Exploratory Grasping: Performance Bounds and Asymptotically Optimal Algorithms for Learning to Robustly Grasp an Unknown Polyhedral Object

Michael Danielczuk^{*1}, **Ashwin Balakrishna**^{*1}, **Daniel Brown**^{2}, **Ken Goldberg**^{1}
* equal contribution
{mdanielczuk, ashwin_balakrishna}@berkeley.edu

Abstract: Learning robust robot grasping policies is an important step for robotics to be viable for commercial automation, such as household assistance, warehousing, and manufacturing. There has been significant prior work on data-driven algorithms for learning general-purpose grasping policies, but these policies can consistently fail to grasp certain objects which are significantly out of the distribution of objects seen during training or which have very few high quality grasps. For such objects, we study strategies for efficient grasp exploration to increase grasp reliability. Precisely, we formalize the problem of efficiently exploring grasps on an unknown polyhedral object through sequential interaction as a Markov Decision Process in a setting where a camera can be used to (1) distinguish stable poses and (2) determine grasp success/failure. We then present a bandit-style algorithm, Exploratory Grasping, which leverages the structure of the grasp exploration problem to rapidly find high performing grasps on new objects through online interaction. We provide vanishing regret guarantees for Exploratory Grasping under certain assumptions on the structure of the grasp exploration problem. Results suggest that Exploratory Grasping can significantly outperform both general-purpose grasping pipelines and two other online learning algorithms and achieves performance near that of the optimal policy on both the Dex-Net adversarial and EGAD! object datasets.

Keywords: Grasping, Online Learning

1 Introduction

Robot grasping systems have a broad array of applications such as warehousing, assistive robotics, and household automation [1, 2, 3, 4]. There has been significant prior work in geometric algorithms for defining grasping policies [5, 6, 7, 8], but these methods can often be difficult to apply when object geometry is unknown. These challenges have motivated a large array of recent work on utilizing large-scale datasets of previously attempted grasps both in simulation [4, 9, 1, 10] and in physical experiments [11, 3, 1, 12] to learn data-driven general-purpose grasping policies which can generalize to objects of varying geometries. To further enable generalization, there has also been work on applying reinforcement learning to learn grasping policies [2, 3, 13, 14]. However, while these techniques have shown significant success in practice, their generality comes at a cost: a single policy which aims to grasp every object may fail to generalize to certain objects [15, 16].

To address this challenge, we study the problem of online grasp exploration to systematically explore sampled grasps on an object to discover robust and stable grasps on a given object with unknown geometry. Specifically, we study a new setting in which a robot is tasked with interacting with an unknown polyhedral object so that it can reliably grasp the object in the future. To this end, we aim to design an algorithm which enables systematic and efficient grasp exploration on objects with sparse grasps and adversarial geometries, which can cause persistent failures in general-purpose grasping systems [16, 15]. The intuition is that while general-purpose grasping policies can be broadly applied to a large set of objects, there is still a large class of objects that cause these systems to struggle [15, 16]. This motivates developing algorithms for grasp exploration to learn

[1] University of California, Berkeley. [2] University of Texas at Austin.

per-object policies for such objects to explore a large set of possible grasps without the burden of generalization to other objects. Then, when these difficult-to-grasp objects are encountered, the resulting object-specific policies can be deployed instead of a general-purpose grasping policy.

We formulate the grasp exploration problem as a Markov Decision Process (MDP) and study how parameters of this MDP affect the fundamental difficulty of grasp exploration. We then present an efficient algorithm which leverages the structure of the grasping problem to efficiently explore grasps across different object stable poses and can quickly learn robust grasping policies even for objects that general-purpose grasping policies routinely fail to grasp. We formalize a new problem setting in which an agent seeks to explore grasps across different stable poses of an object by repeatedly attempting grasps on the object and dropping it when a grasp is successful to randomize its stable pose and provide efficient algorithms for grasp exploration in this setting. This paper contributes (1) a novel formulation of the grasp exploration problem as an MDP and intuitive parameter-dependent performance bounds on a family of existing tabular reinforcement learning algorithms for grasp exploration, (2) an efficient bandit-inspired algorithm, Exploratory Grasping, for grasp exploration in this MDP with associated no-regret guarantees, and (3) experiments suggesting that Exploratory Grasping is able to significantly outperform baseline algorithms which explore via tabular reinforcement learning or select actions greedily with respect to general-purpose grasping policies on both the Dex-Net adversarial and EGAD! object datasets.

2 Related Work

Work in analytic robot grasping assumes that object geometry and pose are known precisely and leverages this knowledge to motivate geometric algorithms for grasp planning [5, 7, 8, 6]. Recently, learning-based algorithms have been used to develop general-purpose algorithms for planning robust grasps on a wide range of objects of varying geometries with data-driven strategies [14, 4, 11, 10, 1, 17, 18, 9] and online exploration through reinforcement learning [2, 3]. While the latter approaches have been very effective in learning end-to-end policies for grasp planning from visual input, these policies can consistently fail on certain objects [15, 16]. In this work, we study the problem of systematic online grasp exploration across different stable poses of specific objects. Thus, in contrast to prior work which attempts to learn a single policy to grasp a wide range of objects, we develop an exploration strategy that can rapidly discover high-quality grasps on specific objects.

There have also been several papers that use a multi-armed bandit framework for online grasp exploration [19, 20, 21, 22, 23, 24, 25]. In contrast to these works, we consider a formulation where the robot must learn grasps across all poses of the object without human supervision. Laskey et al. [20] consider the setting where some prior geometric knowledge is known, but present an algorithm which is limited to 2D objects and cannot operate directly on visual inputs. Li et al. [22] and Oberlin and Tellex [21] relax these assumptions by exploring grasps in one fixed stable pose of a 3D object with RGB or depth observations. We extend these ideas by repeatedly dropping the object and exploring grasps in all encountered stable poses. Thus, Exploratory Grasping naturally explores grasps over the distribution of likely stable poses [26], yielding a robust policy which can reliably grasp the object when randomly dropped or placed in front of the robot. Additionally, we provide formal guarantees establishing asymptotic convergence of Exploratory Grasping to an oracle policy which knows the best quality grasp in each object stable pose in advance.

A key requirement for successful exploration of grasps on an object is discovering the object's resting stable poses, since the object will necessarily be in one of these poses when a grasp is planned. There has been significant prior work on orienting parts into specific stable poses through a series of parallel jaw gripper movements [27], toppling actions [28], and squeezing actions [29]. However, these approaches require knowledge of an object's geometry apriori to plan motions to achieve specific stable poses. When an object's geometry is not known, but assumed to be polyhedral, prior work [26] has established that repeatedly dropping the object from a known initial distribution of poses onto a flat workspace results in an a stationary distribution over stable poses. Thus, this dropping procedure provides a convenient method to explore grasps in new stable poses of an object. We leverage this insight to (1) formulate the grasp exploration problem for an unknown object and (2) develop a grasp exploration algorithm which can discover high-quality grasps across different object stable poses.

3 Grasp Exploration Problem Formulation

Given a single unknown polyhedral object on a planar workspace, the objective is to learn a grasping policy that maximizes the likelihood of grasp success over stable poses of the object [26, 30]. We formulate the grasp exploration problem as a Markov Decision Process (Section 3.1), define assumptions on the environment (Section 3.2) and formulate the policy learning objective (Section 3.3).

3.1 Grasp Exploration as an MDP

We consider exploring grasps on an unknown, rigid, polyhedral object $\mathcal{O}$ which rests in one of a finite set of N distinguishable stable poses with associated drop probabilities $\{\lambda_s\}_{s=1}^{N}$. We study polyhedral objects since they admit a finite number of stable resting poses, but assume that the robot does not initially know any of these stable poses or the number of stable poses N. Note that any object for which a triangular mesh provides a high fidelity model can be well-modeled as polyhedral [31]. The robot must discover new stable poses from experience by attempting grasps on the object in its current stable pose and re-dropping the object when grasps are successful. We assume an overhead digital camera that cannot reliably determine the 3D shape of the object, but can be used to recognize distinguishable 3D poses by performing planar translations and rotations of the image into a canonical orientation and translation. We also assume that the camera can be used after each grasp attempt to determine if the grasp is successful after the gripper is moved out of the camera's field of the view.

We formulate this problem as a Markov decision process (MDP) [32] by defining a *grasp MDP*, $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R)$, as follows:

1. **State:** We define a one-to-many mapping from the full set of object stable poses Σ to the set of overhead point clouds $\mathcal{I}$ that are scale-invariant, translation-invariant, and rotationally-invariant about the vertical axis. Then, we define the state space as the set of *distinguishable* stable poses $\mathcal{S}$. We define $\mathcal{S}$ as the set of equivalence classes within Σ, where two poses are equivalent if they map to the same set of overhead point clouds $\mathcal{I}$. We assume the point cloud $\mathcal{I}$ is obtained from a depth image observation $o \in \mathbb{R}^{+}_{H\times W}$ from an overhead camera with known intrinsics.
2. **Grasp Actions:** We define a set $\mathcal{A}_s$ of K grasp actions, such as parallel-jaw or suction grasps, in each stable pose $s \in \mathcal{S}$ of the object. The K grasps are sampled from the depth image observation for each stable pose as in Mahler et al. [33]. The full action space is the union of the actions available at each pose: $\mathcal{A} = \bigcup\limits_{s\in\mathcal{S}} \mathcal{A}_s$.
3. **Transition Function:** An unknown transition probability distribution $P(s' \mid s, a)$ defining the probability of the object transitioning to stable pose s' if grasp action a is executed in stable pose s. If a is a successful grasp, then the object is dropped into the workspace to sample a new stable pose s' based on unknown drop probabilities $\{\lambda_s\}_{s=1}^{N}$, while if a is a failed grasp, the object topples into some new pose s' with unknown probability $\delta_{s,s'}$.
4. **Reward Function:** Rewards are drawn from a Bernoulli distribution with unknown parameter $\phi_{s,a}$: $R(s, a) \sim \text{Ber}(\phi_{s,a})$ for $a \in \mathcal{A}_s$. $R(s, a) = 1$ if executing a in stable pose s results in the object being successfully grasped and lifted, and 0 otherwise.

3.2 Assumptions

To study policy learning in a grasp MDP, we first establish assumptions on the system dynamics to ensure that all poses are reachable and which describe how the object pose can evolve when the object is (1) dropped or (2) when a grasp is attempted.

Assumption 3.1. Grasp Dynamics: If a grasp succeeds, we assume that the robot can randomize the pose of the object before dropping it to sample subsequent stable poses from the associated unknown stable pose drop probabilities $\{\lambda_s\}_{s=1}^{N}$ for $\mathcal{O}$. If a grasp fails, we assume that the object's pose will either remain unchanged or topple into some other pose s' with unknown probability $\delta_{s,s'}$.

Assumption 3.2. Drop Dynamics: The categorical distribution over stable poses defined by drop probabilities $\{\lambda_s\}_{s=1}^{N}$ is a stationary distribution that is independent of prior actions and poses when $\mathcal{O}$ is dropped from a fixed height with its orientation randomized as in Goldberg et al. [26].

Assumption 3.3. Irreducibility: We assume that there exists a policy π such that the Markov chain over stable poses induced by executing π in grasp MDP $\mathcal{M}$ is irreducible, and thus can reach all stable poses with nonzero probability for any initialization.

Note that Assumption 3.3 is satisfied if, for all poses $s \in \mathcal{S}$, $\lambda_s > 0$ and there exists a grasp with success probability $\epsilon > 0$. As a result, we assume that these conditions hold for analysis. However, since object toppling is also possible, note that these conditions are sufficient but not necessary for ensuring irreducibility.

3.3 Learning Objective

The objective is to learn a policy $\pi : \mathcal{S} \to \mathcal{A}$ that maximizes the expected average reward over an infinite time horizon under the state distribution induced by π. Let $\tau = \{(s_t, \pi(s_t))\}_{t=1}^{T}$ be a trajectory of all states and actions when executing policy π over some time horizon T and let $r(\tau) = \sum_{t=1}^{T} r(s_t, \pi(s_t))$ be the sum of rewards for all states and actions in τ, and let $p(\tau|\pi)$ be the trajectory distribution induced by policy π. Then the expected average reward obtained from policy π in grasp MDP $\mathcal{M}$ is given as:

$$J(\mathcal{M}, \pi, T) = \frac{1}{T}\mathbb{E}_{\tau \sim p(\tau|\pi)}\left[r(\tau)\right] \tag{1}$$

The objective is to find the policy which maximizes expected reward over an infinite time horizon:

$$\pi^* = \underset{\pi}{\operatorname{argmax}} \lim_{T \to \infty} J(\mathcal{M}, \pi, T) \tag{2}$$

4 Reinforcement Learning for Grasp MDPs

We first study the performance of reinforcement learning algorithms for grasp exploration and leverage the structure of the grasp MDP described in Section 3 to establish a bound on the cumulative regret for a variety of existing tabular reinforcement learning algorithms when applied to grasp MDPs. See Section A of the supplementary material for all proofs.

4.1 Analyzing Grasp MDPs

A common metric with which to measure policy performance in online-learning settings is *regret*, which has been analyzed in the reinforcement learning setting by a variety of prior work [34, 35, 36, 37]. Intuitively, regret quantifies the difference in accumulated reward within T timesteps between a given policy π and optimal infinite horizon policy π^* for MDP $\mathcal{M}$. More precisely, we define average regret based on the definition in [34]:

$$\text{Regret}(\mathcal{M}, \pi, T) = \max_{\pi'}\left[\lim_{T \to \infty} J(\mathcal{M}, \pi', T)\right] - J(\mathcal{M}, \pi, T) \tag{3}$$

Recent theoretical work [34] on reinforcement learning for tabular MDPs yielded algorithms which can attain average regret proportional to the diameter of the MDP, a measure of the furthest distance between pairs of states under an appropriate policy. However, for general MDPs, this diameter can be arbitrarily large, making these regret bounds difficult to interpret in practical settings. We leverage the structure of grasp MDPs to derive an intuitive upper bound on the MDP diameter, which helps precisely quantify the difficulty of grasp exploration based on the parameters of the grasp MDP.

We begin by defining the Markov chain over $\mathcal{S}$ induced by a stationary deterministic policy π.

Definition 4.1. Pose Evolution under π: Given stationary policy π, the transitions between pairs of states in $\mathcal{M}$ is defined by a Markov chain. Precisely, the transition probabilities under π, denoted by P^π where $P^\pi[s, s'] = P(s' \mid s, a = \pi(s))$, are given as follows:

$$P^\pi[s, s'] = \phi_{s,\pi(s)}\lambda_{s'} + (1 - \phi_{s,\pi_s(s)})\delta_{s,s'} \tag{4}$$

Given this Markov Chain over poses for a given policy π, we can now analyze the diameter of the grasp MDP, denoted $D(\mathcal{M})$, by considering the hitting time between stable poses in $\mathcal{M}$ as defined in [38]. Note that by Assumption 3.3, the Markov chain corresponding to P^π is irreducible, and we can compute the hitting time between any pair of states in closed form [38].

Definition 4.2. Let $T^{\pi}_{s \to s'}$ denote the hitting time between states s and s' under policy π under the Markov chain defined in Definition 4.1. Then the diameter of $\mathcal{M}$ is defined as follows [34]:

$$D(\mathcal{M}) = \max_{s \neq s'} \min_{\pi} \mathbb{E}\left[T^{\pi}_{s \to s'}\right] \tag{5}$$

Intuitively, $D(\mathcal{M})$ measures the temporal distance between the furthest apart states in an MDP under the policy which minimizes this distance. We now leverage the structure of the grasp MDP to establish an upper bound on $D(\mathcal{M})$.

Lemma 4.1. *The diameter of the grasp MDP $\mathcal{M}$ can be bounded above as follows:*

$$D(\mathcal{M}) \leq \frac{1}{\epsilon \lambda_1}, \tag{6}$$

where ϵ is a lower bound on the success probability of the highest quality grasp over all stable poses and λ_1 is the drop probability for the least likely stable pose.

Lemma 4.1 captures the intuition that the diameter of the MDP should be large if the best grasp in each stable pose has a low success probability (ϵ is small), or if there exists a stable pose with very low drop probability (λ_1 is small).

Now we can establish regret bounds for a variety of tabular reinforcement learning algorithms when applied to $\mathcal{M}$ by combining diameter dependent regret bounds from prior work and the bound on grasp MDP diameter established in Lemma 4.1.

Theorem 4.1. *UCRL2 [34], KL-UCRL [36] and PSRL [39] admit average regret bounded above as follows for any grasp MDP $\mathcal{M}$:*

$$\textit{Regret}(\mathcal{M}, \pi, T) \sim \tilde{O}\left(\frac{N}{\epsilon \lambda_1}\sqrt{\frac{K}{T}}\right), \tag{7}$$

while UCRLV [35] admits the following average regret bound for grasp MDP $\mathcal{M}$.

$$\textit{Regret}(\mathcal{M}, \pi, T) \sim \tilde{O}\left(\sqrt{\frac{NK}{\epsilon \lambda_1 T}}\right), \tag{8}$$

with N, K, T, ϵ and λ_1 defined as in Section 3.

While the diameter can be uninterpretable and difficult to directly compute or bound for general MDPs [34], Theorem 4.1 leverages the specific structure of grasp MDPs to relate the accumulated regret of common RL algorithms to intuitive parameters of the grasp MDP. This result also serves to shed light on the fundamental difficulty of grasp exploration in the context of reinforcement learning.

5 A Bandit-Style Algorithm for Efficient Learning in Grasp MDPs

One interesting feature of the grasp MDP $\mathcal{M}$ is that most objects have a small, finite set of stable poses [26]. This motivates learning a set of N bandit policies, each of which explore their grasps in a particular object stable pose. However, the grasp exploration problem in each pose is not necessarily decoupled. For example, there may exist a pose s with no available high quality grasps but a high likelihood of a failed grasp causing the object to topple into another pose with high quality grasps. Then, the optimal policy may deliberately fail to grasp the object in pose s in order to obtain access to grasps in the more favorable stable pose, leading it to avoid grasp exploration in poses without high quality grasps. To avoid this behavior, we introduce Assumption 5.1.

Assumption 5.1. We assume that $\delta_{s,s'} \leq \epsilon \lambda_{s'}$ for all $s \neq s'$ where $\delta_{s,s'}$ is the probability of toppling into pose s' given a failed grasp in pose s, ϵ is a lower bound on the success probability of the highest quality grasp over all stable poses, and $\lambda_{s'}$ is the drop probability of pose s'.

Assumption 5.1 ensures that there exists a grasp in all stable poses s such that the probability of transitioning to new pose s' via a grasp attempt is higher than that of toppling from pose s to pose s'. Given this assumption, the optimal grasp exploration policy in $\mathcal{M}$ reduces to selecting the grasp with highest success probability in each encountered pose, as this policy maximizes both reward at the current timestep and exploration of other stable poses when the object is re-dropped. In other words, the global optimal policy is the *greedy* policy. Given this structure, we can view the grasp exploration problem as N independent multi-armed bandit problems corresponding to grasp exploration in each pose. However, although grasp exploration can be performed independently in each pose, the success of a grasp exploration policy in one pose affects the time available to explore grasps in another pose.

We propose a simple and intuitive grasp exploration algorithm, Exploratory Grasping: maintain the parameters of N independent bandit policies $(\pi_s^{\mathcal{B}})_{s=1}^N$ for $\pi_s^{\mathcal{B}} : s \to \mathcal{A}_s$ where $\pi_s^{\mathcal{B}}$ is only *active* in stable pose s. We let $\pi^{\mathcal{B}}$ denote the meta policy induced by executing bandit policy $\pi_s^{\mathcal{B}}$ in pose s and assume that $\pi_s^{\mathcal{B}}$ is learned by running a no-regret online learning algorithm $\mathcal{B}$ for grasp exploration in pose s. Some examples of no-regret algorithms for the stochastic multi-armed bandit problem include the UCB-1 algorithm [40] and Thompson Sampling [41]. We then formulate a new notion of regret capturing the gap between $\pi^{\mathcal{B}}$ and the optimal policy on their respective distributions, and show that Exploratory Grasping achieves vanishing average regret despite the interdependence between pose exploration times.

Let $p_T^{\mathcal{B}}$ denote the distribution of poses seen under the sequence of policies $\pi_{1:T}^{\mathcal{B}}$ at each round of learning up to time T and let p_T^* denote the distribution of poses seen when executing the optimal policy (π^*) in $\mathcal{M}$ up to time T. We define the average regret accrued by Exploratory Grasping in grasp MDP $\mathcal{M}$ after T rounds as the difference in accumulated reward of the optimal policy on pose distribution p_T^* and the accumulated reward of $\pi^{\mathcal{B}}$ on pose distribution $p_T^{\mathcal{B}}$.

Definition 5.1. The average regret accumulated by running $\mathcal{B}$ in each stable pose is defined as the difference between the average regret for each pose visited by the optimal policy π^* weighted by the probability of it visiting each pose and the corresponding quantity for the executed policy $\pi^{\mathcal{B}}$:

$$\mathbb{E}\left[\mathcal{R}^{\mathcal{B}}(T)\right] = \sum_{s=1}^{N} p_T^*(s)\mathbb{E}\left[\frac{1}{T_s^*}\sum_{t=1}^{T_s^*} R(s,\pi^*(s))\right] - \sum_{s=1}^{N} p_T^{\mathcal{B}}(s)\mathbb{E}\left[\frac{1}{T_s^{\mathcal{B}}}\sum_{t=1}^{T_s^{\mathcal{B}}} R(s,\pi_t^{\mathcal{B}}(s))\right]$$

where T_s^* is the time spent by the optimal policy in pose s and $T_s^{\mathcal{B}}$ is the time spent by $\pi^{\mathcal{B}}$ in pose s.

In Section A of the supplementary material, we show that the average regret as defined in Definition 5.1 vanishes to 0 in the limit as $T \to \infty$ as stated in Theorem 5.1.

Theorem 5.1. *The average regret accrued by Exploratory Grasping, when using any no-regret bandit algorithm $\mathcal{B}$ for grasp exploration in each encountered stable pose, vanishes in the limit:*

$$\lim_{T\to\infty} \mathbb{E}\left[\mathcal{R}^{\mathcal{B}}(T)\right] = 0$$

This result leverages the precise structure of the grasp MDP, namely that the optimal policy is the greedy policy, to provide sublinear regret guarantees for Exploratory Grasping, which executes any standard no-regret online learning algorithm for grasp exploration in each stable pose.

6 Simulation Experiments

In simulation experiments, we study three questions: (1) Does Exploratory Grasping facilitate grasp exploration on objects for which general-purpose grasping policies, such as Dex-Net 4.0 [4] and GG-CNN [1], perform poorly? (2) Does learning separate policies for each stable pose as in Exploratory Grasping accelerate grasp exploration? (3) Can Exploratory Grasping be applied in realistic settings in which objects may topple due to failed grasps? Unfortunately due to the pandemic, we were unable to run physical experiments for the paper. However, since Exploratory Grasping samples grasps on image based observations of the object and is evaluated in a similar simulation environment as used in Mahler et al. [42], which transferred learned grasping policies to physical bin picking environments, we expect that Exploratory Grasping can likely be extended to grasp exploration in the real world.

6.1 Experimental Setup

To evaluate each policy, we choose 7 objects from the set of Dex-Net 2.0 adversarial objects [33] as well as 39 evaluation objects from the EGAD! dataset [43]. We use these objects in experiments because general-purpose grasping policies for parallel-jaw grippers (Dex-Net 4.0 and GG-CNN) perform poorly on them, yet they contain high-quality grasps in multiple poses. We sample a set of K parallel-jaw grasps on the image observation of pose s of each object as in [33], and calculate the ground-truth quality of each grasp, $\phi_{s,a}$, using a wrench resistance metric that measures the ability of the grasp to resist gravity [42]. We remove poses that have no grasps with nonzero ground-truth quality and renormalize the stable pose distribution. Then, we randomize the initial pose of each object and execute Exploratory Grasping and baselines, sampling rewards from Ber($\phi_{s,a}$). If the grasp succeeds, we randomize the pose, choosing a stable pose according to the stable pose distribution of the object. Otherwise, we leave the object in the same stable pose. We rollout each policy for 10 rollouts of 10 trials, where new grasps are sampled for each trial and each rollout evaluates the policy over 10,000 timesteps of grasp exploration. Since there is stochasticity in both the grasp sampling and the policies themselves, we average policy performance across the 10 rollouts and 10 trials. In addition, we smooth policy performance across a sliding window of 20 timesteps and report average reward for each timestep.

6.2 Policies

We compare 3 variants of Exploratory Grasping against 3 baselines to evaluate whether Exploratory Grasping is able to (1) substantially outperform general-purpose grasping policies [4] on challenging objects and (2) learn more efficiently than other online learning algorithms which update a grasp quality convolutional network (GQCNN) online or explore grasps via reinforcement learning. We also instantiate Exploratory Grasping with different algorithms for $\mathcal{B}$ to study how this choice affects grasp exploration efficiency and implement an oracle baseline to study whether Exploratory Grasping is able to converge to the optimal policy. We compare the following baselines and Bandit Grasping variants: **GQCNN**, which selects grasps greedily with respect to the Grasp Quality Convolutional Network from Dex-Net 4.0 [4] with probability 0.95 and selects a grasp uniformly at random with probability 0.05, **GQCNNFT**, which additionally fine tunes the GQCNN policy online from agent experience, an implementation of **UCRL2** from [44], a tabular RL algorithm discussed in Section 4, and instantiations of Exploratory Grasping with the UCB-1 algorithm [40] (**Exploratory Grasping (UCB)**), Thompson sampling (**Exploratory Grasping (TS)**) with a uniform Beta distribution prior, and Thompson sampling with a GQCNN prior of strength 5 (**Exploratory Grasping (TS-5)**) as in [22]. Finally, we implement an oracle baseline that chooses grasps with the best ground-truth metric at each timestep to establish an upper bound on performance.

6.3 Policy Learning without Toppling

We first evaluate the above baselines and Exploratory Grasping variants in a setting in which toppling is not possible ($\delta_{s,s'} = 0,\ \forall s \neq s$). This exacerbates the difficulty of grasp exploration since an object must be successfully grasped in a given pose for policies to be able to explore grasps in other poses. As shown in Figure 1, the GQCNN policy typically performs very poorly, achieving an average reward of less than 0.1 per timestep. While the online learning policies also start poorly, they quickly improve, and Exploratory Grasping (TS) eventually converges to the optimal policy. We also find that Exploratory Grasping (TS-5), which leverages GQCNN as a prior using the method presented in [22], further speeds convergence to the optimal policy. This result is promising, as it suggests that the exploration strategy in Exploratory Grasping can be flexibly combined with general-purpose grasping policies to significantly accelerate grasp exploration on unknown objects. We find that the GQCNNFT policy performs very poorly even though it continues to update the weights of the network online with the results of each grasp attempt and samples a random grasp with probability 0.05, which aids exploration. We hypothesize that this is due to the initial poor performance of GQCNN—since the vast majority of fine-tuning grasps attain zero reward, the network is unable to explore enough high-quality grasps on the object. Overall, these results suggest that Exploratory Grasping can greatly increase grasp success rates on objects for which GQCNN performs poorly.

We also find that Exploratory Grasping policies greatly outperform the tabular RL policy UCRL2 by leveraging the structure of the grasp MDP. Both the UCB and Thompson sampling implementations maintain separate policies in each pose, thereby leveraging the fact that the optimal policy at each timestep is the greedy policy. This allows the Exploratory Grasping policies to not waste timesteps exploring the possible transitions to other states with low rewards. Thompson sampling additionally leverages the fact that the rewards are distributed as Bernoulli random variables, which may explain the significant performance gap between the Thompson sampling and UCB implementations. Thus, the Exploratory Grasping policies quickly learn to choose high-quality grasps in each pose and transition quickly to new, unexplored poses.

We also perform sensitivity analysis of Exploratory Grasping to the grasp MDP parameters ϵ and λ_1 and find that Exploratory Grasping quickly converges to the optimal policy unless ϵ or λ_1 is low. In particular, we find that ϵ has an outsized effect on performance, which is intuitive given that ϵ affects the ability of Exploratory Grasping to both achieve immediate grasp successes *and* explore new poses, while λ_1 only influences the latter. See Section C of the supplementary material for details.

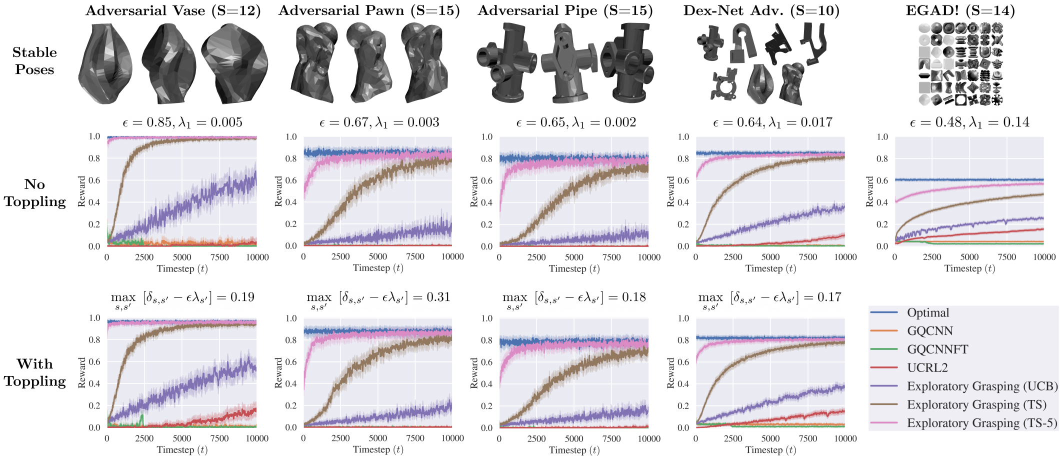

Figure 1: **Simulation Experiments:** Performance of each policy across the vase, pawn and pipe objects from the Dex-Net 2.0 adversarial object set [33] (first three columns), as well as aggregated performance over the 7 Dex-Net objects (fourth column) and the 39 EGAD! objects (fifth column). We report the number of distinguishable stable poses (S) for each object, as well as its ϵ (ground truth value for the lowest quality best grasp across poses) and λ_1 (least likely stable pose probability) values, and show a top-down view of its three most likely stable poses. We report mean values of each metric for the datasets. The first row of plots shows reward over time for each policy in the original setting, where the object stays in the same pose until successfully grasped. The second row shows policy performance in the more realistic setting where the object may topple into an alternate pose when a grasp is unsuccessful. The value above each plot with toppling indicates the maximum difference in transitioning via toppling and transitioning via grasping to a new pose. In both cases, Exploratory Grasping quickly converges to optimal performance, while the other algorithms fail to reach optimal performance even after 10,000 grasp attempts. This result holds even when Assumption 5.1 is violated.

6.4 Policy Learning with Toppling

We repeat the same experiments as in the previous section, with the additional condition that each object may now topple into another pose when a grasp is unsuccessful (Section 3.2). We use the toppling analysis from Correa et al. [28] to determine the toppling transition matrix for a given object. Specifically, we generate the distribution of next states from a given state by sampling non-colliding pushes across vertices on the object, finding their distribution of next states given perturbations around the nominal push point, and average the distribution from all of the pushes. Then, during policy rollouts, if a grasp fails, we choose the next state according to the corresponding topple transition probabilities. See the last row of Figure 1 for learning curves for Exploratory Grasping and baselines. Results suggest that even in cases where values of $\delta_{s,s'}$ are considerably larger than $\epsilon\lambda_{s'}$, Exploratory Grasping policies still achieve significantly better performance than baselines. These results suggest that although Assumption 5.1 is needed for analysis, in practice Exploratory Grasping can also efficiently explore grasps on objects where Assumption 5.1 is violated.

7 Conclusion

We study a new problem setting in which a robot is tasked with exploring grasps across stable poses of an object by repeatedly attempting grasps and dropping the object to explore grasps on new object poses. We formalize this problem as an MDP and study how the difficulty of grasp exploration depends on intuitive parameters of this MDP. We then present Exploratory Grasping, an efficient, bandit-inspired, algorithm for exploring grasps on an unknown polyhedral object. Exploratory Grasping can be flexibly used with any no-regret online learning algorithm such as UCB or Thompson sampling and results suggest that Exploratory Grasping is able to explore grasps on unknown polyhedral objects significantly more efficiently than online learning algorithms which do not effectively leverage the structure of the grasping problem. In future work, we will explore applying Exploratory Grasping to grasp exploration in a physical bin picking setup. One nice property of Exploratory Grasping is that in the process of exploring grasps, it also explores different stable poses of the object. Thus, we are excited to explore further applications of Exploratory Grasping beyond grasp exploration. For example, Exploratory Grasping could be used to explore different object poses to construct accurate 3D models of unknown objects or inspect parts for defects.

References

[1] D. Morrison, P. Corke, and J. Leitner. Learning robust, real-time, reactive robotic grasping. *Int. Journal of Robotics Research (IJRR)*, 39(2-3):183–201, 2020.

[2] 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.

[3] 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)*, 37 (4-5):421–436, 2018.

[4] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg. Learning ambidextrous robot grasping policies. *Science Robotics*, 4(26):eaau4984, 2019.

[5] A. Bicchi and V. Kumar. Robotic grasping and contact: A review. In *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, volume 1, pages 348–353, 2000.

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

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

[8] J. Kim, K. Iwamoto, J. J. Kuffner, Y. Ota, and N. S. Pollard. Physically based grasp quality evaluation under pose uncertainty. *IEEE Trans. Robotics*, 29(6):1424–1439, 2013.

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

[10] U. Viereck, A. ten 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.

[11] 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)*, pages 3406–3413, 2016.

[12] C. Choi, W. Schwarting, J. DelPreto, and D. Rus. Learning object grasping for soft robot hands. *IEEE Robotics & Automation Letters*, 3(3):2370–2377, 2018.

[13] M. Breyer, F. Furrer, T. Novkovic, R. Siegwart, and J. Nieto. Comparing task simplifications to learn closed-loop object picking using deep reinforcement learning. *IEEE Robotics & Automation Letters*, 4(2): 1549–1556, 2019.

[14] O. Kroemer, S. Niekum, and G. Konidaris. A review of robot learning for manipulation: Challenges, representations, and algorithms. *arXiv preprint arXiv:1907.03146*, 2019.

[15] D. Wang, D. Tseng, P. Li, Y. Jiang, M. Guo, M. Danielczuk, J. Mahler, J. Ichnowski, and K. Goldberg. Adversarial grasp objects. In *Proc. IEEE Conf. on Automation Science and Engineering (CASE)*, pages 241–248, 2019.

[16] K. Sanders, M. Danielczuk, J. Mahler, A. Tanwani, and K. Goldberg. Non-markov policies to reduce sequential failures in robot bin picking. In *Proc. IEEE Conf. on Automation Science and Engineering (CASE)*, 2020.

[17] I. Lenz, H. Lee, and A. Saxena. Deep learning for detecting robotic grasps. *Int. Journal of Robotics Research (IJRR)*, 34(4-5):705–724, 2015.

[18] A. Saxena, J. Driemeyer, and A. Y. Ng. Robotic grasping of novel objects using vision. *Int. Journal of Robotics Research (IJRR)*, 27(2):157–173, 2008.

[19] M. Laskey, Z. McCarthy, J. Mahler, F. T. Pokorny, S. Patil, J. Van Den Berg, D. Kragic, P. Abbeel, and K. Goldberg. Budgeted multi-armed bandit models for sample-based grasp planning in the presence of uncertainty. In *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2015.

[20] 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 *Proc. IEEE Conf. on Automation Science and Engineering (CASE)*, pages 572–579, 2015.

[21] J. Oberlin and S. Tellex. Autonomously acquiring instance-based object models from experience. In *Int. S. Robotics Research (ISRR)*, pages 73–90. Springer, 2015.

[22] K. Li, M. Danielczuk, A. Balakrishna, V. Satish, and K. Goldberg. Accelerating grasp exploration by leveraging learned priors. In *Proc. IEEE Conf. on Automation Science and Engineering (CASE)*, 2020.

[23] O. B. Kroemer, R. Detry, J. Piater, and J. Peters. Combining active learning and reactive control for robot grasping. *Robotics and Autonomous systems*, 58(9):1105–1116, 2010.

[24] C. Eppner and O. Brock. Visual detection of opportunities to exploit contact in grasping using contextual multi-armed bandits. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, pages 273–278, 2017.

[25] Q. Lu, M. Van der Merwe, and T. Hermans. Multi-fingered active grasp learning. *arXiv preprint arXiv:2006.05264*, 2020.

[26] K. Goldberg, B. V. Mirtich, Y. Zhuang, J. Craig, B. R. Carlisle, and J. Canny. Part pose statistics: Estimators and experiments. *IEEE Trans. Robotics and Automation*, 15(5):849–857, 1999.

[27] K. Goldberg. Orienting polygonal parts without sensors. *Algorithmica*, 1993.

[28] C. Correa, J. Mahler, M. Danielczuk, and K. Goldberg. Robust toppling for vacuum suction grasping. In *Proc. IEEE Conf. on Automation Science and Engineering (CASE)*, 2019.

[29] K. Y. Goldberg and M. T. Mason. Bayesian grasping. In *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, pages 1264–1269, 1990.

[30] M. Moll and M. A. Erdmann. Manipulation of pose distributions. *Int. Journal of Robotics Research (IJRR)*, 21(3):277–292, 2002.

[31] A. S. Rao and K. Y. Goldberg. Manipulating algebraic parts in the plane. *IEEE Transactions on Robotics and Automation*, 11(4):598–602, 1995.

[32] M. L. Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[33] 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.

[34] T. Jaksh, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. In *Journal of Machine Learning Research*, 2010.

[35] A. Tossou, D. Basu, and C. Dimitrakakis. Near-optimal optimistic reinforcement learning using empirical bernstein inequalities. In *Proceedings of Machine Learning Research*, 2019.

[36] S. Filippi, O. Cappé, and A. Garivier. Optimism in reinforcement learning and kullback-leibler divergence. In *Annual Allerton Conference on Communication, Control, and Computing*, 2010.

[37] I. Osband and B. Van Roy. Near-optimal reinforcement learning in factored mdps. In *Proc. Advances in Neural Information Processing Systems*, 2014.

[38] H. Chen and F. Zhang. The expected hitting times for finite markov chains. *Linear Algebra and its Applications*, 428(11-12):2730–2749, 2008.

[39] Z. Xu and A. Tewari. Near-optimal reinforcement learning in factored mdps: Oracle-efficient algorithms for the non-episodic setting. In *Proc. Int. Conf. on Machine Learning*, 2018.

[40] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[41] S. Agrawal and N. Goyal. Further optimal regret bounds for thompson sampling. In *Int. Conf. on Artificial Intelligence and Statistics (AISTATS)*, 2013.

[42] 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.

[43] D. Morrison, P. Corke, and J. Leitner. Egad! an evolved grasping analysis dataset for diversity and reproducibility in robotic manipulation. *IEEE Robotics & Automation Letters*, 2020.

[44] R. Fruit. Exploration-exploitation in reinforcement learning. `https://github.com/RonanFR/UCRL`, 2018.

Exploratory Grasping: Performance Bounds and Asymptotically Optimal Algorithms for Learning to Robustly Grasp an Unknown Polyhedral Object Supplementary Material

A Proofs

Here we provide the proofs for all results in Section 4 and Section 5 in the main text.

A.1 Proof of Lemma 4.1

Consider grasp MDP $\mathcal{M}'$ for which the object stable pose does not change when a grasp fails. Thus, it must be the case that $\delta_{l,m} = 0$ when $l \neq m$ and $\delta_{l,l} = 1 \; \forall l$. Note that for any grasp MDP $\mathcal{M}$, it must be the case that $D(\mathcal{M}) \leq D(\mathcal{M}')$ since the probability of transition between any pair of distinct states in $\mathcal{M}'$ is at most the probability of transition in $\mathcal{M}$. Now we establish a bound on $D(\mathcal{M})$ by bounding $D(\mathcal{M}')$.

Without loss of generality, let $\lambda_1 \leq \lambda_s \; \forall s \in \mathcal{S}$. Additionally, define π^* as the policy which selects the grasp with highest success probability on all poses, with associated probability transition matrix P^{π^*} and with hitting time $T^{\pi^*}_{s \to s'}$ defined as in Definition 4.1. Then, the diameter of $\mathcal{M}'$ can be computed as follows.

For MDP $\mathcal{M}'$, it must be the case that

$$\min_{\pi} T^{\pi}_{s \to s'} = T^{\pi^*}_{s \to s'} \tag{9}$$

since π^*, the policy which always picks the highest quality grasp on each pose, minimizes the hitting time between any pair of poses s, s'. Furthermore, note that

$$\max_{s \neq s'} T^{\pi^*}_{s \to s'} = \max_{s} T^{\pi^*}_{s \to 1} \tag{10}$$

since for any starting pose s, the hitting time between s and s' will always be highest for $s' = 1$ (the pose with lowest drop probability) for any policy π. Thus, we see that

$$D(\mathcal{M}') = \max_{s} T^{\pi^*}_{s \to 1} \tag{11}$$

Finally, we leverage equation 11 to compute an upper bound on $D(\mathcal{M}')$ as follows.

Let π_ϵ be any policy which selects a grasp with success probability ϵ on each stable pose l. Note that by Assumption 3.3 we assume that there exists a grasp with success probability at least ϵ on each pose. Without loss of generality, we can consider the case that there exists a grasp with success probability exactly ϵ on each pose and the policy which selects these grasps since the hitting times under this policy will only be lower than those under a policy which selects grasps with success probability greater than ϵ. Then, it follows that

$$\min_{\pi} T^{\pi}_{s \to s'} \leq T^{\pi_\epsilon}_{s \to s'} \; \forall s, s' \tag{12}$$

Now note that since we are considering pose evolution under π_ϵ, which selects a grasp of the same quality ϵ on any pose, the starting pose s does not affect the hitting time. Combining this with the fact that the hitting time to the least likely pose (pose 1) will always be the highest for any policy π and inequality (12) yields that

$$D(\mathcal{M}') \leq \max_{s \neq s'} T^{\pi_\epsilon}_{s \to s'} = T^{\pi_\epsilon}_{2 \to 1} \tag{13}$$

Since the choice of s does not matter under π_ϵ, we use $s = 2$ above without loss of generality. Now, note that the hitting time to pose 1 under π_ϵ is distributed as a geometric random variable with parameter $\epsilon\lambda_1$, which has mean $\frac{1}{\epsilon\lambda_1}$, yielding the desired result. □

A.2 Proof of Theorem 4.1

The result immediately follows from combining the diameter bound from Lemma 4.1 and the regret bounds established for UCRL2 [1], KL-UCRL [2] and PSRL [3] (average regret $\tilde{O}(DS\sqrt{A/T})$) and for UCRLV [4] (average regret $\tilde{O}(\sqrt{DSA/T})$) where D is the MDP diameter and S and A are the cardinalities of the state space and action space respectively. □

A.3 Proof of Theorem 5.1

We bound the expected regret of Exploratory Grasping by decomposing the regret into two terms, one which depends on the divergence between the distribution of poses seen by the optimal policy and $\pi^{\mathcal{B}}$, and the other of which depends on the difference in rewards attained by $\pi^{\mathcal{B}}$ and π^* when evaluated on the distribution of poses seen by $\pi^{\mathcal{B}}$. For simplicity, we refer to $\pi^{\mathcal{B}}$ as π for the proof.

$$\mathbb{E}\left[\mathcal{R}^{\mathcal{B}}(T)\right] = \sum_{s=1}^{S} p_T^*(s)\mathbb{E}\left[\frac{1}{T_s^*}\sum_{t=1}^{T_s^*} R(s,\pi^*(s))\right] - \sum_{s=1}^{S} p_T^\pi(s)\mathbb{E}\left[\frac{1}{T_s^\pi}\sum_{t=1}^{T_s^\pi} R(s,\pi_t(s))\right] \tag{14}$$

$$= \sum_{s=1}^{S}\left(p_T^*(s) - p_T^\pi(s)\right) g_s^*(T_s^*) + \sum_{s=1}^{S} p_T^\pi(s)\left(g_s^*(T_s^*) - g_s^\pi(T_s^\pi)\right) \tag{15}$$

$$= \mathbb{E}\left[\mathcal{R}_\pi^{\mathcal{B}}(T)\right] + \sum_{s=1}^{S}\left(p_T^*(s) - p_T^\pi(s)\right) g_s^*(T_s^\pi) \tag{16}$$

where (15) follows from letting $g_s^*(T_s^*) = \mathbb{E}\left[\frac{1}{T_s^*}\sum_{t=1}^{T_s^*} R(s,\pi^*(s))\right]$ and $g_s^\pi(T_s^\pi) = \mathbb{E}\left[\frac{1}{T_s^\pi}\sum_{t=1}^{T_s^\pi} R(s,\pi_t(s))\right]$ and (16) follows from letting $\mathbb{E}\left[\mathcal{R}_\pi^{\mathcal{B}}(T)\right]$ denote the expected regret on the distribution of poses visited by policy π and noting that the average reward for the optimal policy, g_s^*, is independent of the timesteps spent in the pose (i.e., $g_s^*(T_s^\pi) = g_s^*(T_s^*)\ \forall s$).

We first focus on the first term in (16). We know that $\mathbb{E}\left[\mathcal{R}_\pi^{\mathcal{B}}(T)\right]$ approaches zero if each pose is visited infinitely often in the limit as $T \to \infty$ provided that $\mathcal{B}$ is a no-regret online learning algorithm:

$$\lim_{T\to\infty} p_T^\pi(s) > 0, \forall s \in \{1, 2, \ldots, S\} \quad \Rightarrow \quad \lim_{T\to\infty} \mathbb{E}\left[\mathcal{R}_\pi^{\mathcal{B}}(T)\right] = 0 \tag{17}$$

Thus, it remains to be shown that under π, all poses are visited infinitely often in the limit as $T \to \infty$. Note that this statement is equivalent to showing that in the limit as $T \to \infty$, bandit algorithm $\mathcal{B}$ selects grasps on each pose with non-zero success probability with non-zero probability. Suppose that this was not the case (i.e., that as $T \to \infty$, $\mathcal{B}$ assigns zero grasp probability to all grasps with non-zero success probability). This would imply that $\mathcal{B}$ only selects grasps with zero success probability, and thus incurs constant regret on its own distribution ($\lim_{T\to\infty} \mathbb{E}\left[\mathcal{R}_\pi^{\mathcal{B}}(T)\right] > 0$). This contradicts the initial assumption that $\mathcal{B}$ is a no-regret online learning algorithm, showing that under π, all poses must be visited infinitely often in the limit as $T \to \infty$.

Now we shift our attention to the second term in (16). Given that $\mathcal{B}$ is a no-regret online learning algorithm, it must be the case that $g_s^\pi(T_s^\pi) \xrightarrow[T_s^\pi \to \infty]{} g_s^*(T_s^\pi)\ \forall s$. This implies that in the limit as $T \to \infty$, π and π^* have the same success rate on all stable poses. Two policies with the same success rate on all stable poses induce the same Markov chain over $\mathcal{S}$, and thus admit the same stable pose distribution. Thus, $p_T^\pi(s) \xrightarrow[T \to \infty]{} p_T^*(s)$, implying that the second term also approaches 0 as $T \to \infty$. □

B Experimental Details

Object Selection: We choose 7 objects from the set of adversarial objects in Dex-Net 2.0 [5] because these objects had empirically been shown to be difficult to grasp for the Dex-Net policy. Similarly, the recently-introduced EGAD! object dataset [6] was created to contain objects with few high-quality parallel-jaw grasps. For this dataset, we select all objects for which there exists at least one sampled grasp of quality $\epsilon = 0.1$ on at least one stable pose of the object. Of the 49 objects in the EGAD evaluation dataset, 39 met this criterion.

Pose Selection: For each of the objects, we remove stable poses from the distribution in simulation if they occur with less than a 0.1% chance or if they do not contain a sampled grasp with quality at least $\epsilon = 0.1$. When a pose is removed, the remaining stable pose distribution is renormalized.

Grasp Sampling: We sample a set of $K = 100$ parallel-jaw grasps on the image observation of each pose of each object as in [5]. This sampling process is done using the depth image grasp sampler from the GQCNN repository and is repeated for up to 10 iterations. At each iteration, the sampled grasps' ground truth qualities are calculated using the robust wrench resistance metric that measures the ability of the grasp to resist gravity [7]. If no grasps are found with quality of at least $\epsilon = 0.1$, then the sampling process is repeated for another iteration where more grasps are sampled. If a grasp is found with quality at least ϵ, then that grasp is selected along with 99 other grasps chosen at random from the sampled grasps. If the maximum number of iterations are exceeded without finding a grasp with quality ϵ, the stable pose is discarded.

C Sensitivity Experiments

C.1 Sensitivity to Grasp MDP Parameters

We perform sensitivity analysis of Exploratory Grasping to the grasp MDP parameters ϵ and λ_1. For these experiments, we evaluate the policy using a set of synthetic objects with $\lambda_1 = \{0.001, 0.01, 0.1, 0.2\}$ and $\epsilon = \{0.1, 0.25, 0.5, 0.75, 1.0\}$ and for simplicity consider the case in which toppling is not possible. In each case, we choose a single grasp on each pose to have reliability ϵ with all other grasps having a mean parameter of 0. The results are shown in Figure 1. These results suggest that unless ϵ or λ_1 are low, Exploratory Grasping quickly converges to the optimal policy. In particular, ϵ has an outsized effect on the accumulated reward; with $\epsilon \leq 0.10$, we observe that the policy fails to approach the optimal policy regardless of λ_1 through the first 10,000 timesteps.

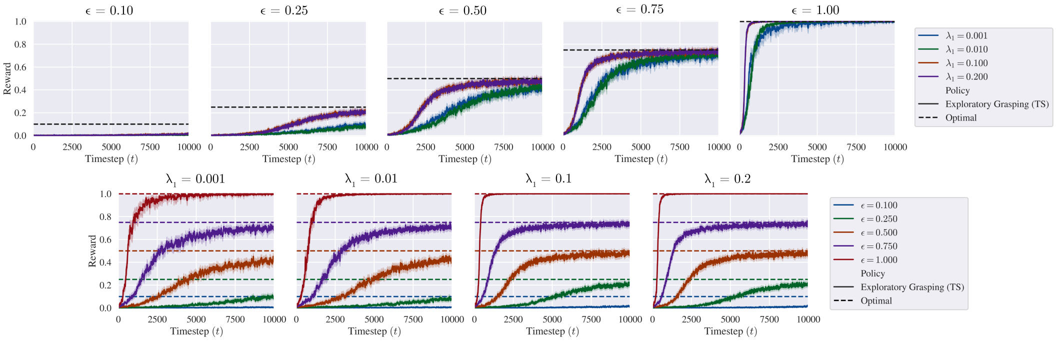

Figure 1: Sensitivity of Exploratory Grasping to the grasp MDP parameters ϵ and λ_1, as shown across 20 synthetic objects. Unless ϵ or λ is low, Exploratory Grasping quickly converges to the optimal policy. However, when either is low, particularly ϵ, the policy converges much more slowly, taking even more than the 10,000 timesteps shown for some combinations.

References

[1] T. Jaksh, R. Ortner, and P. Auer. Near-optimal regret bounds for reinforcement learning. In *Journal of Machine Learning Research*, 2010.

[2] S. Filippi, O. Cappé, and A. Garivier. Optimism in reinforcement learning and kullback-leibler divergence. In *Annual Allerton Conference on Communication, Control, and Computing*, 2010.

[3] Z. Xu and A. Tewari. Near-optimal reinforcement learning in factored mdps: Oracle-efficient algorithms for the non-episodic setting. In *Proc. Int. Conf. on Machine Learning*, 2018.

[4] A. Tossou, D. Basu, and C. Dimitrakakis. Near-optimal optimistic reinforcement learning using empirical bernstein inequalities. In *Proceedings of Machine Learning Research*, 2019.

[5] 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.

[6] D. Morrison, P. Corke, and J. Leitner. Egad! an evolved grasping analysis dataset for diversity and reproducibility in robotic manipulation. *IEEE Robotics & Automation Letters*, 2020.

[7] 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.