Efficient Planar Sample-Based Grasp Planning with Uncertainty Using Budgeted Multi-Armed Bandit Models for [v3 Jan 5, 2015]

Michael Laskey[1],Jeff Mahler[1], Zoe McCarthy[1], Florian T. Pokorny[3], Sachin Patil[1], Jur Van Den Berg[4], Danica Kragic[3], Pieter Abbeel[1], Ken Goldberg[2]

***Abstract*—[TODO: NOT SURE IF WE NEED TO CHANGE THE AUTHORS LIST OR NOT]Sampling perturbations in shape, state, and control can facilitate grasp planning in the presence of uncertainty arising from noise, occlusions, and surface properties such as transparency and specularities. Monte-Carlo sampling is computationally demanding, even for planar models. We consider an alternative based on the multi-armed bandit (MAB) model for making sequential decisions, which can apply to a variety of uncertainty models. We formulate grasp planning as a "budgeted multi-armed bandit model" (BMAB) with finite stopping time to minimize "simple regret", the difference between the expected quality of the best grasp and the expected quality of the grasp evaluated at the stopping time. To evaluate MAB-based sampling, we compare it with Monte-Carlo sampling for grasping an uncertain planar object with shape uncertainty defined by a Gaussian process implicit surface (GPIS), but the method is also applicable to other models of uncertainty. We derive distributions on contact points, surface normal, and center of mass under shape uncertainty and use these to formulate the associated MAB model, finding that it computes grasps of similar quality to Monte-Carlo sampling and can reduce computation time by an order of magnitude.**

***Note to Practitioners*—Planning for a grasp in an unknown environment can be difficult due to uncertainties. For example, a given object may have transparency which makes modern kinect-like sensor unable to accurately determine shape or an object may have a built up of mildew or dust, which makes the friction coefficient unknown. Monte-Carlo integration is often used to exhaustively evaluate the distribution on a grasp quality metric given these uncertainties. We apply algorithms from the sequential decision making literature to intelligently allocate samples for grasp quality evaluation to grasps likely to have higher quality given the previous samples, reducing plan time by roughly an order of magnitude.**

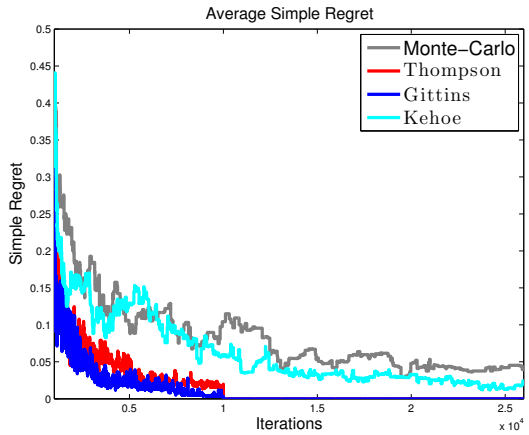


Fig. 1: [TODO: NEW TEASER PHOTO, NOT SURE IF WE SHOULD HAVE SOME SORT OF REAL OBJECT OR NOT](Top Left) Image of a common household measuring cup. (Top Right) Image from a PR2 Primesense camera that reflects the uncertainty that can be induced in objects from transparency. (Bottom) Comparision of Multi-Armed Bandit Techniques (Bayes UCB, Thompson, Gittins) vs. Monte-Carlo sampling to determine the best grasp in a set of 1000 grasps. As you can see the bandit techniques converge in simple regret Eq. 2 a magnitude faster than the traditional approach of Monte-Carlo sampling to determine the highest quality grasp.

I. Introduction

Consider a robot processing orders in a warehouse, where it frequently encounters new consumer products and must process them quickly. The robot may need to rapidly plan grasps for these objects without prior knowledge of object shape, pose and material properties like friction coefficient or center of mass. Furthermore, the robot may not be able to measure these quantities exactly due to sensor imprecision and missing data, which could result from occlusions or surface properties such as transparency or reflective surfaces.

Analytic grasp quality metrics have been developed to determine the stability of a grasp when all the parameters of the object and robot manipulator are exactly known. One common measure of stability is force closure, the ability to resist external forces and torques in arbitrary directions []. Grasps in force closure can be further ranked by the relative magnitude of forces and torques that must be exerted by the gripper to resist external perturbations [12]. Recent works have explored computing the probability of a grasp achieving the force closure condition given uncertainty in parameters such as pose [10], [42], [22] and object shape [20], [28]. Many methods for evaluating grasp quality in the presence of uncertainty use exhaustive Monte-Carlo sampling over the possible values of uncertain quantities [10], [22], [42], [20], [**?**] which can be very time consuming. We are interested in ruling out grasps with a low probability in only a few samples and allocate more sampling effort to grasps that are likely to be high quality.

The multi-armed bandit (MAB) model for sequential decision making problems [5], [24], [**?**] provides a formal way to reason about allocating sampling effort to grasps such that the grasp(s) with highest quality can be determined in as few

[1]Department of Electrical Engineering and Computer Sciences; {mdlaskey, zmccarthy, jmahler, sachinpatil, pabbeel}@berkeley.edu

[2]Department of Industrial Engineering and Operations Research and Department of Electrical Engineering and Computer Sciences; goldberg@berkeley.edu

[1−2] University of California, Berkeley; Berkeley, CA 94720, USA

[3]Computer Vision and Active Perception Lab, Centre for Autonomous Systems, School of Computer Science and Communication, KTH Royal Institute of Technology, Stockholm, Sweden {fpokorny, dani}@kth.se

[4]Google; Amphitheatre Parkway, Mountain View, CA 94043, USA jurvandenberg@gmail.com

samples as possible. In a standard MAB there are a set of possible options (or 'arms' in the literature [5]) that each return a numeric reward from a stationary distribution. The goal in a MAB problem is to sequentially select ones of the possible options such that a measure of expected reward is maximized. Solutions to the MAB model are particularly useful in applications where it is too expensive to fully evaluate a set of options; for example, in optimal design of clinical trials [37], market pricing [36], and choosing strategies for games [39]. The budgeted multi-armed bandit model [27] is a specialization of the MAB model where an agent has a fixed number of decisions to make before choosing the best option. The objective is to maximize the expected reward of the decision made at the stopping time, or equivalently to minimize "simple regret", which is the difference between the true expected reward of an optimal arm and the true expected reward of the arm pulled at the stopping time.

Our main contribution is formulating the problem of ranking a set of candidate grasps according to a quality metric in the presence of uncertainty as a budgeted multi-armed bandit problem. We study this formulation using probability of force closure [10], [42], [21] as a quality metric under uncertainty in pose, shape, robot motion, and friction coefficient. We model shape uncertainty using Gaussian process implicit surfaces (GPISs), a Bayesian representation of shape uncertainty that has been used in various robotic applications [11], [16]. Uncertainty in pose is represented as a normal distributions around the orientation and translation of the object. Uncertainity in motion is represented as a normal distribution around the end point of a planned gripper trajectory and uncertainty in friction coefficient is a normal distribution around an expected friction coefficient. [TODO: JEFF: I PERSONALLY THINK THE ABOVE 3 STATEMENTS MAY BE TOO MUCH DETAIL FOR AN INTRO] Our approach of using BMAB is applicable to any distributions that can be sampled from, however performance could potentially vary. We compare the performance of several popular algorithms for solving the BMAB problem: Bayes UCB, Thompson sampling, and Gittins indices. [TODO: GOING TO FILL BOTTOM PARAGRAPH IN LATER] Initial results suggest a promising avenue with a Xx improvement over the traditional Monte-Carlo integration and Xx improvement in the experiments that we tried.

II. Related Work

Past work on grasping under uncertainty has considered shape uncertainty [14], [40], uncertainty in contact locations with an object [44], and uncertainty in object pose [10], [42], [22]. The effect of uncertainty in object geometry on grasp selection has been studied for spline representations of objects [10], extruded polygonal mesh models [20], [21], and point clouds [17]. A common method for evaluating a probabilistic grasp quality measure is to use Monte-Carlo sampling [10], [20], [21], which involves exhaustively sampling from distributions on random quantities and averaging the quality over these samples to approximate an expected value [8]. Exhaustive sampling can be computationally expensive, which motivated the use of Cloud Computing to distribute sampling effort [21].

To further address the computational complexity, Kehoe et al. [20] demonstrated a procedure for finding a minimum bound on expected grasp quality given shape uncertainty, which reduced the number of samples needed in Monte-Carlo sampling to choose the highest quality grasps. However, the proposed adaptive sampling approach pruned grasps using only the sample mean and did not utilize any estimates of how accurate the current sample mean is, which in practice could lead to good grasps being thrown away. Laaksonen et al. [23] used Markov Chain Monte-Carlo (MCMC) sampling to estimate grasp quality and object pose under shape and pose uncertainty when the robot is able to obtain new information via tactile sensor. MCMC provides a Bayesian framework for inference with a hidden state, but it can be slow to converge to the correct distribution due to burn in and mixing conditions [2].

We chose to study our MAB sampling method on distributions for friction coefficient, pose, motion and shape. For shape uncertainty we decided to use a Gaussian process implicit surface representation. Our decision to use this uncertainty model is based on GPIS's ability to combine different modes of noise observations such as tactile, laser and visual [33], [43], [11] and its recent use in modeling uncertainty for a number of robotic applications. Hollinger et al. used GPIS as a model of uncertainty to perform active sensing on the hulls in underwater boats [16]. Dragiev et al. showed how GPIS can enable a grasp controller on the continuous signed distance function [11]. Mahler et al. used the GPIS representation to find locally optimal anti-podal grasps by framing grasp planning as an optimization problem [28]. However, this relied on an approximation to grasp quality without guarantees on accuracy. We propose an adaptive sampling approach known as the Multi-Armed Bandit Model. This paper is a substantially revised and expanded version on [**?**] and [28].

III. Preliminaries and Problem Definition

We consider selecting a grasp on object from a given set of candidate grasps. We formulate this problem for grasping a planar object from above using parallel-jaw grippers. We assume that the object is rigid and remains stationary throughout the grasp. In this work we consider uncertainty in shape, pose, robot motion, and friction coefficient and we assume that distributions on these quantiites are given. In this section, we formally define our grasping model, formalize our sources of uncertainty, introduce our metric for grasp quality, and finally define our grasp planning objective.

A. Candidate Grasp Model

Our candidate grasp model is illustrated in Fig. 2. Let d denote the dimensionality of the shape representations we are using to select grasps and m denote the number of jaws on the robotic gripper. In this work we formulate the MAB problem for 2-dimensional shapes using parallel-jaw grippers, so $d = 2$ and $m = 2$. In this work we also assume a finite width for each parallel jaw $w_j \in \mathbb{R}$.

Similar to [10], we parameterize a grasp using a *line of action* for each gripper jaw, where each line of action is a 1D

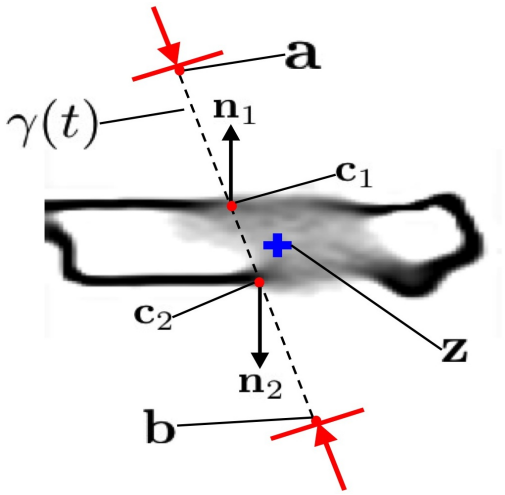


Fig. 2: llustration of our grasping model for parallel jaw grippers on a GPIS model of a marker with shape uncertainty near the object center. Jaw placements are illustrated by a red direction arrow and line. The grasp plan consists of a line of action $\gamma(t)$ with endpoints $\mathbf{a}$ and $\mathbf{b}$. When following the grasp plan, the jaws contact the shape at locations $\mathbf{c}_1$ and $\mathbf{c}_2$ with outward pointing unit surface normals $\mathbf{n}_1$ and $\mathbf{n}_2$. Together with the center of mass of the object $\mathbf{z}$, these values can be used to determine the forces and torques that a grasp can apply to an object. [TODO: JEFF: UP TO THIS POINT I DON'T BELIEVE THE READER HAS SEEN A GPIS. WE WILL REPLACE THE MARKER WITH A DETERMINISTIC SHAPE, BUT THIS FIGURE GIVES A GOOD IDEA OF WHAT WE WANT TO ILLUSTRATE HERE.]

curve $\gamma(t) : [0, 1] \rightarrow \mathbb{R}^d$ with endpoints $\gamma(0) = \mathbf{a}$ and $\gamma(1) = \mathbf{b}$. The line of action defines the trajectory that a jaw follows as the gripper closes around a shape. A *grasp plan* is the set of lines of action for each of the jaws, $\Gamma = \{\gamma_1(\cdot), ..., \gamma_m(\cdot)\}$ []. For parallel-jaw grippers, $\Gamma = \{\gamma(t), \gamma(1-t)\}$, since the two jaws approach the shape in opposite directions. Furthermore, in this work we consider only straight lines of action. The line of action model allows us to compute perturbations to the contact points and normals that occur due to shape uncertainty.

Given a grasp plan and a deterministic shape, we define the *contact points* as the spatial locations at which the jaws come into contact with the object when following the given plan, $\mathbf{c}_1, ..., \mathbf{c}_m \in \mathbb{R}^d$. We also refer to the unit outward pointing surface normals at the contact points as $\mathbf{n}_1, ..., \mathbf{n}_m \in \mathbb{R}^d$ and the object center of mass as $\mathbf{z} \in \mathbb{R}^d$. Together these form the set of grasp parameters $g = (\mathbf{c}_1, ..., \mathbf{c}_m, \mathbf{n}_1, ...\mathbf{n}_m, \mathbf{z})$ that enable us to evaluate the forces and torques that a given grasp can apply to an object.

B. Sources of Uncertainty

In this work we consider uncertainty in shape, pose, robot motion, and friction coefficient. Fig. 3 illustrates a graphical model of the relationship between these sources of uncertainty. In this section we describe each source of uncertainty and our model of the uncertainty. We wish to emphasize that the models of uncertainty described in this work are not the only models that may be used with MAB algorithms and that the reader may choose parameters and distributions that best fit their needs, however performance is subject to vary on empirical results.

1) Shape Uncertainty

Uncertainty in object shape results from sensor noise and missing sensor data, which can occur due to transparency, specularity, and occlusions [28]. Following [**?**], [28] we represent the distribution over possible surfaces given sensing noise using a Gaussian process implicit surface (GPIS). A GPIS represents a distribution over signed distance functions (SDFs), a surface representation commonly used in 3D reconstruction and SLAM []. A SDF is a real-valued function $f : \mathbb{R}^d \rightarrow \mathbb{R}$ that is greater than 0 outside the object, 0 on the surface and less than 0 inside the object. A GPIS is a gaussian distribution over SDF values at a fixed set of query points $\mathcal{X} = \{\mathbf{x}_1, ...\mathbf{x}_n\}, \mathbf{x}_i \in \mathbb{R}^d$, $f(\mathbf{x}_i) \sim \mathcal{N}(\mu_f(\mathbf{x}_i), \Sigma_f(\mathbf{x}_i))$, where $\mu_f(\cdot)$ and $\Sigma_f(\cdot)$ are the mean and covariance functions of the GPIS. See Appendix A for detail on how to train a mean and covariance function for a GPIS. In this work, we set $\mathcal{X}$ to a uniform $M \times M$ grid of points with square cells. For convenience, in later sections we will refer to the GPIS parameters as $\theta = (\mu_f(x), \Sigma_f(x))$.

[TODO: JEFF: DESCRIBE ALTERNATIVE REPRESENTATION OF SHAPE UNCERTAINTY THAT WE RUN EXPERIMENTS ON. LIKELY THIS WILL BE BEN'S POLYGONAL MODEL]

2) Pose Uncertainty

In typical robotics applications the pose of objects in the envinroment is determined by registering the object frame-of-reference to the control frame-of-reference used for grasp execution. Therefore pose uncertainty may come from either a) uncertainty about the registration of the robot's grasping frame-of-reference to its sensing frame-of-reference and b) uncertainty about the pose of known object models in the robot's sensor data. In practice this uncertainty may be quantified based on the algorithms used to register these frames to one another. The effects of pose uncertainty on robotic grasping has been studied by [42], [22].

In 2-dimensional space, the pose of an object T is a member of the Lie Algebra $SE(2)$. This matrix is defined by a rotation angle ϕ and two translation coordinates $\mathbf{t} = (t_x, t_y)$, summarized in parameter vector $\xi = (\phi, \mathbf{t})^T \in \mathbb{R}^3$:

$$T = \begin{bmatrix} \cos(\phi) & -\sin(\phi) & t_x \\ \sin(\phi) & \cos(\phi) & t_y \\ 0 & 0 & 1 \end{bmatrix}.$$

One challenge with pose is that the pose matrix T is used to apply pose perturbations to an object in practice, but uncertainty is mathematically more easily quantified in terms of the pose parameters ξ. Folowing Barfoot and Furgale, we assume that we are given a mean pose matrix $\bar{T} \in SE(3)$ and zero-mean Gaussian uncertainty on the pose parameters $\xi \sim \mathcal{N}(\mathbf{0}, \Sigma_\xi)$, and we define the pose random variable T as

$$T = \exp(\xi^\wedge)\,\bar{T}$$

where $\xi^\wedge$ is the twist operator as defined in [4].

3) Motion Uncertainty

In practice a robot may not be able to execute a desired grasp plan Γ exactly due to slight errors in actuation or feedback measurements used for trajectory following [20]. In this work, we model motion uncertainty as Gaussian uncertainty

around the angle of approach and centroid of a straight line grasp plan Γ. Formally, let $\hat{\mathbf{y}} = \frac{1}{2}(\mathbf{a}+\mathbf{b})$ denote the center of a planned line of action $\gamma(t)$ and $\hat{\psi}$ denote the angle that the planned line $\mathbf{b}-\mathbf{a}$ makes with the x-axis of the 2D coordinate system on our shape representation. Then the random center $\mathbf{y} \sim \mathcal{N}(\hat{\mathbf{y}}, \Sigma_y)$ and the random angle $\psi \sim \mathcal{N}(\hat{\psi}, \sigma_\psi^2)$. For shorthand in the remainder of this paper we will refer to the random motion parameters as $\rho = \{\mathbf{y}, \psi\}$. In practice Σ_y^2 and σ_ψ^2 might be set from repeatibility measurements for a robot [32].

4) Friction Uncertainty

As shown in [44], uncertainty in friction coefficient can cause grasp quality to significantly vary. The expected friction coefficient $\hat{\mu}$ could be derived by means of object classification and a look up table []. However, friction coefficients may be uncertain due to factors such as material between a gripper and an object (e.g. dust, water, moisture), variations in the gripper material due to manufacturing tolerances, or misclassification of the object surface to be grasped. We model uncertainty in friction coefficient as Gaussian noise, $\mu \sim \mathcal{N}(\hat{\mu}, \sigma_\mu^2)$.

C. Grasp Quality

We measure the quality of grassp using the probability of force closure [42], [22], [20], [21] given a grasp plan Γ, which we denote $P_F(\Gamma)$. Force closure measures the ability to resist external wrenches, or force and torques vectors, assuming the grasp can apply infinite force.

Formally, force closure is a binary-valued quantity F that is 1 if a grasp can resist wrenches in arbitrary directions and 0 otherwise. Let $\mathcal{W} \in \mathbb{R}^6$ denote the contact wrenches derived from contact locations $\mathbf{c}_1, ...\mathbf{c}_m$, normals $\mathbf{n}_1, .., \mathbf{n}_m$, friction coefficient μ, and center of mass $\mathbf{z}$ for a given grasp and shape. If the origin lies within the convex hull of $\mathcal{W}$, then the grasp is in force closure[26]. In this work we rank grasps using the probability of force closure given uncertainty in shape, pose, robot motion, and friction coefficient [10], [21]:

$$P_F(\Gamma) = P(F = 1|\Gamma, \theta, \xi, \rho, \mu).$$

[TODO: JEFF: REVISE ABOVE SYMBOLS FOR CLARITY]

To estimate $P_F(\Gamma)$, we generate samples from each of the above distributions in sequence using the relationships defined by the graphical model in Fig. 3.To sample from $p_F(\Gamma)$, we need to sample from the distributions associated with a line of action $p(\mathbf{n}_i, \mathbf{c}_i|\gamma_i(t), \xi, \theta, \rho)$. Using Bayes rule we can rewrite this as

$$p(\mathbf{n}_i, \mathbf{c}_i|\gamma_i(t), \theta, \xi, \rho) =$$
$$p(\mathbf{n}_i|\mathbf{c}_i, \theta)p(\mathbf{c}_i|\gamma_i(t), \theta, \rho, \xi)$$

Appendix A, describes how to draw shape sample from a GPIS model, which is used to compute $p(\mathbf{c}_i|\gamma_i(t), \theta, \rho, \xi)$ along with the other sampled distribution on pose (ξ) and motion (ρ). Appendix C, describes how to sample from $p(\mathbf{n}_i|\mathbf{c}_i, \theta)$ and presents a novel visualization technique for the distribution on surface normals. Appendix D, describes a way to calculate the expected center of mass assuming a uniform mass distribution. Then we use these quantities to determine the grasp parameters g. Finally, we compute the forces and torques that can be applied by g to form the contact wrench set $\mathcal{W}$ and evaluate the force closure condition.

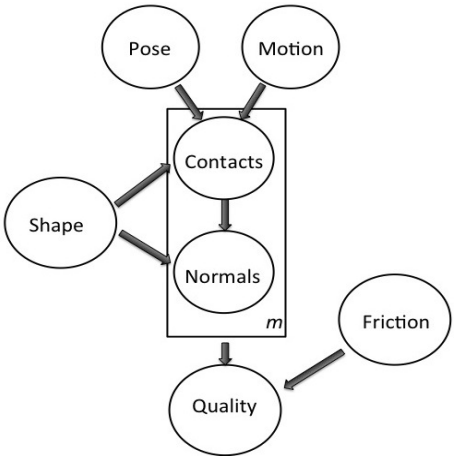


Fig. 3: A graphical model that illustrates the relationship between the different types of uncertainty in an object. Center of Mass uncertainty is dependent on the pose and shape of the object, however friction coefficient is independent of all other types. [TODO: SHOULD WE USE WORDS OR SYMBOLS FOR THE GRAPHICAL MODEL?]

D. Objective

Given the sources of uncertainty and their relationships as described above, our goal is to find the grasp that maximizes the probability of force closure from a set of P prespecified candidate grasps $\mathcal{G} = \{\Gamma_1, ..., \Gamma_P\}$:

$$\Gamma^* = \underset{\Gamma \in \mathcal{G}}{\text{argmax}}\ P\left(F = 1|\Gamma, \theta, \xi, \rho, \mu\right) \quad (1)$$

One method to solve Equation 1 is to exhaustively evaluate $P_F(\Gamma)$ for all grap plans in $\mathcal{G}$ using Monte-Carlo integration and then sort the plans by this quality metric. This method has been evaluated for shape uncertainty [10], [20] and pose uncertainty [42] but is computationally expensive since it may require many samples for each of a large set of candidates to converge to the true value. More recent works have considered adaptive sampling to discard grasp plans that are not likely to be optimal without fully evaluating their quality [21] and searching for locally optimal grasp plans over a continuous set using Sequential Convex Programming [**?**]. [TODO: JEFF: DIFFERENTIATE FROM THESE] In this work we show that this objective can be framed as a budgeted multi-armed bandit (BMAB) problem.

IV. Multi-Armed Bandits for Grasp Selection

We propose framing the problem in the multi-armed bandit (MAB) model and forming a policy for iteratively selecting which grasp to evaluate based on the probability of force closure estimate from the samples so far. The goal of the MAB approach is to allocate more sampling effort to grasps that appear to have higher quality based on the evaluation performed so far.

The multi-armed bandit model, originally described by Robbins [35], is a statistical model of an agent attempting to make a sequence of correct decisions while concurrently

gathering information about each possible decision. In a traditional MAB problem, a gambler has K independent slot machines, or "arms" to play. When an arm is played (or "pulled" in the literature), it returns an amount of money from a fixed reward distribution, $P_k, k = 1, ..., K$, that is unknown to the gambler. The goal of the gambler is to come up with a method for determining which arms to pull, how many times to pull each arm, and what order to pull them in such that the average cumulative rewards is maximized over many pulls. If the gambler knew the machine with the highest expected reward, the gambler would only pull that arm. However, since the reward distributions are unknown, a successful gambler needs to trade off exploiting the arms that currently yields the highest reward and exploring new arms to see if they give better rewards on average. Developing a policy that successfully trades between exploration and exploitation to maximize average reward has been the focus of extensive research since the problem formulation [7], [35], [6].

Success in MAB problems is commonly measured in terms of *regret*, the difference between the expected optimal reward and the expected reward of the selected arm on a single pull. Traditional bandit algorithms minimize cumulative regret, the sum of regret over the entire sequence of arm choices. Lai and Robbins showed that an optimal solution to the bandit problem is bounded by a logarithmic function of the number of arm pulls [24]. They presented an algorithm called (Upper Confidence Bound) UCB that obtains this bound asymptotically [24]. The algorithm maintains a confidence bound on the distribution of reward based on prior observations and pulls the arm with the highest upper confidence bound. Many variants of UCB have thus been proposed []. [TODO: JEFF: CITE A LOT OF THE LITERATURE HERE] Since then a several other algorithms have been shown to achieve this bound, such as the Gittins index policy [] and Thompson sampling for certain reward distributions [].

Our grasp selection problem is similar in that we would like to find the grasp with the highest quality while conserving computational resources. However, in our problem one only usually cares about the quality when the final grasp is selected for execution (the difference between the expected quality of the optimal arm and the arm selected is known as "simple regret" in the literature) rather than the cumulative regret as studied in the standard MAB problem. This is an instance of the Pure Exploration multi-armed bandit problem[7], in which an agent attempts to recommend the best arm at the end of an exploration phase but is not penalized for making exploratory arm pulls during this phase. Hence, the exploration and exploitation stage are decoupled. The end of the exploration phase, or "stopping time," may not be known to the agent ahead of time. Therefore, solutions to the Pure Exploration problem can be thought of as anytime algorithms, meaning they must return a valid solution whenever they are stopped.

Formally, given arms $\{\Gamma_1, ..., \Gamma_K\}$ with respective mean rewards $\mu_1, ..., \mu_K$, the goal in the Pure Exploration problem is to find the optimal arm $\mu^* = \max_{k \in \{1,...,K\}} \mu_k$. The expected simple regret, or suboptimality, at time t is given by

$$E[r_t] = \mu^* - \mu_t \tag{2}$$

where μ_t is the estimate of the best arm at time t from the previous observations.

Several recent works have studied Pure Exploration algorithms that lead to upper bounds on the expected regret. Audibert el al. demonstrated an algorithm called Successive Rejects that divides up the total budget into successively shorter phases and discards the worst arm left at the end of each phase. This algorithm can return the best arm with near-optimal probability depending on the hardness of the problem [3]. In addition, UCB-like methods have been proposed that measure a confidence gap and then pull the arm with the highest confidence interval [13]. For example, Bubeck et al. showed that the simple regret at any time for the UCB algorithm is bounded by a polynomial function of the number of timesteps [7].

In this work, we frame the grasp selection problem of Section **??** as a Pure Exploration multi-armed bandit problem. Each arm corresponds to a different grasp plan and pulling an arm corresponds to sampling from the graphical model in Fig. 3 and evaluating the force closure condition. Since force closure is a binary value, we can think of each grasp plan Γ_i as having a Bernoulli reward distribution with probability of success $P_F(\Gamma_i)$. Thus, the expected value of the force closure condition is equal to $P_F(\Gamma_i)$, and algorithms that seek to minimize the expected simple regret will also minimize the suboptimiality of the grasp plan chosen.

In Section V, we will discuss some of the most popular algorithms for solving the multi-armed bandit problem in detail.

V. Bandit Algorithms

[TODO: JEFF: I FIND THIS TRANSITION A BIT AWKWARD BUT I'M NOT SURE HOW TO REORG RIGHT NOW]. The algorithms discussed in Section Section **??** are frequentist, meaning the algorithms treat unknown parameters as fixed and use confidence intervals to score the arms. However, recently Bayesian approachs have gained interest in the MAB community due to their empirical success on real world problems such as ad suggestions [19] [1]. In Bayesian algorithms, the agent maintains a belief distribution over the reward distribution on the arms. The belief distribution enables a trade off between exploration and exploitation by naturally increasing confidence around the true reward distribution as more rewards are observed for an arms. Theoretical results have shown that these methods are capable of achieving the lower bound described by Lai and Robbin [1] [19]. Furthermore in an empirical study Bayesian methods have been shown to outperform the UCB family [9]. See [] for a review of bayesian methods in multi-armed bandit problems.

In the context of grasping, the reward for achieving force close on a sample from the uncertain parameters is a Bernoulli random variable with probabilty of success θ. In the Bayesian setting we treat θ as belonging to a distribution. A common choice for such a distribution is the Beta distribution, which is the conjugate prior of the Bernoulli distribution. One benefit of a conjugate prior is that the posterior update of the belief distribution is of the same form as the original prior, simplifying sampling and analysis. Beta distributions are specified

by shape parameters α and β, where $\alpha > 0$ and $\beta > 0$. The mean of the Beta distribution is given by $\alpha/(\alpha+\beta)$. To update the prior Beta distribution one adds the count of observed successes of the event to α and the count of the observed failures to β Often the prior $\alpha = 1$ and $\beta = 1$ is used before any rewards are observed, which leads to a uniform distribution on θ.

Given a proposed grasp plan Γ , we draw samples from the shape distribution $p(\theta)$, the distribution on pose $p(\xi)$, distribution on motion $p(\rho)$ and the distribution on friction coefficient $p(\mu)$. The distribution on force closure can then be estimated as Beta- Bernoulli Process with shape parameters α and β. Thus, we can write the expected probability of force closure as follows

$$P_F(\Gamma) = \frac{\alpha}{\alpha+\beta} \tag{3}$$

Whats interesting in the context of a Bayesian BMAB problem our graphical model in Fig. 3, is now equivalent in terms of inference to 4 and we only need to estimate α and β to determine grasp quality. Regardless of the distributions in Fig. 3, we are still able to maintain the theoretical guarantees given by the Bayesian MAB algorithms because the probability of force closure is a Beta-Bernoulli Process [1] [19] [41], which we will now discuss for each of the three proposed algorithms in detail. [TODO: MICHAEL: THE ABOVE STATEMENT IS TRUE AND I WANT TO CONVEY THIS MESSAGE, HOWEVER IT MIGHT TAKE SOME WORK TO CONVEY]

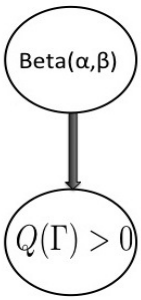


Fig. 4: A graphical model that illustrates the relationship between the Bernoulli distribution of the probability force closure and its conjugate prior Beta distribution that has two shape parameters α and β

A. Bayes-UCB

Bayes UCB, detailed in Algorithm 1, is a version of the UCB algorithm that uses a Bayesian posterior update to get a confidence bound instead of the frequentist concentration inequality approach. To get an upper confidence bound from the posterior, the Bayes UCB algorithm uses the quantile of the Beta distribution up to a specified probability which depends on the timestep and horizon. TAt each time step the arm with the highest quantile is chosen and the Beta distribution is updated based on the observed reward. For binary rewards (i.e. Bernoulli distributions) the expected number of pulls of suboptimal arms is bounded and experimentally this algorithm has been shown to outperform UCB for binary rewards [19].

Algorithm 1: Bayes-UCB for Beta-Bernoulli Process

Result: Current Best Arm, Γ^*
For Beta(1,1) prior, Stopping Horizon n:
for *t=1,2,...,n* **do**
  for *j = 1,...,K* **do**
    Compute: $q_j(t) = Q\left(1-\frac{1}{t}, p_j^{t-1}\right)$
  Draw arm$I_t = \text{argmax}_{j=1...K} q_j(t)$
  Observe reward $X_{I_t,t} \in \{0,1\}$
  Update posterior:
  Set $S_{I_t,t+1} = S_{I_t,t} + X_{I_t,t}$
  Set $F_{I_t,t+1} = F_{I_t,t} + 1 - X_{I_t,t}$
  Set $\theta_j^t \sim \text{Beta}(S_{I_t,t+1}, F_{I_t,t+1})$

B. Thompson Sampling

Thompson Sampling is a Bayesian method for the multi-armed bandit problem. We will describe it now in detail for the Beta-Bernoulli process. All arms are initialized with a prior Beta distributions, which is normally Beta($\alpha = 1$,$\beta = 1$) to reflect a uniform prior on the θ of the Bernoulli distribution. Then for each arm draw $\theta_{j,t} \sim \text{Beta}(\alpha, \beta)$ and pull the arm with the highest $\theta_{j,t}$ drawn. The reward, $X_{i,t}$ is observed from that arm, j, and the corresponding Beta distribution is updated. This is repeated until a stopping time is reached. The full algorithm is shown in Algorithm 1.

The randomness of Thompson sampling allows for it to quickly explore the more arms then Bayes-UCB, which makes conservative arm pulls based on confidence bounds. It is also less prone to local solutions because of its stochastic nature. Thus, for cases where exploration and exploitation are decoupled Thompson sampling can find a better arm faster. Thompson sampling has recently been shown to approach the Lai and Robbins bound [1] and has empirically been shown to outperform frequentist methods like UCB in certain settings [9]. Variants of it are even used commercially in products like Microsoft's adPredictor, which is used by Bing, the search engine, [15].

Algorithm 2: Thompson Sampling for Beta-Bernoulli Process

Result: Current Best Arm, Γ^*
For Beta(1,1) prior:
for *t=1,2,...* **do**
  Draw $\theta_{j,t} \sim \text{Beta}(S_{j,t}+1, F_{j,t}+1)$ for $j = 1, ..., k$
  Play $I_t = j$ for j with maximum $p_{j,t}$
  Observe reward $X_{I_t,t} \in \{0,1\}$
  Update posterior:
  Set $S_{I_t,t+1} = S_{I_t,t} + X_{I_t,t}$
  Set $F_{I_t,t+1} = F_{I_t,t} + 1 - X_{I_t,t}$

C. The Gittins Index Method

[TODO: LET ME KNOW HOW CLEAR THIS IS, GITTINS CAN BE HARD TO DESCRIBE IN GENERAL] One possible solution to solve the MAB problem is to treat it as an Markov Decision

Process (MDP) and use Markov Decision theory. This solution makes a lot of sense when the distribution is known because the all elements in the standard MDP tuple, $\{S, A, T, R, \gamma\}$, would be known and it is optimal with respect γ [41].

However, the curse of dimensionality effects performance because if you have K arms, a finite horizon of T and a Beta-Bernoulli distribution on your arms then your state space is on the order of T^{2*K}. Hence the complexity of solving MAB using Markov Decision theory increases exponentially with the number of bandit processes. A key insight though was given by Gittins, who showed that instead of solving the k-dimensional MDP one can instead solve k 1-dimensional optimization problems: for each arm i, $i = 1, ..., k$, and for each state of $x^i = \{\alpha_0 + S_t, \beta_0 + F_t\}^i$, where S_t and F_t correspond to the number of success and failures at pull t.

$$v^i(x^i) = \max_{\tau > 0} \frac{\mathcal{E}[\sum_{t=0}^{\tau} \gamma^t r^i(X_t^i) | X_0^i = x_i]}{\mathcal{E}[\sum_{t=0}^{\tau} \gamma^t | X_0^i = x_i]} \tag{4}$$

The indices can be considered as a computation of the value in choosing an arm conditioned on the fact that you will give up an choose another arm at some point. Once you know the state of your k arms, the algorithm is to select the one with the highest index. For Best Arm Identification you want your discount factor γ to approach 1, since you should never stop pulling the best arm. Generally the computation of the Gittins indices is too expensive, however in the Beta-Bernoulli case it is actually possible [19]. We computed the Gittins indices offline using the restart method proposed by Katehakis et al. [18].

Algorithm 3: The Gittins Index Method for Beta-Bernoulli Process

Result: Current Best Arm, Γ^*
For Beta(1,1) prior, Table of Indices v, Discount Factor γ:
for *t=1,2,...* **do**
  Pull arm $k = \underset{x_k \in X}{\text{argmax}} v(x_k)$
  Observe reward $R_{I_t,t} \in \{0, 1\}$
  Update posterior:
  Set $S_{I_t,t+1} = S_{I_t,t} + R_{I_t,t}$
  Set $F_{I_t,t+1} = F_{I_t,t} + 1 - R_{I_t,t}$
  Set $x_k = \{1 + SI_t, t + 1, 1 + F_{I_t,t+1}\}$

.

VI. Experiments

For the experiments we used the Brown Vision Lab 2D dataset [**?**], the same used in [10]. We downsampled the image by a factor of 2 to create a 40 x 40 occupancy map, which holds 1 if the point cloud was observed and 0 if it was not observed, and a measurement noise map, which holds the variance 0-mean noise added to the SDF values. The parameters of the GPIS were selected using maximum likelihood on a held-out set of validation shapes. The noise of the motion, position and friction coefficient was set to the following variances $\sigma_{mu} = 0.4$, $\sigma_{rot} = 0.3$ rads,$\sigma_{trans} = 3$. Our visualization technique follows the approach of [28] and consisted of drawing many shape samples from the distribution and blurring accordingly to a histogram equalization scheme.

We did experiments for the case of two hard contacts in 2-D. We drew random lines of actions $\gamma_1(t)$ and $\gamma_2(t)$ by sampling around a circle with radius $\sqrt{2}n$ and sampling the circles origin, then projecting onto the largest inscribing circle in the workspace.

A. Multi-Armed Bandit Experiments

[TODO: ADD BAYES AND KEHOE RESULTS] We consider the problem of selecting the best grasp plan, Γ^* out of a set G. For our experiments we look at selecting the best grasp out of a size of $|G| = 1000$. We initialize all algorithms by sampling each grasp 1 time. We draw samples from our graphical model using the technique described in Sec. III-C. In Fig. 5, we plotted the simple regret averaged over 100 randomly selected shapes in the Brown Vision Lab 2D dataset and compare the different methods (Bayes-UCB, Thompson, Gittins, Kehoe et al. [21] and Monte-Carlo Integration). In Fig. 6, we plot the current grasp quality the algorithm selects vs. samples drawn. We then show the grasps selected for five shapes at a stopping time of $T = 9000$ for all the methods listed in Fig. 13.

Interestingly in Fig 6, Gittins and Thompson suggest on average higher quality grasps with than Monte-Carlo integration and the method listed in Kehoe et al [21] with the same number of samples drawn . In Fig. 7 we plot samples per grasp quality, Gittins and Thompson appear to allocate significantly more grasp samples to grasps of high quality, thus they are quickly able to ignore the low quality grasps. Kehoe et al. and Monte-Carlo integration take a more evenly distributed approach to sample allocation, which could explain the performance difference in Fig. 5 and Fig. 6.

[TODO: MICHAEL: NEED TO DISCUSS HOW TO IMPLEMENT WORST CASE SCENARIO, ADVERSIAL APPLIES CHANGING DISTRIBUTIONS ONLINE (WHICH WOULD NOT MODEL THE PROBLEM).]

B. Sensitivity Analysis

We now will show how well the top two algorithms Thompson Sampling and the Gittens Index Method perform under a variation in noise from friction coefficient uncertainty, shape uncertainty, rotational pose and translation pose. The experiments are performed with the same setup as before but now we increase the variance parameters across a set range for each parameter to simulate low, medium and high levels of noise. All experiments were averaged across 10 shapes randomly selected with from the Brown dataset with a set size $|G| = 1000$.

For friction coefficient we varied the variance across the following values $\sigma_\mu = \{0.05, 0.2, 0.4\}$. As illustrated in Fig. 9, the performance of the bandit algorithm remains largely unchanged, with typical convergence to zero in simple regret less than 2000 evaluations.

For rotational uncertainty in pose, we varied σ_{rot} over the set of $\{0.03, 0.12, 0.24\}$ radians. As illustrated in Fig. 10, the performance of the bandit algorithms is effected by the change

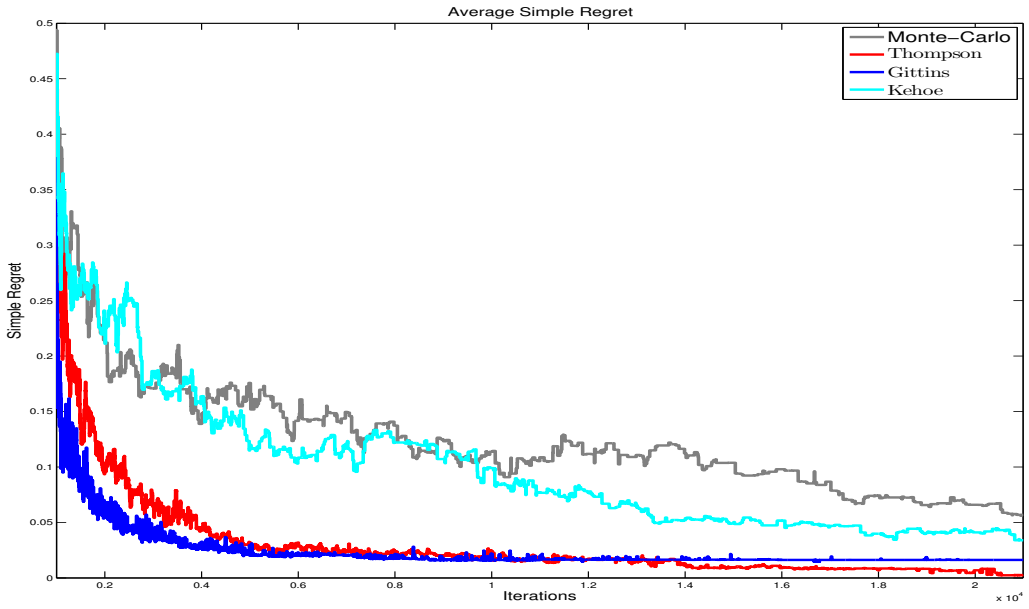


Fig. 5: Comparison of Simple Regret convergence for the four sequential decision methods (Monte-Carlo, Bayes -UCB, Thompson and Gittins). Graph is averaged over 100 shapes from the Brown Sillohoute Dataset [**?**] with a set $|G| = 1000$ for each shape. As you can see the BMAB methods converge almost a magnitude faster than random allocation. It is worth noting that Gittins outperform the other two algorithms, which is useful when choosing which one to implement

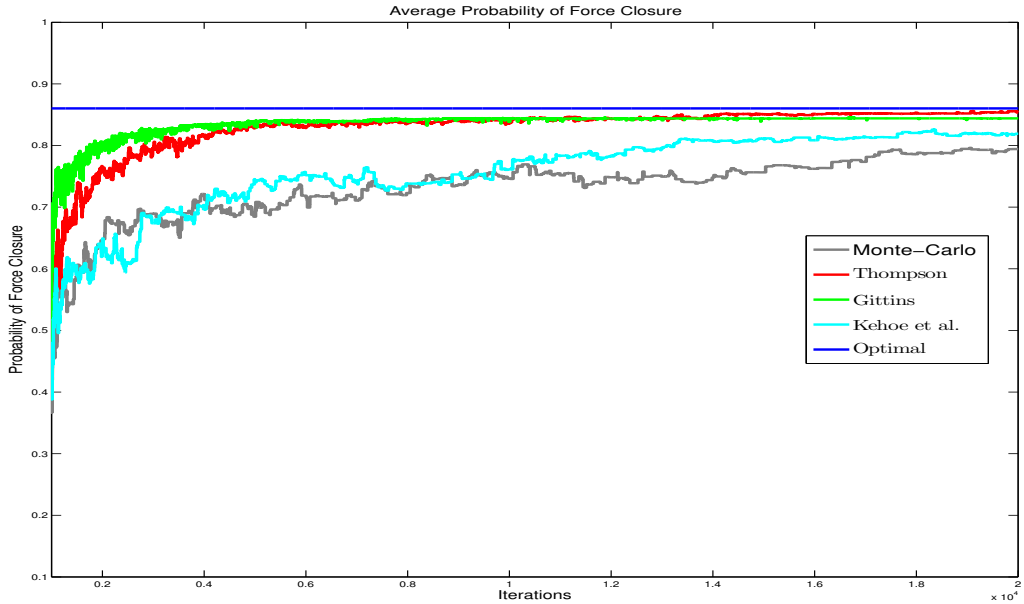


Fig. 6: Comparison of the current average probability of force closure vs. the number of samples pulled.Graph is averaged over 100 shapes randomly drawn from the Brown Vision 2D Lab Dataset [**?**] with a set $|G| = 1000$ for each shape. We demonstrate this for Thompson, Gittins, Monte-Carlo and the approach taking in Kehoe et al [21]. We also demonstrate what the average probability of force closure for the approximate optimal policy. Emperically, it appears that Thompson and Gittins converge at a faster rate to the optimal solution, which is desired for an anytime algorithm

in rotation, increase in variance to 0.24 radians or 13^{deg} causes the convergence in simple regret to not be reached until 5500 samples or an average of 5.5 samples per grasp. This can be explained because such a large variance causes a a drop in quality across all grasps and makes it harder to separate the outliers [**?**]. The quality of the best grasp along with the grasp for each round is shown in 11.

For translational uncertainty in pose, we varied σ_{trans} in the range of $\{3, 12, 24\}$ units (on a 40 x 40 unit workspace). As you can see in Fig. 10, the performance of the bandit algorithms is effected by the change in rotation, increase noise of $\sigma_{trans} = 24$ causes the convergence to not be reached until around 5000 samples for the Gittens Method and 8200

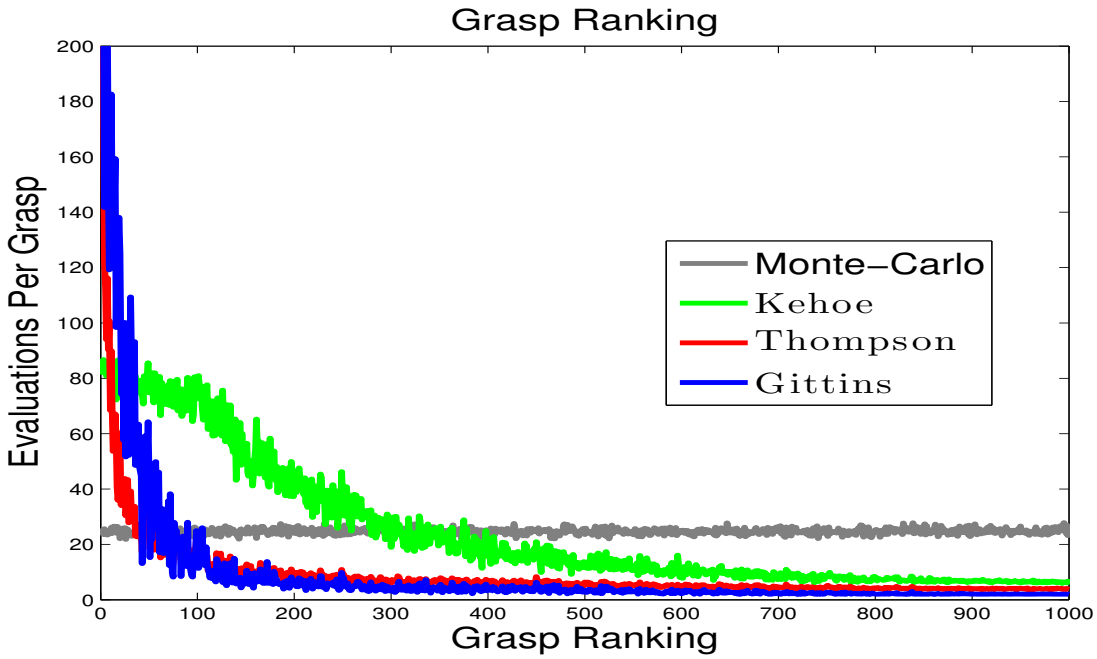


Fig. 7: Comparison of sample per grasp for the four sequential decision methods (Random, Bayes - UCB, Thompson, Gittins). Graph is averaged over 100 shapes from the Brown Silhouette Dataset [**?**] with a set $|G| = 1000$ for each shape. The best grasps are ranked 1 and worst are 1000. As you can see the MAB algorithm intelligently allocate samples towards high quality grasps based on past observations, where Monte-Carlo Integration takes a uniform approach to allocation.

evaluations for Thompson Sampling.

C. Worst Case Scenario

[TODO: RUNNING EXPERIMENTS TODAY ON THIS]

VII. Limitations

Our budgeted multi-armed bandit approach appears promising, but we still do not know how well it will perform on 3D shapes and large scale grids. Future work will be building an efficient construction of GPIS to scale to 3D and test the bandit method there.

Of the BMAB algorithms we showed, Thompson Sampling is guaranteed to find the best grasp as the stopping time approaches infinity [1] but when do you terminate the algorithm is still an open question. Fixed confidence methods do exist that terminate when a certain confidence interval is reached [29] [31]. However, a known problem is that if two grasps have very similar quality it could greatly increase the time for needed to reach the statistical confidence interval [3]. We proposed treating the BMAB as anytime algorithm and initial results in Fig. 6 suggest that on average grasps at a given stopping time are better than prior methods of Monte-Carlo sampling or the approach of Kehoe et al. However, as shown in [TODO: WORST CASE RESULTS HERE] there can exists pathological cases that can change this. Fortunately, though these cases occur with a small probability, however it is important for a practitioner to be aware of them.

Another problem that was revealed in our analysis was that our current grasp metric, probability of force closure, is not dependent on the center of mass [12]. It only measure the probability that a grasp controller can resist any force provided it can exert an infinite force. One can assume that the grasp controller on a robot hand is powerful enough to apply the proper resistance, but that assumption might be invalid in some

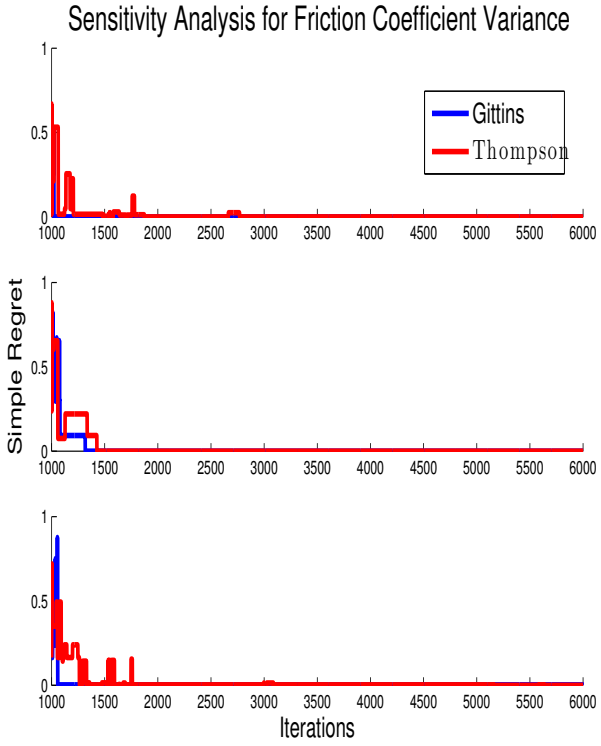


Fig. 9: Sensitivity Analysis for Thompson Sampling and the Gittens Index Method under friction coefficient uncertainty $\sigma_{fric} = \{0.05, 0.2, 0.4\}$ from top to bottom on a 40 x 40 unit workspace averaged over 10 shapes from the Brown Vision Lab Data set. The increase in noise has little effect on the convergence of the two algorithms in simple regret.

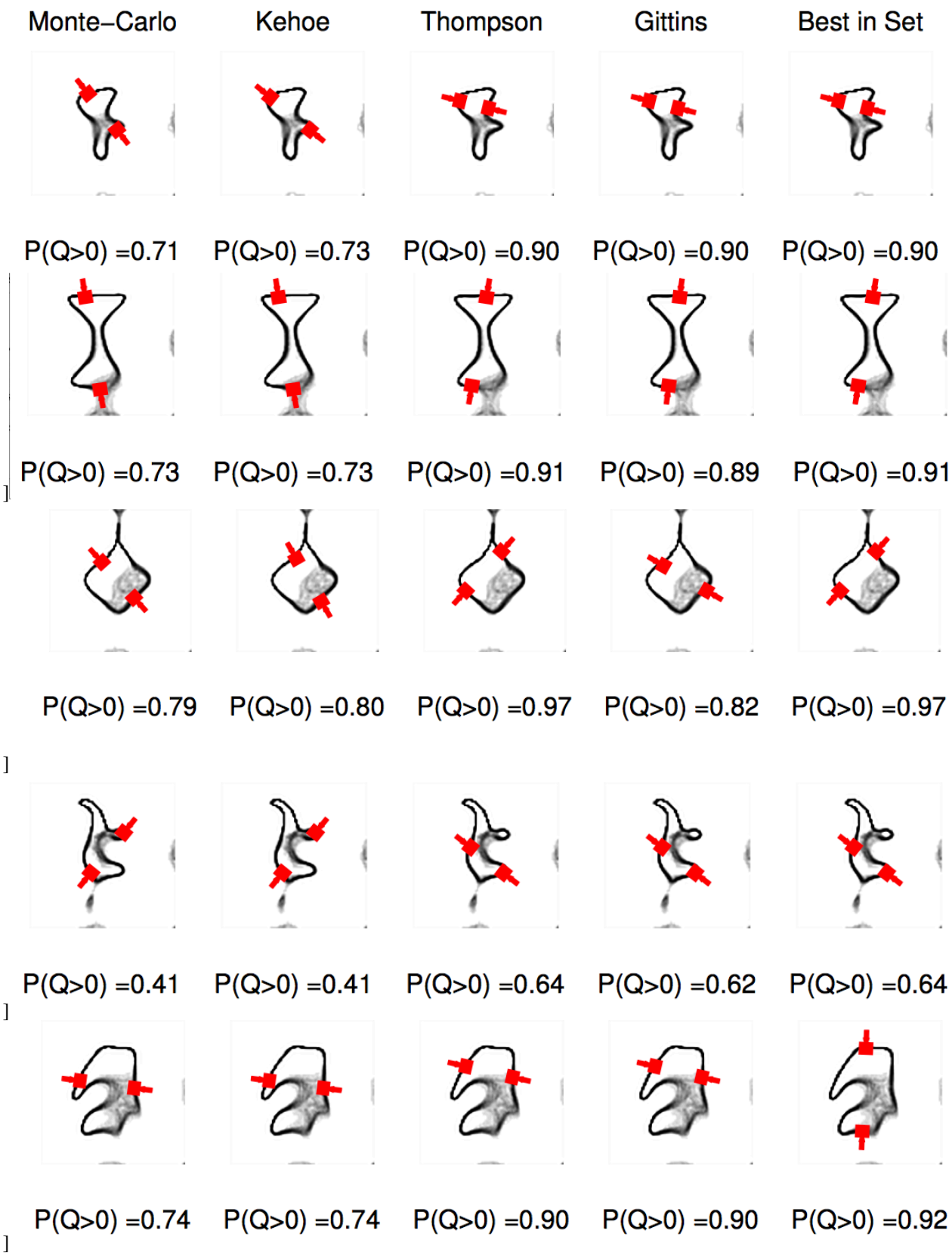
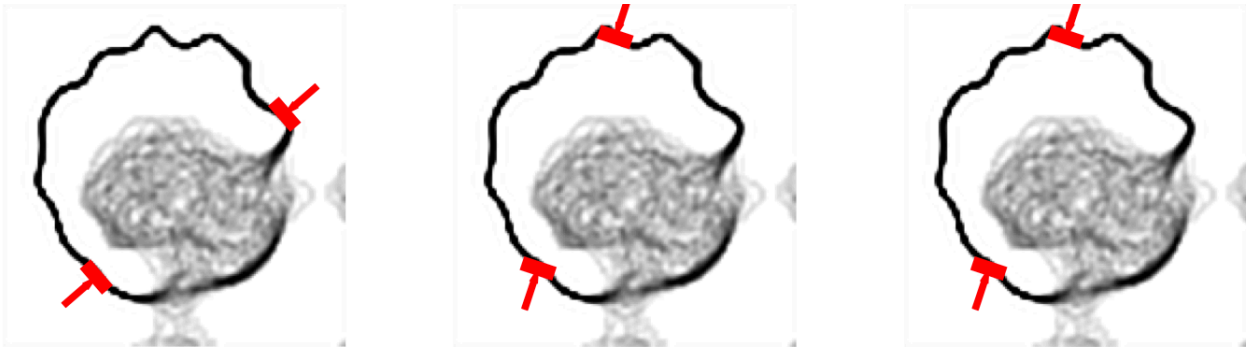


Fig. 8: Five shapes shown from the Brown Visual Lab Dataset with induced shape uncertainity and visualized according to the method described in [28]. The four methods (Monte-Carlo, Kehoe's, Thompson and Gittins) were all run until a stopping time of 9000 evaluations with a uniformly initialized grasp set of $|G| = 1000$. We also showed the estimated best grasp in the set. The grasps and the quality each one found is shown above. In the four out of five cases, Thompson sampling is able to find the best grasp in the set at the stopping interval of 9000, however at the last shape there is a difference of 2% grasp quality.

(a) lTop Grasps Shown for Shape 1



(b) Top Grasps Shown for Shape 2

Fig. 11: Two shapes shown from the Brown Visual Lab Dataset,from Left to Right the variance on rotation σ_{rot} is increased 0.03, 0.12 to 0.24 radians. As you can see the overall variance increase effects the quality of the top grasp in the set of possible grasps. Furthermore for Shape 2, the grasp with low rotational variance is different than that for higher variance because the original grasp is more likely to touch the area of higher shape uncertainty when subjected to high variance in rotation. [TODO: FIND BETTER SHAPES IN THE BROWN DATASET]

applications. A similar metric that is still from Beta-Bernoulli and takes center of mass into account, would be ideal for both accurate grasp quality prediction under uncertainty and the utilization of MAB algorithms. Recent work by Kim et al. developed a physics based simulator that could potentially achieve this goal [22].

VIII. Conclusion

Assessing grasp quality under uncertainty is computationally expensive as it often requires repeated evaluations of the grasp metric over many random samples. In this work, we proposed a multi-armed bandit approach to efficiently identify high-quality grasps under uncertainty in shape, pose, friction coeffiecient and motion. A key insight from our work is that uniformly allocating samples to grasps is inefficient, and we found that a MAB approach prioritizes evaluation of high-quality grasps while quickly pruning-out obviously poor grasps. A pre-requisite for applying a bandit approach is to formulate a representation of how uncertainty affects grasp parameters and thus grasp quality. We purpose treating this as a graphical model and use model the parameters as stochastic noise. Our choice of distributions though is not the focus of the paper. The MAB algorithm is applicable in any context of bounded reward distributions and all the theoretical results we mentioned will transfer to the Beta-Bernoulli case, or the estimation of probability of force closure.

Our anytime BMAB approach is guaranteed to find the best grasp in a given proposal set in the limit of an infinite time and has empirically been shown to outperform the methods of prior work Monte-Carlo and the method purposed by Kehoe et al. [20].

IX. Future Work

Our results are promising and they suggest many avenues of future work. By utilizing the BMAB model, we can encode uncertainty in the grasp parameters and then leverage the existing algorithms to efficiently find the best grasp.

In principle, our method can be applied to other representations of shape uncertainty such as perturbations on polygonal vertices [20] or splines [10]. It can further be applied to other grasp quality metrics or simulation based evaluation methods [25].

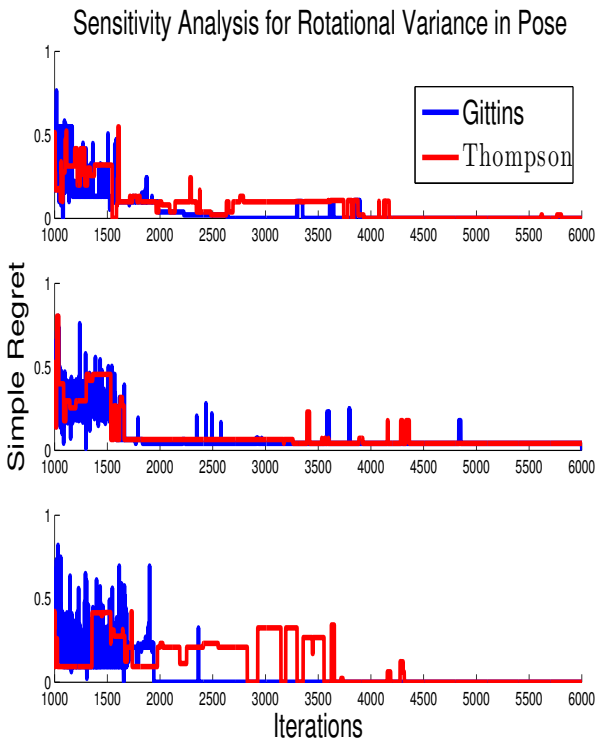


Fig. 10: Sensitivity Analysis for Thompson Sampling and the Gittens Index Method under coefficient of friction uncertainty $\sigma_{rot} = \{0.03, 0.12, 0.24\}$ radians from top to bottom on a 40 x 40 unit workspace averaged over 10 shapes from the Brown Vision Lab Data set. The increase in noise has little effect on the convergence of the two algorithms in simple regret. [TODO: GOING TO RERUN WITH MORE SHAPES TO BE SURE ABOUT THIS]

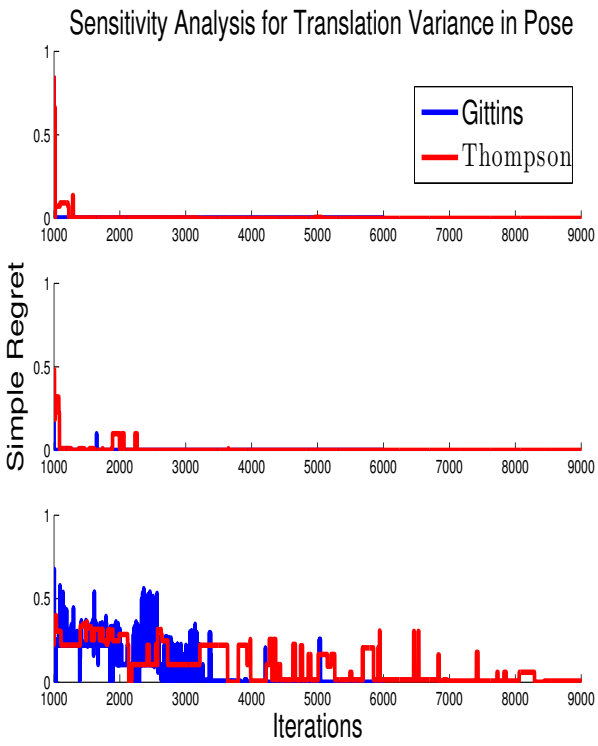


Fig. 12: Sensitivity Analysis for Thompson Sampling under translation uncertainty $\sigma_{trans} = \{3, 12, 24\}$ units from top to bottom on a 40 x 40 unit workspace. As you can see the increase in noise effects performance, however the 5000 samples needed for Gittins to converge at the the highest level of noise (which is a variance of over half the workspace) is much less that the samples needed for uniform allocation to converge in Fig. 5

Future work will also consider applying BMAB approach to grasp planners like GraspIt! [30] to see if our method can handle uncertainty while working under the time constraints needed for most real time applications. While our results are promising, it remains to be seen how well it deals with the increased complexity of 3D models over 2D models and larger scale experiments. However, the BMAB model has a large amount of literature to draw from as we encounter new and more challenging problems [6].

References

[1] S. Agrawal and N. Goyal, "Analysis of thompson sampling for the multi-armed bandit problem," *arXiv preprint arXiv:1111.1797*, 2011.

[2] C. Andrieu, N. De Freitas, A. Doucet, and M. I. Jordan, "An introduction to mcmc for machine learning," *Machine learning*, vol. 50, no. 1-2, pp. 5–43, 2003.

[3] J.-Y. Audibert, S. Bubeck *et al.*, "Best arm identification in multi-armed bandits," *COLT 2010-Proceedings*, 2010.

[4] T. Barfoot and P. Furgale, "Associating uncertainty with three-dimensional poses for use in estimation problems," *Robotics, IEEE Transactions on*, vol. 30, no. 3, pp. 679–693, June 2014.

[5] A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 1998.

[6] D. Bergemann and J. Välimäki, "Bandit problems," Cowles Foundation for Research in Economics, Yale University, Tech. Rep., 2006.

[7] S. Bubeck, R. Munos, and G. Stoltz, "Pure exploration in multi-armed bandits problems," in *Algorithmic Learning Theory*. Springer, 2009, pp. 23–37.

[8] R. E. Caflisch, "Monte carlo and quasi-monte carlo methods," *Acta numerica*, vol. 7, pp. 1–49, 1998.

[9] O. Chapelle and L. Li, "An empirical evaluation of thompson sampling," in *Advances in Neural Information Processing Systems*, 2011, pp. 2249–2257.

[10] V. N. Christopoulos and P. Schrater, "Handling shape and contact location uncertainty in grasping two-dimensional planar objects," in *Intelligent Robots and Systems, 2007. IROS 2007. IEEE/RSJ International Conference on*. IEEE, 2007, pp. 1557–1563.

[11] S. Dragiev, M. Toussaint, and M. Gienger, "Gaussian process implicit surfaces for shape estimation and grasping," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 2011, pp. 2845–2850.

[12] C. Ferrari and J. Canny, "Planning optimal grasps," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*, 1992, pp. 2290–2295.

[13] V. Gabillon, M. Ghavamzadeh, and A. Lazaric, "Best arm identification: A unified approach to fixed budget and fixed confidence," in *Advances in Neural Information Processing Systems*, 2012, pp. 3212–3220.

[14] K. Y. Goldberg and M. T. Mason, "Bayesian grasping," in *Robotics and Automation, 1990. Proceedings., 1990 IEEE International Conference on*. IEEE, 1990, pp. 1264–1269.

[15] T. Graepel, J. Q. Candela, T. Borchert, and R. Herbrich, "Web-scale bayesian click-through rate prediction for sponsored search advertising in microsoft's bing search engine," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 13–20.

[16] G. A. Hollinger, B. Englot, F. S. Hover, U. Mitra, and G. S. Sukhatme, "Active planning for underwater inspection and the benefit of adaptivity," *Int. J. Robotics Research (IJRR)*, vol. 32, no. 1, pp. 3–18, 2013.

[17] K. Hsiao, M. Ciocarlie, and P. Brook, "Bayesian grasp planning," in *ICRA 2011 Workshop on Mobile Manipulation: Integrating Perception and Manipulation*, 2011.

[18] M. N. Katehakis and A. F. Veinott Jr, "The multi-armed bandit problem: decomposition and computation," *Mathematics of Operations Research*, vol. 12, no. 2, pp. 262–268, 1987.

[19] E. Kaufmann, O. Cappé, and A. Garivier, "On bayesian upper confidence bounds for bandit problems," in *International Conference on Artificial Intelligence and Statistics*, 2012, pp. 592–600.

[20] B. Kehoe, D. Berenson, and K. Goldberg, "Estimating part tolerance bounds based on adaptive cloud-based grasp planning with slip," in *Automation Science and Engineering (CASE), 2012 IEEE International Conference on*. IEEE, 2012, pp. 1106–1113.

[21] ——, "Toward cloud-based grasping with uncertainty in shape: Estimating lower bounds on achieving force closure with zero-slip push

grasps," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, 2012, pp. 576–583.
[22] J. Kim, K. Iwamoto, J. J. Kuffner, Y. Ota, and N. S. Pollard, "Physically-based grasp quality evaluation under uncertainty," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, 2012, pp. 3258–3263.
[23] J. Laaksonen, E. Nikandrova, and V. Kyrki, "Probabilistic sensor-based grasping," in *Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ International Conference on*. IEEE, 2012, pp. 2019–2026.
[24] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.
[25] B. León, S. Ulbrich, R. Diankov, G. Puche, M. Przybylski, A. Morales, T. Asfour, S. Moisio, J. Bohg, J. Kuffner, and R. Dillmann, *OpenGRASP: A Toolkit for Robot Grasping Simulation*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2010, vol. 6472, pp. 109–120.
[26] Z. Li and S. S. Sastry, "Task-oriented optimal grasping by multifingered robot hands," *Robotics and Automation, IEEE Journal of*, vol. 4, no. 1, pp. 32–44, 1988.
[27] O. Madani, D. J. Lizotte, and R. Greiner, "The budgeted multi-armed bandit problem," in *Learning Theory*. Springer, 2004, pp. 643–645.
[28] J. Mahler, S. Patil, B. Kehoe, J. van den Berg, M. Ciocarlie, P. Abbeel, and K. Goldberg, "Gp-gpis-opt: Grasp planning under shape uncertainty using gaussian process implicit surfaces and sequential convex programming."
[29] O. Maron and A. W. Moore, "Hoeffding races: Accelerating model selection search for classification and function approximation," *Robotics Institute*, p. 263, 1993.
[30] A. T. Miller and P. K. Allen, "Graspit! a versatile simulator for robotic grasping," *Robotics & Automation Magazine, IEEE*, vol. 11, no. 4, pp. 110–122, 2004.
[31] V. Mnih, C. Szepesvári, and J.-Y. Audibert, "Empirical bernstein stopping," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 672–679.
[32] B. Mooring and T. Pack, "Determination and specification of robot repeatability," in *Robotics and Automation. Proceedings. 1986 IEEE International Conference on*, vol. 3. IEEE, 1986, pp. 1017–1023.
[33] C. Rasmussen and C. Williams, *Gaussian processes for machine learning*. MIT Press, 2006.
[34] C. E. Rasmussen and H. Nickisch, "Gaussian processes for machine learning (gpml) toolbox," *The Journal of Machine Learning Research*, vol. 9999, pp. 3011–3015, 2010.
[35] H. Robbins, "Some aspects of the sequential design of experiments," in *Herbert Robbins Selected Papers*. Springer, 1985, pp. 169–177.
[36] M. Rothschild, "A two-armed bandit theory of market pricing," *Journal of Economic Theory*, vol. 9, no. 2, pp. 185–202, 1974.
[37] R. Simon, "Optimal two-stage designs for phase ii clinical trials," *Controlled clinical trials*, vol. 10, no. 1, pp. 1–10, 1989.
[38] E. Solak, R. Murray-Smith, W. E. Leithead, D. J. Leith, and C. E. Rasmussen, "Derivative observations in gaussian process models of dynamic systems," 2003.
[39] D. L. St-Pierre, Q. Louveaux, and O. Teytaud, "Online sparse bandit for card games," in *Advances in Computer Games*. Springer, 2012, pp. 295–305.
[40] F. Stulp, E. Theodorou, J. Buchli, and S. Schaal, "Learning to grasp under uncertainty," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*. IEEE, 2011, pp. 5703–5708.
[41] R. Weber *et al.*, "On the gittins index for multiarmed bandits," *The Annals of Applied Probability*, vol. 2, no. 4, pp. 1024–1033, 1992.
[42] J. Weisz and P. K. Allen, "Pose error robust grasping from contact wrench space metrics," in *Robotics and Automation (ICRA), 2012 IEEE International Conference on*. IEEE, 2012, pp. 557–562.
[43] O. Williams and A. Fitzgibbon, "Gaussian process implicit surfaces," *Gaussian Proc. in Practice*, 2007.
[44] Y. Zheng and W.-H. Qian, "Coping with the grasping uncertainties in force-closure analysis," *Int. J. Robotics Research (IJRR)*, vol. 24, no. 4, pp. 311–327, 2005.

APPENDIX
GAUSSIAN PROCESS IMPLICIT SURFACE FOR REPRESENTING SHAPE UNCERTAINTY

In order to solve our problem definition, we must estimate $P(Q(\Gamma) > 0)$ for a given grasp . We will first discuss how the GPIS is constructed, then which grasp metric Q we chose and lastly proceed into evaluating $P(Q(\Gamma) > 0)$ efficiently.

A. Gaussian Process (GP) Background

We refer the reader to [**?**] for a more detailed explanation of the GP construction, which we summarize here. Given the training data $\mathcal{D} = \{\mathcal{X}, \mathbf{y}\}$ and covariance function $k(\cdot, \cdot)$, the posterior density $p(sd_*|\mathbf{x}_*, \mathcal{D})$, or the distribution on signed distance field, at a test point $\mathbf{x}_*$ is shown to be [34]:

$$\begin{aligned} p(sd_*|\mathbf{x}_*, \mathcal{D}) &\sim \mathcal{N}\big(\mu(\mathbf{x}_*), \Sigma(\mathbf{x}_*)\big) \\ \mu(\mathbf{x}_*) &= k(\mathcal{X}, \mathbf{x}_*)^\intercal (K + \sigma^2 I)^{-1} \mathbf{y} \\ \Sigma(\mathbf{x}_*) &= k(\mathbf{x}_*, \mathbf{x}_*) - k(\mathcal{X}, \mathbf{x}_*)^\intercal (K + \sigma^2 I)^{-1} k(\mathcal{X}, \mathbf{x}_*)\big) \end{aligned}$$

where $K \in \mathbb{R}^{l \times l}$ is a matrix with entries $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ and $k(\mathcal{X}, \mathbf{x}_*) = [k(\mathbf{x}_1, \mathbf{x}_*), \ldots, k(\mathbf{x}_l, \mathbf{x}_*)]^\intercal$. This derivation can also be used to predict the mean and variance of the function gradient by extending the kernel matrices using the identities [38]:

$$\operatorname{cov}\left(sd(\mathbf{x}_i), sd(\mathbf{x}_j)\right) = k(\mathbf{x}_i, \mathbf{x}_j) \tag{5}$$

$$\operatorname{cov}\left(\frac{\partial sd(\mathbf{x}_i)}{\partial x_k}, sd(\mathbf{x}_j)\right) = \frac{\partial}{\partial x_k} k(\mathbf{x}_i, \mathbf{x}_j) \tag{6}$$

$$\operatorname{cov}\left(\frac{\partial sd(\mathbf{x}_i)}{\partial x_k}, \frac{\partial sd(\mathbf{x}_j)}{\partial x_l}\right) = \frac{\partial^2}{\partial x_k \partial x_l} k(\mathbf{x}_i, \mathbf{x}_j) \tag{7}$$

For our kernel choice we decided to use the square exponential kernel, similar to [11]. Other kernels relevant to GPIS are the thin-plate splines kernel and the Matern kernel [43].

We construct a GPIS by learning a Gaussian process to fit measurements of a signed distance field of an unknown object. Precisely, $x_i \in \mathbb{R}^2$ in 2D and $x_i \in \mathbb{R}^3$ in 3D, and $y_i \in \mathbb{R}$ is a noisy signed distance measurement to the unknown object at x_i.

B. Sampling Shape from GPIS Distribution

To compute the above distribution we must draw samples from $p(\theta)$. In order to draw shape samples from a GPIS, one needs to sample from signed distance function, sd, over the joint on all points in the workspace $\mathcal{W}$ or $p(sd(\mathcal{W}))$. Since this is a GPIS, we know the following

$$p(S) = p(sd(\mathcal{W})) \sim N(\mu(\mathcal{W}), \Sigma(\mathcal{W})) \tag{8}$$

Thus if the workspace is an $n \times n$ grid, the joint distribution is an n^2 multi-variate Gaussian, due to $sd : \mathbb{R}^2 \to \mathbb{R}$. Shape samples drawn from the distribution appear in Fig. 13.

C. Distribution on Surface Normals

Using Eq. 6 and Eq. 7, we can compute the mean of the gradient $\mu_\nabla(x)$ and the covariance of the gradient $\Sigma_\nabla(x)$ respectively. Thus we can compute the distribution around the surface normal for a given point in $\mathcal{W}$. We can now write

One interesting effect of this technique is that we can now marginalize out the line of action model and visual what the surface normal distribution is along a given line of action. To our knowledge this is the first attempt to visual surface normals along a grasp plan. Marginalization can be performed as follows:

$$p(\mathbf{n}_i) = \int_a^b p\big(\mathbf{n}_i = \mathbf{v} | \mathbf{c}_i = \gamma(t)\big) p\big(\mathbf{c}_i = \gamma(t)\big) dt \tag{9}$$

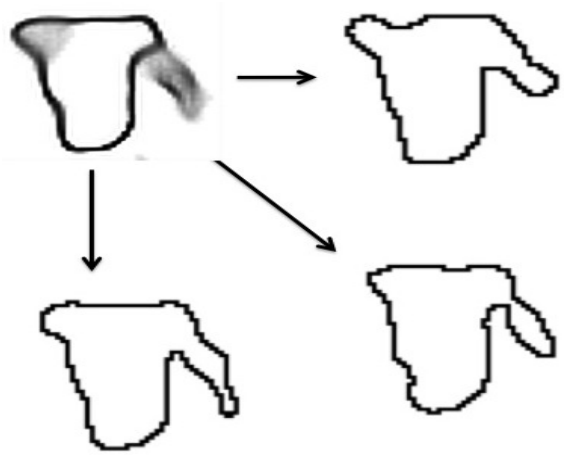


Fig. 13: Shape samples drawn from Eq. 8 on the object in the upper left corner. Given a shape sample we highlight the zero-crossing of the level set in black

Grasp metrics such as Ferrari-Canny require $\mathbf{n}_i$ be normalized, or, equivalently, a member of the sphere $\mathcal{S}^{d-1}$ [12]. To account for this we densely sample from the distribution $p(\mathbf{n}_i))$ and project onto $\mathcal{S}^{d-1}$. In Fig.**??**, we visualize the distribution on $\mathbf{n}_i$ calculated for a given GPIS and approach line of action.

D. Expected Center of Mass

We recall the quantity $P(sd(x) < 0) = \int_{-\infty}^{0} p(sd(x) = s \mid \mu(x), \Sigma(x))ds$ is equal to the probability that x is interior to the surface under the current observations. We assume that the object has uniform mass density and then $P(sd(x) < 0)$ is the expected mass density at x. Then we can find the expected center of mass as:

$$\bar{z} = \frac{\int_{\mathcal{W}} xP(sd(x) < 0)dx}{\int_{\mathcal{W}} P(sd(x) < 0)dx} \tag{10}$$

which can be approximated by sampling $\mathcal{W}$ in a grid and approximating the spatial integral by a sum. Since this operation involves the entire SDF, one would want to use a low resolution grid for computational efficiency. We show the computed density and calculated expected center of mass for a marker in Fig. 14.

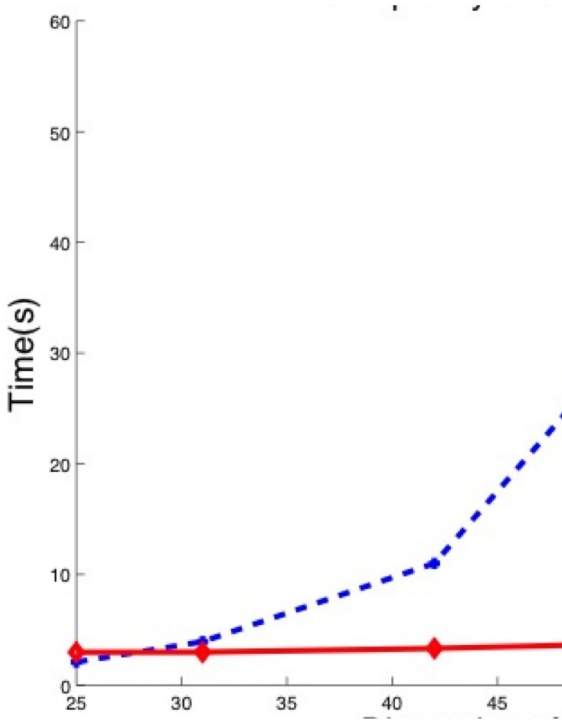


Fig. 14: Left: A surface with GPIS construction and expected center of mass (black X) Right: The distribution on the density of each point assuming uniform density