Multi-Arm Bandit Models for 2D Grasp Planning with Uncertainty [v5 Feb 18, 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— **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 a popular approach to grasp planning under uncertainty, but it may require a large number of samples to converge, 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 MAB with finite stopping time to efficiently determine high quality grasp with respect to a probability of force closure metric. To evaluate MAB-based sampling, we compare with two methods previously proposed for grasping an planar object under shape uncertainty represented as a Gaussian process implicit surface (GPIS) and Gaussian uncertainty in pose, grasp approach direction, and coefficient of friction. Our approach of using MAB is applicable to any distributions that can be sampled from, however performance could potentially vary.**

I. Introduction

Consider a robot fulfilling orders in a warehouse, where it encounters new consumer products and must handle them quickly. The robot may need to rapidly plan grasps, but may not be able to measure these quantities exactly due to sensor imprecision and missing data, which could result from occlusions, transparency, or highly reflective surfaces.

Analytic grasp quality metrics have been developed to plan grasps 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 [26]. 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 force closure given uncertainty in pose [10], [22], [42] and object shape [20], [28]. One option is using Monte-Carlo Integration over the possible values of uncertain quantities to evaluate the probability of force closure for a grasp [10], [22], [42], [20], [**?**], but this can be computationally expensive. In this work, we show that it is possible to rule out grasps with low probability of force closure in only a few samples and to allocate more sampling effort to grasps that are likely to be high quality.

[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

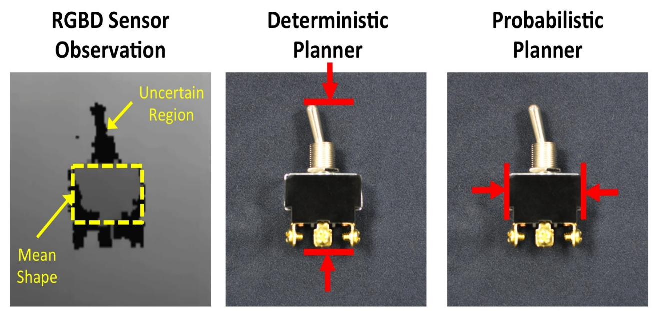

Fig. 1: Illustration of the effects of pose and shape uncertainty on parallel-jaw grasps planned for a mechanical switch. Left: An RGBD camera observing the switch cannot detect the specular metal toggle and screw regions (black) and the average shape segmentation is rectangular. Middle: Optimizing the Ferrari-Canny quality on the estimated mean shape and pose results in an unstable grasp (red arrows) on the true object. Right: Selecting the grasp with the highest probability of force closure given shape and pose distributions chooses a stable grasp for the true object. [TODO: KEEP THIS NEW IMAGE OR REVERT TO OLD? FLORIAN FAVORED STRONGER MOTIVATION FOR UNCERTAINTY]

The multi-armed bandit (MAB) model for sequential decision making [5], [24], [35] provides a formal way to reason about allocating sampling effort. The MAB model is 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]. In a standard MAB there are a set of possible options, or 'arms' [5], that each return a numeric reward from a stationary distribution. The goal in a MAB problem is to select a sequence of options to maximize expected reward. [TODO: EXPLAIN WHY] The MAB algorithms can be interpreted as an anytime algorithms, where at a provided stopping time the algorithm terminates and returns the estimated best grasp with respect to a chosen metric.

We formulate the problem of ranking a set of candidate grasps according to a quality metric in the presence of uncertainty as a MAB problem. We study this formulation using probability of force closure [10], [42], [21] as a quality metric under uncertainty in pose, shape, grasp approach direction, 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]. We model uncertainty in pose as a normal distributions around the orientation and translation of the object. Uncertainty 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.

We compare the performance of Thompson sampling and Gittins indices, two popular algorithms for solving the MAB problem, for selecting grasps with high probabilty of force closure on the Brown Vision 2D Dataset [**?**], [10], a dataset of over 1000 planar objects. [TODO: GET WITHIN 5 OF BEST VALUE NUMBERS] Initial results over 100 shapes suggest that on average the grasp return after a given stopping time on the number of samples is of higher quality than that of previous methods.

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]. One method to represent pose uncertainty is to place a normal distribution around the location of the center of the object as well as the rotational component of the object [4]. This approach is easy to sample from in the planar case, since each sample is from a fixed 3D Multivariate Gaussian.

For shape uncertainty we used a Gaussian process implicit surface representation. 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].

One common method for evaluating a probabilistic grasp quality measure is to use Monte-Carlo sampling [10], [20], [21], which involves sampling from distributions on random quantities and averaging the quality over these samples to empirically estimate a probability distribution[8]. It can be computationally expensive though to sample all proposed grasps to convergence. To address the computational cost, Kehoe et al. [20] proposed an adaptive sampling 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 variance around the current estimate, which in practice could lead to good grasps being thrown away. We propose treating the problem as an Multi-Armed Bandit, which instead of pruning grasps away decide based on past observations what grasp to sample next [5].

A. MAB Model

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. Solutions to the multi-armed bandit model have been used in applications for which evaluating all possible options is expensive or impossible, such as the optimal design of clinical trials [37], market pricing [36], and choosing strategies for games [39].

In a traditional MAB problem, a gambler has K independent one-armed bandits, also known as slot machines. 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 are 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].

At each time step the MAB algorithm incurs *regret*, the difference between the expected reward of the best arm and that of the arm selected. Bandit algorithms minimize cumulative regret, the sum of regret over the entire sequence of arm choices. Lai and Robbins showed that the cumulative regret of the 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. Since then a several other algorithms have been shown to achieve near this bound in terms of convergence, such as the Gittins index policy [41] and Thompson sampling for certain reward distributions [1].

B. Algorithms for MAB

We consider Bayesian MAB algorithms that use previous samples to form a belief distribution on the likelihood of the parameters specifying the distribution of each arm [41], [1], as these methods have been shown empirically to outperform frequentist algorithms (e.g., UCB) [9], [**?**]. Theoretical results have shown that several algorithms for Beta-Bernoulli reward distributions are capable of achieving near the lower bound on the asymptotic rate of convergence in regret described by Lai and Robbins [**?**], [1], [19].

Bayesian algorithms maintain a belief distribution on the grasp quality distributions for each of the candidate grasps to rank. For instance a Bernoulli random variable, θ, can be used to represent a binary grasping metric like force closure. The prior traditional placed on a Bernoulli variable is its conjugate prior the Beta distribution. Beta distributions are specified by shape parameters α and β, where ($\alpha > 0$ and $\beta > 0$).

One benefit of the Beta prior on Bernoulli reward distributions is that updates to the belief distribution after observing rewards from arm pulls can be derived in closed form. Let n_i denote the number of times arm k has been sampled. Then after observing S_i rewards of 1 for arm k, the posterior of the Beta are $\alpha_{i,n_i} = \alpha_{i,0} + S_i, \beta_{i,n_i} = \beta_{i,0} + n_i - S_i$, where $\alpha_{i,0}$ and $\beta_{i,0}$ are the prior shape parameters for arm k before any samples are evaluated. Given the current belief $\alpha_{k,n_i}, \beta_{k,n_i}$ on θ_k for an arm k, the algorithm can predict the probability of an event occurring on the next iteration by taking the expected value:

$$\theta_k = \frac{\alpha}{\alpha + \beta} \tag{1}$$

We describe two popular Bayesian MAB algorithms below for the Beta belief distribution.

1) The Gittins Index Method

One MAB method is to treat the problem as an Markov Decision Process (MDP) and use Markov Decision theory. Formally, a MDP is defined as a set of possible of states, a set of actions, a set of transition probabilities between states, a reward function, and a discount factor [5]. In the Beta-Bernoulli MAB case, the set of actions is the K options and the states are the Beta prior on each option.

Methods such as Value Iteration can compute optimal policies for an MDP with respect to the discount factor γ when all states, actions, and expected rewards can be enumerated [41], [5]. 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 options then your state space is exponential in K. 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 option i, $i = 1, ..., k$, and for each state $x_t^i = \{\alpha_0 + S_t, \beta_0 + F_t\}^i$, where S_t and F_t correspond to the number of success and failures at pull t and the state x_t^i is the Beta prior for option i, as shown in Algorithm 1.

$$v^i(x^i) = \max_{\tau > 0} \frac{\mathcal{E}[\sum_{t=0}^{\tau} \gamma^t r^i(X_t^i) | X_0^i = x_i]}{\mathcal{E}[\sum_{t=0}^{\tau} \gamma^t | X_0^i = x_i]} \tag{2}$$

The indices $v^i(x^i)$, computed in Equation 2,can then be used to form a policy, where at each timestep the agent selects the option with the highest $v^i(x^i)$. Traditionally, the indices are computed offline using a variety of methods [41], we chose to use the restart method proposed by Katehakis et al. [18] due to its ability to be implemented in a dynamic programming fashion.

Algorithm 1: 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 $x_k = \{1 + SI_t, t + 1, 1 + F_{I_t,t+1}\}$

2) Thompson Sampling

Computation of the Gittins indices can increase exponentially in as the discount factor approaches 1, and therefore it is not ideal to use in most cases. Thompson sampling is a less computationally expensive alternative. 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 $\theta_{j,t} \sim$ Beta(α, β) 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 2.

The intuition for Thompson sampling is that the random samples of $\theta_{j,t}$ allow the method to explore. However as it receives more samples it hones in on promising arms, since the Beta distributions approach delta distributions as number of samples drawn goes towards infinity. Chapelle et al. demonstrated empirically that Thompson sampling achieved lower cumulative regret than traditional bandit algorithms like UCB for the Beta-Bernoulli case [9]. Theoretically, Agrawal et al. recently proved an upper bound on the asymptotic complexity of cumulative regret for Thompson sampling that was sub-linear for k -arms and in the case of 2 arms logarithmic [1].

Algorithm 2: Thompson Sampling for Beta-Bernoulli Process

Result: Current Best Arm, Γ^*
For Beta(1,1) prior:
for *t=1,2,...* **do**
    Draw $\theta_{j,t} \sim$ 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 $F_{I_t,t+1} = F_{I_t,t} + 1 - X_{I_t,t}$

III. Grasp Planning Problem Definition

We consider grasping a rigid, planar object from above using parallel-jaw grippers. We assume that the interaction between the gripper and object is quasi-static [20], [21]. In this work we consider uncertainty in shape, pose, robot motion, and friction coefficient and we assume that distributions on these quantities are given.

A. Candidate Grasp Model

[TODO: MAKE ANITPODAL] The grasp model is illustrated in Fig. 2. We formulate the MAB problem for 2-dimensional objects using parallel-jaw grippers. Similar to [28], we parameterize a grasp using a *grasp axis* for a parallel jaw gripper with jaws of width $w_j \in \mathcal{R}$ and a maximum width $w_g \in \mathcal{R}$. The two location of the jaws can be specified as $\mathbf{g_1}, \mathbf{g_2} \in \mathcal{R}^2$, where $||\mathbf{g}_1 - \mathbf{g}_2|| \leq w_g$. We define a grasp plan consisting of the tuple $\Gamma = \{g_1, g_2\}$.

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 along the

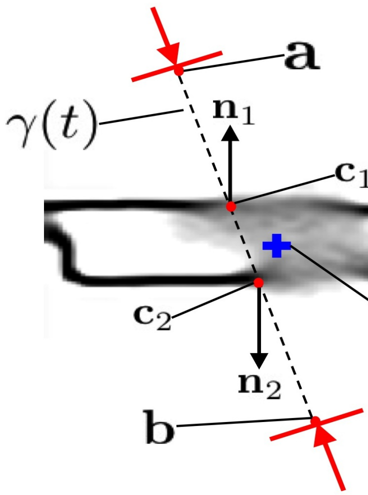

Fig. 2: llustration of our grasping model for parallel jaw grippers on a GPIS model of a marker with shape uncertainty near the object center. Jaw placements are illustrated by a red direction arrow and line. The grasp plan consists of a line of action $\gamma(t)$ with endpoints $\mathbf{a}$ and $\mathbf{b}$. 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.

grasp axis, $\mathbf{c}_1, \mathbf{c}_2 \in \mathbb{R}^2$. We also refer to the unit outward pointing surface normals at the contact points as $\mathbf{n}_1 \mathbf{n}_2 \in \mathbb{R}^d$, the object center of mass as $\mathbf{z} \in \mathbb{R}^d$ and the friction coefficient as μ. The center of mass may be computed based on a known mass density of the object. Together these form the set of grasp parameters $g = (\mathbf{c}_1, \mathbf{c}_2, \mathbf{n}_1, \mathbf{n}_2, \mathbf{z}, \mu)$ 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.

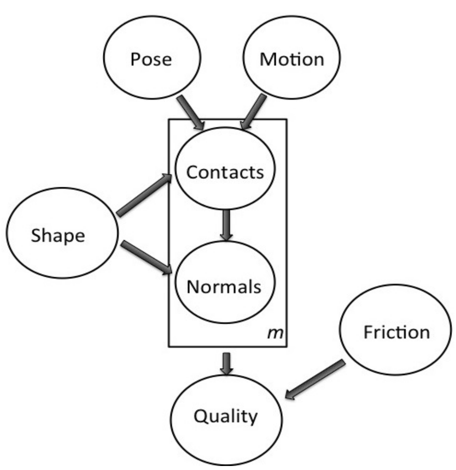

Fig. 3: A graphical model that illustrates the relationship between the different types of uncertainty in an object. As illustrated uncertainty in pose, motion and shape effect the the contacts and surface normals that makes up the grasp. Friction coefficient is independent of this relationship. The box around contact and normals means there are repeated nodes, in this case we have $m = 2$ corresponding to the two jaws in the gripper.

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 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 [33]. See Mahler et al. for details on how to train a mean and covariance function and sample shapes from a GPIS [28]. 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 $\theta = (\mu_f(x), \Sigma_f(x))$.

2) *Pose Uncertainty*

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

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

Following 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 $\xi \sim \mathcal{N}(\mathbf{0}, \Sigma_\xi)$. [4].

3) *Approach Uncertainty*

In practice a robot may not be able to execute a desired grasp plan Γ exactly due to errors in actuation or feedback measurements used for trajectory following [20]. We model approach 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. We model uncertainty in the approach center as $\mathbf{y} \sim \mathcal{N}(\hat{\mathbf{y}}, \Sigma_y)$ and uncertainty in the approach angle as $\psi \sim \mathcal{N}(\hat{\psi}, \sigma_\psi^2)$. For shorthand in the remainder of this paper we will refer to the uncertain approach parameters as $\rho = \{\mathbf{y}, \psi\}$. In practice Σ_y^2 and σ_ψ^2 might be set from repeatability measurements for a robot [32].

4) *Friction Uncertainty*

As shown in [44], uncertainty in friction coefficient can cause grasp quality to significantly vary. 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 grasp 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.

Formally, force closure is a binary-valued quantity F that is 1 if the 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]. We rank grasps using the probability of force closure given uncertainty in shape, pose, robot motion, and friction coefficient [10], [21]:

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

To estimate $P_F(\Gamma)$, we generate samples from each of the distributions in sequence using the relationships defined by the graphical model in Fig. 3. After sampling a shape, pose, approach direction, and friction coefficient, we compute the contact locations $\mathbf{c}_i$ and the surface normals $\mathbf{n}_i$ using Bayes rule:

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

Mahler et al. [28] describe how to draw shape sample from a GPIS model. We then also sample from $p(\xi)$ and $p(\rho)$ to compute $p(\mathbf{c}_i|\gamma_i(t), \theta, \rho, \xi)$ via ray tracing along the line of action $\gamma_i(t)$ [**?**]. Then $p(\mathbf{n}_i|\mathbf{c}_i, \theta)$ corresponds to the normal vector at the sampled contact point. 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.

D. Objective

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

$$\Gamma^* = \underset{\Gamma \in \mathcal{G}}{\text{argmax}}\ P\left(F = 1|\Gamma, \theta, \xi, \rho, \mu\right) \tag{3}$$

One method to approximately solve Equation 3 is to exhaustively evaluate $P_F(\Gamma)$ for all grasp 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 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]. We formulate the problem as a MAB model and compare two different Bayesian MAB algorithms, Thompson Sampling and Gittins Indices in solving it.

IV. Grasp Planning as a Multi-Armed Bandit

In this work, we frame the grasp selection problem of Section III-D as a multi-armed bandit problem. Each arm corresponds to a different grasp plan, Γ_i, 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, each grasp plan Γ_i has a Bernoulli reward distribution with probability of success $P_F(\Gamma_i)$. The expected value of the force closure condition for grasp Γ_i is $\mu_i = E\left[F|\Gamma_i, \theta, \xi, \rho, \mu\right] = P_F(\Gamma_i)$, and therefore minimizing cumulative regret up to a stopping time T_s is equivalent to the objective of Equation **??**.

[TODO: THINK OF STOPPING CRITERIA, THATS NOT ANYTIME] One can think of the proposed algorithm as an anytime algorithm, it can be stopped at anytime during its computation and return the current estimate of the best grasp or wait until its estimate of the best grasp is within some set confidence. The confidence interval can be determined via the variance of the Beta distribution, which is defined as

$$\text{Var}(P_F(\Gamma_i)) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \tag{4}$$

We then run the algorithm until the estimated best grasp at time t has $\text{Var}(P_F(\Gamma_i))$

.

V. Simulation Experiments

We used the Brown Vision Lab 2D dataset [**?**], the same used in [10].Examples of the images can be seen in 6. We down sampled 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. [TODO: FIGURE OUT WHY THESE PARAMETERS WERE SET, OR THINK OF MORE LOGICAL SETTINGS] The noise of the motion, position and friction coefficient was set to the following variances $\sigma_{mot} = 0.2$ rads, $\sigma_{mu} = 0.4$, $\sigma_{rot} = 0.3$ rads and $\sigma_{trans} = 3$ grid cells. We thought these were reasonable values based on the size of the 40x40 workspace and the amount of noise expected in industrial applications. We us the GPIS-Blur visualization technique [28]. We performed experiments for the case of two hard contacts in 2-D. We drew random grasp plans Γ by sampling the angle of grasp axis around a circle with radius $\sqrt{2}M$, where M is the dimension of the workspace, and then sampling the circle's origin.

A. Multi-Armed Bandit Experiments

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. 4, we plotted the probability of force closure,P_F , of the grasp chose vs. stopping time T_s averaged over 100 randomly selected shapes in the Brown Vision Lab 2D dataset and compare the different methods (Thompson, Gittins, Adaptive Sampling [21] and Monte-Carlo Integration). We set the discount factor $\gamma = 0.98$ for Gittins because that was the highest we could compute the indices for, since it scales exponentially in computation. The adaptive sampling method of Kehoe et al. prunes grasps every 1000 iterations based on lowest sample mean and removes 10% of the current grasp set. To illustrate the difference in grasp quality returned at a stopping time of $T = 9000$, we show the grasps selected for three randomly selected objects for all methods listed in Fig. 6. Interestingly in Fig 4, Gittins and Thompson select at stopping time T_s on average higher quality grasps than Monte-Carlo integration and the method listed in Kehoe et al [21] with the same number of samples drawn .

In Fig. 5, we plot samples per grasp quality, Gittins and Thompson appear to allocate more samples to grasps of high quality. In contrast to Monte-Carlo, which uniformly allocate grasp samples. The ability to find grasps with a higher probability of force faster can be explain by the fact that MAB methods hone in their sampling efforts on promising grasps. The adaptive sampling method proposed by Kehoe et al. also allocates more samples to promising grasps,however it is not as efficient as the MAB methods shown. As illustrated by the difference in allocated samples from grasps ranked 200 to 800 in Fig. 5.

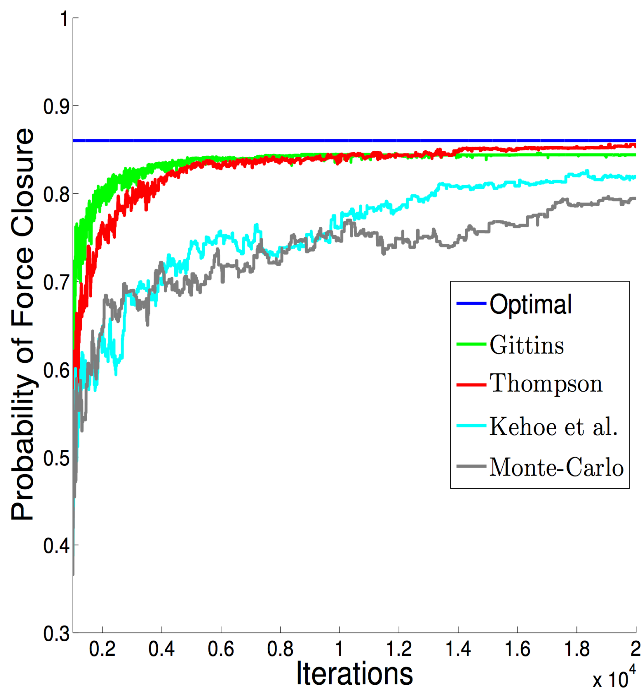

Fig. 4: Comparison of the current average probability of force closure vs. the stopping time T_s.Graph is averaged over 100 shapes randomly drawn from the Brown Vision 2D Lab Dataset [?] with a set $|G| = 1000$ for each shape. We demonstrate this for Thompson, Gittins, Monte-Carlo and the approach taking in Kehoe et al [21]. We also demonstrate what the average probability of force closure for the approximate optimal policy. Empirically, it appears that Thompson and Gittins converge at a faster rate to the optimal solution, which is desired for an anytime algorithm

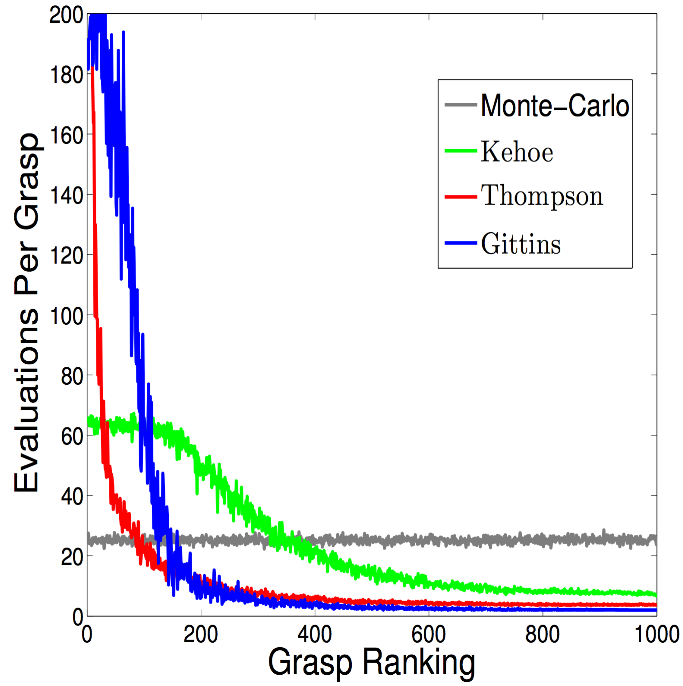

Fig. 5: Comparison of sample per grasp for the four sequential decision methods (Monte-Carlo, Thompson, Gittins). Graph is averaged over 100 shapes from the Brown Silhouette Dataset [?] with a set $|G| = 1000$ for each shape. The best grasps are ranked 1 and worst are 1000. As illustrated the MAB algorithms intelligently allocate samples towards high quality grasps based on past observations, where Monte-Carlo Integration takes a uniform approach to allocation.

B. Sensitivity Analysis

[TODO: NEED TO REDO WITH LARGER NUMBER OF SHAPES] We now will show how well Thompson Sampling 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 $|G| = 1000$, or 1000 grasp plans Γ.

For friction coefficient we varied the variance across the following values $\sigma_\mu = \{0.05, 0.2, 0.4\}$. As illustrated in Table 1, 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 Table 1, the performance of the bandit algorithms is effected by the change in rotation, increase in variance to 0.24 radians or $13^{\deg}$ causes the convergence in simple regret to not be reached until around 4432 samples or an average of 5.5 samples per grasp.

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. **??**, 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 8763 evaluations for Thompson Sampling.

C. Worst Case

The MAB algorithms use the observations of samples drawn to decide which grasp to sample next from. To show worst case performance under such a model, we sorted the quality of all 1000 grasps offline and arranged the order of samples, so that the top 500 grasps have samples drawn in the order of worst to best and the bottom 500 grasps have samples drawn in

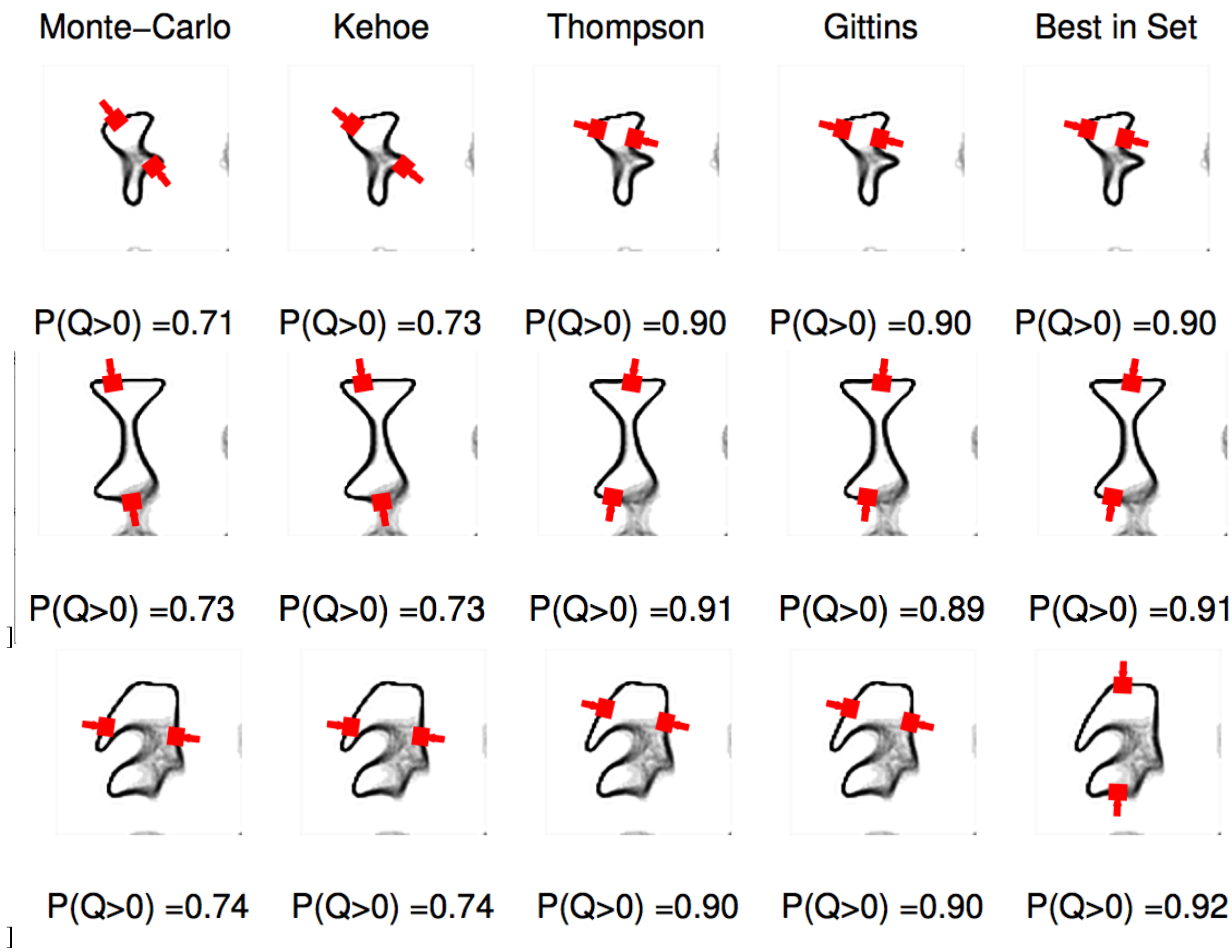

Fig. 6: [TODO: NEED TO CHANGE THIS FIGURE, WONDERING WHAT SHOULD BE CONVEYED HERE]Three shapes shown from the Brown Visual Lab Dataset with induced shape uncertainty and visualized according to the method described in [28]. The four methods (Monte-Carlo, Kehoe's, Thompson and Gittins) were all run until 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 two out of three 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.

order of best to worst. The intent here is to provide misleading observations to the bandit algorithms. We demonstrate in Fig. 7 a case where the observations are misleading.

As illustrated in Fig. 7, all the methods are affected by worst case performance. It would appear that when the observations are misleading the best thing to do is simply uniform allocation of grasp samples. One interesting aspect that is illustrated is that Thompson sampling is able to make improvement while the other two methods are stuck in local optimal. This is because Thompson sampling is guaranteed to find the best grasp in the limit of infinite iterations [1].

VI. Limitations

Our 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. Since currently the bandit algorithm doesn't utilize correlations between grasps, we are throwing away a lot of information that could be useful for choosing what to grasp to sample next.

We proposed treating the MAB as anytime algorithm and initial results in Fig. 4 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 Fig. 7 there can exists pathological cases that can mislead bandit algorithms to focus samples on the wrong grasps. Fortunately, though these cases occur with a small very probability, however it is important for a practitioner to be aware of them.

VII. 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 coefficient 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

Sensitivity Analysis for Convergence to Best Grasp for Thompson Sampling			
Uncertainty Type	Low Uncertainty (Iterations)	Medium Uncertainty (Iterations)	High Uncertainty (Iterations)
Translation Variance in Pose, σ_{trans}	1210	2207	8763
Friction Coefficient Variance, σ_{fric}	1985	1456	1876
Rotational Variance in Pose, σ_{rot}	4230	4431	4432

TABLE I: Sensitivity Analysis for convergence to estimated best grasp for Thompson Sampling under rotational variance $\sigma_{rot} = \{0.03, 0.12, 0.24\}$ radians, translation uncertainty $\sigma_{trans} = \{3, 12, 24\}$ units friction coefficient uncertainty $\sigma_{fric} = \{0.05, 0.2, 0.4\}$ from left to right on a 40 x 40 unit workspace averaged over 10 shapes from the Brown Vision Lab Data set. The sensitivity analysis shows that large variance in translational uncertainty in pose can increase the amount of iterations needed for the bandit algorithm to converge to the highest quality grasp in the set.

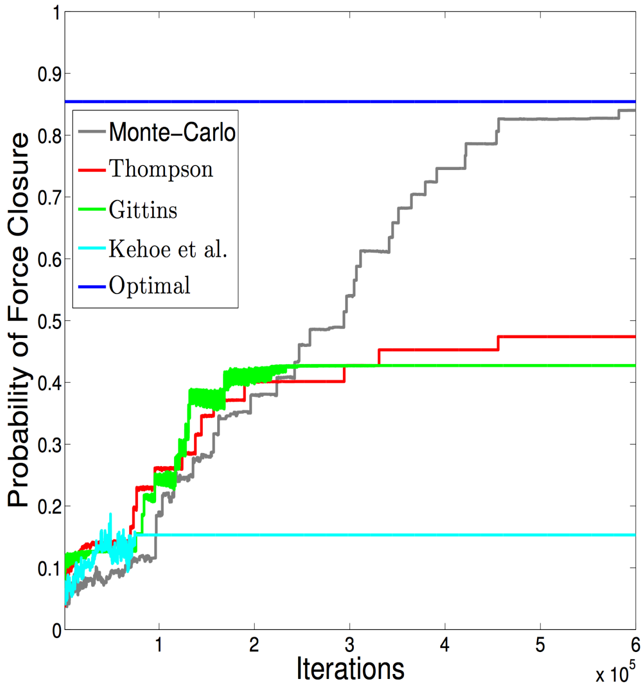

Fig. 7: Comparison of the current average probability of force closure vs. the stopping time T_s for the pathological case where the samples are sorted to be misleading. The graph is averaged over 100 shapes randomly drawn from the Brown Vision 2D Lab Dataset [**?**] with a set $|G| = 1000$ for each 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. Empirically, it appears that when samples are misleading the best policy is uniform allocation. However, it also illustrated that Thompson sampling can recover from this situation where Gittins and Kehoe et al. are not able to and get stuck in local solutions.

formulate a representation of how uncertainty affects grasp parameters and thus grasp quality. We propose treating this as a graphical model and model the parameters as stochastic noise. The MAB algorithm is then applicable in the context of bounded reward distributions. Initial results have shown our MAB approach it to outperform the methods of prior work Monte-Carlo and the method purposed by Kehoe et al. [20] in terms of finding a higher quality grasp faster.

As we scale to larger 3D objects the possible number of grasp candidates will be substantially larger, the more options a MAB algorithm has leads to generally a harder problem [7]. The MAB algorithms presented above assumed that each option is independent. In the grasping context though spatial information about grasps could provide information about the quality of similar grasps. We can estimate a correlated Beta-Bernoulli distribution using an Kernel density estimation on the α and β parameter [**?**].However we have to find an appropriate feature representation for a grasp, which is an exciting project for on going work.

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.