# Accelerating Grasp Exploration by Leveraging Learned Priors

Han Yu Li*, Michael Danielczuk*, Ashwin Balakrishna*, Vishal Satish, Ken Goldberg

***Abstract*— The ability of robots to grasp novel objects has industry applications in e-commerce order fulfillment and home service. Data-driven grasping policies have achieved success in learning general strategies for grasping arbitrary objects. However, these approaches can fail to grasp objects which have complex geometry or are significantly outside of the training distribution. We present a Thompson sampling algorithm that learns to grasp a given object with unknown geometry using online experience. The algorithm 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 the novel object. We find that seeding the policy with the Dex-Net prior allows it to more efficiently find robust grasps on these objects. 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 when evaluated over 300,000 training runs across a set of 3000 object poses.**

## I. 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 [1, 22, 25, 26, 30] and data-driven methods [9, 13, 14] 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 [6, 18, 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 [18] 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 [8, 14, 24], 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 [12, 19]. However, unlike Laskey *et al.* [12] which focuses on grasping 2D objects where some rough geometric knowledge is known and Mahler *et al.* [19] 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. In addition, the algorithm learns to grasp a specific object through online interaction, unlike Mahler *et al.* [19] which learns a general grasping policy for arbitrary objects. 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 [18] to guide online grasp exploration on unknown 3D objects with only depth-image observations.

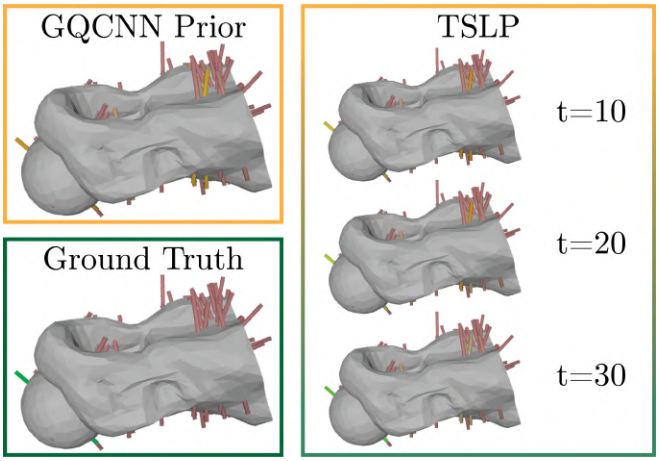


Fig. 1: For adversarial objects, state-of-the-art grasp planning algorithms may incorrectly predict the distribution over grasp qualities (left column), where each whisker represents a grasp candidate colored by the likelihood of success (red indicates a poor grasp, green indicates a robust grasp). We 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.

The contributions of this paper 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 [16].
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.

* equal contribution

The AUTOLAB at UC Berkeley (automation.berkeley.edu) {katherine.li,mdanielczuk,ashwin_balakrishna,vsatish,goldberg}@berkeley.edu

## II. Related Work

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 [1, 15, 22, 25, 26] or use geometric similarities between known and unknown objects to infer grasps on unknown objects [19] 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 [9, 13, 21, 28], self-supervision across many physical trials [2, 8, 14, 24], simulated grasp attempts [7, 29], or sim-to-real transfer methods such as domain randomization [4] or domain adaptation [6]. Hybrid approaches generate simulated grasp labels using analytical grasp metrics such as force closure or wrench resistance [16–18]. 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. A recent paper in sim-to-real transfer learning correct for inaccurate gripper poses predicted by the neural network by combining domain adaptation and visual servoing in the grasp planning process [23]. However, for adversarial objects [30], 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. 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.

Past works have formulated grasp planning as a Multi-Armed Bandit problem for grasping 2D objects where some geometric knowledge is known [12] or for transferring grasps to unknown 3D objects using a dataset of grasps on 3D objects with known geometry. Laskey *et al.* [12] found that Thompson sampling with a uniform prior significantly outperformed uniform allocation or iterative pruning in 2D grasp planning in terms of convergence rate to within 3% of the optimal grasp, but their policy is limited to 2D grasps and cannot operate directly on visual inputs. Mahler *et al.* [19] extend [12] 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 [19]. For objects with geometrically similar neighbors in Dex-Net 1.0, the algorithm converges to the optimal grasp approximately 2 times faster than Thompson sampling without priors [19]. In contrast, 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 [18] as a learned prior to guide active grasp exploration. Instead of learning a general grasping strategy for arbitrary objects as [19], the algorithm learns to grasp a specific object through online interactions with the object.

## III. 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 (Sections III-A and III-B), cast grasp exploration in the Bayesian bandits framework (Section III-C), and formally define the policy learning objective (Section III-D).

### A. Assumptions

We make the following assumptions about the environment.

1) **Pose Consistency:** We assume that the object remains in the same pose during all rounds of learning. In simulation, this can be achieved by using ground-truth knowledge of physics and object geometry. In physical experiments, the pose consistency assumption will not hold generally. We discuss methods to approximately enforce pose consistency in physical experiments in Section VIII.
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 [18].

### B. Definitions

1) **Observation:** An overhead depth image observation of the object at time $t = 0$ before policy learning has begun, given by $o \in \mathbb{R}_+^{H \times W}$.
2) **Arms:** We define a set of $K$ arms, $\{a_k\}_{k=1}^K$.
3) **Actions:** Given a selected arm $k$ we define a corresponding grasp action $u_k \in \mathcal{U}$.
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 $\theta$ 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 $t$.

### C. 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.

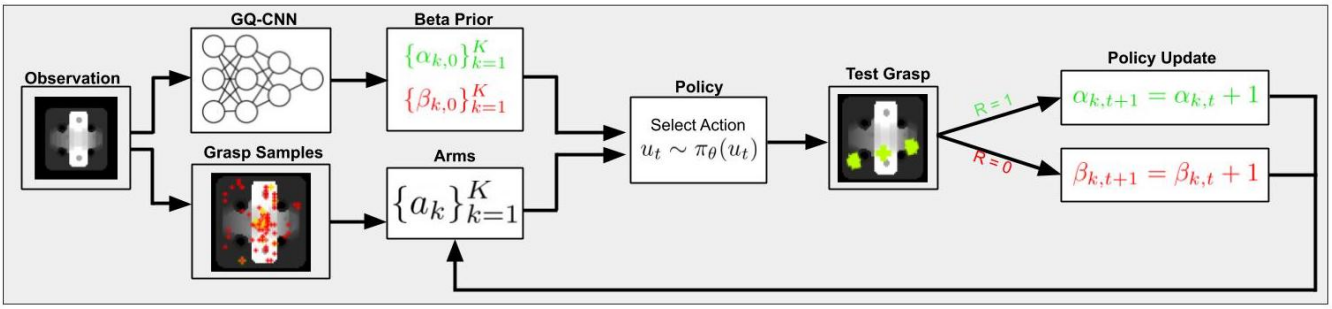


Fig. 2: **Method Overview:** A pre-trained GQ-CNN is used to set the priors on the reward parameters for each arm given the initial observation $o$ and arms are sampled on observation $o$. Then, at each timestep the learned policy selects an arm and executes the corresponding action in the environment. The Thompson sampling parameters are updated based on the reward received as described in Section IV-A.

### D. 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 $\theta$ to maximize the following:

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

## IV. 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 from a Bernoulli distribution as defined in Section III, we represent the prior with a Beta distribution, the conjugate prior for a Bernoulli distribution. As noted in [12], 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 2 for a full method overview.

### A. 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 [12], 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:

$$\mathbb{E}\left[\hat{p}_{k,t}\right] = \frac{\alpha_{k,t}}{\alpha_{k,t} + \beta_{k,t}} \tag{IV.1}$$

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 $o$.

### B. Leveraging Neural Network Priors

We use a pre-trained Grasp Quality Convolutional Neural Network (GQ-CNN) from [18] 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 [30], 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 IV-A, 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 IV.1 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 IV-A, results in a stochastic policy $\pi_\theta(u_k)$ parameterized by $\theta = \left(\{(\alpha_k, \beta_k)\}_{k=1}^{K}, \phi\right)$, the learned Beta distribution shape parameters across all arms and the fixed parameters of the GQ-CNN used for initialization.

### C. Prior Mismatch

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.* [5], termed the *prior mismatch*. However, unlike Chapelle *et al.* [5], 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 $q_p = \left(Q_\phi(o,u_k)\right)_{k=1}^K$ 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 [10, 11]. 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 $\tau$ 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.

## V. Practical Implementation

We implement the method from Section IV in a simulated environment using 3D object models from the Dex-Net 4.0 dataset [18]. 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 [16]. 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. Each parallel jaw grasp is represented by a center point $\mathbf{p} = (x,y,z) \in \mathbb{R}^3$ and a grasp axis $\mathbf{v} \in \mathbb{R}^3$ [19]. They are visualized as whiskers in Figures 1 and 5. 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 camera intrinsics. Note that TSLP can also be easily be applied with different types of grasps such as Suction grasps [17] provided that the actions corresponding to the arms are parameterized accordingly. 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 IV 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 [17]. 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 $T$, Pre-trained 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 α_{k,0}, β_{k,0} ∀ k ∈ {1,2,...K} using
Q_φ(o,u_k) using method from Section IV-B.
for t = 1,...,T do
    Select action u_k using Thompson sampling as in Section IV-A
    Execute u_k and observe r(u_k)
    Update α_{k,t}, β_{k,t} ∀k ∈ {1,2,...K} as in Section IV-A
end for
```

## VI. Experiments

### A. 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.* [18], 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 (III.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 all, and all learning is performed online. 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.

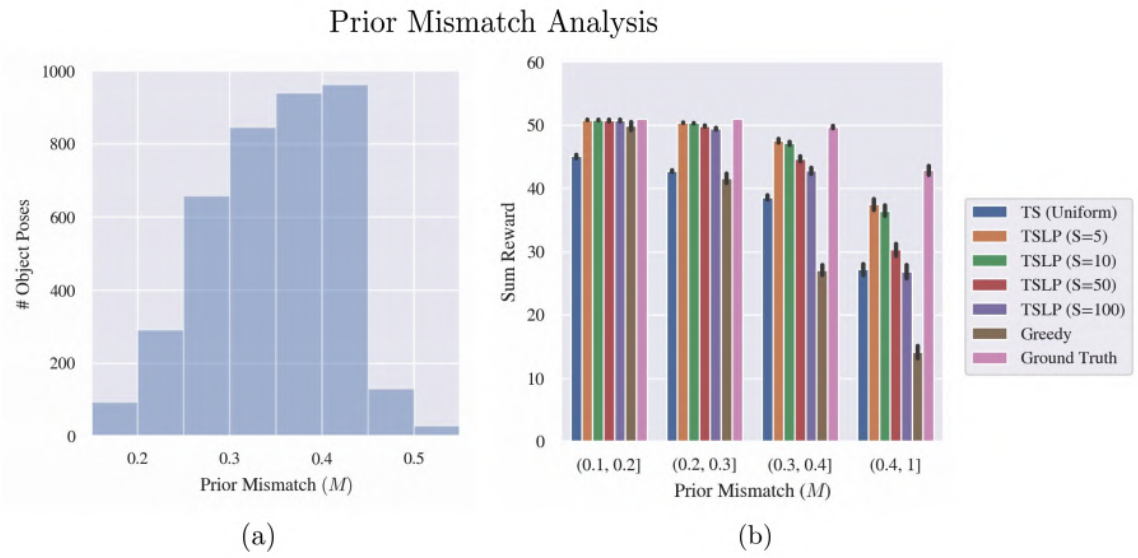


Fig. 3: (a) The distribution of prior mismatch $M$ for the 3946 total object poses in the dataset used by [18]. We find that $M$ ranges from 0.16 to 0.64 and a value between 0.4 and 0.45 is the most common, accounting for about 25% of the object poses. All object poses with $M$ above 0.55 are placed 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.

TABLE I: **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 $\pm$ 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.

| **Greedy** | **TS (Uniform)** | **TSLP (S=5)** | **TSLP (S=10)** | **TSLP (S=50)** | **TSLP (S=100)** | **Ground Truth** |
|---|---|---|---|---|---|---|
| $54.33 \pm 33.02$ | $72.08 \pm 20.67$ | $\mathbf{89.37 \pm 17.88}$ | $88.43 \pm 18.53$ | $82.63 \pm 23.51$ | $78.88 \pm 26.29$ | $94.53 \pm 13.49$ |

We additionally compare the learned policy to a greedy policy that repeatedly selects the grasp with highest quality under GQ-CNN as in [18] and a ground truth oracle policy, which repeatedly selects the grasp with the highest quality under the ground truth grasp quality metrics computed in simulation. The former gives an idea of policy performance if no online exploration is performed, while the latter provides an upper bound on possible performance since it can access the true grasp success probabilities, which are not available to our algorithm.

### *B. Simulation Experiments*

We conduct simulation experiments across object poses with a wide range of prior mismatches $M$, as shown in Figure 3(a), which plots the frequency of prior mismatch values 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(b) and Table I. Figure 3(b) shows policy performance as a function of prior mismatch, given the distribution of objects over prior mismatch values shown in Figure 3(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.

Table I 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 4, 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 ($M = 0.29$),

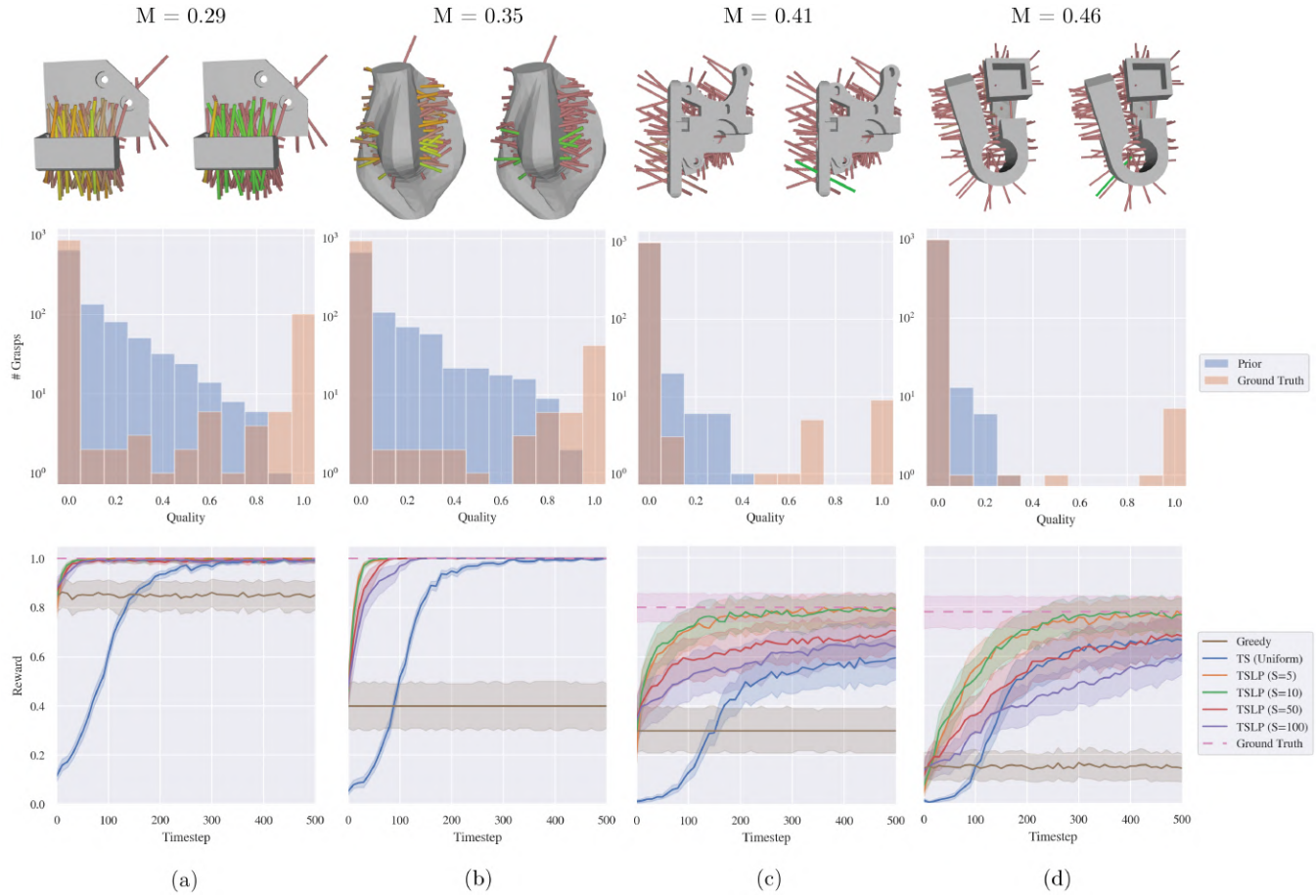


Fig. 4: Visualization of policy performance for all baselines and TSLP policies (labeled with their prior strength). The first row visualizes grasp qualities as measured by the GQ-CNN prior (left) and the ground truth grasp success probabilities (right) for a single stable pose of each of the four objects (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.

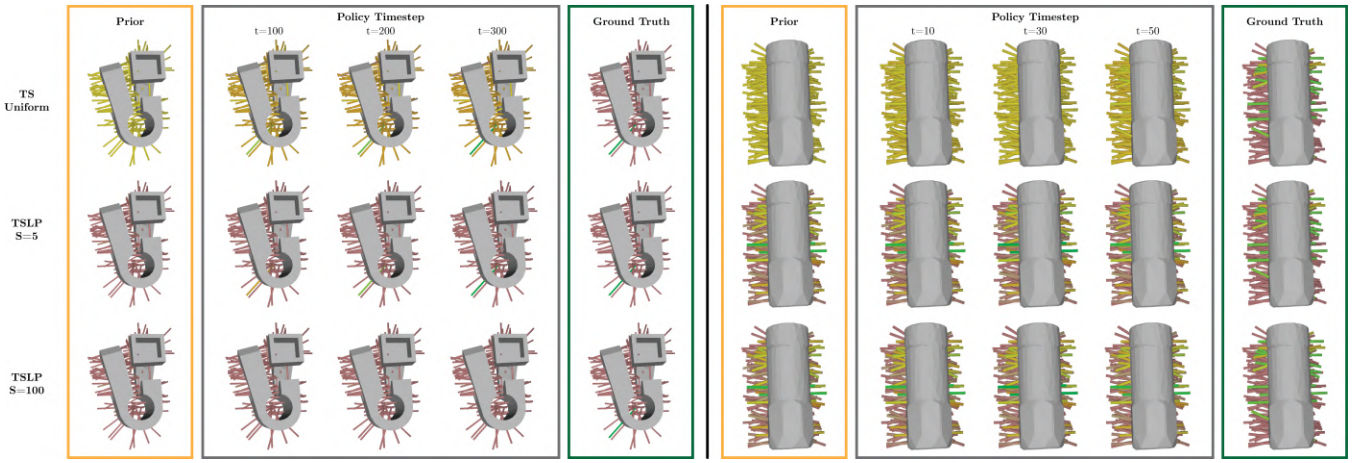


Fig. 5: We visualize the evolution of the mean Bernoulli parameter (defined in Equation (IV.1)) inferred by TSLP with varying prior strengths on sampled arms over learning steps for two different objects. Grasps with high estimated success probabilities or ground truth quality values are colored green, 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.

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 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 5 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.

## VII. Discussion and Conclusion

In this paper, 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.

## VIII. Future Work

In future work, we will design new online learning algorithms to explore grasps across different object stable poses and extend experiments to new grasping modalities, such as suction. In addition, we will explore ways to approximately enforce pose consistency in physical experiments. For example, we can use a string to lift the object after each grasp and put it into pose. Additionally, we can detect stable pose changes by evaluating whether the observed depth image changes in a way that cannot be described by a planar rotation and translation. Using the Super4PCS algorithm [20], we can compute the registration of the new point cloud with respect to the original point cloud and restrict the range of output to planar transformations. If the algorithm cannot find such a planar transformation, we resample grasps on the new pose.

## Acknowledgments

This research was performed at the AUTOLAB at UC Berkeley in affiliation with the Berkeley AI Research (BAIR) Lab, Berkeley Deep Drive (BDD), the Real-Time Intelligent Secure Execution (RISE) Lab, and the CITRIS "People and Robots" (CPAR) Initiative. Authors were also supported by the Scalable Collaborative Human-Robot Learning (SCHooL) Project, a NSF National Robotics Initiative Award 1734633, and in part by donations from Google and Toyota Research Institute. Ashwin Balakrishna and Michael Danielczuk are supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. 1752814. This article solely reflects the opinions and conclusions of its authors and do not reflect the views of the sponsors. We thank our colleagues who provided feedback and suggestions, in particular Brijen Thananjeyan.

## References

[1] 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.

[2] C. Bodnar, A. Li, K. Hausman, P. Pastor, and M. Kalakrishnan, "Quantile qt-opt for risk-aware vision-based robotic grasping", *ArXiv*, vol. abs/1910.02787, 2019.

[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] 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.

[7] 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.

[8] 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.

[9] 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.

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

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

[12] 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.

[13] 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.

[14] 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.

[15] M. Liu, Z. Pan, K. Xu, K. Ganguly, and D. Manocha, "Deep differentiable grasp planner for high-dof grippers", *ArXiv*, vol. abs/2002.01530, 2020.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] N. Mellado, D. Aiger, and N. J. Mitra, *Super 4pcs library*, https://github.com/nmellado/Super4PCS, 2017.

[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] O.-M. Pedersen, E. Misimi, and F. Chaumette, "Grasping Unknown Objects by Coupling Deep Reinforcement Learning, Generative Adversarial Networks, and Visual Servoing", in *ICRA 2020 - IEEE International Conference on Robotics and Automation*, Paris, France: IEEE, May 2020, pp. 1–8.

[24] 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.

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

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

[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] 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.

[30] 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.