[\sum_{i,j \in [k]} \frac{\partial^2 \phi_t}{\mathrm{d}(\mathbf{v}_t)_i \mathrm{d}(\mathbf{v}_t)_j} \frac{\mathrm{d}(\mathbf{v}_t)_i}{\mathrm{d}B_t} \frac{\mathrm{d}(\mathbf{v}_t)_j}{\mathrm{d}B_t}\right] \mathrm{d}t \\
&= \mu \nabla f(\mathbf{x}_t) \cdot \mathbf{v}_t \mathrm{d}t + \mathbf{v}_t \cdot (-\mu \nabla f(\mathbf{x}_{\lfloor \frac{t}{\eta} \rfloor \eta})\mathrm{d}t - \gamma \mathbf{v}_t \mathrm{d}t + \sqrt{2\gamma\mu}\mathrm{d}B_t) + 2\gamma\mu k \cdot \mathrm{d}t \\
&= \mu(\nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}_0)) \cdot \mathbf{v}_t \mathrm{d}t - \gamma \left|\left|\mathbf{v}_t\right|\right|_2^2 \mathrm{d}t + \sqrt{2\gamma\mu}(v \cdot \mathrm{d}B_t) + 2\gamma\mu k \cdot \mathrm{d}t \\
&\leq \mu M \left|\left|\mathbf{x}_t - \mathbf{x}_{\lfloor \frac{t}{\eta} \rfloor \eta}\right|\right|_2 \left|\left|\mathbf{v}_t\right|\right|_2 \mathrm{d}t - \gamma \left|\left|\mathbf{v}_t\right|\right|_2^2 \mathrm{d}t + \sqrt{2\gamma\mu}(v \cdot \mathrm{d}B_t) + 2\gamma\mu k \cdot \mathrm{d}t \\
&= \mu M \left|\left|\int_{\lfloor \frac{t}{\eta} \rfloor \eta}^{t} \mathbf{v}_s \mathrm{d}s\right|\right|_2 \left|\left|\mathbf{v}_t\right|\right|_2 \mathrm{d}t - \gamma \left|\left|\mathbf{v}_t\right|\right|_2^2 \mathrm{d}t + \sqrt{2\gamma\mu}(v \cdot \mathrm{d}B_t) + 2\gamma\mu k \cdot \mathrm{d}t \\
&\leq \mu M \left(\int_{\lfloor \frac{t}{\eta} \rfloor \eta}^{t} \left|\left|\mathbf{v}_s\right|\right|_2 \left|\left|\mathbf{v}_t\right|\right|_2 \mathrm{d}s\right) \mathrm{d}t - \gamma \left|\left|\mathbf{v}_t\right|\right|_2^2 \mathrm{d}t + \sqrt{2\gamma\mu}(v \cdot \mathrm{d}B_t) + 2\gamma\mu k \cdot \mathrm{d}t \\
&\leq \frac{\mu M}{2} \left(\int_{\lfloor \frac{t}{\eta} \rfloor \eta}^{t} \left|\left|\mathbf{v}_s\right|\right|_2^2 + \left|\left|\mathbf{v}_t\right|\right|_2^2 \mathrm{d}s\right) \mathrm{d}t - \gamma \left|\left|\mathbf{v}_t\right|\right|_2^2 \mathrm{d}t + \sqrt{2\gamma\mu}(v \cdot \mathrm{d}B_t) + 2\gamma\mu k \cdot \mathrm{d}t.
\end{aligned}$$

Integrating, we get:

$$\begin{aligned}
\phi_t &\leq \phi_0 - (\gamma - \frac{\mu M \eta}{2}) \int_0^t \left|\left|\mathbf{v}_s\right|\right|_2^2 \mathrm{d}s + \sqrt{2\gamma\mu} \int_0^t \left|\left|\mathbf{v}_s\right|\right|_2 \frac{\mathbf{v}_s}{\left|\left|\mathbf{v}_s\right|\right|_2} \cdot \mathrm{d}B_s + 2\gamma\mu k t \\
&\leq \phi_0 - \frac{\gamma}{2} \int_0^t \left|\left|\mathbf{v}_s\right|\right|_2^2 \mathrm{d}s + \sqrt{2\gamma\mu} \int_0^t \left|\left|\mathbf{v}_s\right|\right|_2 \frac{\mathbf{v}_s}{\left|\left|\mathbf{v}_s\right|\right|_2} \cdot \mathrm{d}B_s + 2\gamma\mu k t.
\end{aligned}$$

At this point we repeat the analysis from the continuous case (only losing a multiplicative constant due to the $\gamma/2$ multiplier not being γ). □

Chapter 4

Public Data-Augmented Stochastic Optimization

4.1 Introduction and Problem Definition

Differentially Private Gradient Descent (DP-GD) [76, 13, 1][1], and its variants [52] have become the de facto standard algorithms for training machine learning models with differential privacy. While DP-GD is known to be optimal in terms of obtaining both optimal excess empirical risk [13], and excess population risk [18] for convex losses, the obtained error guarantees suffer from an explicit polynomial dependence on the model dimension (k). This polynomial dependence significantly impacts the privacy/utility trade-off when $k \geq n_{priv}$, where n_{priv} is the number of private training samples. Even empirically, when DP-GD is used to train large deep learning models, there is a significant drop in accuracy compared to the non-private counterpart [72].

In this chapter, we revisit the problem of using public data (i.e., data without privacy concerns) to improve the privacy/utility trade-offs for differentially private model training. *Specifically, we design differentially private variants of mirror descent [69] that use the loss function generated by the public data as the mirror map and differentially private gradients on the private data as the linear term.* For linear regression as well as a class of more general convex optimization settings, we show that the excess population risk *asymptotically* improves over the best known bounds under differential privacy (without access to public data samples) [13, 17] when n_{pub} is sufficiently large (i.e., a small polynomial in k), and the public and private feature vectors are drawn from the same non-isotropic sub-Gaussian distribution. Here, n_{pub} is the number of public data samples. Even if n_{pub} is small, our algorithm generalizes DP-GD, so it never performs worse than DP-GD. Furthermore, we show empirically that our differentially private variant of mirror descent, assisted with public

[1] Again, we recall that the popular algorithm is *stochastic* gradient descent, i.e. we subsample examples to get an estimate of the gradient, and not gradient descent, which uses all examples. To simplify the presentation, in this chapter we ignore the distinction between the two consider only gradient descent.

data, can improve the privacy-utility trade-offs by effectively reducing the variance in the noise added to the gradients in differentially private model training.

Learning Geometry with Mirror Maps: Common to most differentially private model training algorithms, including DP-GD, DP-FTRL [52], and our algorithm, is a differentially private estimator of the gradient of the loss $\nabla_\theta \mathcal{L}(\theta_t; D_{priv}) = \sum_{d \in D_{priv}} \nabla_\theta \ell(\theta_t; d)$ generated by the private dataset D_{priv} at a given model state $\theta_t \in \mathbb{R}^p$. This estimator adds isotropic Gaussian noise $N(0, \sigma^2 \mathbb{I}_k)$ to $\nabla_\theta \mathcal{L}(\theta_t; D_{priv})$, where σ depends on the privacy parameters (ϵ, δ) and the maximum allowable value of $||\nabla_\theta \ell(\theta_t; d)||_2$ (a.k.a. the clipping norm [1]). It is well known that for most learning tasks, the set of gradients for $\mathcal{L}(\theta_t; D_{priv})$ is seldom isotropic [44, 2]. Hence, it is natural to wonder if the Gaussian noise in the differentially private estimator can be made to respect the geometry of the gradients.

Prior works [90, 7, 50] have used public data (D_{pub}) to *explicitly* learn this geometry, mostly in the form of preconditioner matrices [29] to be multiplied to the estimated noisy gradients. In this chapter, we take an *implicit* approach towards respecting this geometry, by using the loss $\mathcal{L}(\theta; D_{pub})$ generated by the public data as the mirror map in classical mirror descent. As a first order approximation, one can view it as doing DP-GD on $\mathcal{L}(\theta; D_{priv})$ while using $\mathcal{L}(\theta; D_{pub})$ as a regularizer. This approach has the following advantages: (i) The information of the geometry is "free", i.e., one does not need to learn the preconditioner explicitly from the public data, (ii) Unlike prior works [90, 50], one does not need to assume that the gradients of $\mathcal{L}(\theta; D_{priv})$ lie in a low rank subspace, and (iii) It is easier to implement since it does not need to maintain an additional data structure for the preconditioner due to the geometry being implicitly defined.

We note that differentially private mirror descent has been considered before by [79, 83]. Their results are not directly comparable to ours because (i) they do not have access to in-distribution public data, (ii) as shown in [13], without public data, it is impossible to achieve the bounds we achieve, and (iii) in our experiments, we solve unconstrained optimization problems whereas those works choose the mirror map based on the constraint set rather than the dataset. The utility bounds we prove in this chapter also apply to a public data-assisted variant of accelerated mirror descent in [83].

In-distribution vs. Out-of-distribution Public Data: Prior works have considered settings where the public data comes from the same distribution as the private data (a.k.a. *in-distribution*) [14, 90, 50, 7, 85], and where they can be different (a.k.a. *out-of-distribution*) [1, 71, 70, 56, 58, 89].

In the in-distribution setting, it is typical that there are fewer public data samples available than private data samples – i.e., $n_{pub} \ll n_{priv}$ – as it is harder to obtain public datasets than ones with privacy constraints attached. In-distribution public data could come from either altruistic *opt-in* users [63, 8] or from users who are incentivized to provide such data (e.g., mechanical turks). Out-of-distribution (OOD) public data may be easier to obtain but can have various degrees of freedom; e.g., the domains of private and public data may not be identical, the representation of some classes may vary, the distributions can be mean shifted, etc. It is usually hard to quantify these degrees of freedom to the extent that we can provide precise guarantees. Hence, we leave this aspect for future exploration, and work with the (idealized) assumption that the public data comes from the same distribution as the private data, or, at least, that the differences between these two distributions are not material. It worth emphasizing that although our utility results are for the in-distribution case, our algorithm can be used *as is* in out-of-distribution settings. In a restricted set of experiments, we do compare with one of the SoTA [7] for training with OOD public data, and demonstrate improvements in privacy/utility trade-off.

Choice of Empirical Benchmark: Mirror descent as a first step optimizes the mirror map function. In our setting, this corresponds to pre-training on the public loss function $\mathcal{L}(\theta; D_{pub})$ before running the differentially private optimization procedure on $\mathcal{L}(\theta; D_{priv})$. Since pre-training on public data is intuitive and easy, we always compare to DP-GD (and its variants) that have been pre-trained to convergence with the public loss. We show that our algorithm *outperforms* even pre-trained DP-GD. To our knowledge, ours is the first empirical work that compares to this strong (but fair) benchmark.

Other Uses of Public Data in Differentially Private Learning: The use of in-distribution public data has been extensively explored both theoretically and empirically. On the theoretical side, it has been shown [3, 16] that a combination of private and public data samples can yield asymptotically better worst-case PAC learning guarantees than either on their own. Another line of work [71, 70, 15, 31, 67] considers public data that is unlabelled, but otherwise comes from the same distribution as the private data; the primary goal is to use the private data to generate labels for the public data, which can then be used arbitrarily. Additionally, [38] showed that for convex ERMs, using $\approx k$ in-distribution public data samples, one can obtain dimension independent population risk guarantees. However, the main tool used to prove differential privacy (i.e., privacy amplification by iteration) heavily relies on convexity. As a result, their algorithm is inapplicable to the deep learning problems we consider in this chapter.

Prior to our work only two papers considered out-of-distribution data from a theory standpoint. [12] assume that whether a data record is public or private depends on its label; e.g., the public data may contain many negative examples, but few positive examples. They show that halfspaces can be learned in this model. [58] consider synthetic data generation and provide guarantees that depend on the Rényi divergences between the public and private distributions. [1, 81] provided techniques to effectively use out-of-distribution public data for pre-training for DP-GD. However, they did not consider techniques to improve a pre-trained model using private and public data, which is the focus of our work. A recent work [89] uses public data to dynamically adjust the privacy budget and clipping norm. Our technique crucially uses the public data to learn the geometry of the gradients; [89] is complementary to ours and can be utilized for potential additional gains from using the public data after pre-training.

Problem Formulation

For convenience we restate the classic differentially private stochastic convex optimization (DP-SCO) [22, 13, 17, 18] setting. Let τ be a distribution over a fixed domain $\mathcal{D}$. Given a dataset $D \in \mathcal{D}^*$ drawn i.i.d. from τ, and a convex loss function $\ell_{priv} : \mathbb{R}^p \times \mathcal{D} \to \mathbb{R}$, the objective is to approximately solve $\arg\min_{\theta \in \mathcal{C}} \mathbb{E}_{d \sim \tau}\left[\ell_{priv}(\theta; d)\right]$, while preserving DP. Here, $\mathcal{C} \subseteq \mathbb{R}^p$ is the constraint set. Usually one solves the SCO problem via empirical risk minimization (ERM), i.e., $\theta_{priv} \in \arg\min_{\theta \in \mathcal{C}} \mathcal{L}(\theta; D)$, where $\mathcal{L}(\theta; D) = \frac{1}{|D|}\sum_{d \in D} \ell_{priv}(\theta; d)$, and then uses θ_{priv} as a proxy. Up to a dependence on dimensionality k, in the differentially private setting, a direct translation from ERM to the SCO setting provides optimal rates [13, 17, 18].

We consider the DP-SCO setting with *heterogeneous data*, where there are two datasets D_{priv} (with n_{priv} samples) and D_{pub} (with n_{pub} samples) drawn i.i.d. from the *same distribution*. The private dataset D_{priv} requires privacy protection, whereas the public dataset D_{pub} does not. Our algorithm allows the usage of a separate public loss function ℓ_{pub}. For example, we give a theoretical analysis where ℓ_{priv} and ℓ_{pub} both correspond to the linear regression loss $\frac{1}{2}(y - \langle \mathbf{x}, \theta \rangle)^2$. In practice too, one will likely choose $\ell_{priv} = \ell_{pub}$, but we may clip the gradients of ℓ_{priv} for privacy. In general, ℓ_{pub} can be arbitrary, and we give a theoretical analysis for this more general setting as well.

4.2 Our Results and Technical Overview

We first analyze our algorithm for public-data augmented DP-SCO, PDA-DPMD, in the special case of linear regression, obtaining the following result:

Theorem 72 (Informal Statement of Theorem 76)**.** *Consider the problem of minimizing the mean-squared error in linear regression of a model θ, given samples $d_i = (\mathbf{x}_i, y_i)$. Suppose $||\mathbf{x}||_2 \leq 1$ for all samples, and $|y - \langle \theta^*, \mathbf{x} \rangle| \leq 1$ for the optimal model θ^*. Let $\bar{H}$ be the Hessian of the empirical loss function, and assume $n_{priv} \geq n_{pub}$ and $n_{pub} = \Omega(\frac{\log(k/\delta)}{\lambda_{\min}(\bar{H})})$. Let*

$$\chi = \max\left\{\frac{1}{\lambda_{\min}(\bar{H})}, \lambda_{\max}(\bar{H}) n_{pub}\right\} \cdot \sum_i \min\left\{1, \frac{\log(1/\delta)}{\lambda_i(\bar{H})^2 n_{pub}}\right\}.$$

Then, PDA-DPMD is (ϵ, δ)-DP and we have the following guarantee on $\mathcal{L}(\theta) := \mathbb{E}_{d \sim \tau}\left[\ell(\theta; d)\right]$:

$$\mathbb{E}\left[\mathcal{L}(\theta_{priv}) - \mathcal{L}(\theta^*)\right] \leq \tilde{O}\left(\frac{\chi \log(1/\delta)}{\epsilon^2 n_{priv}^2} + \frac{1}{\lambda_{\min}(\bar{H}) n_{priv}}\right).$$

We note that in this setting, DP-GD obtains an error of roughly $\frac{p}{\lambda_{\min}(\bar{H})\epsilon^2 n_{priv}^2} + \frac{1}{\lambda_{\min}(\bar{H}) n_{priv}}$. If we use PDA-DPMD instead, we can show that given a sufficient number of public samples, the first term depends on the *average* rather than the *minimum* eigenvalue. For example, if $\bar{H}$ has one eigenvalue being $1/k^{1.5}$ and the remaining eigenvalues being $1/p$, then with

$n_{pub} = \tilde{\Omega}(k^{2.5})$ public samples, PDA-DPMD obtains an error of $\frac{k^2}{\epsilon^2 n_{priv}^2} + \frac{k^{1.5}}{n_{priv}}$, whereas DP-GD gets $\frac{k^{2.5}}{\epsilon^2 n_{priv}^2} + \frac{k^{1.5}}{n_{priv}}$. Since PDA-DPMD generalizes DP-GD, unsurprisingly, it still recovers the error bound of DP-GD in the isotropic case.

To prove Theorem 76, we show that the public sample Hessian, private sample Hessian, and population Hessian are all good approximations of each other. This lets us argue the following: for an ellipsoid that is approximately the same shape as $\mathcal{C}$, we can bound the strong convexity parameter of Ψ with respect to this ellipsoid's Minkowski norm. Then, by the concentration of the public sample Hessian, the strong convexity parameters of the population loss and private sample loss with respect to this ellipsoid's Minkowski norm are within a constant factor of the public loss' strong convexity parameter. This lets us use the framework of [79] to obtain the desired excess empirical loss bound, which gives a population loss bound as well by uniform stability.

Theorem 73 (Informal Statement of Theorem 86)**.** *Suppose the private loss is convex and L-Lipschitz, the public loss is* 1*-strongly convex and* 1*-strongly convex with respect to the Minkowski norm of a convex body Q with Gaussian width G_Q, the public gradients have "variance" V^2, and the private and public losses share a minimizer. Then PDA-DPMD has excess empirical loss:*

$$O\left(\frac{VLG_Q\sqrt{\log(1/\delta)}}{\epsilon n_{priv}\sqrt{n_{pub}}}\right)$$

Note that in contrast with the excess empirical loss for DP-GD of $\frac{L||\mathcal{C}||_2\sqrt{k\log(1/\delta)}}{\epsilon n_{priv}}$, (i) We have a dependence on the variance V instead of $||\mathcal{C}||_2$, (ii) We replace $\sqrt{k}$ with G_Q, and (iii) Our loss is further decreased by $1/\sqrt{n_{pub}}$. In particular, G_Q is at most $\sqrt{k}$, but can be constant if e.g. the public loss functions have a much larger strong convexity parameter in all but a constant number of basis directions. Note that if we have $n_{pub} = G_Q^2 \leq k$ public samples, then the dependences on n_{pub} and G_Q cancel out, i.e. this error bound has no explicit dependence on dimension. Using standard techniques, we can turn this into a dimension-independent excess population loss bound (see Theorem 86), again assuming $n_{pub} \geq G_Q^2$. To the best of our knowledge, ours is the first work on augmenting private training with public data to show a theoretical improvement over DP-SGD (here the dependence $\frac{1}{\sqrt{n_{pub}}}$) due to pre-training on public data. In particular, we show pre-training improves the standard DP-SGD bounds even under a totally isotropic geometry. We again use the framework of [79], which shows that the excess empirical loss is a function of the Bregman divergence (defined in the following section) between the initial model θ_0, which we choose to be the minimize of the public loss, and the optimal solution θ^*. By using the bounded variance of the public gradients, we are able to bound this initial Bregman divergence, giving our result.

4.3 Preliminaries

Mirror Maps: A mirror map is a differentiable function $\Psi : \mathbb{R}^k \to \mathbb{R}$ that is strictly convex. Since Ψ is strictly convex and differentiable, $\nabla\Psi : \mathbb{R}^k \to \mathbb{R}^k$ provides a bijection from $\mathbb{R}^k$ to itself. One can view θ as lying in a primal space and $\nabla\Psi(\theta)$ as lying in a dual space. In turn, we could now consider optimizing over the value $\nabla\Psi(\theta)$ in the dual space instead of optimizing over θ in the primal space. Mirror descent does exactly that, performing gradient descent in the dual space by computing the gradient $\mathbf{g}_t = \nabla\ell(\theta_t)$ (where θ_t lies in the primal space), taking a step in the opposite direction in the dual space, and then using the inverse of the mirror map to determine θ_{t+1}. Mirror descent is essentially motivated as minimizing a (linearized) loss plus a Bregman divergence (induced by Ψ) as the regularizer [69]. More formally, similar to proximal gradient descent, mirror descent is equivalent to taking the gradient $\mathbf{g}_t$ and performing the update $\theta_{t+1} = \arg\min_{\theta\in\mathcal{C}}[\eta\langle\mathbf{g}_t, \theta\rangle + B_\Psi(\theta, \theta_t)]$ where $B_\Psi(\theta_1, \theta_2) = \Psi(\theta_1) - \Psi(\theta_2) - \langle\nabla\Psi(\theta_2), \theta_1 - \theta_2\rangle$ is the Bregman divergence generated by Ψ. Note that, if $\Psi(\theta) = \|\theta\|_2^2$, then the Bregman divergence is simply $B_\Psi(\theta_1, \theta_2) = \|\theta_1 - \theta_2\|_2^2$ and mirror descent is equivalent to the usual gradient descent.

Gaussian Width: Given a bounded set $Q \subset \mathbb{R}^k$, the Gaussian width of Q, G_Q, is a measure of how isotropic the set is. G_Q is defined as $\mathbb{E}_{\mathbf{g}\sim N(0,\mathbb{I}_k)} \max_{\mathbf{x}\in Q}\langle\mathbf{g}, \mathbf{x}\rangle$. Although the Gaussian width is well-defined for any bounded set, to gain intuition it suffices to consider defining the Gaussian width of convex sets containing the origin such that $\max_{\mathbf{x}\in Q} ||\mathbf{x}||_2 = 1$; rescaling any such set by a constant changes the Gaussian width by the same constant. If Q is just the unit ℓ_2-ball, the "most isotropic" set satisfying this condition, then we have $G_Q = \Theta(\sqrt{k})$; in particular, since every set Q satisfying $\max_{\mathbf{x}\in Q} ||\mathbf{x}||_2 = 1$ is contained in the ℓ_2-ball, this is the maximum Gaussian width of any such set. On the other hand, if Q is just the line from the origin to a single unit vector, we have $G_Q = \Theta(1)$. More generally, for any ellipsoid centered at the origin whose axes have radii $0 \le r_i \le 1, 1 \le i \le k$, we have that the Gaussian width of this ellipsoid is $\Theta(\sqrt{\sum_{i=1}^k r_i^2})$. As other examples, the Gaussian width of the ℓ_1-ball of radius 1 is roughly $\log k$, and the Gaussian width of the ℓ_∞ ball of radius $1/\sqrt{k}$ is roughly $\sqrt{k}$.

4.4 Algorithm Description

In this section, we present our algorithm Public Data-Assisted Differentially Private Mirror Descent (PDA-DPMD). Given in Algorithm 1, it is a variant of mirror descent using noisy gradients, but we also pre-train on public data and use the public loss as our mirror map Ψ.

Note that Line 5 of PDA-DPMD is equivalent to the following: Choose $\theta_{t+1/2}$ to be the point such that $\nabla\Psi(\theta_{t+1/2}) = \nabla\Psi(\theta_t) - \eta(\mathbf{g}_t + \mathbf{b}_t)$, and then use the Bregman projection $\theta_{t+1} = \arg\min_{\theta\in\mathcal{C}} B_\Psi(\theta, \theta_{t+1/2})$. Intuitively, PDA-DPMD is similar to DP-GD, with the main difference being we apply the gradient steps to $\nabla\Psi(\theta)$ rather than to θ itself. Note that PDA-DPMD reshapes the gradient and noise *automatically* given ℓ_{pub} and D_{pub}. In

Algorithm 1 Public Data-Assisted Differentially Private Mirror Descent (PDA-DPMD)

Input: Public/private datasets D_{pub}, D_{priv} of sizes n_{pub}, n_{priv}, private/public loss functions ℓ_{priv}, ℓ_{pub}, privacy parameters (ϵ, δ), number of iterations T, learning rate $\eta : \{0, 1, \ldots, T-1\} \to \mathbb{R}^+$, constraint set: $\mathcal{C}$, clipping norm L: an upper bound on $\max_{\theta \in \mathcal{C}} ||\nabla \ell_{priv}(\theta)||_2$

```
$\Psi(\theta) := \frac{1}{n_{pub}} \sum\limits_{d \in D_{pub}} \ell_{pub}(\theta; d)$
$\theta_0 \leftarrow \arg\min_{\theta \in \mathcal{C}} \Psi(\theta)$, $\sigma^2 \leftarrow \frac{8L^2 T \log(1/\delta)}{(\epsilon n_{priv})^2}$
for $t = 0, \ldots, T-1$ do
    $\mathbf{g}_t \leftarrow \frac{1}{n_{priv}} \sum_{d \in D_{priv}} \text{clip}\left(\nabla \ell_{priv}(\theta; d), L\right)$, where $\text{clip}\left(\mathbf{v}, L\right) = \mathbf{v} \cdot \min\left\{1, \frac{L}{||\mathbf{v}||_2}\right\}$
    $\theta_{t+1} \leftarrow \arg\min_{\theta \in \mathcal{C}} \left[\eta_t \langle \mathbf{g}_t + \mathbf{b}_t, \theta \rangle + B_\Psi(\theta, \theta_t)\right]$, where $\mathbf{b}_t \sim \mathcal{N}(0, \sigma^2 \cdot \mathbb{I}_k)$
end for
return $\theta_{priv} := \frac{1}{T} \sum\limits_{t=1}^{T} \theta_t$
```

contrast, e.g., private AdaGrad implementations [51, 7] assume knowledge of the geometry of the loss function has already been learned prior to running their algorithms. Also, for an appropriate choice of Ψ, one can recover an algorithm that projects the private gradients to a low-dimensional subspace as in the algorithms of [90] and [51]. From Fact 15, Theorem 16, and Theorem 14 we have the privacy guarantee for Algorithm 1:

Theorem 74. *Algorithm 1 (PDA-DPMD) is (ϵ, δ)-DP with respect to the private dataset D_{priv}.*

4.5 Error Bounds for Linear Regression

In this section, we apply Algorithm 1 (PDA-DPMD) to linear regression – an important example that is still amenable to theoretical analysis. We prove utility guarantees, with supporting simulation.

Problem setup: Given a data sample $d_i = (\mathbf{x}_i, y_i)$, the loss of a model θ is defined as $\ell(\theta; d_i) := \frac{1}{2}(y_i - \langle \theta, \mathbf{x}_i \rangle)^2$. Consider two datasets drawn i.i.d. from a fixed distribution τ: i) The public dataset D_{pub} with n_{pub} data samples, and ii) The private dataset D_{priv} with n_{priv} data samples. In this section, we will denote both the public and private loss functions (ℓ_{pub} and ℓ_{priv} respectively in Algorithm 1) by ℓ.

Assumption 75. *We assume that we are given an initial constraint set*[2] *$\mathcal{C}_0 = \{\theta : ||\theta||_2 \leq r\}$ with $r = O(1)$ that contains the population minimizer, i.e., $\theta^* = \arg\min_{\theta \in \mathbb{R}^k} \mathbb{E}_{d \sim \tau} \ell(\theta; d) \in \mathcal{C}_0$. We further assume that for each feature vector $||\mathbf{x}||_2 \leq 1$, and for each response $|y - \langle \theta^*, \mathbf{x} \rangle| \leq 1$. Let $\bar{H}$ be the Hessian of the loss function $\mathbb{E}_{d \sim \tau}\left[\ell(\theta; d)\right]$. In terms of data set sizes, we assume that $n_{priv} \geq n_{pub}$ and $n_{pub} = \Omega\left(\frac{\log(k/\delta)}{\lambda_{\min}(\bar{H})}\right)$.*

[2] The assumption that $\mathcal{C}_0$ is centered at the origin is without loss of generality.

Excess population risk guarantees: In Theorem 76 we first provide the excess population risk guarantee for Algorithm 1 (PDA-DPMD) under Assumption 75. Furtheremore, in some regimes we demonstrate asymptotic improvement over standard privacy/utility trade-offs for algorithms without access to public data samples.

Theorem 76. *Consider Assumption 75. We run Algorithm 1 (PDA-DPMD) using $L = O(1)$, constraint set $\mathcal{C} = \left\{\theta \in \mathcal{C}_0 : ||\nabla\Psi(\theta)||_2 = O\left(\sqrt{\frac{\log(1/\delta)}{n_{pub}}}\right)\right\}$, and an appropriate choice of η_t and T. Let*

$$\chi = \max\left\{\frac{1}{\lambda_{\min}(\bar{H})}, \lambda_{\max}(\bar{H})n_{pub}\right\} \cdot \sum_i \min\left\{1, \frac{\log(1/\delta)}{\lambda_i(\bar{H})^2 n_{pub}}\right\}.$$

Then, Algorithm 1 is (ϵ, δ)-DP and we have the following guarantee on $\mathcal{L}(\theta) := \mathbb{E}_{d\sim\tau}\left[\ell(\theta; d)\right]$:

$$\mathbb{E}\left[\mathcal{L}(\theta_{priv}) - \mathcal{L}(\theta^*)\right] \leq \tilde{O}\left(\frac{\chi\log(1/\delta)}{\epsilon^2 n_{priv}^2} + \frac{1}{\lambda_{\min}(\bar{H})n_{priv}}\right).$$

The expectation is over D_{pub}, D_{priv}, and the algorithm. $\widetilde{O}(\cdot)$ hides polylog factors in n_{priv}, n_{pub} and $\lambda_{\min}(\bar{H})$.

We note that the idea of using public data to shrink the constraint set $\mathcal{C}$ is similar to the idea used by [19] for mean estimation, though their result iteratively uses each private mean estimate to shrink the constraint set before re-estimating the mean, as opposed to our one-shot approach to shrinking the constraint set using public data.

To interpret χ in Theorem 76, note that a natural setting of parameters to consider would be where the feature vectors (i.e., the $\mathbf{x}$'s) are coming from a mean-zero truncated Gaussian distribution with covariance $\frac{1}{k} \cdot \mathbb{I}$. In this case, all λ_i are $1/k$. If $n_{pub} = \widetilde{\Omega}(p)$, then χ evaluates to k^2, and so we get a bound of $\widetilde{O}\left(\frac{k^2}{\epsilon^2 n_{priv}^2} + \frac{k}{n_{priv}}\right)$, matching the excess population risk of DP-GD. Note that one can still recover DP-GD's loss bound with Algorithm 1 even if $n_{pub} = O(1)$ by instead setting Ψ to be $\frac{1}{2}||\theta||_2^2$ and $\mathcal{C} = \mathcal{C}_0$.

One can also consider a non-isotropic setting, where $\lambda_{\min}(\bar{H})$ is $1/k^{1.5}$ instead of $1/k$, but all other eigenvalues remain roughly $1/k$. In this setting, DP-GD would give an error bound of $\widetilde{O}\left(\frac{k^{2.5}}{\epsilon^2 n_{priv}^2} + \frac{k^{1.5}}{n_{priv}}\right)$. If $n_{pub} = \widetilde{\Omega}(k^{3/2})$, we again match the DP-GD bound. If instead we have $n_{pub} = \widetilde{\Omega}(k^c)$ for $2 \leq c \leq 2.5$, then χ in our loss bound becomes $k^{4.5-c}$, and our loss bound becomes $\widetilde{O}\left(\frac{k^{4.5-c}}{\epsilon^2 n_{priv}^2} + \frac{k^{1.5}}{n_{priv}}\right)$. Once $c = 2.5$, the first term becomes $\frac{k^2}{\epsilon^2 n_{priv}^2}$, matching the corresponding term for the isotropic setting. This shows that PDA-DPMD asymptotically improves over DP-GD under a non-isotropic geometry when given sufficiently many public data samples, with the improvement increasing as the number of public samples increases.

We now prove Theorem 76. We first show that the set $\mathcal{C}$ contains θ^* with high probability. To do this, we need a bound on the gradients of ℓ at θ^*.

Lemma 77. *Under Assumption 75, for all $d \in supp(\tau)$ we have $||\nabla\ell(\theta^*; d)||_2 \leq 1$.*

Proof. The loss function for the pair $d = (\mathbf{x}, y)$ is $||\mathbf{x}||_2^2$-smooth, and minimized (i.e. has gradient $\mathbf{0}$) at the point $\theta^* + \frac{y - \langle \theta^*, \mathbf{x} \rangle}{||\mathbf{x}||_2^2}\mathbf{x}$. In turn, by smoothness and Assumption 75 we have:

$$||\nabla \ell(\theta^*; d)||_2 = \left|\left|\nabla \ell(\theta^*; d) - \nabla \ell(\theta^* + \frac{y - \langle \theta^*, \mathbf{x} \rangle}{||\mathbf{x}||_2^2}\mathbf{x}; d)\right|\right|_2 \leq ||\mathbf{x}||_2^2 \cdot \left|\frac{y - \langle \theta^*, \mathbf{x} \rangle}{||\mathbf{x}||_2^2}\right| \cdot ||\mathbf{x}||_2 \leq 1.$$

□

We can now show that the gradient of the public loss evaluated at θ^* is bounded with high probability, implying it is in $\mathcal{C}$.

Lemma 78. *With probability at least* $1 - \delta$*, for* Ψ *as defined in Algorithm 1, we have* $||\nabla \Psi(\theta^*)||_2 \leq O(\frac{\sqrt{\log(1/\delta)}}{\sqrt{n_{pub}}})$.

Proof. Since θ^* is the population minimizer of ℓ in $\mathbb{R}^p$, and $\mathbb{E}_{d \sim \tau}[\ell(\theta; d)]$ is strongly convex, we have $\mathbb{E}_{d \sim \tau}[\nabla \ell(\theta^*; d)] = \mathbf{0}$. The lemma now follows from a vector Azuma inequality (see e.g. [48]) applied to the vector sum $\nabla \Psi(\theta^*)$, and Lemma 77, which gives that $||\nabla \ell(\theta^*; d)||_2 \leq 1$ for all $d \in supp(\tau)$. □

We can also use the bound on the gradients $\nabla \ell(\theta^*; d)$ to show every loss function is Lipschitz within the constraint set.

Lemma 79. *For all* d*,* $\ell(\theta; d)$ *is* L*-Lipschitz within* $\mathcal{C}_0$ *for* $L = O(1)$*.*

Proof. Each $\ell(\theta; d)$ is 1-smooth, we have $\theta^* \in \mathcal{C}$. In turn, by smoothness and Lemma 77, each $\ell(\theta; d)$ is L-Lipschitz for $L = 1 + 2\,||\mathcal{C}_0||_2$, which is $O(1)$ under Assumption 75, giving the lemma. □

We now show that the sample Hessian approximates the population Hessian for both D_{priv} and D_{pub}, i.e. the geometry of Ψ matches the population loss' geometry and the private sample loss' geometry.

Lemma 80. *Let* $\hat{H}_{pub}$ *be the Hessian of the public loss function* Ψ*, and* $\hat{H}_{priv}$ *be the Hessian of the private loss function* $\frac{1}{n_{priv}} \sum_{d \in D_{priv}} \ell(\theta; d)$*. Then under Assumption 75 with probability* $1 - \delta$*, we have*

$$\frac{1}{2}\bar{H} \preccurlyeq \bar{H} - \frac{\lambda_{\min}(\bar{H})}{2}\mathbb{I} \preccurlyeq \hat{H}_{pub} \preccurlyeq \bar{H} + \frac{\lambda_{\min}(\bar{H})}{2}\mathbb{I} \preccurlyeq 2\bar{H},$$
$$\frac{1}{2}\bar{H} \preccurlyeq \bar{H} - \frac{\lambda_{\min}(\bar{H})}{2}\mathbb{I} \preccurlyeq \hat{H}_{priv} \preccurlyeq \bar{H} + \frac{\lambda_{\min}(\bar{H})}{2}\mathbb{I} \preccurlyeq 2\bar{H}.$$

Proof. The outer inequalities $\frac{1}{2}\bar{H} \preccurlyeq \bar{H} - \frac{\lambda_{\min}(\bar{H})}{2}\mathbb{I}$ and $\bar{H} + \frac{\lambda_{\min}(\bar{H})}{2}\mathbb{I} \preccurlyeq 2\bar{H}$ follow from the $\lambda_{\min}(\bar{H})$-strong convexity of the population loss, i.e. $\lambda_{\min}(\bar{H})\mathbb{I} \preccurlyeq \bar{H}$. So it suffices to prove the inner inequalities.

Let H_d be the Hessian of $\ell(\theta; d)$. By 1-smoothness of H and $\lambda_{\min}(\bar{H})$-strong convexity of $\bar{H}$, we have:

$$\mathbf{0} \preccurlyeq H_d \preccurlyeq \mathbb{I} \qquad \forall d,$$
$$\mathbf{0} \preccurlyeq \bar{H} \preccurlyeq \mathbb{I}.$$

And so:

$$-\mathbb{I} \preccurlyeq H_d - \bar{H} \preccurlyeq \mathbb{I} \qquad \forall d.$$

The inner inequalities now follow from a matrix Bernstein inequality, and the sample complexity lower bounds given in Assumption 75. □

We can now prove our main result.

Proof of Theorem 76. Algorithm 1 is (ϵ, δ)-DP by Theorem 74.

For the utility guarantee, with probability at most 3δ, one of the high probability events described in Lemmas 78 and 79 fails to hold. In this case, by e.g., Lemma 79 we can use a naive bound of $O(||\mathcal{C}_0||_2)$ on the loss. Since δ is negligible, the contribution of this case to the expected excess loss is negligible, so we ignore it here. We now wish to follow the analysis of Theorem A.1 in [79]. To do so, we need to calculate various parameters in that theorem statement:

- By $\lambda_{\min}(\bar{H})$-strong convexity of Ψ, $||\mathcal{C}||_2 = O(\min\{1, \frac{\sqrt{\log(1/\delta)}}{\lambda_{\min}(\bar{H})\sqrt{n_{pub}}}\})$.
- We can assume without loss of generality that $||\theta||_2 \leq r/2$. This is because if we replace r with $2r$ in the definition of $\mathcal{C}_0$, the parameters of the problem do not change asymptotically, but this condition is now enforced. Under this assumption, any line passing through θ^* has an intersection with $\mathcal{C}_0$ of length $\Omega(1)$. Now, by strong convexity and the definition of $\mathcal{C}$, this implies $\mathcal{C}$ is contained within an ellipsoid $\tilde{Q}$ whose axes are the eigenvectors of $\hat{H}_{pub}$, and whose axis lengths are $\Theta(\min\{1, \frac{\sqrt{\log(1/\delta)}}{\lambda_i\sqrt{n_{pub}}}\})$. Furthermore, $\mathcal{C}$ contains $\tilde{Q}$ rescaled in all dimensions by a constant. This means the symmetric convex hull Q of $\mathcal{C}$ is also contained in $\tilde{Q}$, and contains $\tilde{Q}$ rescaled by a constant. So the strong convexity of Ψ with respect to the Q-norm is within a constant factor of the strong convexity of Ψ with respect to the $\tilde{Q}$-norm.

  Now, let $||\cdot||_{\tilde{Q}}$ be the Minkowski $\tilde{Q}$-norm $||\mathbf{x}||_{\tilde{Q}} = \min\{a \in \mathbb{R}_{\geq 0} : \mathbf{x} \in a\tilde{Q}\}$. In the direction of the ith eigenvector, Ψ is $\frac{1}{\lambda_i(\hat{H}_{pub})}$-strongly convex with respect to the norm

$||\cdot||_{Q'}$ for the set $Q' = \{\theta \in \mathbb{R}^p : ||\nabla\Psi(\theta)||_2 \leq 1\}$, so it is $\Theta(\frac{\min\{\lambda_i(\hat{H}_{pub})^2, \log(1/\delta)/n_{pub}\}}{\lambda_i(\hat{H}_{pub})})$-strongly convex with respect to the $\tilde{Q}$-norm, and thus the Q-norm, in this direction. So Δ, the strong convexity parameter of Ψ with respect to the Q-norm is:

$$\Delta = \Theta\left(\min_i\left\{\frac{\min\{\lambda_i(\hat{H}_{pub})^2, \log(1/\delta)/n_{pub}\}}{\lambda_i(\hat{H}_{pub})}\right\}\right) =$$

$$\Theta\left(\min\left\{\lambda_{\min}(\hat{H}_{pub}), \frac{\log(1/\delta)}{\lambda_{\max}(\hat{H}_{pub})n_{pub}}\right\}\right).$$

By Lemma 80, conditioned on the event in that lemma Δ is

$$\Theta\left(\min\left\{\lambda_{\min}(\bar{H}), \frac{\log(1/\delta)}{\lambda_{\max}(\bar{H})n_{pub}}\right\}\right).$$

- By a similar argument to the previous item, we get that the squared Gaussian width $G_{\mathcal{C}}^2$ is at most $G_{\tilde{Q}}^2$, which is

$$O\left(\sum_i \min\left\{1, \frac{\log(1/\delta)}{\lambda_i(\bar{H})^2 n_{pub}}\right\}\right)$$

- By Lemma 80, conditioned on the event in that lemma, the Hessians of the public sample loss, private sample loss, and population loss are constant-approximations of each other.. From the definition of strong convexity with respect to a function (see Section 2.2 of [79]), any quadratic function is 1-strongly convex with respect to itself, and in turn $\Theta(1)$-strongly convex with respect to another quadratic function whose Hessian is within a constant factor of its own, since this implies the Bregman divergences induced by the two functions are also within a constant factor. So the sample private loss $\frac{1}{n_{priv}}\sum_{d\in D_{priv}} \ell(\theta; d)$ is $\Theta(1)$-strongly convex with respect to Ψ.

We will view Algorithm 1 as equivalently using $\Psi' = \frac{1}{\Delta}\Psi$ in place of Ψ, and $\eta_t' = \Delta\eta_t$ in place of η_t. Ψ' is 1-strongly convex with respect to the Q-norm, and the sample private loss is now $\Theta(\Delta)$-strongly convex with respect to Ψ'. Now, following the proof of Theorem A.1 in [79], setting $\eta_t' = 1/\Delta t$ and $T = \frac{||\mathcal{C}||_2^2(\epsilon n_{priv})^2}{||\mathcal{C}||_2^2 + G_{\mathcal{C}}^2}$, conditioned on the high probability events we get an excess empirical loss bound of:

$$\tilde{O}\left(\frac{\log(1/\delta)\max\{\frac{1}{\lambda_{\min}(\bar{H})}, \lambda_{\max}(\bar{H})n_{pub}\} \cdot \sum_i \min\left\{1, \frac{\log(1/\delta)}{\lambda_i(\bar{H})^2 n_{pub}}\right\}}{\epsilon^2 n_{priv}^2}\right).$$

For an excess population loss bound, we need to show uniform stability. Note that since the Hessian of Ψ, $\bar{H}_{pub}$, is fixed everywhere then PDA-DPMD just applies $\bar{H}_{pub}^{-1} \preccurlyeq$

$O(1/\lambda_{\min}(\bar{H})) \cdot \mathbb{I}_k$ to the noisy gradients. This implies that each step of PDA-DPMD is contractive, and thus that the uniform stability parameter of PDA-DPMD is $O(1/\lambda_{\min}(\bar{H}))$ times that of DP-GD using the same settings of η_t, T. The uniform stability of DP-GD on a convex L-Lipschitz loss is $O(\frac{L^2 \sum_t \eta_t}{n})$ (see e.g. Appendix A of [17] for a proof). Plugging in the parameters for our setting, this is $O(\log(\epsilon n_{priv}) n_{priv})$, so PDA-DPMD has uniform stability parameter $O(\log(\epsilon n_{priv})/(\lambda_{\min}(\bar{H}) n_{priv}))$. The expected excess population loss is at most the uniform stability parameter plus the expected excess empirical loss, giving the theorem. □

Local Stability Properties: Since in linear regression the public loss function has the same Hessian $\hat{H}_{pub}$ everywhere, mirror descent effectively is DP-GD, but applying $\hat{H}_{pub}^{-1}$ to the noisy gradient. This allows us to readily characterize the effective noise being added, and show that the noise causes each iterate θ_t to be moved by an amount proportional to $1/\lambda_{\mathbf{v}}$ in a direction where the strong convexity parameter is $\lambda_{\mathbf{v}}$:

Theorem 81. *Let the Hessian of Ψ be $\hat{H}_{pub} = \sum_i \lambda_i \mathbf{v}_i \mathbf{v}_i^\top$, where $\mathbf{v}_i$ are the unit eigenvectors of $\hat{H}_{pub}$. Fix an iteration t as well as starting point θ_t and private gradient $\mathbf{g}_t$ in PDA-DPMD. Let $\bar{\theta}$ be the value of θ_{t+1} after performing the mirror descent update with $\mathbf{b}_t = \mathbf{0}$ at iteration t, and let where $\hat{\theta}$ be the value of the next iterate θ_{t+1} if noise is added. Then for any (unit) direction $\mathbf{v} = \sum_i a_i \mathbf{v}_i$,*

$$\mathbb{E}\left[|\langle \hat{\theta} - \bar{\theta}, \mathbf{v}\rangle|\right] = \eta\sigma\sqrt{\frac{2}{\pi} \cdot \sum_i \left(\frac{a_i}{\lambda_i}\right)^2}.$$

In contrast, for DP-GD, $\mathbb{E}\left[|\langle \hat{\theta} - \bar{\theta}, \mathbf{v}\rangle|\right] = \eta\sigma\sqrt{\frac{2}{\pi}}$ for all $\mathbf{v}$.

Proof of Theorem 81. Let $\mathbf{b}_t$ be the noise added for privacy. Without noise, mirror descent would set θ^* to be such that:

$$-\eta \mathbf{g}_t = \nabla\Psi(\theta^*) - \nabla\Psi(\theta_t).$$

Similarly, given the noisy gradient $\mathbf{g}_t + \mathbf{b}_t$, mirror descent would set $\hat{\theta}$ to be such that:

$$-\eta(\mathbf{g}_t + \mathbf{b}_t) = \nabla\Psi(\hat{\theta}) - \nabla\Psi(\theta_t).$$

We then have:

$$-\eta \mathbf{b}_t = \nabla\Psi(\hat{\theta}) - \nabla\Psi(\theta^*).$$

In turn, recalling that Ψ has a fixed Hessian we have:

$$\hat{\theta} - \theta^* = -\eta \bar{H}_{pub}^{-1} \mathbf{b}_t$$

We can now directly prove the theorem:

$$\mathbb{E}\left[|\langle\hat{\theta}-\theta^*,\mathbf{v}\rangle|\right] = \eta\mathbb{E}\left[|\langle\bar{H}_{pub}^{-1}\mathbf{b}_t,\mathbf{v}\rangle|\right]$$

$$= \eta\mathbb{E}\left[|\langle(\sum_i \frac{1}{\lambda_i}\mathbf{v}_i\mathbf{v}_i^\top)\mathbf{b}_t,\mathbf{v}\rangle|\right] = \eta\mathbb{E}\left[|\sum_i \frac{a_i}{\lambda_i}\langle\mathbf{b}_t,\mathbf{v}_i\rangle|\right]$$

$$= \eta\mathbb{E}\left[|\sum_i N(0,(a_i/\lambda_i)^2)|\right] = \eta\mathbb{E}\left[|N(0,\sum_i(a_i/\lambda_i)^2)|\right] = \sqrt{\frac{2}{\pi}}\cdot\eta\sigma\sqrt{\sum_i\left(\frac{a_i}{\lambda_i}\right)^2}.$$

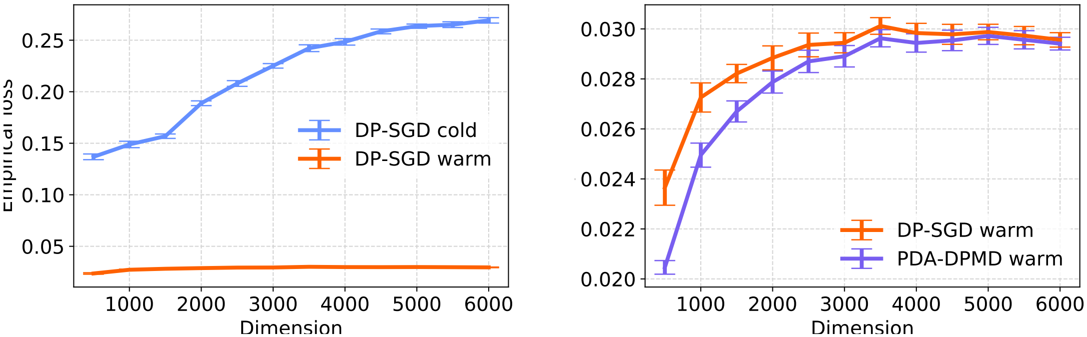
□

Simulation Results: To corroborate our theoretical results with empirical validation, we run PDA-DPMD on synthetic data for the linear regression problem with mean squared error loss. We vary the dimensionality of the problem k from 500 to 6000. For each k, we generate 10,000 private samples and $1.5k$ public samples. The optimal θ^* is sampled from $\mathcal{N}(0,\mathbb{I}_k)$. To introduce a non-isotropic geometry, we sample the feature vector $\mathbf{x}_i$ such that 40 of the first $k/5$ features and 80 of the last $4k/5$ features, chosen uniformly at random, are set to 0.05, and the rest of the features are set to 0. In this way, the expected ℓ_2-norm of each feature vector (and in turn each gradient) is $O(1)$, and thus the effects of clipping should not vary with k. The predicted variable y_i is sampled from $\mathcal{N}(\theta^*\cdot\mathbf{x}_i, 0.01)$ so that the population mean squared error loss is always 0.01, i.e. independent of dimension. We set $\epsilon = 1$, $\delta = 10^{-5}$.

We consider three algorithms: (i) DP-GD with a "cold start", i.e. using a random initialization, (ii) DP-GD with a "warm start" on the model pre-trained with public data, and (iii) PDA-DPMD after pre-training on public data. Note that the exact optimum on the public data can be computed exactly as $\theta^*_{pub} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$. The mirror descent step can also be solved exactly by applying the inverse of the Hessian $\mathbf{X}^\top\mathbf{X}$ to the gradient, since the Hessian is the same everywhere. For numerical stability, we add a small constant times the identity matrix to the Hessian before computing its inverse. We also normalize the Hessian of the loss function so its inverse (which is applied to the gradient before taking a step in PDA-DPMD) has maximum eigenvalue of one. This ensures that if the Hessian were a multiple of the identity matrix, DP-GD and PDA-DPMD would behave exactly the same for the same hyperparameter choice.

We perform a grid search over the learning rate, clipping norm, and number of epochs used and report the best empirical loss. We perform 20 trials for each algorithm and dimension.

Figure 4.1a shows the empirical loss of cold- and warm-start DP-GD. Our results show that pre-training with a number of public samples linear in the dimension allows DP-GD to achieve nearly dimension-independent error. Figure 4.1b compares warm-start DP-GD and PDA-DPMD. The loss of PDA-DPMD is never worse than that of warm-start DP-GD, and can be substantially lower for smaller dimensions. We observed that the ratio of the maximum and minimum eigenvalues of the Hessian $\mathbf{X}^\top\mathbf{X}$ decreases as p increases, which means that the Hessian has poorly-concentrated eigenvalues at small k but gets closer to



(a) Cold start DP-GD vs. warm start DP-GD.

(b) Warm start DP-GD vs. PDA-DPMD.

Figure 4.1: The empirical loss on synthetic linear regression data. The mean and error bars for a 95% confidence interval over 20 runs are plotted. The optimal population loss is 0.01.

the identity matrix as k increases. Since PDA-DPMD recovers warm start DP-GD when the Hessian is the identity, we can expect that PDA-DPMD obtains less of an advantage over DP-GD as the Hessian gets closer to the identity.

4.6 Error Bounds for General Convex Optimization

In this section, we give excess loss bounds for a more general class of loss functions.

We will use the following "bounded variance" assumption on the distribution of the datasets and the public loss function:

Assumption 82. *For some minimizer $\theta^* \in \arg\min_{\theta \in \mathcal{C}} \mathbb{E}_{d \sim \tau}\left[\ell_{priv}(\theta; d)\right]$ we have that θ^* is also the minimizer of $\mathbb{E}_{d \sim \tau}\left[\ell_{pub}(\theta; d)\right]$ in $\mathcal{C}$ and $\mathbb{E}_{d \sim \tau}\left[||\nabla \ell_{pub}(\theta^*; d) - \mathbb{E}_{d \sim \tau}[\nabla \ell_{pub}(\theta^*; d)]||_2^2\right] \leq V^2$. In particular, this implies*
$$\mathbb{E}_{D \sim \tau^{n_{pub}}}\left[\left|\left|\frac{1}{n_{pub}} \sum_{d \in D} \nabla \ell_{pub}(\theta^*; d) - \mathbb{E}_{d \sim \tau}[\nabla \ell_{pub}(\theta^*; d)]\right|\right|_2^2\right] = O\left(\frac{V^2}{n_{pub}}\right).$$

We note that while Assumption 82 to capture the most general setting under which our analysis holds, and thus captures scenarios where ℓ_{pub} and ℓ_{priv} could potentially be very different loss functions, the reader can think of them as differing only slightly. Indeed, Assumption 82 captures several scenarios we might see in practice, such as (i) $\ell_{pub} = \ell_{priv}$ (which can occur if $||\mathcal{C}||_2$ is small), (ii) ℓ_{priv} is the clipped version of ℓ_{pub} (see e.g., [77] for a discussion on the effects of clipping on the loss function), and (iii) ℓ_{pub} is ℓ_{priv} but with a regularizer added.

We first bound the excess empirical loss on the public loss function ℓ_{pub}, compared to the private *population* minimizer θ^* *rather than the empirical minimizer.* This is because the empirical minimizer θ_{emp} of the private loss function could be far away from θ^*, and in turn $\nabla\Psi(\theta_{emp})$ could have much larger norm in expectation than $\nabla\Psi(\theta^*)$. Our PDA-DPMD excess empirical loss bound will be in terms of $\nabla\Psi(\theta)$, where θ is the point we are comparing to, so it is preferable to use $\theta = \theta^*$ for this reason. Our empirical loss bound now follows by using the bounded variance assumption to control the Bregman divergence between the initial iterate and the population minimizer:

Theorem 83. *Suppose the private loss function $\mathcal{L} := \frac{1}{n_{priv}} \sum_{d \in D_{priv}} \ell_{priv}(\theta; d)$ is L-Lipschitz and convex. Suppose ℓ_{pub} is m-strongly convex, and let Q be the minimal convex body containing the origin such that each $\ell_{pub}(\theta; d)$ is 1-strongly convex with respect to the Minkowski norm $||\cdot||_Q$ (defined as $||\mathbf{x}||_Q = \min\{c \in \mathbb{R}_{\geq 0} | \mathbf{x} \in cQ\}$). Then PDA-DPMD is (ϵ, δ)-differentially private with respect to the private database D_{priv} and choosing $\eta_t = \eta$ for all t we have:*

$$\mathbb{E}_{D_{pub} \sim \tau^{n_{pub}}}[\mathcal{L}(\theta_{priv})] - \mathcal{L}(\theta^*) \leq \frac{V^2}{2m\eta T n_{pub}} + \eta \cdot O(L^2 \, ||Q||_2^2 + \sigma^2(G_Q^2 + ||Q||_2^2)).$$

Proof. The privacy guarantee follows as before. Following the analysis of Theorem 3.2 of [79], we have:

$$\mathbb{E}[\mathcal{L}(\theta_{priv})] - \mathcal{L}(\theta^*) \leq \frac{B_\Psi(\theta^*, \theta_0)}{\eta T} + \eta \cdot O(L^2 \, ||Q||_2^2 + \sigma^2(G_Q^2 + ||Q||_2^2), \tag{4.1}$$

Let θ^* in particular be the minimizer satisfying Assumption 82. By m-strong convexity, we have:

$$B_\Psi(\theta^*, \theta_0) = \Psi(\theta^*) - \Psi(\theta_0) - \nabla\Psi(\theta_0) \cdot (\theta^* - \theta_0) \leq \frac{1}{2m} \, ||\nabla\Psi(\theta^*) - \nabla\Psi(\theta_0)||_2^2 \, .$$

Plugging this into Eq. (4.1) and noting that any Ψ we sample is 1-strongly convex with respect to $||\cdot||_Q$, we get:

$$\mathbb{E}[\mathcal{L}(\theta_{priv})] - \mathcal{L}(\theta^*) \leq \frac{\mathbb{E}\left[||\nabla\Psi(\theta^*) - \nabla\Psi(\theta_0)||_2^2\right]}{2m\eta T} + \eta \cdot O(L^2 \, ||Q||_2^2 + \sigma^2(G_Q^2 + ||Q||_2^2)$$

We will show that without loss of generality, we can assume

$$\nabla\Psi(\theta_0) = \mathbf{0}, \mathbb{E}_{d \sim \tau}[\nabla\ell_{pub}(\theta^*; d)] = \mathbf{0}.$$

Once we have this assumption, Assumption 82 completes the proof.

The assumption follows since by convexity of $\mathcal{C}$ we have

$$\langle \nabla\Psi(\theta_0), \theta_0 - \theta^* \rangle \leq 0, \quad \langle \mathbb{E}_{d \sim \tau}[\nabla\ell_{pub}(\theta^*; d)], \theta^* - \theta_0 \rangle \leq 0 \tag{4.2}$$

Then for any choice of Ψ and $\mathcal{C}$ where either θ_0 or θ^* is on the boundary of $\mathcal{C}$, suppose we extend $\mathcal{C}$ infinitesmally along the line $\{\theta_0 + c(\theta^* - \theta) | c \in \mathbb{R}\}$ (i.e., take a point on this line infinitesmally outside of $\mathcal{C}$ and update $\mathcal{C}$ to be the convex hull of itself and this point). Then by (4.2) we have that θ^*, θ_0, defined as the minimizers in $\mathcal{C}$, move apart from each other along this line and in turn by strong convexity the quantity $||\nabla\Psi(\theta^*) - \nabla\Psi(\theta_0)||_2^2$ cannot decrease. This implies that for any fixed ℓ_{pub} and τ, the quantity $\mathbb{E}\left[||\nabla\Psi(\theta^*) - \nabla\Psi(\theta_0)||_2^2\right]$ is maximized for a choice of $\mathcal{C}$ such that $\nabla\Psi(\theta_0) = \mathbf{0}$ and $\mathbb{E}_{d\sim\tau}[\nabla\ell_{pub}(\theta^*; d)] = \mathbf{0}$. □

The above bound is scale-invariant, so to simplify the presentation of this section, we assume, without loss of generality, that $m = 1$ (this also implies Q is contained within the unit ℓ_2-ball, i.e. $||Q||_2 \leq 1$). By rescaling Ψ and η appropriately, we do not affect the behavior of PDA-DPMD, but get that Ψ is 1-strongly convex.

By chaining the following lemma with Theorem 83, we get an excess empirical loss bound with respect to the sample minimizer rather than the population minimizer as desired.

Lemma 84. *Let τ be a distribution over $\mathcal{D}$, $\ell : \mathcal{C}\times\mathcal{D} \to \mathbb{R}$ be a function such that $\ell(\theta; d)$ is L-Lipschitz and convex in θ for any fixed $d \in supp(\tau)$. Let θ^* be the minimizer of $\mathbb{E}_{d\sim\tau}[\ell(\theta; d)]$. Then, we have $\mathbb{E}_{D\sim\tau^n}\left[\mathcal{L}(\theta^*; D) - \min_{\theta\in\mathcal{C}} \mathcal{L}(\theta; D)\right] \leq \frac{L||\mathcal{C}||_2}{\sqrt{n}}$.*

Proof. By convexity, for all $\theta \in \mathcal{C}$ we have $\mathcal{L}(\theta; D) \geq \mathcal{L}(\theta^*; D) + \langle\nabla\mathcal{L}(\theta^*; D), \theta - \theta^*\rangle$. Note that by optimality of θ^* and convexity, for all $\theta \in \mathcal{C}$ we have $\langle\mathbb{E}_{d\sim\tau}[\nabla\ell(\theta^*; d)], \theta - \theta^*\rangle \geq 0$. In turn, by the Cauchy-Schwarz inequality we can conclude that $\mathcal{L}(\theta^*; D) - \min_{\theta\in\mathcal{C}} \mathcal{L}(\theta; D)$ is always upper bounded by $||\mathcal{C}||_2 \cdot ||\nabla\mathcal{L}(\theta^*; D) - \mathbb{E}_{d\sim\tau}[\nabla\ell(\theta^*; d)]||_2$. By L-Lipschitzness of each $\ell(\theta; d)$ we have:

$$\mathbb{E}_{D\sim\tau^n}\left[||\nabla\mathcal{L}(\theta^*; D) - \mathbb{E}_{d\sim\tau}[\nabla\ell(\theta^*; d)]||_2\right] \leq \frac{L}{\sqrt{n}},$$

Which completes the proof. □

Excess Population Risk of PDA-DPMD: We now translate our excess empirical loss bound to a excess population loss. We use Lemma F.5 of [13], restated in Lemma 85 for convenience, which provides a black-box translation from empirical loss to population loss:

Lemma 85. *For any (ϵ, δ)-DP algorithm for minimizing $\frac{1}{n_{priv}} \sum_{d\in D_{priv}} \ell(\theta; d)$ over $\mathcal{C}$, the expected excess population loss exceeds the expected excess empirical loss by $O(L\,||\mathcal{C}||_2\,\epsilon + ||\mathcal{C}||_2^2\,\delta)$.*

Given this lemma, it is straightforward to derive excess population loss bounds:

Theorem 86. *For $\eta = \frac{V}{L\sqrt{Tn_{pub}}}$, $T = \frac{\epsilon^2 n_{priv}^2}{G_Q^2 \log(1/\delta)}$, and setting $\epsilon = \frac{\sqrt{VG_Q}\log^{1/4}(1/\delta)}{\sqrt{n_{priv}}n_{pub}^{1/4}||\mathcal{C}||_2}$, the expected population loss of PDA-DPMD is*

$$O\left(\frac{L\sqrt{V\,||\mathcal{C}||_2}\sqrt{G_Q}\log^{1/4}(1/\delta)}{\sqrt{n_{priv}}n_{pub}^{1/4}} + \frac{L\,||\mathcal{C}||_2}{\sqrt{n_{priv}}} + ||\mathcal{C}||_2^2\,\delta\right).$$

Theorem 86 follows immediately from Theorem 83, Lemma 84, and Lemma 85. Note that if $n_{pub} \geq G_Q^2$, which is at most k, then the above bound has no explicit dependence on dimension. For comparison, if we were to only train on public data, the standard non-private excess population loss bound has dependence $O(1/\sqrt{n_{pub}})$ (and no dependence on dimension). So in the regime where $n_{pub} \approx G_Q^2$ and $n_{priv} \gg n_{pub}$, our bound is asymptotically much better than the baseline of training only on public examples.

Local Stability Properties of PDA-DPMD: If the public loss function has a Hessian everywhere that has the same eigenvectors regardless of location (but perhaps different eigenvalues), we can generalize Theorem 81:

Theorem 87. *Suppose for the public loss function Ψ, its Hessian is defined everywhere, and for a fixed orthonormal basis $\{\mathbf{v}_i\}_i$, the Hessian at every point can be written as $\sum_i w_i(\theta)\mathbf{v}_i\mathbf{v}_i^\top$ for scalar functions $w_i : \mathbb{R}^k \to \mathbb{R}^+$ such that for all i, θ, we have $w_i(\theta) \geq m$. Fix an iteration t as well as private gradient $\mathbf{g}_t$ in PDA-DPMD. Let θ^* be the value of θ_{t+1} after performing the mirror descent update with $\mathbf{b}_t = \mathbf{0}$ at iteration t, and let $\{\tilde{w}_i\}_i, c \geq 0$ be such that for each i the smallest value of $w_i(\theta)$ in the ellipsoid $E := (\sum_i \frac{1}{\tilde{w}_i}\mathbf{v}_i\mathbf{v}_i^\top)B_R$ (where B_R is the ℓ_2 ball of radius $R := \eta(1+c)\sqrt{k}\sigma$), centered at θ^*, is at least $\tilde{w}_i$. Then for any (unit) direction $\mathbf{v} = \sum_i a_i\mathbf{v}_i$,*

$$\mathbb{E}\left[|\langle\hat{\theta} - \theta^*, \mathbf{v}\rangle|\right] \leq \eta\sigma\left[\sqrt{\frac{2}{\pi}} \cdot \sqrt{\sum_i \left(\frac{a_i}{\tilde{w}_i}\right)^2} + \frac{3(1+c)^2}{2m} \cdot e^{-c^2k/2}\right],$$

where $\hat{\theta}$ is the value of the next iterate θ_{t+1} if noise is added.

Proof. Let $\mathbf{b}_t$ be the noise added for privacy. Without noise, mirror descent would set θ^* to be such that:

$$-\eta\mathbf{g}_t = \nabla\Psi(\theta^*) - \nabla\Psi(\theta_t).$$

Similarly, given the noisy gradient $\mathbf{g}_t + \mathbf{b}_t$, mirror descent would set $\hat{\theta}$ to be such that:

$$-\eta(\mathbf{g}_t + \mathbf{b}_t) = \nabla\Psi(\hat{\theta}) - \nabla\Psi(\theta_t).$$

We then have:

$$-\eta\mathbf{b}_t = \nabla\Psi(\hat{\theta}) - \nabla\Psi(\theta^*).$$

Since we assume the Hessian of Ψ is defined everywhere, we have that $\nabla\Psi(\hat{\theta}) - \nabla\Psi(\theta^*) = \nabla^2\Psi(\tilde{\theta})(\hat{\theta} - \theta^*)$ for some $\tilde{\theta}$ on the line between $\hat{\theta}$ and θ^*. In turn, we have:

$$\hat{\theta} - \theta^* = -\eta(\nabla^2\Psi(\tilde{\theta}))^{-1}\mathbf{b}_t$$

The norm x of $\mathbf{b}_t$ sampled from $N(0, \sigma^2\mathbb{I}_k)$ has the chi distribution, i.e. pdf proportional to $(x/\sigma)^{k-1}e^{-(x/\sigma)^2/2}$. In particular, this gives the following standard tail bound:

$$\Pr[\|X\|_2 > (1+c)\sqrt{k}\sigma] \leq \exp(-c^2 p/2).$$

So with probability at least $1 - e^{-c^2k/2}$, we have the event $\mathcal{E}$ that $\|\mathbf{b}_t\|_2$ is at most $(1+c)\sqrt{k}\sigma$. Conditioned on this event, we have that $\hat{\theta}$, and thus $\tilde{\theta}$, is in E, i.e. the lower bounds $\tilde{w}_i$ apply to $\nabla^2\Psi(\tilde{\theta})$. Since conditioning on $\mathcal{E}$ only decreases the expectation of $|\langle\hat{\theta}-\theta^*, \mathbf{v}\rangle|$, we have:

$$\mathbb{E}\left[\eta|\langle\hat{\theta}-\theta^*,\mathbf{v}\rangle||\mathcal{E}\right]\cdot\Pr[\mathcal{E}] = \eta\mathbb{E}\left[|\langle(\nabla^2\Psi(\tilde{\theta}))^{-1}\mathbf{b}_t,\mathbf{v}\rangle||\mathcal{E}\right]\cdot\Pr[\mathcal{E}]$$

$$\leq \eta\mathbb{E}\left[|\langle(\sum_i \frac{1}{\tilde{w}_i}\mathbf{v}_i\mathbf{v}_i^\top)\mathbf{b}_t,\mathbf{v}\rangle||\mathcal{E}]\right]\cdot\Pr[\mathcal{E}] \leq \eta\mathbb{E}\left[|\langle(\sum_i \frac{1}{\tilde{w}_i}\mathbf{v}_i\mathbf{v}_i^\top)\mathbf{b}_t,\mathbf{v}\rangle|\right] = \mathbb{E}\left[|\sum_i \frac{a_i}{\tilde{w}_i}\langle\mathbf{b}_t,\mathbf{v}_i\rangle|\right]$$

$$= \eta\mathbb{E}\left[|\sum_i N(0,(a_i/\tilde{w}_i)^2)|\right] = \eta\mathbb{E}\left[|N(0,\sum_i(a_i/\tilde{w}_i)^2)|\right] = \sqrt{\frac{2}{\pi}}\cdot\eta\sigma\sqrt{\sum_i\left(\frac{a_i}{\tilde{w}_i}\right)^2}.$$

When $\mathcal{E}$ does not happen, we have $w_i(\theta) \geq m$ everywhere. So we have:

$$\mathbb{E}\left[\eta|\langle\hat{\theta}-\theta^*,v\rangle||\neg\mathcal{E}\right]\cdot\Pr[\neg\mathcal{E}] = \eta\mathbb{E}\left[|\langle(\nabla^2\Psi(\tilde{\theta}))^{-1}b,v\rangle||\neg\mathcal{E}\right]\cdot\Pr[\neg\mathcal{E}]$$

$$\leq \eta\cdot\frac{1}{m}\mathbb{E}\left[|\langle b,v\rangle||\neg\mathcal{E}\right]\cdot e^{-c^2k/2}$$

To determine $\mathbb{E}\left[|\langle\mathbf{b}_t,\mathbf{v}\rangle||\neg\mathcal{E}\right]$, note that the distribution of $\langle\mathbf{b}_t,\mathbf{v}\rangle$ conditioned on $\neg\mathcal{E}$ is equivalent to the distribution of the first coordinate of $\mathbf{b}_t$ conditioned on $\neg\mathcal{E}$. We can sample $\mathbf{b}_t$ by first sampling its norm $\|\mathbf{b}_t\|_2$ conditioned on $\neg\mathcal{E}$, and then sampling a point on the sphere with radius $\|\mathbf{b}_t\|_2$ (no conditioning is required here). The expected absolute value of any coordinate $(\mathbf{b}_t)_i$ given $\|\mathbf{b}_t\|_2$ can be bounded as:

$$\mathbb{E}\left[|(\mathbf{b}_t)_i|\right] \leq \sqrt{\mathbb{E}\left[(\mathbf{b}_t)_i^2\right]} = \|\mathbf{b}_t\|_2/\sqrt{k}.$$

The inequality is Jensen's inequality, and the equality uses the fact that the coordinates $\mathbf{b}_i$ on the sphere are identically distributed, and so we have:

$$p\cdot\mathbb{E}\left[(\mathbf{b}_t)_i^2\right] = \mathbb{E}\left[\sum_i(\mathbf{b}_t)_i^2\right] = \|\mathbf{b}_t\|_2^2.$$

We now just need to bound the expectation of $\|\mathbf{b}_t\|_2$, given that it is at least R. Since the distribution of $\|\mathbf{b}_t\|_2/\sigma$ has pdf proportional to $x^{k-1}e^{-x^2/2}$, this expectation is σ times:

$$\frac{\int_{(1+c)\sqrt{k}}^\infty x^k e^{-x^2/2}}{\int_{(1+c)\sqrt{k}}^\infty x^{k-1}e^{-x^2/2}} = \frac{\Gamma((k+1)/2)(1-P((k+1)/2,(1+c)^2k/2))}{\sqrt{2}\Gamma(k/2)(1-P(k/2,(1+c)^2k/2))}.$$

Where Γ is the gamma function and P is the regularized gamma function. Analytically, we can verify that $\frac{\Gamma((k+1)/2)}{\Gamma(k/2)} \leq \sqrt{k/2}$ for all $k \geq 1$, and $\frac{(1-P((k+1)/2,(1+c)^2k/2))}{(1-P(k/2,(1+c)^2k/2))} \leq 3(1+c)^2$ for all $k \geq 1$. So we get:

$$\mathbb{E}\left[||\mathbf{b}_t||_2 \,|\neg\mathcal{E}\right] \leq \frac{3(1+c)^2}{2}\sqrt{k}\sigma$$

Putting it all together, we get:

$$\mathbb{E}\left[\eta|\langle\hat{\theta}-\theta^*, v\rangle||\neg\mathcal{E}\right] \cdot \Pr[\neg\mathcal{E}] \leq \eta \cdot \frac{1}{m} \cdot \frac{3(1+c)^2}{2} \cdot e^{-c^2k/2} \cdot \sigma$$

Now applying the law of total expectation gives the theorem statement. □

Note that the condition $w_i(\theta) \geq m$ can be enforced by adding an ℓ_2-regularizer to the public loss function (since mirror descent only cares about differences in the gradients of the public loss function, the private training phase of PDA-DPMD behaves the same regardless of where this regularizer is centered).

Bibliography

[1] Martın Abadi et al. "Deep Learning with Differential Privacy". In: *Proc. of the 2016 ACM SIGSAC Conf. on Computer and Communications Security (CCS'16)*. 2016, pp. 308–318.

[2] Naman Agarwal et al. "Efficient full-matrix adaptive regularization". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 102–110.

[3] Noga Alon, Raef Bassily, and Shay Moran. "Limits of private learning with access to public data". In: *arXiv preprint arXiv:1910.11519* (2019).

[4] Ehsan Amid et al. "Public Data-Assisted Mirror Descent for Private Model Training". In: *CoRR* abs/2112.00193 (2021). arXiv: 2112.00193. URL: https://arxiv.org/abs/2112.00193.

[5] David Applegate and Ravi Kannan. "Sampling and Integration of near Log-Concave Functions". In: *Proceedings of the Twenty-Third Annual ACM Symposium on Theory of Computing*. STOC '91. New Orleans, Louisiana, USA: Association for Computing Machinery, 1991, pp. 156–163. ISBN: 0897913973. DOI: 10.1145/103418.103439. URL: https://doi.org/10.1145/103418.103439.

[6] Hilal Asi and John C. Duchi. "Near Instance-Optimality in Differential Privacy". In: *arXiv preprint arXiv:2005.10630* (2020).

[7] Hilal Asi et al. "Private adaptive gradient methods for convex optimization". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 383–392.

[8] Brendan Avent et al. "{BLENDER}: Enabling local search with a hybrid differential privacy model". In: *26th {USENIX} Security Symposium ({USENIX} Security 17)*. 2017.

[9] Jordan Awan et al. "Benefits and Pitfalls of the Exponential Mechanism with Applications to Hilbert Spaces and Functional PCA". In: *Proceedings of the 36th International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR, 2019, pp. 374–384. URL: http://proceedings.mlr.press/v97/awan19a.html.

[10] Dominique Bakry and Michel Émery. "Diffusions hypercontractives". fre. In: *Séminaire de probabilités de Strasbourg* 19 (1985), pp. 177–206. URL: http://eudml.org/doc/113511.

[11] M. Balcan, T. Dick, and E. Vitercik. "Dispersion for Data-Driven Algorithm Design, Online Learning, and Private Optimization". In: *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*. 2018, pp. 603–614.

[12] Raef Bassily, Shay Moran, and Anupama Nandi. "Learning from mixtures of private and public populations". In: *arXiv preprint arXiv:2008.00331* (2020).

[13] Raef Bassily, Adam Smith, and Abhradeep Thakurta. "Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds". In: *Proc. of the 2014 IEEE 55th Annual Symp. on Foundations of Computer Science (FOCS)*. 2014, pp. 464–473.

[14] Raef Bassily, Om Thakkar, and Abhradeep Thakurta. "Model-Agnostic Private Learning". In: *NeurIPS*. 2018.

[15] Raef Bassily, Abhradeep Guha Thakurta, and Om Dipakbhai Thakkar. "Model-agnostic private learning". In: *Advances in Neural Information Processing Systems* (2018).

[16] Raef Bassily et al. "Private Query Release Assisted by Public Data". In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 2020, pp. 695–703. URL: https://proceedings.mlr.press/v119/bassily20a.html.

[17] Raef Bassily et al. "Private stochastic convex optimization with optimal rates". In: *Advances in Neural Information Processing Systems*. 2019, pp. 11279–11288.

[18] Raef Bassily et al. "Stability of Stochastic Gradient Descent on Nonsmooth Convex Losses". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 4381–4391. URL: https://proceedings.neurips.cc/paper/2020/file/2e2c4bf7ceaa4712a72dd5ee136dc9a8-Paper.pdf.

[19] Sourav Biswas et al. "CoinPress: Practical Private Mean and Covariance Estimation". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 14475–14485. URL: https://proceedings.neurips.cc/paper/2020/file/a684eceee76fc522773286a895bc8436-Paper.pdf.

[20] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013. ISBN: 9780199535255. URL: https://books.google.com/books?id=koNqWRluhP0C.

[21] Nicholas Carlini et al. "Extracting Training Data from Large Language Models". In: *USENIX Security Symposium*. 2021.

[22] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. "Differentially private empirical risk minimization". In: *Journal of Machine Learning Research* 12.Mar (2011), pp. 1069–1109.

[23] Kamalika Chaudhuri, Anand Sarwate, and Kaushik Sinha. "A Near-Optimal Algorithm for Differentially-Private Principal Components". In: *Journal of Machine Learning Research* 14.53 (2013), pp. 2905–2943. URL: http://jmlr.org/papers/v14/chaudhuri13a.html.

[24] Rishav Chourasia, Jiayuan Ye, and Reza Shokri. "Differential Privacy Dynamics of Langevin Diffusion and Noisy Gradient Descent". In: *CoRR* abs/2102.05855 (2021). arXiv: 2102.05855. URL: https://arxiv.org/abs/2102.05855.

[25] Yuval Dagan and Gil Kur. *A bounded-noise mechanism for differential privacy*. 2020. arXiv: 2012.03817 [cs.DS].

[26] Arnak Dalalyan. "Theoretical guarantees for approximate sampling from smooth and log-concave densities". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79 (June 2017). DOI: 10.1111/rssb.12183.

[27] Christos Dimitrakakis et al. "Robust and Private Bayesian Inference". In: *Algorithmic Learning Theory*. Ed. by Peter Auer et al. Cham: Springer International Publishing, 2014, pp. 291–305. ISBN: 978-3-319-11662-4.

[28] Alexey Dosovitskiy et al. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations*. 2021. URL: https://openreview.net/forum?id=YicbFdNTTy.

[29] John Duchi, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization." In: *Journal of machine learning research* 12.7 (2011).

[30] Alain Durmus and Eric Moulines. "High-dimensional Bayesian inference via the unadjusted Langevin algorithm". In: *Bernoulli* 25 (Nov. 2019), pp. 2854–2882. DOI: 10.3150/18-BEJ1073.

[31] Cynthia Dwork and Vitaly Feldman. "Privacy-preserving prediction". In: *Conference On Learning Theory*. PMLR. 2018, pp. 1693–1702.

[32] Cynthia Dwork and Aaron Roth. "The Algorithmic Foundations of Differential Privacy". In: *Found. Trends Theor. Comput. Sci.* 9.3–4 (Aug. 2014), pp. 211–407. ISSN: 1551-305X. DOI: 10.1561/0400000042. URL: https://doi.org/10.1561/0400000042.

[33] Cynthia Dwork et al. "Calibrating Noise to Sensitivity in Private Data Analysis". In: *Proc. of the Third Conf. on Theory of Cryptography (TCC)*. 2006, pp. 265–284. URL: http://dx.doi.org/10.1007/11681878%5C_14.

[34] Cynthia Dwork et al. "On the Complexity of Differentially Private Data Release: Efficient Algorithms and Hardness Results". In: *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*. STOC '09. Bethesda, MD, USA: Association for Computing Machinery, 2009, pp. 381–390. ISBN: 9781605585062. DOI: 10.1145/1536414.1536467. URL: https://doi.org/10.1145/1536414.1536467.

[35] Murat A. Erdogdu and Rasa Hosseinzadeh. “A Brief Note on the Convergence of Langevin Monte Carlo in Chi-Square Divergence”. In: *CoRR* abs/2007.11612 (2020). arXiv: 2007.11612. URL: https://arxiv.org/abs/2007.11612.

[36] Tim van Erven and Peter Harremos. “Rényi Divergence and Kullback-Leibler Divergence”. In: *IEEE Transactions on Information Theory* 60.7 (2014), pp. 3797–3820. DOI: 10.1109/TIT.2014.2320500.

[37] Dan Feldman et al. “Private Coresets”. In: *Proceedings of the Forty-First Annual ACM Symposium on Theory of Computing*. STOC ’09. Bethesda, MD, USA: Association for Computing Machinery, 2009, pp. 361–370. ISBN: 9781605585062. DOI: 10.1145/1536414.1536465. URL: https://doi.org/10.1145/1536414.1536465.

[38] Vitaly Feldman et al. “Privacy amplification by iteration”. In: *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2018, pp. 521–532.

[39] James Foulds et al. “On the Theory and Practice of Privacy-Preserving Bayesian Data Analysis”. In: *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*. UAI’16. Jersey City, New Jersey, USA: AUAI Press, 2016, pp. 192–201. ISBN: 9780996643115.

[40] Arun Ganesh and Kunal Talwar. “Faster Differentially Private Samplers via Rényi Divergence Analysis of Discretized Langevin MCMC”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 7222–7233. URL: https://proceedings.neurips.cc/paper/2020/file/50cf0fe63e0ff857e1c9d01d827267ca-Paper.pdf.

[41] Arun Ganesh and Jiazheng Zhao. “Privately Answering Counting Queries with Generalized Gaussian Mechanisms”. In: *2nd Symposium on Foundations of Responsible Computing (FORC 2021)*. Ed. by Katrina Ligett and Swati Gupta. Vol. 192. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2021, 1:1–1:18. ISBN: 978-3-95977-187-0. DOI: 10.4230/LIPIcs.FORC.2021.1. URL: https://drops.dagstuhl.de/opus/volltexte/2021/13869.

[42] Badih Ghazi, Ravi Kumar, and Pasin Manurangsi. *On Avoiding the Union Bound When Answering Multiple Differentially Private Queries*. 2020. arXiv: 2012.09116 [cs.DS].

[43] Anupam Gupta et al. “Differentially Private Combinatorial Optimization”. In: *Proceedings of the Twenty-First Annual ACM-SIAM Symposium on Discrete Algorithms*. SODA ’10. Austin, Texas: Society for Industrial and Applied Mathematics, 2010, pp. 1106–1125. ISBN: 9780898716986.

[44] Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. “Gradient Descent Happens in a Tiny Subspace”. In: *CoRR* abs/1812.04754 (2018). URL: http://arxiv.org/abs/1812.04754.

[45] Moritz Hardt, Benjamin Recht, and Yoram Singer. "Train Faster, Generalize Better: Stability of Stochastic Gradient Descent". In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. ICML'16. New York, NY, USA: JMLR.org, 2016, pp. 1225–1234.

[46] Moritz Hardt and Guy N. Rothblum. "A Multiplicative Weights Mechanism for Privacy-Preserving Data Analysis". In: *2010 IEEE 51st Annual Symposium on Foundations of Computer Science*. 2010, pp. 61–70. DOI: 10.1109/FOCS.2010.85.

[47] Moritz Hardt and Kunal Talwar. "On the Geometry of Differential Privacy". In: *Proceedings of the Forty-Second ACM Symposium on Theory of Computing*. STOC '10. Cambridge, Massachusetts, USA: Association for Computing Machinery, 2010, pp. 705–714. ISBN: 9781450300506. DOI: 10.1145/1806689.1806786. URL: https://doi.org/10.1145/1806689.1806786.

[48] Thomas P. Hayes. "A large-deviation inequality for vector-valued martingales". In: 2003.

[49] N.L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*. John Wiley & Sons Incorporated, 1995. ISBN: 9780471127918. URL: https://books.google.com/books?id=q03oAAAACAAJ.

[50] Peter Kairouz et al. "(Nearly) Dimension Independent Private ERM with AdaGrad Rates
via Publicly Estimated Subspaces". In: *COLT*. 2021.

[51] Peter Kairouz et al. "Dimension Independence in Unconstrained Private ERM via Adaptive Preconditioning". In: *CoRR* abs/2008.06570 (2020). arXiv: 2008.06570. URL: https://arxiv.org/abs/2008.06570.

[52] Peter Kairouz et al. "Practical and private (deep) learning without sampling or shuffling". In: *ICML*. 2021.

[53] Michael Kapralov and Kunal Talwar. "On differentially private low rank approximation". In: *Proceedings of the 2013 Annual ACM-SIAM Symposium on Discrete Algorithms*. 2013, pp. 1395–1414. DOI: 10.1137/1.9781611973105.101. eprint: https://epubs.siam.org/doi/pdf/10.1137/1.9781611973105.101. URL: https://epubs.siam.org/doi/abs/10.1137/1.9781611973105.101.

[54] Michael Kearns. "Efficient Noise-Tolerant Learning from Statistical Queries". In: *J. ACM* 45.6 (1998), pp. 983–1006. ISSN: 0004-5411. DOI: 10.1145/293347.293351. URL: https://doi.org/10.1145/293347.293351.

[55] D.A. Klain, G.C. Rota, and L.A.R. di Brozolo. *Introduction to Geometric Probability*. Lezioni Lincee. Cambridge University Press, 1997. ISBN: 9780521596541. URL: https://books.google.com/books?id=Q1ytkNM6BtAC.

[56] Xuechen Li et al. "Large Language Models Can Be Strong Differentially Private Learners". In: *arXiv preprint arXiv:2110.05679* (2021).

[57] Fang Liu. "Generalized Gaussian Mechanism for Differential Privacy". In: *IEEE Transactions on Knowledge and Data Engineering* 31 (2019), pp. 747–756.

[58] Terrance Liu et al. "Leveraging Public Data for Practical Private Query Release". In: *arXiv preprint arXiv:2102.08598* (2021).

[59] L. Lovász and S. Vempala. "Fast Algorithms for Logconcave Functions: Sampling, Rounding, Integration and Optimization". In: *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*. 2006, pp. 57–68. DOI: 10.1109/FOCS.2006.28.

[60] László Lovász and Santosh Vempala. "The geometry of logconcave functions and sampling algorithms". In: *Random Structures & Algorithms* 30.3 (2007), pp. 307–358. DOI: 10.1002/rsa.20135. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/rsa.20135. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/rsa.20135.

[61] Yi-An Ma et al. "Is There an Analog of Nesterov Acceleration for MCMC?" In: *CoRR* abs/1902.00996 (2019). arXiv: 1902.00996. URL: http://arxiv.org/abs/1902.00996.

[62] Frank McSherry and Kunal Talwar. "Mechanism Design via Differential Privacy". In: *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*. 2007, pp. 94–103. DOI: 10.1109/FOCS.2007.66.

[63] Chris Merriman. "Microsoft reminds privacy-concerned Windows 10 beta testers that they're volunteers". In: *The Inquirer. http://www.theinquirer.net/2374302* (2014). URL: http://%20www.theinquirer.net/%202374302.

[64] Kentaro Minami et al. "Differential Privacy without Sensitivity". In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee et al. Curran Associates, Inc., 2016, pp. 956–964. URL: http://papers.nips.cc/paper/6050-differential-privacy-without-sensitivity.pdf.

[65] Darakhshan J. Mir. "Differential privacy: an exploration of the privacy-utility landscape". PhD thesis. Rutgers University, 2013.

[66] Ilya Mironov. "Rényi differential privacy". In: *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*. IEEE. 2017, pp. 263–275.

[67] Anupama Nandi and Raef Bassily. "Privately Answering Classification Queries in the Agnostic PAC Model". In: *Proceedings of the 31st International Conference on Algorithmic Learning Theory*. Ed. by Aryeh Kontorovich and Gergely Neu. Vol. 117. Proceedings of Machine Learning Research. PMLR, 2020, pp. 687–703. URL: https://proceedings.mlr.press/v117/nandi20a.html.

[68] Arvind Narayanan and Vitaly Shmatikov. "Robust De-anonymization of Large Sparse Datasets". In: *2008 IEEE Symposium on Security and Privacy (S&P 2008), 18-21 May 2008, Oakland, California, USA*. IEEE Computer Society, 2008, pp. 111–125. DOI: 10.1109/SP.2008.33. URL: https://doi.org/10.1109/SP.2008.33.

[69] A. Nemirovsky and D. Yudin. *Problem Complexity and Method Efficiency in Optimization*. Wiley & Sons, New York, 1983.

[70] Nicolas Papernot et al. “Scalable private learning with pate”. In: *arXiv preprint arXiv:1802.08908* (2018).

[71] Nicolas Papernot et al. “Semi-supervised knowledge transfer for deep learning from private training data”. In: *arXiv preprint arXiv:1610.05755* (2016).

[72] Nicolas Papernot et al. “Tempered Sigmoid Activations for Deep Learning with Differential Privacy”. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 35.10 (2021), pp. 9312–9321. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17123.

[73] Matthew Reimherr and Jordan Awan. “KNG: The K-Norm Gradient Mechanism”. In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 10208–10219. URL: http://papers.nips.cc/paper/9210-kng-the-k-norm-gradient-mechanism.pdf.

[74] Alfréd Rényi. *On measures of entropy and information*. English. Proc. 4th Berkeley Symp. Math. Stat. Probab. 1, 547-561 (1961). 1961.

[75] Gareth O. Roberts and Richard L. Tweedie. “Exponential convergence of Langevin distributions and their discrete approximations”. In: *Bernoulli* 2.4 (Dec. 1996), pp. 341–363. URL: https://projecteuclid.org:443/euclid.bj/1178291835.

[76] Shuang Song, Kamalika Chaudhuri, and Anand D Sarwate. “Stochastic gradient descent with differentially private updates”. In: *2013 IEEE Global Conference on Signal and Information Processing*. IEEE. 2013, pp. 245–248.

[77] Shuang Song et al. “Evading the Curse of Dimensionality in Unconstrained Private GLMs”. In: *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*. Ed. by Arindam Banerjee and Kenji Fukumizu. Vol. 130. Proceedings of Machine Learning Research. PMLR, 2021, pp. 2638–2646. URL: https://proceedings.mlr.press/v130/song21a.html.

[78] Thomas Steinke and Jonathan Ullman. “Between Pure and Approximate Differential Privacy”. In: *Journal of Privacy and Confidentiality* 7.2 (2017). DOI: 10.29012/jpc.v7i2.648.

[79] Kunal Talwar, Abhradeep Thakurta, and Li Zhang. “Private empirical risk minimization beyond the worst case: The effect of the constraint set geometry”. In: *arXiv preprint arXiv:1411.5417* (2014).

[80] Jun Tang et al. “Privacy Loss in Apple’s Implementation of Differential Privacy on MacOS 10.12”. In: *CoRR* abs/1709.02753 (2017). arXiv: 1709.02753. URL: http://arxiv.org/abs/1709.02753.

[81] Florian Tramer and Dan Boneh. "Differentially Private Learning Needs Better Features (or Much More Data)". In: *International Conference on Learning Representations*. 2020.

[82] Santosh Vempala and Andre Wibisono. "Rapid Convergence of the Unadjusted Langevin Algorithm: Isoperimetry Suffices". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 8092–8104.

[83] Di Wang, Minwei Ye, and Jinhui Xu. "Differentially Private Empirical Risk Minimization Revisited: Faster and More General". In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper/2017/file/f337d999d9ad116a7b4f3d409fcc6480-Paper.pdf.

[84] Feng-Yu Wang and Jian Wang. "Functional inequalities for convolution probability measures". In: *Ann. Inst. H. Poincaré Probab. Statist.* 52.2 (May 2016), pp. 898–914. DOI: 10.1214/14-AIHP659. URL: https://doi.org/10.1214/14-AIHP659.

[85] Jun Wang and Zhi-Hua Zhou. "Differentially Private Learning with Small Public Data". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.04 (2020). DOI: 10.1609/aaai.v34i04.6088.

[86] Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. "Privacy for Free: Posterior Sampling and Stochastic Gradient Monte Carlo". In: *Proceedings of the 32nd International Conference on Machine Learning*. Ed. by Francis Bach and David Blei. Vol. 37. Proceedings of Machine Learning Research. Lille, France: PMLR, 2015, pp. 2493–2502. URL: http://proceedings.mlr.press/v37/wangg15.html.

[87] Larry Wasserman and Shuheng Zhou. "A Statistical Framework for Differential Privacy". In: *Journal of the American Statistical Association* 105.489 (2010), pp. 375–389. DOI: 10.1198/jasa.2009.tm08651. eprint: https://doi.org/10.1198/jasa.2009.tm08651. URL: https://doi.org/10.1198/jasa.2009.tm08651.

[88] Andre Wibisono. "Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem". In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, 2018, pp. 2093–3027. URL: http://proceedings.mlr.press/v75/wibisono18a.html.

[89] Da Yu et al. "Do not Let Privacy Overbill Utility: Gradient Embedding Perturbation for Private Learning". In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL: https://openreview.net/forum?id=7aogOj%5C_VYO0.

[90] Yingxue Zhou, Zhiwei Steven Wu, and Arindam Banerjee. "Bypassing the Ambient Dimension: Private SGD with Gradient Subspace Identification". In: *ICLR*. 2020.