Accelerating Grasp Exploration by Leveraging Learned Priors

Hanyu Li
Michael Danielczuk
Ashwin Balakrishna
Ken Goldberg

Electrical Engineering and Computer Sciences
University of California at Berkeley

Technical Report No. UCB/EECS-2020-129
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2020/EECS-2020-129.html

May 30, 2020

Accelerating Grasp Exploration by Leveraging Learned Priors

by

Han Yu Li

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Master of Science, Plan II

in

Department of Electrical Engineering and Computer Sciences

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Ken Goldberg, Research Advisor
Professor Pieter Abbeel, Second Reader

Spring 2020

The dissertation of Han Yu Li, titled Accelerating Grasp Exploration by Leveraging Learned Priors, is approved:

Research Advisor

Digitally signed by Ken Goldberg
DN: cn=Ken Goldberg, o, ou,
email=goldberg@berkeley.edu,
c=US
Date: 2020.05.30 18:24:11 -08'00'

Date

PAbbeel

Date 29 - MAY - 2020

University of California, Berkeley

Accelerating Grasp Exploration by Leveraging Learned Priors

Abstract

Accelerating Grasp Exploration by Leveraging Learned Priors

by

Han Yu Li

Master of Science, Plan II in Department of Electrical Engineering and Computer Sciences

University of California, Berkeley

Professor Ken Goldberg, Research Advisor

The ability of robots to grasp novel objects has industry applications in e-commerce order fulfillment and home service. Despite the recent success of data-driven grasping policies, they can fail to grasp objects which have complex geometry or are significantly outside of the training distribution. For such objects, we present a Thompson sampling algorithm that leverages learned priors from the Dexterity Network robot grasp planner to guide grasp exploration and provide probabilistic estimates of grasp success for each stable pose of novel objects. In addition, we introduce a new formulation of the mismatch between a prior distribution on grasp qualities and the ground truth grasp quality distribution, and perform empirical analysis studying the effect of this mismatch on policy performance.

In the first part of the thesis, simulation experiments suggest that the best learned policy attains an average total reward 64.5% higher than a greedy baseline and achieves within 5.7% of an oracle baseline. Total reward is defined as the average sum reward over $300,000$ training runs across a set of 3000 object poses or approximately 1600 objects. In addition, we find that Thompson sampling without a neural network prior attains an average total reward 43.4% higher than a greedy baseline and achieves within 4.6% of the best learned policy when evaluated over $20,000$ training runs across a set of 200 object poses.

In the second part of the thesis, we change the object's stable pose during learning. Simulation experiments suggest that the best learned policy attains an average total reward at least 150.1% higher than a greedy baseline and achieves within at most 12.15% of an oracle baseline when evaluated over 5000 training runs per object across a total of 25 stable poses across all 4 objects.

To Nancy, David and Eric

Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank Professor Ken Goldberg for giving me the opportunity to do research in AUTOLab for the past two years, for inspiring me and having confidence in me.

This report is the result of collaboration with members of AUTOLab. I would like to thank Michael Danielczuk, Ashwin Balakrishna and Vishal Satish for their contributions to the project, and for their mentorship and support during difficult times.

Also, I would like to thank Professor Pieter Abbeel for introducing me to artificial intelligence and robotics, and for taking the time to be my second reader.

The people I met at UC Berkeley made my experience memorable. I want to thank Jackson Sipple for being my closest companion and always cheering me up. I would also like to thank my friends at Cal for the wonderful time we have shared together.

Last but not least, my journey at Cal would not be possible without the support of my family. I would like to thank my parents for being my advocates and confidants, and for the sacrifices they made to give me the best education possible.

Chapter 1

Introduction

Robotic grasping has a wide range of industry applications such as warehouse order fulfillment, manufacturing, and assistive robotics. However, grasping is a difficult problem due to uncertainty in sensing and control, and there has been significant prior work on both analytical [2, 22, 24, 25, 31] and data-driven methods [11, 15, 16] for tackling these challenges. Recently, data-driven grasping algorithms have shown impressive success in learning grasping policies which generalize across a wide range of objects [8, 19, 21]. However, these techniques can fail to generalize to novel objects that are significantly different from those seen during training. Precisely, we investigate learning grasping policies for objects where general purpose grasping systems such as [19] produce relatively inaccurate grasp quality estimates, resulting in persistent failures during policy execution.

This motivates algorithms which can efficiently learn from on-policy experience by repeatedly attempting grasps on a new object and leveraging grasp outcomes to adjust the sampling distribution. Deep reinforcement learning has been a popular approach for online learning of grasping policies from raw visual input [10, 16, 23], but these approaches often take prohibitively long to learn robust grasping policies. These approaches typically attempt to learn *tabula rasa*, limiting learning efficiency. In this work, we introduce a method which leverages information from a general purpose grasping system to provide a prior for the learned policy while using geometric structure to inform online grasp exploration. We cast grasp exploration in the multi-armed bandits framework as in [14, 20]. However, unlike Laskey *et al.* [14] which focuses on grasping 2D objects where some rough geometric knowledge is known and Mahler *et al.* [20] which presents a method to transfer grasps to new 3D objects using a dataset of grasps on 3D objects with known geometry, we focus on efficiently learning grasping policies for 3D objects directly from depth image observations. Specifically, we present a method which leverages prior grasp success probabilities from the state-of-the-art Dex-Net 4.0 grasp quality network GQ-CNN [19] to guide online grasp exploration on unknown 3D objects with only depth-image observations.

This project is done in collaboration with Ashwin Balakrishna, Michael Danielczuk and Vishal Satish from UC Berkeley Automation Lab. I describe below the overall contributions and my individual contributions to the project.

The contributions of this report are:

1. A new problem formulation for leveraging learned priors on grasp quality to accelerate online grasp exploration.
2. An efficient algorithm, Thompson Sampling with Learned Priors (TSLP), for learning grasping policies on novel 3D objects from depth images by leveraging priors from the Dex-Net 4.0 robot grasping system [17].
3. A new formulation of the mismatch between a prior distribution on grasp qualities and the ground truth grasp quality distribution and empirical analysis studying the effect of this mismatch on policy performance.
4. Simulation experiments suggesting that TSLP attains an average total reward 64.5% higher than a greedy baseline when evaluated over $300,000$ training runs across 3000 object poses and is able to effectively leverage information from a GQ-CNN prior.

My individual contributions are:

5. Simulation experiments suggesting that Thompson sampling without a GQ-CNN prior attains an average total reward 43.4% higher than a greedy baseline and achieves within 4.6% of the best TSLP policy when evaluated over $20,000$ training runs across a set of 200 object poses.
6. Simulation experiments that simulate repeated dropping of the object and results suggesting that the best learned policy attains an average total reward at least 150.1% higher than a greedy baseline and achieves within at most 12.15% of an oracle baseline when evaluated over 5000 training runs across a total of 25 stable poses across all 4 objects.

Chapter 2

Related Work

2.1 Grasp Planning Algorithms

Robot grasping methods develop policies that execute grasps on novel objects, and can be divided into analytical methods and data-driven methods. Analytic methods assume knowledge of the geometry of the object to be grasped [2, 22, 24, 25] or use geometric similarities between known and unknown objects to infer grasps on unknown objects [20] However, the generalization of these methods is limited for objects dissimilar to the known objects, or when geometric information is unknown [3], as in the case we consider.

Data-driven methods rely on labels from humans [11, 15, 21, 28], self-supervision across many physical trials [10, 16, 23], simulated grasp attempts [9, 30] or sim-to-real transfer methods such as domain randomization [4] or domain adaptation [8]. Pinto *et al.* [23] presents an end-to-end self-supervised learning system that trains a deep Convolutional Neural Network (CNN) to select grasps through thousands of trials. Similar to TSLP, the algorithm in [23] works with raw visual inputs. However, it requires over 700 hours of training to learn robust policies [23]. Kalashnikov *et al.* [10] introduces a self-supervised reinforcement learning framework for grasping that generalizes to 96% success rate on objects not seen in training. The algorithm in [10] works directly with visual inputs; however, it requires 800 robot hours of training. Both Pinto *et al.* [23] and Kalashnikov *et al.* [10] learn to grasp in a one-shot setting. In contrast, the method presented aims to grasp novel objects through on-policy experience and multiple trials, without requiring a large amount of training beforehand. Dex-Net 4.0 grasp quality network GQ-CNN [19], which is used by TSLP, requires 24 hours of training on a single NVIDIA TITAN Xp graphics processing unit (GPU).

Hybrid approaches to grasping generate simulated grasp labels using analytical grasp metrics such as force closure or wrench resistance [17–19]. These data-driven and hybrid approaches train a deep neural network on the labeled data to predict grasp quality or directly plan reliable grasps on novel objects. However, for adversarial objects [31], for which very few high quality grasps exist, or for objects significantly out of the training distribution, grasps planned by these methods may still fail. In contrast, the presented method aims to leverage learned grasp quality estimates to enable efficient online learning for difficult-to-grasp objects through physical exploration of one pose of one object at a time, without previous knowledge of the object's geometry.

2.2 Multi-Armed Bandit Problem and Thompson Sampling

The Multi-Armed Bandit (MAB) is a statistical model in which the agent attempts to choose actions that maximize rewards, and the reward of each action is drawn from a fixed distribution unknown to the player [26]. A concrete example of MAB is a casino slot machine with k arms. The player does not know the reward of each arm; hence, there is a trade-off between exploring new arms and exploiting the arm that currently yields the highest reward.

Bayesian approaches for solving MAB use previous samples to form a belief distribution on the reward of each arm. Thompson sampling is a Bayesian algorithm for choosing actions that maximize rewards with respect to a randomly drawn belief [29]. Thompson sampling has been shown to be effective in addressing the exploration-exploitation trade-off in MAB, since it empirically achieves lower regret than other popular bandit algorithms such as Upper Confidence Bound [5].

2.3 Grasp Planning as Multi-Armed Bandit

Past works have formulated grasp planning as a Multi-Armed Bandit problem for grasping 2D objects where some geometric knowledge is known [14] or for transferring grasps to unknown 3D objects using a dataset of grasps on 3D objects with known geometry.

To improve efficiency in grasp exploration, Laskey *et al.* [14] compare the convergence times of two multi-armed bandit algorithms, Thompson sampling and Gittins index, against two baselines, uniform allocation and iterative pruning. The uniform allocation baseline selects each grasp with equal probability, whereas the iterative pruning method prunes 10% of grasps with the lowest sampled means every 1000 iterations. The Gittins index method treats the grasp planning problem as a Markov Decision Process (MDP), with the disadvantage that computational time can become exponential as the discount factor approaches 1, which is needed for long-term grasp planning. In comparison, Thompson sampling is much more computationally efficient. Its asymptotic complexity is sub-linear for k arms where $k > 2$ [1]. Experiments on 100 object shapes with 1000 grasps per object suggest that Thompson sampling with a uniform prior and the Gittins index MAB method converged to within 3% of optimal grasp (as measured by normalized probability of force closure) up to 10 times faster than uniform allocation, and 5 times faster than iterative pruning [14]. However, the policy presented is limited to 2D grasps and cannot operate directly on visual inputs. In addition, the Thompson sampling algorithm uses a uniform prior distribution, but a prior informed by a pre-trained neural network can further increase efficiency.

Mahler *et al.* [20] extend [14] to 3D and incorporate prior information from Dex-Net 1.0, a dataset of over 10,000 3D object models and a set of associated robust grasps. The algorithm then uses Thompson sampling, in which the prior belief distribution for each grasp is calculated based on its similarity to grasps and objects from the Dex-Net 1.0 database [20]. For objects with geometrically similar neighbors in Dex-Net 1.0, the algorithm converges to the optimal grasp approximately 2 times faster (in terms of number of iterations) than Thompson sampling without priors [20]. However, Dex-net 1.0 has has 29.4% longer runtime per iteration compared to Thompson sampling, due to the time to query the database [20]. In addition, for objects that have no geometrically similar neighbors in the dataset, the Dex-Net 1.0 algorithm performs no better than Thompson sampling without priors [20]. We present a Bayesian multi-armed bandit algorithm for robotic grasping with depth image inputs that does not require a database to compute priors but instead leverages the Dex-Net 4.0 grasping system from [19] as a learned prior to guide active grasp exploration.

Gordon *et al.* [7] applies the contextual bandit framework to robot-assisted feeding. The algorithm uses online learning to generalize to previously-unseen food items [19]. Similar to TSLP, Gordon *et al.* [7] uses a pre-trained neural network as a reward estimator and uses a contextual bandit algorithm to do grasp exploration. Unlike TSLP, Gordon *et al.* [7] updates the neural network after each subsequent grasp attempt, and uses Linear Upper Confidence bound (LinUCB) to select grasp. In contrast, TSLP only uses the neural network for prior distribution and does not update the neural network during experiments.

Chapter 3

Accelerating Grasp Exploration by Leveraging Learned Priors

3.1 Problem Statement

Given a single unknown object on a planar workspace, the objective is to effectively leverage prior estimates on grasp qualities to learn a grasping policy that maximizes the likelihood of grasp success. We first define the parameters and assumptions on the environment (Assumptions and Definitions), cast grasp exploration in the Bayesian bandits framework (Bayesian Bandits), and formally define the policy learning objective (Learning Objective).

Assumptions

We make the following assumptions about the environment.

1. **Pose Consistency:** We assume that the object remains in the same pose for all rounds of learning. In simulation, this can be achieved by using ground-truth knowledge of physics and object geometry, and for a physical system this can be realized by only lifting the object slightly above the plane when a grasp succeeds and placing it back down again.

2. **Evaluating Grasp Success:** We assume that the robot can evaluate whether a grasp has succeeded. In simulation, grasp success can be computed by using ground-truth knowledge of physics and object geometry. In physical experiments, success or failure can be determined using load cells, as in [19].

Definitions

1. **Observation:** An overhead depth image observation of the object at time $t = 0$ before policy learning has begun, given by $o \in \mathbb{R}_+^{H \times W}$.

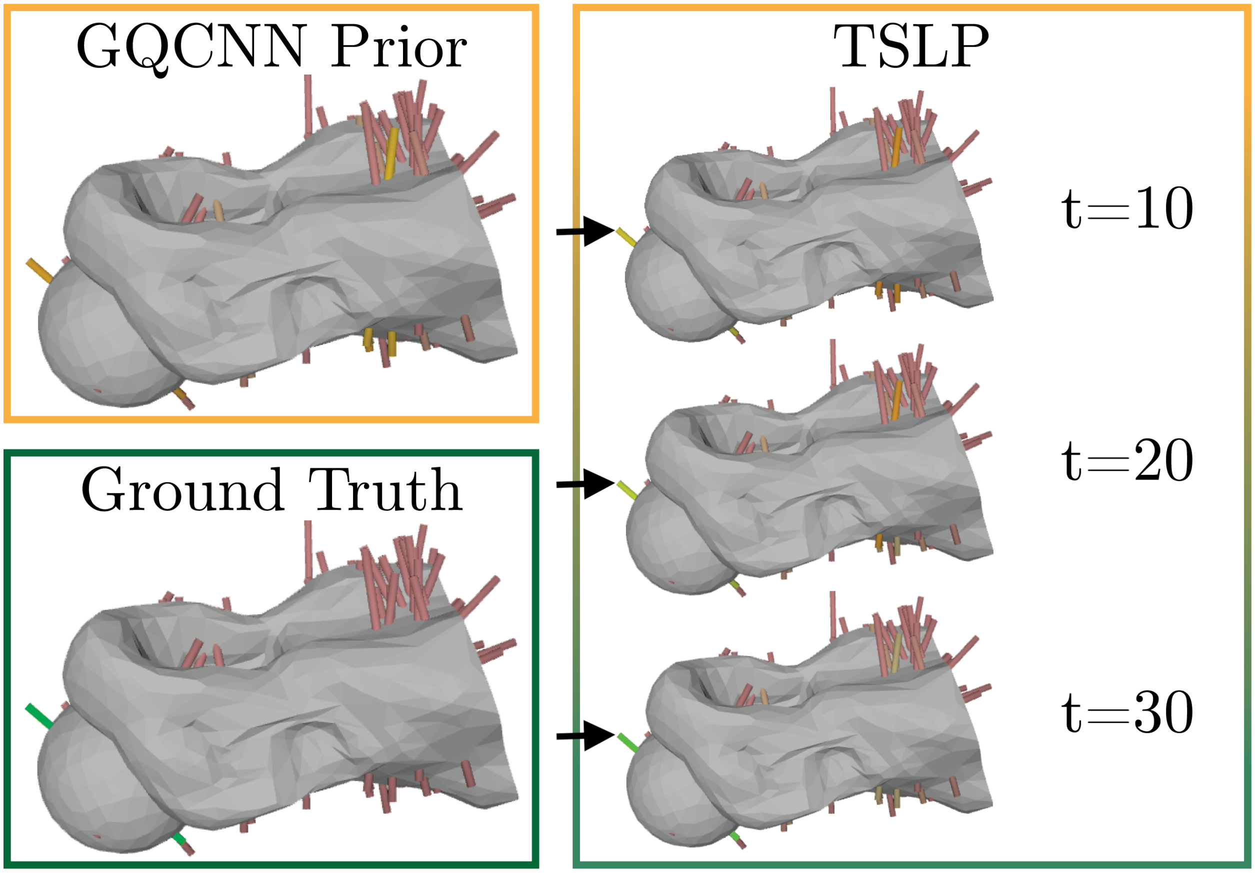


Figure 3.1: For adversarial objects, state-of-the-art grasp planning algorithms may incorrectly predict the distribution over grasp qualities (left column), where each whisker represents a grasp candidate colored by the likelihood of success (red indicates a poor grasp, green indicates a robust grasp). We find that TSLP can use the prior to efficiently discover the best grasp on the object (right column). Here, the policy discovers the only robust grasp despite a poor initial estimate of its quality from the GQ-CNN prior.

2. **Arms:** We define a set of K arms, $\{a_k\}_{k=1}^K$, by sampling a set of K antipodal grasps from observation o using the same method as in [17].

3. **Actions:** Given a selected arm k we define a corresponding action $u_k \in \mathcal{U}$ as a parallel jaw grasp represented by a center point $\mathbf{p} = (x, y, z) \in \mathbb{R}^3$ and a grasp axis $\mathbf{v} \in \mathbb{R}^3$ [20]. These grasp actions are visualized as whiskers in Figures 3.1 and 3.7.

4. **Reward Function:** Rewards for each arm are drawn from a Bernoulli distribution with unknown parameter p_k: $r(u_k) \sim \text{Ber}(p_k)$. Here $r(u_k) = 1$ if executing u_k results in the object being successfully grasped, and 0 otherwise.

5. **Priors:** We assume access to priors on the Bernoulli parameter p_k for each arm k.

6. **Policy:** Let $\pi_\theta(u_k)$ denote a policy parameterized by θ which selects an arm k and executes the action u_k. Thus, $\pi_\theta(u_k)$ defines a distribution over $\mathcal{U}$ at any given timestep t.

Bayesian Bandits

A multi-armed bandits problem is defined by an agent which must make a decision at each timestep $t \in \{1, 2, \ldots T\}$ by selecting an arm $k \in \{1, 2, \ldots K\}$ to pull. After each arm pull, the agent receives a reward which is sampled from an unknown reward distribution. In the Bayesian bandits framework [27], the agent maintains a belief over the parameters of the reward distribution for each arm, which can optionally be seeded with a known prior. The objective is to learn a policy with a distribution over arms that maximizes the cumulative expected reward over T rounds.

Learning Objective

The objective in policy learning is to maximize the total accumulated reward, which corresponds to maximizing the frequency with which the object is grasped. Let u_t denote the action selected at timestep t. Then the objective is to learn policy parameters θ to maximize the following:

$$J(\theta) = \mathbb{E}_{u_t \sim \pi_\theta(u_t)} \left[\sum_{t=1}^{T} r(u_t) \right] \tag{3.1}$$

3.2 Grasp Exploration Method

We discuss how to leverage learned priors from GQ-CNN to guide grasp exploration by using Thompson sampling, to learn a vision-based grasping policy. Since rewards are drawn

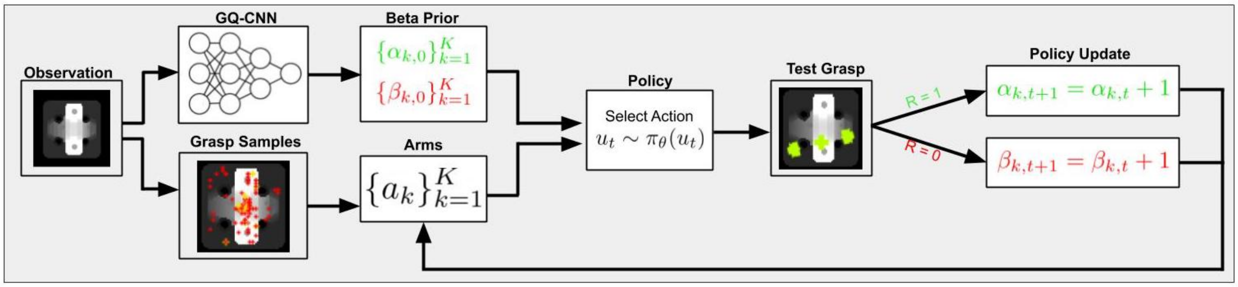

Observation
GQ-CNN
Beta Prior
$\{\alpha_{k,0}\}_{k=1}^{K}$
$\{\beta_{k,0}\}_{k=1}^{K}$
Grasp Samples
Arms
$\{a_k\}_{k=1}^{K}$
Policy
Select Action
$u_t \sim \pi_\theta(u_t)$
Test Grasp
R = 1
R = 0
Policy Update
$\alpha_{k,t+1} = \alpha_{k,t} + 1$
$\beta_{k,t+1} = \beta_{k,t} + 1$

Figure 3.2: **Method Overview:** A pre-trained GQ-CNN is used to set the priors on the reward parameters for each arm given the initial observation o and arms are sampled on observation o. Then, at each timestep the learned policy selects an arm and executes the corresponding action in the environment. The Thompson sampling parameters are updated based on the reward received as described in Section 3.2.

from a Bernoulli distribution as defined in Section 3.1, we represent the prior with a Beta distribution, the conjugate prior for a Bernoulli distribution. As noted in [14], this choice of prior is convenient since we can update the belief distribution over an arm k after executing corresponding action u_k in closed form given the sampled reward. See Figure 3.2 for a full method overview.

Thompson Sampling with a Beta-Bernoulli Process

Given that we pull arm k at time t and receive reward $r(u_k) \in \{0, 1\}$, as shown in [14], we can form the posterior of the Beta distribution by updating the shape parameters $\alpha_{k,t}$ and $\beta_{k,t}$:

$$\alpha_{k,t+1} = \alpha_{k,t} + r(u_k)$$
$$\beta_{k,t+1} = \beta_{k,t} + (1 - r(u_k))$$

For Thompson sampling, at time t, the policy samples $\hat{p}_{k,t} \sim \text{Beta}(\alpha_{k,t}, \beta_{k,t})$ for all arms $k \in \{1, 2, \ldots K\}$, selects arm $k^* = \text{argmax}_k\, \hat{p}_{k,t}$, and executes the corresponding action u_{k^*} in the environment. Note that the expected Bernoulli parameter for arm k can be computed from the current shape parameters $\alpha_{k,t}$ and $\beta_{k,t}$ as follows:

$$\mathbb{E}\left[\hat{p}_{k,t}\right] = \frac{\alpha_{k,t}}{\alpha_{k,t} + \beta_{k,t}} \tag{3.2}$$

However, it remains to appropriately initialize $\alpha_{k,0}$ and $\beta_{k,0}$. Note that setting $\alpha_{k,0} = \beta_{k,0} = 1 \ \forall \ k \in \{1, 2, \ldots K\}$ corresponds to a prior which is uniform on $[0, 1]$ for Bernoulli parameter $p_{k,t}$. We instead set $\alpha_{k,0}, \beta_{k,0}$ according to a learned prior by using the initial depth image observation o.

Leveraging Neural Network Priors

We use a pre-trained Grasp Quality Convolutional Neural Network (GQ-CNN) from [19] to obtain an initial estimate of the probability of grasp success. GQ-CNN learns a Q-function, $Q_\phi(\cdot, \cdot)$, which given an overhead depth image of an object and a proposed parallel jaw grasp, estimates the probability of grasp success. However, as explored in [31], there exist many objects for which the analytical methods used for training GQ-CNN are relatively inaccurate, resulting in significant errors. Thus, we refine the initial GQ-CNN grasp quality estimates with online exploration.

We first compute $Q_\phi(o, u_k)\ \forall\ k \in \{1, 2, \ldots K\}$ and use these estimates as each arm's initial mean Bernoulli parameter. Note that $\alpha_{k,t}$ and $\beta_{k,t}$, as defined in Section 3.2, correspond to the cumulative number of grasp successes and grasp failures respectively for action u_k up to time t. Thus, $(\alpha_{k,0}, \beta_{k,0})$ can be interpreted as pseudo-counts of grasp successes and failures respectively for action u_k before policy learning has begun, while prior strength $S = \alpha_{k,0} + \beta_{k,0}$ can be interpreted as the number of pseudo-rounds before policy learning. If S is large, the prior induced by $(\alpha_{k,0}, \beta_{k,0})$ will significantly influence the expected Bernoulli parameter given in 3.2 for many rounds, while if S is small, the resulting prior will be quickly washed out by samples from online exploration. We enforce the following initial conditions for $(\alpha_{k,0}, \beta_{k,0})$, given the GQ-CNN prior:

$$\frac{\alpha_{k,0}}{\alpha_{k,0} + \beta_{k,0}} = Q_\phi(o, u_k)$$
$$\frac{\beta_{k,0}}{\alpha_{k,0} + \beta_{k,0}} = 1 - Q_\phi(o, u_k)$$

For a desired prior strength $S = \alpha_{k,0} + \beta_{k,0}$, we set:

$$\alpha_{k,0} = S \cdot Q_\phi(o, u_k)$$
$$\beta_{k,0} = S \cdot (1 - Q_\phi(o, u_k))$$

This prior enforcement technique in conjunction with online learning with Thompson Sampling, as discussed in Section 3.2, results in a stochastic policy $\pi_\theta(u_k)$ parameterized by $\theta = \big(\{(\alpha_k, \beta_k)\}_{k=1}^{K}, \phi\big)$, the learned Beta distribution shape parameters across all arms and the fixed parameters of the GQ-CNN used for initialization.

3.3 Prior Mismatch

Kendall rank correlation coefficient

To measure the quality of the GQ-CNN prior, we define a notion of dissimilarity between the prior and ground truth grasp probabilities, as in Chapelle *et al.* [6], termed the *prior mismatch*. *Prior mismatch* describes the difference between the prior distribution and the ground truth distribution. We define *prior mismatch* in the context of grasping. Unlike

Chapelle *et al.* [6], which primarily focuses on mismatch between the mean of the prior distribution and true Bernoulli parameter, we present a new metric based on the discrepancy between how arms are ranked under the prior and under the ground truth distribution.

Given the grasp quality estimates of the GQ-CNN prior $q_p = (Q_\phi(o, u_k))_{k=1}^K$ and the ground truth grasp probabilities $q_g = (p_k)_{k=1}^K$ on all K arms, let $\mathcal{P} = \{(q_p[k], q_g[k])\}_{k=1}^K$. We then compute Kendall's tau coefficient, defined as:

$$\tau = \frac{N_c - N_d}{\sqrt{(N_c + N_d + T_p)(N_c + N_d + T_g)}}$$

where N_c and N_d are the number of concordant and discordant pairs in $\mathcal{P}$, respectively, and T_p and T_g are the number of pairs for which $q_p[i] = q_p[j]$ and $q_g[i] = q_g[j]$, respectively [12, 13]. As a rank correlation coefficient, $\tau \in [-1, 1]$, where 1 denotes a perfect match in the rankings and -1 denotes perfectly inverse rankings. We define the prior mismatch M as a dissimilarity measure that maps τ to $[0, 1]$:

$$M = \frac{1 - \tau}{2}$$

In practice, to control for stochasticity when sampling arms on the initial observation o, we average M over 10 independently sampled sets of K arms.

We use a rank correlation coefficient metric instead of L2 distance between q_p and q_g, because capturing the relative ordering of arms is often more useful than continuous distance measures such as L2 distance.

Relationship Between Metrics

In this section, we provide intuition for the ground truth grasp probabilities q_g, the prior mismatch M, and the L2 distance $\|q_p - q_g\|_2$ by exploring their respective distributions and how they correlate with each other. We define $\|q_p - q_g\|_2$ as the L2 distance between the GQ-CNN prior q_p and ground truth grasp probabilities q_g. We define N as the number of nonzero grasps in ground truth grasp probabilities q_g.

Figure 3.3 shows the frequency of $\|q_p - q_g\|_2$ and N for 1000 object-pose pairs. While the distribution of L2 distance $\|q_p - q_g\|_2$ is relatively spread out from 0.2 to 9.5, the number of nonzero grasps, N, is close to 0 for most object-pose pairs. This shows that in most cases, there are very few good grasps, which presents an exploration challenge that Thompson Sampling with Learned Priors (TSLP) aims to solve.

Figure 3.4 shows the relationship between L2 distance $\|q_p - q_g\|_2$ and number of nonzero grasps N for 2400 object poses and the relationship between prior mismatch M and number of nonzero grasps N for 2000 object poses. While there is a strong positive correlation between $\|q_p - q_g\|_2$ and N, the relationship between prior mismatch M and number of nonzero grasps N is a weak negative correlation. Since most of the values in ground truth grasp probabilities q_g and GQ-CNN prior q_p are close to zero (Figure 3.3), the L2 distance between them is

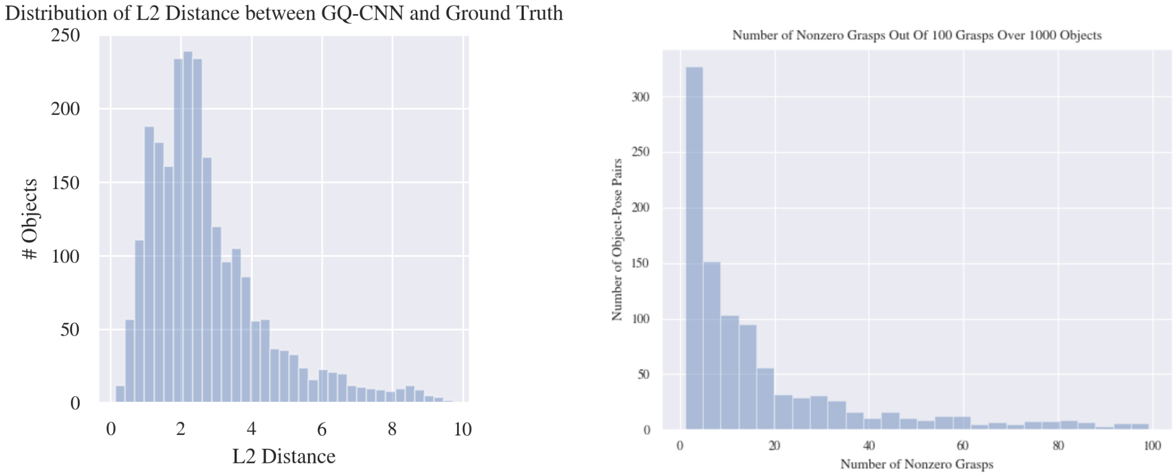


Figure 3.3: (Left) The distribution of L2 Distance $\|q_p - q_g\|_2$ for 1000 object poses in the dataset used by [19]. We find that $\|q_p - q_g\|_2$ ranges from 0.2 to 9.5 and a value between 2.00 and 2.75 is the most common. (Right) The distribution of number of nonzero grasps N for 1000 object poses in the dataset used by [19]. We find that N ranges from 0 to 100 and most of the values are very close to 0.

similar to the number of nonzero entries. However, the Kendall rank correlation coefficient captures the rank difference, which includes both cases in which there are few good grasps, as well as cases in which there are many good grasps. As a result, prior mismatch M is less correlated with the number of nonzero grasps N compared to the L2 distance $\|q_p - q_g\|_2$. This is desirable because in Thompson sampling capturing the relative ordering of arms is much more relevant than the number of nonzero grasps.

3.4 Practical Implementation

We implement the method from Section 3.2 in a simulated environment using 3D object models from the Dex-Net 4.0 dataset [19]. We render a simulated depth image of the object using camera parameters that are selected to be consistent with a Photoneo PhoXi S industrial depth camera. Arms are selected by sampling parallel-jaw antipodal grasp candidates on the observation o using the antipodal image grasp sampling technique from Dex-Net 2.0 [17]. The antipodal grasp sampler thresholds the depth image to find areas with high gradients, then uses rejection sampling over pairs of pixels to find antipodal grasp points. Once the arms are sampled from the image, we calculate the prior grasp probabilities using GQ-CNN, then deproject each grasp from image space into a 3D grasp using the known

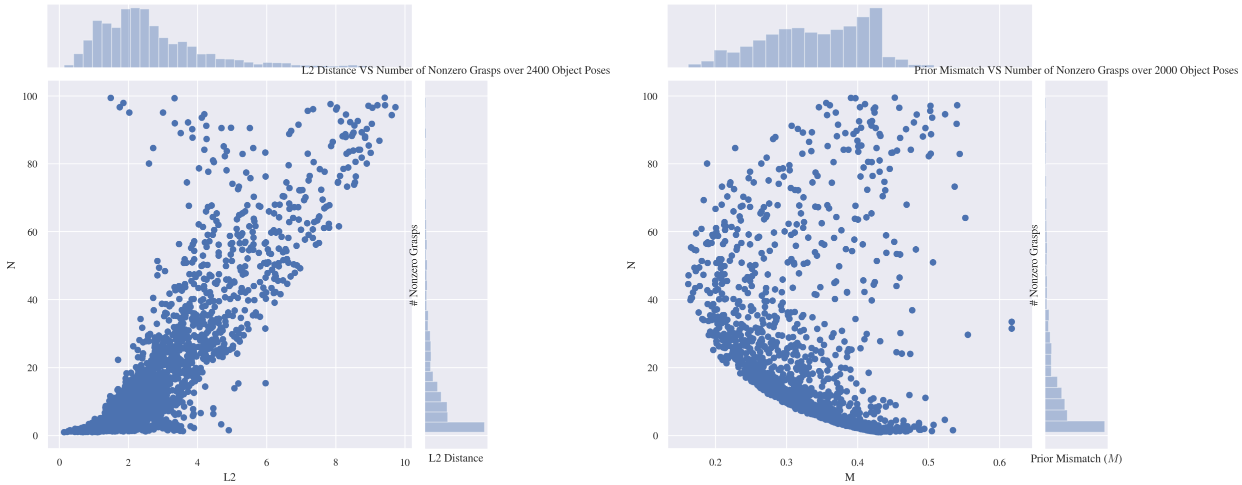


Figure 3.4: (Left) The relationship between L2 Distance $\|q_p - q_g\|_2$ and number of nonzero grasps N for 2400 object poses in the dataset used by [19]. We see a strong positive correlation between $\|q_p - q_g\|_2$ and N. The top and right show the distributions of L2 Distance $\|q_p - q_g\|_2$ and number of nonzero grasps N for 2400 object poses. (Right) The relationship between prior mismatch M and number of nonzero grasps N for 2000 object poses in the dataset used by [19]. We see a weak negative correlation between M and N. The top and right show the distribution of L2 Distance M and number of nonzero grasps N for 2000 object poses. The top and right show the distributions of prior mismatch M and number of nonzero grasps N for 2000 object poses.

camera intrinsics. We then iteratively choose grasps according to the policy for a set number of timesteps and collect the reward for each grasp.

Algorithm 1 summarizes the full approach discussed in Section 3.2 along with implementation details. If we are unable to sample K arms or if none of the corresponding grasps has ground truth quality greater than zero, we do not consider the object pose. In simulation, we evaluate the probability of grasp success for each arm using the robust wrench resistance metric, which measures the grasp's ability to resist the gravity wrench under perturbations in the grasp pose, as in [18]. Then, rewards during policy learning and evaluation are sampled from a Bernoulli distribution with parameter defined by this metric. Note that while computing this metric requires knowledge of the object geometry, this metric is simply used to simulate grasp success on a physical robotic system and is not exposed to TSLP.

Algorithm 1 Thompson Sampling with Learned Priors (TSLP) for Image-Space Grasp Exploration

```
Input: Number of arms ($K$), Maximum Iterations $T$, Pretrained GQ-CNN $Q_\phi(\cdot,\cdot)$, Prior Strength $S$
Output: Grasp exploration policy: $\pi_\theta(u_k)$
  Capture observation $o$, sample $K$ antipodal grasps $\{a_k\}_{k=1}^K$, and compute prior beliefs $\alpha_{k,0}, \beta_{k,0}\ \forall\ k \in \{1, 2, \dots K\}$ using $Q_\phi(o, u_k)$ using method from Section 3.2.
  for $t = 1, ..., T$ do
    Select action $u_k$ using Thompson sampling as in Section 3.2
    Execute $u_k$ and observe $r(u_k)$
    Update $\alpha_{k,t}, \beta_{k,t}\ \forall k \in \{1, 2, \dots K\}$ as in Section 3.2
  end for
```

Baselines

To evaluate the benefits of online learning in grasp exploration, we compare TSLP policies against three static policies that do not perform any updates. The policies we choose are greedy, ground truth and softmax.

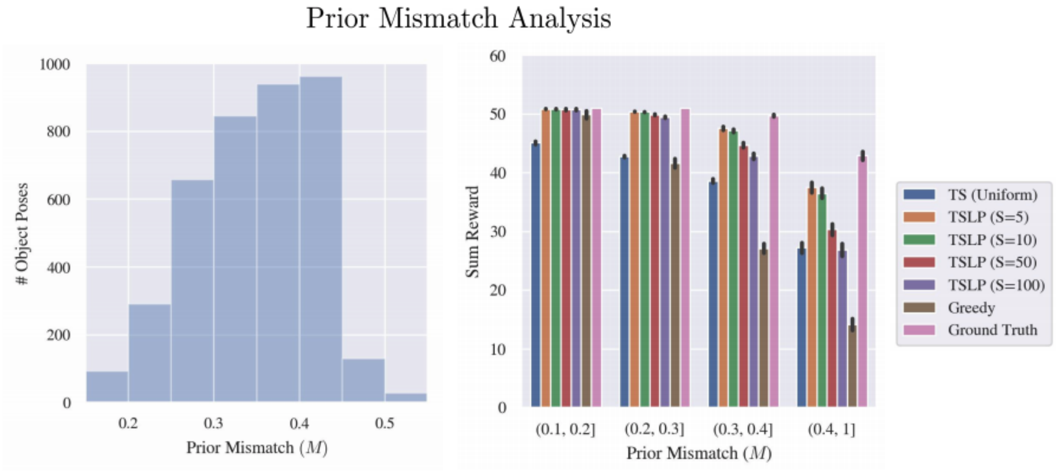
The greedy policy repeatedly selects the grasp with highest quality under GQ-CNN as in [19]. This gives an idea of policy performance if no online exploration is performed. We expect the greedy policy to do well if the best grasp in ground truth coincides with the best grasp under the GQ-CNN prior.

The ground truth oracle policy repeatedly selects the grasp with the highest quality under the ground truth grasp quality metrics computed in simulation. This provides an upper bound on possible performance since it can access the true grasp success probabilities, which are not available to our algorithm.

The softmax policy samples directly from the GQ-CNN prior q_p. This gives an idea of whether the GQ-CNN prior estimates allocate the highest scores to the best grasps. We expect the softmax policy to do well when there are multiple grasps with high success probabilities, but the best ground truth grasp is not selected by the greedy policy. In this case, the softmax policy would outperform the greedy policy, since the softmax policy samples grasps proportional to their GQ-CNN scores. To create the sampling distribution for the softmax policy, we first compute the GQ-CNN prior $q_p = (Q_\phi(o, u_k))_{k=1}^K$. Then each arm k is sampled with probability P_k, a softmax over q_p.

$$P_k = \frac{e^{Q_\phi(o,u_k)}}{\sum_{j=1}^K e^{Q_\phi(o,u_j)}}$$

In most cases, we expect the softmax policy to perform poorly because the GQ-CNN prior q_p is often an accurate representation of the ground truth grasp probabilities q_g.



(a)

(b)

Figure 3.5: (a) The distribution of prior mismatch M for the 3946 total object poses in the dataset used by [19]. We find that M ranges from 0.16 to 0.64 and a value between 0.4 and 0.45 is the most common, accounting for about 25% of the object poses. All object poses with M above 0.55 are placed into the highest bin. (b) The sum of rewards over policy evaluation computed over 3000 object poses randomly selected from the dataset. This plot suggests that the chosen metric accurately describes the mismatch between the prior and ground truth grasp quality distribution, as performance of all TSLP policies decreases with increased prior mismatch.

3.5 Simulation Experiments on Single Poses

Setup

In simulation experiments, we evaluate both the accuracy of the prior mismatch metric and the ability of TSLP to increase grasp exploration efficiency. We assess whether TSLP can discover higher quality grasps than baselines which do not explore online or which explore online but do not leverage learned priors for the grasp selection policy. In both experiments, we make use of the dataset from Mahler *et al.* [19], which contains approximately 1,600 object meshes.

We evaluate the learned policies every 10 steps of learning, and perform 500 learning steps in total for all experiments. To evaluate the learned policies, we sample 100 grasps from the current policy without policy updates and compute the metric defined in Equation (3.1). We evaluate TSLP with a variety of different prior strengths to evaluate how important the GQ-CNN prior is for policy performance. We also compare to Thompson sampling with a uniform prior over the arms. Thus, this policy does not utilize the GQ-CNN prior at

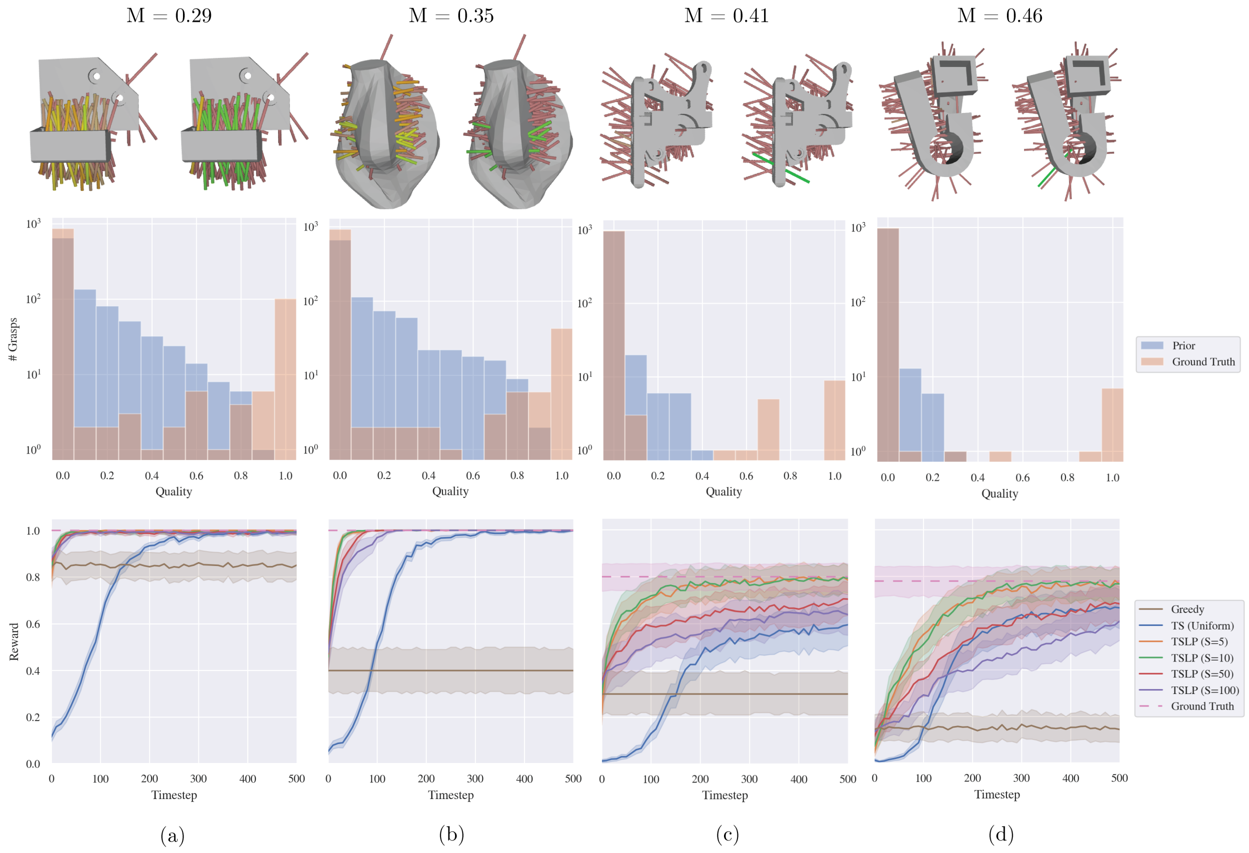


Figure 3.6: Visualization of policy performance for all baselines and TSLP policies (labeled with their prior strength). The first row visualizes grasp qualities as measured by the GQ-CNN prior (left) and the ground truth grasp success probabilities (right) for a single stable pose of each of the four objects (shown top down). Green whiskers indicate high estimated or ground truth grasp quality, while red whiskers indicate low estimated or ground truth grasp quality. In the second row, we visualize the distributions of GQ-CNN prior and ground truth grasp qualities. (a) With a low prior mismatch ($M = 0.29$), the greedy policy performs well and all Thompson sampling policies with non-zero prior strengths converge quickly to the ground truth. (b-c) For objects with higher prior mismatch, the Thompson sampling policies with non-zero prior strength rapidly improve on the prior for object poses with higher prior mismatch ($M = 0.35$ and $M = 0.40$, respectively). (d) For objects with very high prior mismatch ($M = 0.46$), the Thompson sampling policies with non-zero prior strength converge more slowly, but still show improvement on the baseline with prior strength 0.

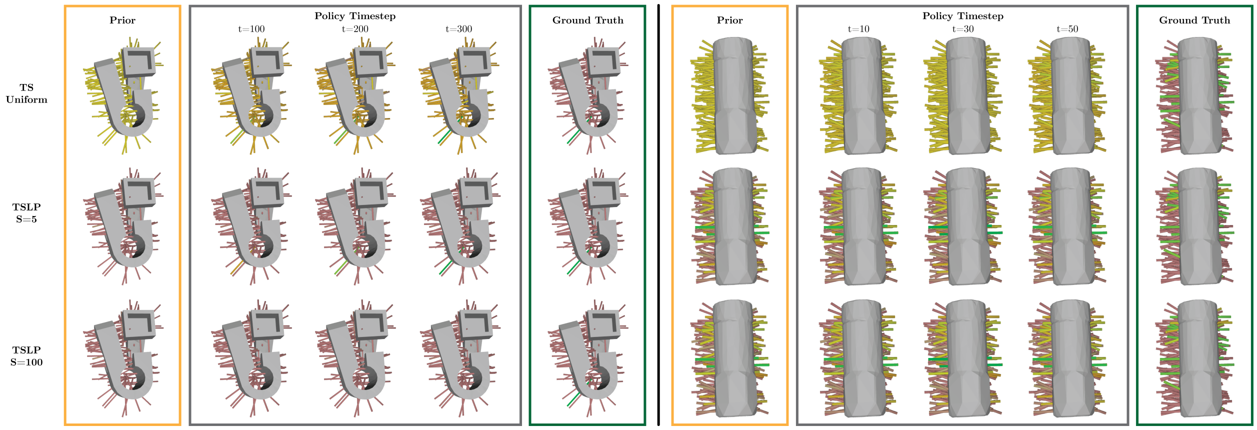


Figure 3.7: We visualize the evolution of the mean Bernoulli parameter (defined in Equation (3.2)) inferred by TSLP with varying prior strengths on sampled arms over learning steps for two different objects. Grasps with high estimated success probabilities or ground truth quality values are colored green, while those with low estimated success probabilities or ground truth qualities are colored orange or red. The inferred mean Bernoulli parameter for TSLP eventually converges to the ground truth probabilities. For the first object, we note that TSLP is able to find the best grasps when the prior strength is relatively weak, but unable to do so when the prior strength is too high since the prior is overly pessimistic ($M = 0.46$). For the second object, the prior is relatively good ($M = 0.31$), so increasing the prior strength accelerates discovery of the best grasps.

all, and all learning is performed online. Uniform prior means that all arms have the same expectation as defined in Equation (3.2). We set $\alpha_{k,0} = \beta_{k,0} = 1$ for all arms $k \in \{1, 2, \ldots K\}$. As a result, $S = 2$.

Note that when evaluating policies, there are two key sources of uncertainty: (1) the variability in the arms sampled on the initial observation o, and (2) the inherent stochasticity during learning given a set of arms. To control for variations in these parameters, when reporting results on a particular pose of an object, 10 different sets of $K = 100$ arms are sampled on the corresponding observation o. Then, for each of these sets of arms, every policy is trained 10 times for a total of 100 rollouts for each object pose.

TSLP Policies

We evaluate TSLP policies with different values for strength S (as defined in Section 3.2) to study the effects of increasing S on policy outcomes.

We conduct simulation experiments across object poses with a wide range of prior mismatches M, as shown in Figure 3.5(a), which plots the frequency of prior mismatch values

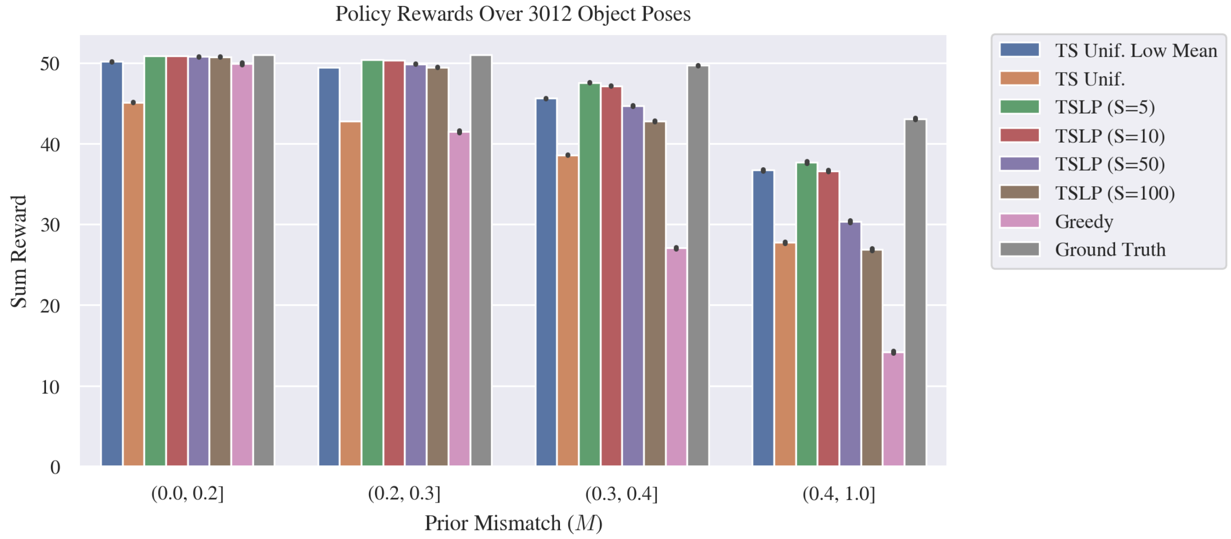


Figure 3.8: The sum of rewards over policy evaluation computed over 3012 object poses randomly selected from the dataset used by [19]. This plot suggests that lowering the mean of uniform prior policy (TS Unif. Low Mean) drastically improves performance compared to TS Unif. TS Unif. Low Mean achieves similar results to the best TSLP policy ($S = 5$). In addition, the chosen metric accurately describes the mismatch between the prior and ground truth grasp quality distribution, as performance of all TSLP policies decreases with increased prior mismatch.

over the 3946 total object poses in the dataset. When the prior mismatch is relatively low, we expect policies which give more weight to the prior to perform well, while if the prior mismatch is high, we expect policies which prioritize online exploration over following the prior to attain higher rewards.

We evaluate each policy on 3000 of these object poses and compute the sum reward of all policies averaged over the $300,000$ total training runs (100 training runs per object pose). The results are shown in Figure 3.5(b) and Table 3.1. Figure 3.5(b) shows policy performance as a function of prior mismatch, given the distribution of objects over prior mismatch values shown in Figure 3.5(a) over 3000 total object poses. These results suggest that the metric introduced here accurately models prior mismatch, as increased prior mismatch causes performance for all online learning policies, as well as the greedy policy, to degrade. A second trend is that object poses with higher prior mismatch also tend to have lower ground truth quality values, suggesting that GQ-CNN especially struggles to identify high quality grasps when very few are present or when the highest quality grasps have comparatively lower quality.

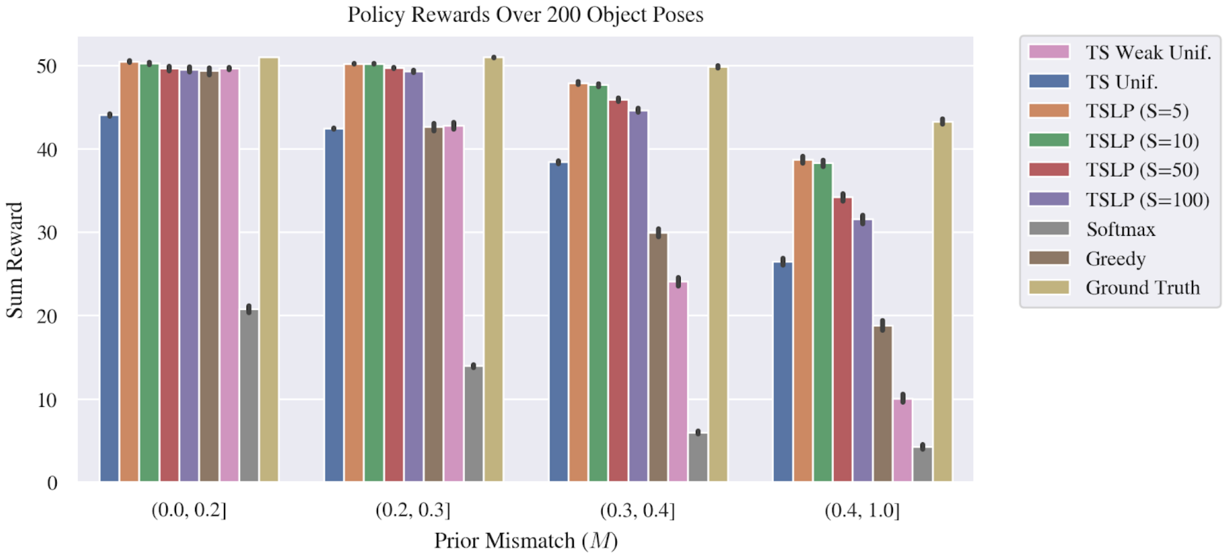


Figure 3.9: The sum of rewards over policy evaluation computed over 200 object poses randomly selected from the dataset used by [19]. This plot suggests that decreasing the strength of uniform prior policy (TS Unif. Weak Uniform) achieves better performance for objects with low prior mismatch < 0.2. However, decreasing the strength of uniform prior policy results in worse performance for objects with high prior mismatch > 0.3. This plot also shows that sampling directly from the GQ-CNN prior (Softmax) results in poor performance across all prior mismatch bins.

Table 3.1 shows that TSLP significantly outperforms the greedy baseline and is able to achieve average total reward that is very close to the ground truth policy. This result suggests that TSLP is able to successfully leverage priors from GQ-CNN to outperform GQ-CNN on a wide variety of objects of varying geometries.

As a further case study, we select a set of 4 objects, as shown in Figure 3.6, which are diverse in their shapes and sizes and vary widely in their prior mismatch M. As expected, the ground truth policy (GT) achieves the best performance since it uses oracle information. We find that for objects with relatively low prior mismatch ($M = 0.29$), the greedy policy and the Thompson sampling policies which place very high weight on the GQ-CNN prior (high prior strength) perform very well. However, for objects with higher prior mismatch ($M = 0.35$, $M = 0.41$), we find that the greedy policy performs much more poorly, and online exploration is critical to finding high quality grasps. However, even with high prior mismatch, the gap in performance between the Thompson sampling policies that use the prior and the uniform prior Thompson sampling policy indicates that the GQ-CNN prior helps accelerate grasp exploration substantially. Finally, for objects with very high prior

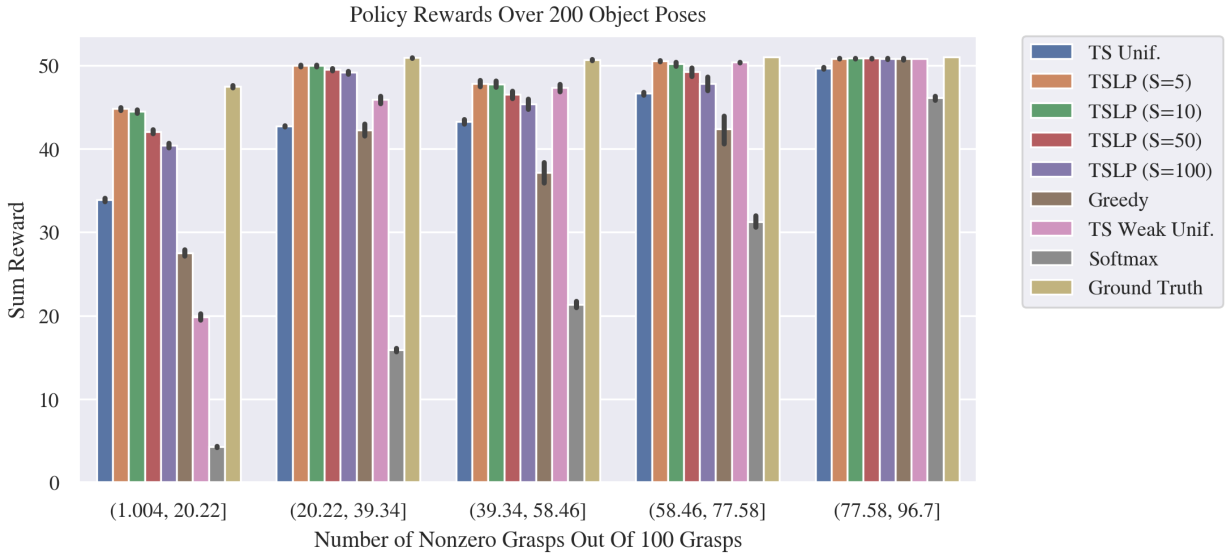


Figure 3.10: The sum of rewards over policy evaluation computed over 200 object poses randomly selected from the dataset used by [19]. This plot shows that the number of nonzero grasps is another accurate metric for the mismatch between the prior and ground truth grasp quality distribution, as performance of all TSLP policies increase with increased number of nonzero grasps. This plot suggests that decreasing the strength of uniform prior policy (TS Unif. Weak Uniform) achieves better performance for objects with high number of nonzero grasps > 58.46, which means that more than half the grasps are nonzero. However, decreasing the strength of uniform prior policy results in worse performance for objects with low number of nonzero grasp < 20.22, which means that less than a quarter of the grasps are nonzero. This plot also shows that the softmax policy does well when the number of nonzero grasps is very high (> 77.58)

mismatch ($M = 0.46$), the greedy policy and Thompson sampling policies with high prior strengths perform poorly, as expected. However, Thompson sampling policies with low prior strength outperform Thomspon sampling with a uniform prior. This result indicates that although the prior is of very low quality, it still provides useful guidance to the Thompson sampling policy if a low prior strength is used.

Figure 3.7 shows how the mean Bernoulli parameter inferred by TSLP evolves over learning steps for each of the sampled arms. TSLP is able to successfully learn grasp qualities close to the ground truth grasp qualities for a wide variety of different objects. Note that the learned policy is generally more accurate for higher quality grasps, which makes sense since Thompson sampling directs exploration towards high reward grasps, allowing it to focus on distinguishing between high quality grasps rather than capturing the quality distribution of low quality grasps. For the first object, TSLP is able to find the best grasp when the prior strength is relatively weak, but performs poorly when the prior strength is set too high. For the second object, the prior mismatch is lower, so increasing the prior strength accelerates discovery of the best grasps on the object. Note that with a uniform prior, Thompson sampling is generally able to discover most of the best grasps, but fails to distinguish them from bad grasps, resulting in poorer policy performance when these grasps are sampled during policy evaluation.

Non-Neural-Network Policies

We evaluate policies that do not use a neural network to study the capacity of online learning in discovering good grasps *tabula rasa*.

Recall that the results for TS (Uniform) in Figures 3.5, 3.6, 3.7 and Table 3.1 have $\alpha_{k,0} = \beta_{k,0} = 1$ for all arms $k \in \{1, 2, \ldots K\}$. So $\mathbb{E}[\hat{p}_{k,t}] = 0.5$ and $S = 2$. Since uniform prior only requires that all arms have the same expectation as defined in Equation (3.2), we explore settings of TS (Uniform) with different values of $\mathbb{E}[\hat{p}_{k,t}]$ and S. We expect the policy with lowered $\mathbb{E}[\hat{p}_{k,t}]$ to do well in cases where most of the ground truth grasp success probabilities are low, because $\mathbb{E}[\hat{p}_{k,t}]$ is closer to ground truth. We expect the policy with lowered S to do well when the GQ-CNN estimates are misleading (give high scores for bad grasps). A lower value for S allows the policy to adapt more quickly based on rewards from grasp attemps.

First, we lower $\mathbb{E}[\hat{p}_{k,t}]$ while keeping S the same. We set $\alpha_{k,0} = 0.02$ and $\beta_{k,0} = 1.98$ for all arms $k \in \{1, 2, \ldots K\}$. So $\mathbb{E}[\hat{p}_{k,t}] = 0.01$ and $S = 2$. A low mean of $\mathbb{E}[\hat{p}_{k,t}] = 0.01$ implies that the algorithm has a pessimistic prior for all grasps, assuming that they are all bad. On the other hand, $\mathbb{E}[\hat{p}_{k,t}] = 0.5$ implies that the algorithm is neutral about the grasp, estimating that it has a 50% chance of success. Keeping strength S the same means that the learning rate is the same as the previous setting. Since the ground truth grasp probabilities q_g are close to zero, as shown in Figure 3.3(b), a low mean of $\mathbb{E}[\hat{p}_{k,t}] = 0.01$ (TS Unif. Low Mean) is a better estimate for q_g compared to $\mathbb{E}[\hat{p}_{k,t}] = 0.5$ (TS Unif.) So we expect TS Unif. Low Mean to outperform TS Unif. Figure 3.8 shows that lowering the mean of uniform prior policy (TS Unif. Low Mean) drastically improves performance compared to TS Unif. TS Unif. Low Mean achieves similar results to the best TSLP policy ($S = 5$).

Secondly, we lower S while keeping $\mathbb{E}[\hat{p}_{k,t}]$ the same. We set $\alpha_{k,0} = \beta_{k,0} = 0.01$ for all arms $k \in \{1, 2, \ldots K\}$. So $\mathbb{E}[\hat{p}_{k,t}] = 0.5$ and $S = 0.02$. We name this policy TS Weak Unif. A weak strength of $S = 0.02$ implies that the algorithm has a low confidence on the prior, so estimated grasp probabilities largely depends on rewards, and the effect of prior distributions is small. Keeping the mean $\mathbb{E}[\hat{p}_{k,t}] = 0.5$ implies that the algorithm is still neutral about each grasp, estimating that it has a 50% chance of success. Since the GQ-CNN prior gets less emphasis in TS Weak Unif., we expect it to perform comparable to other policies when there is a high number of good grasps, since it has a high likelihood to discover good grasps by chance. When there is a low number of good grasps, we expect TS Weak Unif. to perform worse than other policies, because it relies less on prior distributions which are useful when there is a low number of good grasps.

Figure 3.9 shows that decreasing the strength of uniform prior policy (TS Unif. Weak Uniform) achieves better performance for objects with low prior mismatch < 0.2. However, decreasing the strength of uniform prior policy results in worse performance for objects with high prior mismatch > 0.3. Figure 3.9 also shows that sampling directly from the GQ-CNN prior (Softmax) results in poor performance across all prior mismatch bins.

Figure 3.10 shows the same result as Figure 3.9, but it is binned by number of nonzero grasps instead of prior mismatch M. Figure 3.10 shows that the number of nonzero grasps N is another accurate metric for the mismatch between the prior and ground truth grasp quality distribution, as performance of all TSLP policies increase with increased number of nonzero grasps. Figure 3.10 suggests that decreasing the strength of uniform prior policy (TS Unif. Weak Uniform) achieves better performance for objects with high number of nonzero grasps > 58.46, which means that more than half the grasps are nonzero. However, decreasing the strength of uniform prior policy results in worse performance for objects with low number of nonzero grasp < 20.22, which means that less than a quarter of the grasps are nonzero. Figure 3.10 also shows that the softmax policy does well when the number of nonzero grasps is very high (> 77.58).

Table 3.2 shows that the TS Uniform Prior with Low Mean significantly outperforms TS Uniform or TS Uniform with Low Strength, and achieves performance within 4.6% of the best performing TSLP policy, TSLP (S=5) and within 10.1% of the ground truth oracle baseline. Table 3.2 shows that the TSLP algorithm is able to do well without a neural network prior.

3.6 Simulation Experiments on Multiple Poses

We consider the case in which the robot is presented with a single object. The robot redrops the object after each grasp and learns to grasp the object in each of its stable poses. This is equivalent to sampling from the stable pose distribution at each timestep.

Problem Statement

Assumptions

Since the object changes its stable pose during learning, we remove the pose consistency assumption defined in Assumption 1. Instead, we assume that the stable pose distribution remains the same throughout the learning process.

Definitions

The definitions of observations, arms, actions and reward functions are the same as Section 3.1. However, the policy now expects grasps on multiple poses.

Table 3.1: **Policy evaluation on large object set:** We evaluate each TSLP policy, the greedy and Thompson sampling with uniform prior baselines, and the ground truth policy on a dataset of 3000 object poses and report the average sum reward over all runs on each of the object poses (300,000 total training runs per policy, 100 training runs per object for each policy) in the format of mean ± standard deviation. Since we evaluate the policy 51 times per episode, the maximum possible sum reward is 51. For readability, we scale all results by a factor of 100/51 for a maximum scaled sum reward of 100. We find that the best performing TSLP policy, TSLP (S=5), outperforms the greedy baseline by 64.5% while achieving performance within 5.7% of the ground truth oracle baseline.

TS (Uniform)	**TSLP (S=5)**	**TSLP (S=10)**	**TSLP (S=50)**	**TSLP (S=100)**
72.08 ± 20.67	89.37 ± 17.88	88.43 ± 18.53	82.63 ± 23.51	78.88 ± 26.29

Greedy	**Ground Truth**
54.33 ± 33.02	94.53 ± 13.49

Table 3.2: **Policy evaluation of additional policies:** We evaluate the softmax policy, the Thompson sampling with uniform prior variations, TSLP policy, the greedy and the ground truth policy on a dataset of 200 object poses and report the average sum reward over all runs on each of the object poses (20,000 total training runs per policy, 100 training runs per object for each policy) in the format of mean ± standard deviation. Since we evaluate the policy 51 times per episode, the maximum possible sum reward is 51. For readability, we scale all results by a factor of 100/51 for a maximum scaled sum reward of 100. We find that the TS Uniform Prior with Low Mean significantly outperforms other variations of TS Uniform and achieves performance within 4.6% of the best performing TSLP policy, TSLP (S=5) and within 10.1% of the ground truth oracle baseline.

Greedy	**Softmax**	**TS Uniform Low Strength**	**TS Uniform Low Mean**	**Ground Truth**
59.84 ± 46.31	72.08 ± 18.57	50.20 ± 46.92 85.82 ± 21.84	94.49 ± 17.12	

TS (Uniform)	**TSLP (S=5)**	**TSLP (S=10)**	**TSLP (S=50)**	**TSLP (S=100)**
70.67 ± 24.22	89.80 ± 21.51	89.31 ± 21.47	85.27 ± 26.12	82.45 ± 29.22

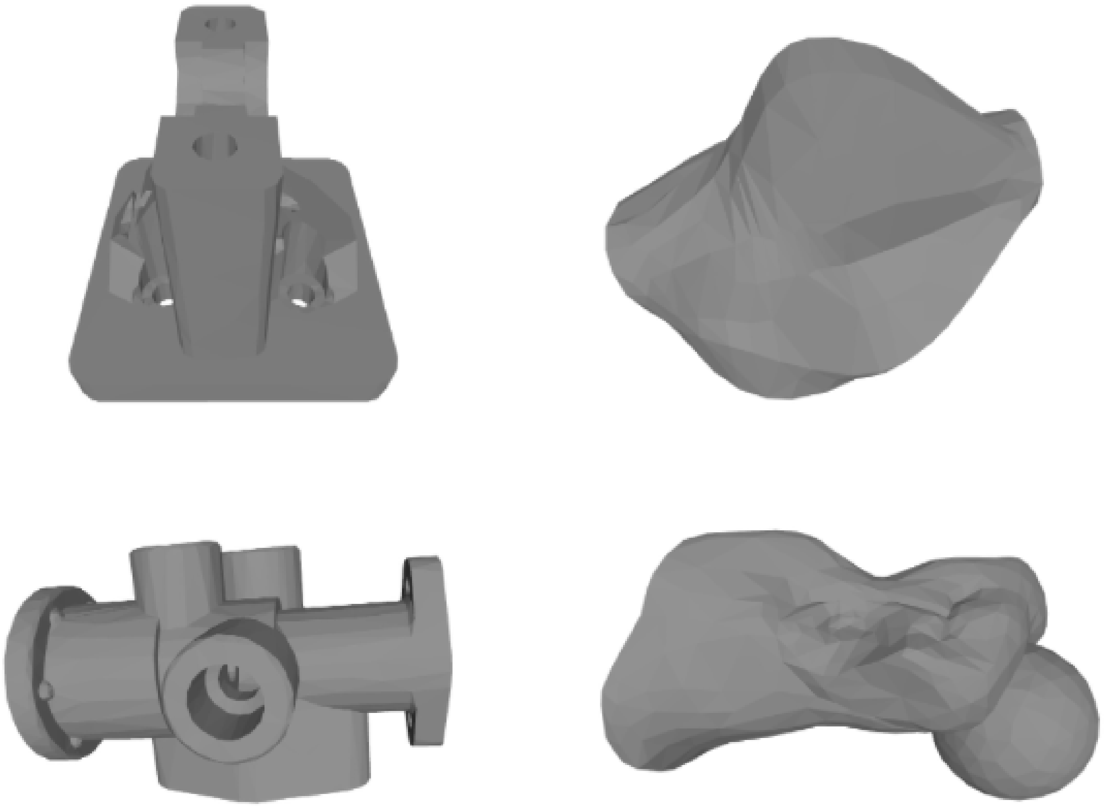


Figure 3.11: Visualization of objects used in Figure 3.12 and Figure 3.13 in their most likely stable pose. They are called engine part, vase, pipe connector and pawn (left to right, top to bottom).

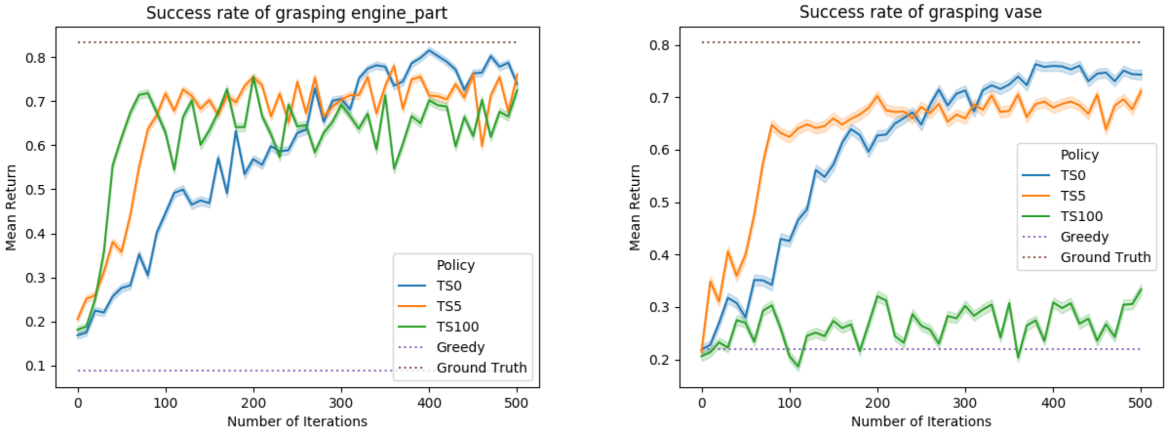
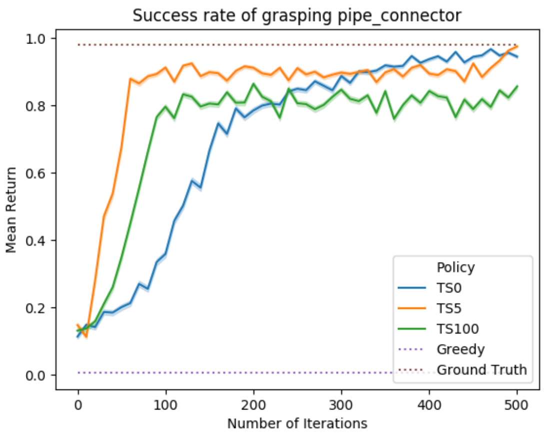


Figure 3.12: (Left) Average grasp success rate versus iteration for engine part ($M' = 0.24$). Engine part has a low prior mismatch value, which means that the GQ-CNN prior is relatively accurate in ranking grasps. Therefore, different strength settings for TSLP all do equally well. (Right) Average grasp success rate versus iteration for vase ($M' = 0.30$). Vase has a medium-low prior mismatch value. In this case, TS with uniform prior does better than TSLP with a low strength prior. TSLP with a high strength prior of 100 fails to discover any good grasps.

- **Policy:** Let $\pi_\theta(u_k^l|P^l)$ denote a policy parameterized by θ which selects an arm k for stable pose P^l and executes the action u_k^l. Thus, $\pi_\theta(u_k^l|P^l)$ maps a stable pose P^l to a distribution over $\mathcal{U}$.

Bayesian bandits

The robot is presented with an object O that has L stable poses. We consider each stable pose as an independent Bayesian bandit problem. This creates a set of L independent Bayesian bandit problems to explore grasps on each of the L poses. The poses are coupled by a probability distribution p^λ that defines the probability of transitioning to each of the L stable poses at any time. At each timestep, a stable pose P_t is sampled from p^λ and a corresponding depth image observation is captured. Grasps are sampled on the resulting image and a learned prior is used to seed the belief over the parameters of the reward distribution for each of the arms (grasps) if the stable pose was previously unseen. Then, grasps are explored on P_t according to the learned policy π. Here, stable pose P_t can be viewed as a filter which selects which of L possible bandit problems is active at timestep t.



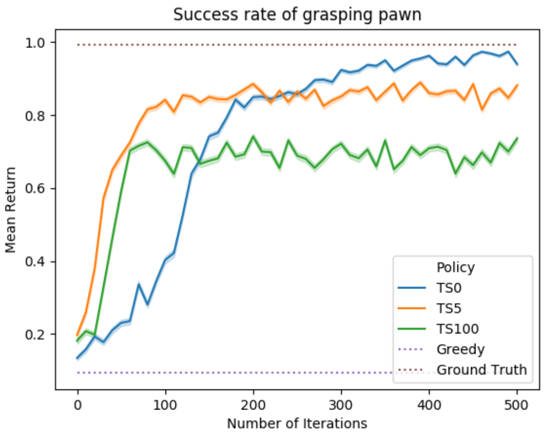


Figure 3.13: (Left) Average grasp success rate versus iteration for pipe connector ($M' = 0.40$). Pipe connector has a medium-high prior mismatch value. As a result, TS with uniform prior achieves comparable performance to TSLP with a low strength of 5, while TSLP with a high strength of 100 does worse. (Right) Average grasp success rate versus iteration for pawn ($M' = 0.42$). Pawn has a high prior mismatch values. As a result, TS with uniform prior does better than TSLP with a low strength prior, while TSLP with a high strength of 100 does significantly worse.

Learning Objective

The objective in policy learning is to maximize the total accumulated reward, and thus maximize the frequency with which the object is grasped over the distribution of possible stable poses. We define the policy learning objective as follows where θ is a vector of the policy parameters.

$$J(\theta) = \max_{\theta} \sum_{t=1}^{T} \mathbb{E}_{P_t \sim p^{\lambda}(P)} \left[\mathbb{E}_{u_t \sim \pi_\theta(u_t|P_t)} \left[r(P_t, u_t) \right] \right] \tag{3.3}$$

Prior Mismatch

We introduce a new definition of prior mismatch that is weighted by the stable pose distribution. Let $M(P)$ be the prior mismatch for stable pose P as defined in Section 3.3. Then the prior mismatch M' of object O is the weighted average of the prior mismatch of its stable poses.

$$M' = \mathbb{E}_{P \sim p^{\lambda}(P)} \left[M(P) \right] \tag{3.4}$$

Experiments

We conduct simulation experiments across 4 objects with a wide range of prior mismatches M' as shown in Figure 3.11. The practical implementation of experiments differs from Section 3.4 only in that a new stable pose is sampled from p^λ at each timestep.

Figures 3.12 and 3.13 show the learning curves for each object. Compared to Figures 3.6, Figure 3.12 and 3.13 appear more jagged because the pose may change after each iteration. As expected, the ground truth policy achieves the best performance since it uses oracle information. We find that for all objects, TS with uniform prior performs the best, closely followed by TSLP with low strength, and TSLP with high strength performs the worst. This presents an interesting contrast to Figure 3.6, in which prior mismatch is a clear predictor for the performance of different strength values. Experiments suggest that TS with uniform prior is the best policy when poses are re-sampled after each grasp, and the policy performs close to ground truth for a wide range of prior mismatch values. In addition, Figures 3.12 and 3.13 show that TSLP ($S = 10$, $S = 100$) learn faster than TS with uniform prior and performs better early on during the training process. This shows that leveraging GQ-CNN prior helps the algorithm discover good grasps when the object changes its stable pose during the learning process.

Tables 3.3, 3.4, 3.5, and 3.6 show the numeric results at iteration $t = 500$ in Figures 3.12 and 3.13. Tables 3.3, 3.4, 3.5, and 3.6 suggest that the best learned policy attains an average total reward at least 150.1% higher than a greedy baseline and achieves within at most 12.15% of an oracle baseline when evaluated over 5000 training runs per object across a total of 25 stable poses across all 4 objects.

Note that compared to Table 3.1 and 3.2, TSLP policies has a much higher gain over the greedy policy in the multiple-pose setting. For engine part and vase, this is because the objects are inherently more difficult to grasp, as seen by lower ground truth success probabilities. For pipe connector and pawn, this is due to the objects having high prior mismatch values; as a result, the greedy policy performs extremely poorly for these objects.

Table 3.3: **Policy evaluation of engine part:** We evaluate TSLP policy, the greedy policy, Thompson sampling with uniform prior, and the ground truth policy on 1 object with 5 stable poses and report the average sum reward over all runs on all of the object's stable poses ($5,000$ total training runs per policy per object) in the format of mean ± standard deviation. Since we evaluate the policy 100 times per episode, the maximum possible sum reward is 100. We find that the best performing TSLP policy, TSLP ($S = 5$), outperforms the greedy baseline by 814.0% while achieving performance within 12.15% of the ground truth oracle baseline.

Greedy	**TS (Uniform)**	**TSLP (S=5)**	**TSLP (S=100)**	**Ground Truth**
8.31 ± 27.61	73.90 ± 43.92	75.95 ± 42.74	72.46 ± 44.67	82.88 ± 37.67

Table 3.4: **Policy evaluation of vase:** We evaluate TSLP policy, the greedy policy, Thompson sampling with uniform prior, and the ground truth policy on 1 object with 6 stable poses and report the average sum reward over all runs on all of the object's stable poses ($5,000$ total training runs per policy per object) in the format of mean ± standard deviation. Since we evaluate the policy 100 times per episode, the maximum possible sum reward is 100. We find that the best performing TSLP policy, TS (Uniform), outperforms the greedy baseline by 150.1% while achieving performance within 10.1% of the ground truth oracle baseline.

Greedy	**TS (Uniform)**	**TSLP (S=5)**	**TSLP (S=100)**	**Ground Truth**
30.18 ± 45.90	75.49 ± 43.01	70.49 ± 45.61	36.60 ± 48.17	83.11 ± 37.47

Table 3.5: **Policy evaluation of pipe connector:** We evaluate TSLP policy, the greedy policy, Thompson sampling with uniform prior, and the ground truth policy on 1 object with 8 stable poses and report the average sum reward over all runs on all of the object's stable poses ($5,000$ total training runs per policy per object) in the format of mean ± standard deviation. Since we evaluate the policy 100 times per episode, the maximum possible sum reward is 100. We find that the best performing TSLP policy, TSLP ($S = 5$), outperforms the greedy baseline by 31360% while achieving performance within 0.94% of the ground truth oracle baseline.

Greedy	**TS**	**TSLP**	**TSLP**	**Ground**
	(Uniform)	**(S=5)**	**(S=100)**	**Truth**
0.31 ± 5.58	94.47 ± 22.86	97.52 ± 15.55	85.60 ± 35.10	98.44 ± 12.40

Table 3.6: **Policy evaluation of pawn:** We evaluate TSLP policy, the greedy policy, Thompson sampling with uniform prior, and the ground truth policy on 1 object with 6 stable poses and report the average sum reward over all runs on all of the object's stable poses ($5,000$ total training runs per policy per object) in the format of mean ± standard deviation. Since we evaluate the policy 100 times per episode, the maximum possible sum reward is 100. We find that the best performing TSLP policy, TS (Uniform), outperforms the greedy baseline by 949.1% while achieving performance within 5.99% of the ground truth oracle baseline.

Greedy	**TS**	**TSLP**	**TSLP**	**Ground**
	(Uniform)	**(S=5)**	**(S=100)**	**Truth**
8.96 ± 28.56	94.00 ± 23.75	88.18 ± 32.29	73.59 ± 44.09	99.63 ± 6.08

Chapter 4

Conclusion and Future Work

In this report, we present Thompson Sampling with Learned Priors (TSLP), a bandit exploration strategy for robotic grasping which facilitates use of expressive neural network-based prior belief distributions and enables efficient online exploration for objects for which this prior is inaccurate. We quantify the notion of prior mismatch as it pertains to the ranking of arms and explore the effect of prior strength on the efficiency and efficacy of online learning. Experiments suggest that across a dataset of 3000 object poses, TSLP outperforms both a greedy baseline as well as a Thompson sampling baseline that uses a uniform prior and is able to leverage a GQ-CNN prior to significantly accelerate grasp exploration. Experiments that explore multiple stable poses of a single object suggest that TSLP outperforms a greedy baseline and is able to grasp the object with high success probability in expectation over all stable poses of the object.

For future work, we will modify the simulation experiments that explore multiple poses to more closely imitate physical experiments. First, we will re-drop the object only after grasp success. This is because in physical experiments, it is not possible to re-drop an object after a failed grasp without human intervention. Therefore, we will sample from the object's stable pose distribution when the object is re-dropped after a successful grasp, but assume that the object's stable pose remains unchanged after a failed grasp. Let P_t and u_t denote the stable pose and corresponding action on that stable pose selected at timestep t. Then the distribution of observed stable poses is defined as follows.

$$p_t^{\lambda}(P) = \begin{cases} P_{t-1} & r(P_{t-1}, u_{t-1}) = 0 \\ P \sim p^{\lambda}(P) & r(P_{t-1}, u_{t-1}) = 1 \end{cases}$$

In simulation experiments, we assume access to the ground truth pose indices, which would not be possible in physical experiements. Therefore, we will estimate the current stable pose P_t from the current observation o_t. Object stable poses are invariant across translations and rotations in the camera's optical axis. To identify whether the pose has previously been seen, we will take the difference between the binary mask of the object in the current depth image observation o_t and binary masks of previous observations, across translations and rotations. If the difference is below some fixed threshold, we match the current pose to the observed pose that achieved such difference and use the cached set of grasps and policy parameters.

Finally, we will implement more advanced baselines, such as a neural network baseline that updates GQ-CNN online with rewards, perform physical experiments and extend experiments to new grasping modalities such as suction.

Bibliography

[1] S. Agrawal and N. Goyal, “Analysis of thompson sampling for the multi-armed bandit problem,” in *Proceedings of the 25th Annual Conference on Learning Theory*, S. Mannor, N. Srebro, and R. C. Williamson, Eds., ser. Proceedings of Machine Learning Research, vol. 23, Edinburgh, Scotland: PMLR, 2012, pp. 39.1–39.26.

[2] A. Bicchi and V. Kumar, “Robotic grasping and contact: A review,” in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, IEEE, vol. 1, 2000, pp. 348–353.

[3] J. Bohg, A. Morales, T. Asfour, and D. Kragic, “Data-driven grasp synthesis—a survey,” *IEEE Transactions on Robotics*, vol. 30, no. 2, pp. 289–309, 2013.

[4] K. Bousmalis, A. Irpan, P. Wohlhart, Y. Bai, M. Kelcey, M. Kalakrishnan, L. Downs, J. Ibarz, P. Pastor, K. Konolige, *et al.*, “Using simulation and domain adaptation to improve efficiency of deep robotic grasping,” in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, IEEE, 2018, pp. 4243–4250.

[5] O. Chapelle and L. Li, “An empirical evaluation of thompson sampling,” in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2011, pp. 2249–2257.

[6] ——, “An empirical evaluation of thompson sampling,” in *Advances in Neural Information Processing Systems 24*, J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, Eds., Curran Associates, Inc., 2011, pp. 2249–2257.

[7] E. K. Gordon, X. Meng, M. Barnes, T. Bhattacharjee, and S. S. Srinivasa, *Adaptive robot-assisted feeding: An online learning framework for acquiring previously-unseen food items*, 2019. arXiv: `1908.07088 [cs.RO]`.

[8] S. James, P. Wohlhart, M. Kalakrishnan, D. Kalashnikov, A. Irpan, J. Ibarz, S. Levine, R. Hadsell, and K. Bousmalis, “Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks,” in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 12 627–12 637.

[9] E. Johns, S. Leutenegger, and A. J. Davison, “Deep learning a grasp function for grasping under gripper pose uncertainty,” in *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, IEEE, 2016, pp. 4461–4468.

[10] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, *et al.*, "Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation," in *Conf. on Robot Learning (CoRL)*, 2018.

[11] D. Kappler, J. Bohg, and S. Schaal, "Leveraging big data for grasp planning," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, IEEE, 2015, pp. 4304–4311.

[12] M. G. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, no. 1/2, pp. 81–93, 1938.

[13] ——, "The treatment of ties in ranking problems," *Biometrika*, vol. 33, no. 3, pp. 239–251, 1945.

[14] M. Laskey, J. Mahler, Z. McCarthy, F. T. Pokorny, S. Patil, J. Van Den Berg, D. Kragic, P. Abbeel, and K. Goldberg, "Multi-armed bandit models for 2d grasp planning with uncertainty," in *2015 IEEE International Conference on Automation Science and Engineering (CASE)*, IEEE, 2015, pp. 572–579.

[15] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *Int. Journal of Robotics Research (IJRR)*, vol. 34, no. 4-5, pp. 705–724, 2015.

[16] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *Int. Journal of Robotics Research (IJRR)*, vol. 37, no. 4-5, pp. 421–436, 2018.

[17] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," in *Proc. Robotics: Science and Systems (RSS)*, 2018.

[18] J. Mahler, M. Matl, X. Liu, A. Li, D. Gealy, and K. Goldberg, "Dex-net 3.0: Computing robust robot vacuum suction grasp targets in point clouds using a new analytic model and deep learning," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2018.

[19] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg, "Learning ambidextrous robot grasping policies," *Science Robotics*, vol. 4, no. 26, eaau4984, 2019.

[20] J. Mahler, F. T. Pokorny, B. Hou, M. Roderick, M. Laskey, M. Aubry, K. Kohlhoff, T. Kröger, J. Kuffner, and K. Goldberg, "Dex-net 1.0: A cloud-based network of 3d objects for robust grasp planning using a multi-armed bandit model with correlated rewards," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2016, pp. 1957–1964.

[21] D. Morrison, P. Corke, and J. Leitner, "Closing the Loop for Robotic Grasping: A Real-time, Generative Grasp Synthesis Approach," in *Proc. of Robotics: Science and Systems (RSS)*, 2018.

[22] R. M. Murray, *A mathematical introduction to robotic manipulation.* CRC press, 2017.

[23] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, IEEE, 2016, pp. 3406–3413.

[24] D. Prattichizzo and J. C. Trinkle, "Grasping," in *Springer handbook of robotics*, Springer, 2016, pp. 955–988.

[25] E. Rimon and J. Burdick, *The Mechanics of Robot Grasping.* Cambridge University Press, 2019.

[26] H. Robbins, "Some aspects of the sequential design of experiments," *Bulletin of the American Mathematical Society*, vol. 58, no. 5, pp. 527–535, 1952.

[27] D. J. Russo, B. V. Roy, A. Kazerouni, I. Osband, and Z. Wen, "Now publishers - a tutorial on thompson sampling," vol. 11, no. 1, pp. 1–96, 2018.

[28] A. Saxena, J. Driemeyer, and A. Y. Ng, "Robotic grasping of novel objects using vision," *Int. Journal of Robotics Research (IJRR)*, vol. 27, no. 2, pp. 157–173, 2008.

[29] W. R. Thompson, "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples," vol. 25, no. 3/4, pp. 285–294, Dec. 1933.

[30] U. Viereck, A. t. Pas, K. Saenko, and R. Platt, "Learning a visuomotor controller for real world robotic grasping using simulated depth images," *arXiv preprint arXiv:1706.04652*, 2017.

[31] D. Wang, D. Tseng, P. Li, Y. Jiang, M. Guo, M. Danielczuk, J. Mahler, J. Ichnowski, and K. Goldberg, "Adversarial grasp objects," in *2019 IEEE 15th International Conference on Automation Science and Engineering (CASE)*, IEEE, 2019, pp. 241–248.