Multi-Arm Bandit Models for 2D Sample Based Grasp Planning with Uncertainty

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*— For applications such as warehouse order fulfillment, robot grasps must be robust to uncertainty arising from sensing, shape, mechanics, and control. One way to achieve this is to evaluate the performance of candidate grasps by sampling perturbations in shape, pose, and control, computing the probability of force closure for each candidate to identify the grasp with the highest expected quality. Prior work has turned to cloud computing because evaluating the quality of each grasp is computationally demanding. To improve efficiency and extend this work, we consider how Multi-Armed Bandit (MAB) models for optimizing decisions can be applied in this context. We formulate robust grasp planning as a MAB problem and evaluate grasp planning convergence time using 100 object shapes randomly selected from the Brown Vision 2D Lab Dataset from 1000 uniformly distributed candidate grasp angles. We consider the case where shape uncertainty is represented as a Gaussian process implicit surface (GPIS) and there is Gaussian uncertainty in pose, gripper approach angle, and coefficient of friction. Uniform allocation and iterative pruning, the non-MAB methods, converge slowly. In contrast, Thompson Sampling and the Gittins index method, the MAB methods we consider, converged to within 3% of the optimal grasp 5x faster than iterative pruning.**

I. Introduction

Consider a robot fulfilling orders in a warehouse, where it encounters new consumer products and must handle them quickly. Planning grasps using analytic methods requires knowledge of contact locations and surface normals. However, a robot 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.

One common measure of quality is force closure, the ability to resist external forces and torques in arbitrary directions [30]. To cope with uncertainty, recent work has explored computing the probability of force closure given uncertainty in pose [11], [28], [42] and object shape [24], [31]. One way to compute the probability of force closure is using Monte-Carlo integration over sample perturbations in the uncertain quantities[11], [28], [42], [25]. However, performing Monte-Carlo integration for each candidate grasp in a set is computationally expensive. Past work has looked at using parallel programming in the cloud to overcome this problem [24], [25], [26]. In this work, we aim to extend these cloud-based methods by reducing the number of samples needed to converge to a high-quality grasp.

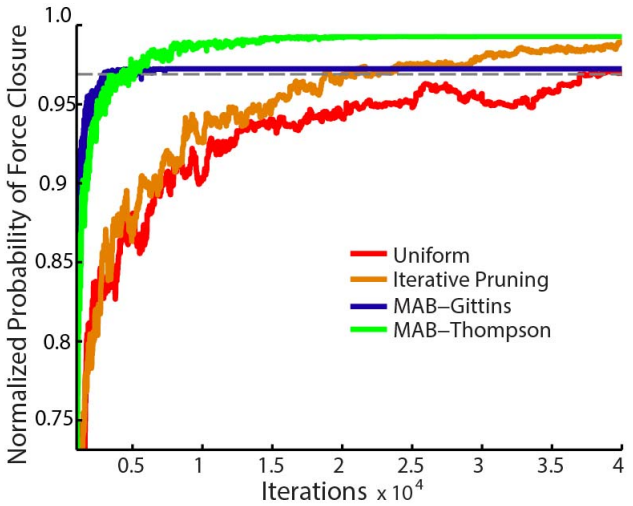

Fig. 1: Number of samples versus the normalized probability of force closure P_F for the best estimated grasp after t samples, $P_F(\Gamma_{\bar{k},t})$, out of 1000 candidate grasps using uniform allocation, iterative pruning (eliminating candidates that perform poorly on initial samples), and our proposed Multi-Armed Bandit (MAB) algorithms (Gittins indices and Thompson Sampling). The normalized P_F is the ratio of $P_F(\Gamma_{\bar{k},t})$ to the highest P_F in the candidate grasp set $P_F(\Gamma^*)$ averaged over 100 independent runs on randomly selected objects from the Brown Vision 2D Dataset [1]. The highest quality grasp was determined by brute force search over all candidate grasps (a method that is 10x more iterations than all methods shown) [25] Uniform allocation and iterative pruning converge to within 3% of the optimal (the dashed grey line) in approximately 40,000 and 20,000 iterations, respectively. In comparison, the MAB methods both converge in approxmately 4,000 iterations.

Our main contribution is formulating this problem as a Multi-Armed Bandit (MAB) and showing it is possible to allocate sampling effort to only grasps that seem promising [4], [29], [35]. In a standard MAB there are a set of possible options, or 'arms' [4], 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. 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 then treat the MAB algorithm as an anytime algorithm that terminates once a user defined confidence level is met or at a given stopping

[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

Preprint submitted to 2015 IEEE International Conference on Automation Science and Engineering. Received March 20, 2015.

time.

We study this formulation using probability of force closure [11], [42], [25] as a quality metric under uncertainty in pose, shape, gripper approach, 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 [13], [21]. We model uncertainty in pose as a normal distribution around the orientation and 2D position of the object. We chose to model uncertainty in grasp approach as a normal distribution around the center and angle of the grasp axis for a parallel jaw gripper. We model uncertainty in friction coefficient as 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, against uniform allocation and an adaptive sampling method known as iterative pruning, which iteratively reduces the set of candidate grasps based on sample mean [25]. In the task of finding grasps with high probability of force closure. We compare on the Brown Vision 2D Dataset, a dataset of 2D planar objects [1], [11].

Our initial results in simulation show that Thompson Sampling, a MAB algorithm, required $5\times$ fewer samples than iterative pruning to plan a grasp within 3% of the estimated highest probability of force closure in the set of 1000 randomly selected grasps averaged over 100 objects. Although our current implementation is local, our methods may be extended to the cloud by solving M MAB problems on disjoint subsets of the arms and aggregating the results [17].

II. Related Work

Most research on grasp planning focuses on finding grasps by maximizing a grasp quality metric. Grasp quality is often measured by the ability to resist external perturbations to the object in wrench space [15], [32]. Analytical quality metrics typically assume precisely known object shape, object pose, and locations of contact [10], [12]. Work on grasping under uncertainty has considered uncertainty in the state of a robotic gripper [19], [40] and uncertainty in contact locations with an object [43] Furthermore, recent work has studied the effects of uncertainty in object pose and gripper positioning[6], [22].

Brook, Ciocarlie, and Hsiao [6], [22] studied a Bayesian framework to evaluate the probability of grasp success given uncertainty in object identity, gripper positioning, and pose by simulating grasps on deterministic mesh and point cloud models. Weisz et al. [42] found that grasps ranked by probability of force closure subject to uncertainty in object pose were empirically more successful on a physical robot than grasps planned using deterministic wrench space metrics. Similarly, Kim et al. [28] planned grasps using dynamic simulations over perturbations in object pose. They also found that the planned grasps were more successful on a physical robot than those planned with classical wrench space metrics.

Recent work has also studied uncertainty in object shape, motivated by the use of uncertain low-cost sensors and tolerances in part manufacturing. Christopoulos et al. [11] sampled spline fits for 2-dimensional planar objects and ranked a set of randomly generated grasps by probability of force closure. Kehoe et al. [24], [25] sampled perturbations in shape for extruded polygonal objects to plan push grasps for parallel-jaw grippers. Several recent works have also studied using Gaussian process implicit surfaces (GPISs) to represent shape uncertainty motivated by its ability to model spatial noise correlations and to integrate multiple sensing modalities [13], [14], [21], [31]. Dragiev et al. [13] uses GPIS to actively explore shapes with tactile sensing to find a hand posture that aligned the gripper fingers to an object's surface normals [14]. Mahler et al. used the GPIS representation to find locally optimal anti-podal grasps by framing grasp planning as an optimization problem [31].

Some probabilistic grasp quality measures, such as probability of force closure, are computed using Monte-Carlo integration [11], [25], [28], [42]. Monte-Carlo involves sampling from distributions on uncertainty quantities and averaging the quality over these samples to empirically estimate a probability distribution [8]. However, it can be computationally expensive to sample all proposed grasps to convergence. To address the computational cost, Kehoe et al. [24] proposed an adaptive sampling procedure called iterative pruning, which periodically discards a subset of the grasps that seem unlikely to be of high probability of force closure. However, the method pruned grasps using only the sample mean, which could discard good grasps. We propose modeling the problem as a Multi-Armed Bandit, which selects the next grasp to sample based on past observations instead [4], [29].

A. MAB Model

The MAB 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 MAB 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].

A traditional MAB example is a gambler with 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 that is unknown to the gambler. The goal of the gambler is to come up with a method to maximize the sum of average rewards over all 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 expected reward and exploring new arms. Developing a policy that successfully trades between exploration and exploitation reward has been the focus of extensive research since the problem formulation [5],[7], [35].

Preprint submitted to 2015 IEEE International Conference on Automation Science and Engineering. Received March 20, 2015.

At each time step the MAB algorithm incurs *regret*, the difference between the expected reward of the best arm and that of the selected arm. Bandit algorithms minimize cumulative regret, the sum of regret over the entire sequence of arm choices. Lai and Robbins [29] 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. They presented an algorithm called (Upper Confidence Bound) UCB that obtains this bound asymptotically [29]. 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.

B. Bayesian 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], [2]. Bayesian methods have been shown empirically to outperform UCB [9], [3]. 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 p can be used to represent a binary grasping metric like force closure. The prior typically 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. At timestep $t = 0$, we pull arm k and observe reward $R_{k,0}$, where $R_{k,0} \in \{0, 1\}$. The posterior of the Beta after this observation is $\alpha_{k,1} = \alpha_{k,0} + R_{k,0}$ $\beta_{k,1} = \beta_{k,0} + 1 - R_{k,0}$, where $\alpha_{k,0}$ and $\beta_{k,0}$ are the prior shape parameters for arm k before any samples are evaluated.

Given the current belief $\alpha_{k,t}, \beta_{k,t}$ for an arm k at time t, the algorithm can calculate the expected Bernoulli parameter, $\bar{p}_{k,t}$, as follows:

$$\bar{p}_{k,t} = \frac{\alpha_{k,t}}{\alpha_{k,t} + \beta_{k,t}} \tag{1}$$

1) The Gittins Index Method: One MAB method is to treat the problem as a Markov Decision Process (MDP) and use Markov Decision theory. Formally, a MDP is defined as a set of states, a set of actions, a set of transition probabilities between states, a reward function, and a discount factor [4]. In the Beta-Bernoulli MAB case, the set of actions is the K arms and the states are the Beta posterior on each arm, or the integer values of $\alpha_{k,t}$ and $\beta_{k,t}$.

Methods such as Value Iteration can compute optimal policies for a discrete MDP with respect to the discount factor γ [4]. However, the curse of dimensionality effects performance because for K arms, a finite horizon of T and a Beta-Bernoulli distribution on each arm then the state space is exponential in K. A key insight 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 k and for each state $x_{k,t} = \{\alpha_{k,t}, \beta_{k,t}\}$ up to a timestep T [41].

The solution to the optimization problem assigns each state an index $v(x_{k,t})$. The indices can then be used to form a policy, where at each timestep the agent selects the arm k_t where $k_t = \underset{1 \le k \le K}{\operatorname{argmax}}\, v(x_{k,t})$. The indices can be computed offline using a variety of methods [41]; we chose to use the restart method proposed by Katehakis et al. [23] because it can be implemented in a dynamic programming fashion. We refer the reader to [16] for a more detailed analysis of the Gittins index method.

2) Thompson Sampling: Computation of the Gittins indices can increase exponentially in time as the discount factor approaches 1. However, in the case of finding the best arm, we want to plan for long term reward and thus want γ as close to 1 as possible. Due to computational constraints we must use a smaller γ, but this can lead to the algorithm pulling only the most promising arm for many iterations [27].

Thompson sampling is an alternative to the Gittins index method that isn't prone to such a problem. In Thompson sampling, for each arm draw $p_{k,t} \sim \text{Beta}(\alpha_{k,t}, \beta_{k,t})$ and pull the arm with the highest $p_{k,t}$ drawn. A reward is then observed, $R_{k,t}$ and the corresponding Beta distribution is updated. All arms are initialized with prior Beta distributions, which is normally Beta($\alpha_{k,0} = 1$, $\beta_{k,0} = 1$) $\forall 1 \le k \le K$ to reflect a uniform prior on the parameter of the Bernoulli distribution, $p_{k,0}$.The full algorithm is shown in Algorithm 1.

Algorithm 1: Thompson sampling for Beta-Bernoulli Process

Result: Current Best Arm, Γ^*
For Beta($\alpha_{k,0} = 1$,$\beta_{k,0} = 1$) $\forall k \in K$ prior:
for *t=1,2,...* **do**
  Draw $p_{k,t} \sim$ Beta($\alpha_{k,t}$,$\beta_{k,t}$) for $k = 1, ..., K$
  Pull $k_t = \underset{k \in K}{\operatorname{argmax}}\, p_{k,t}$
  Observe reward $R_{k,t} \in \{0, 1\}$
  Update posterior:
  if $k = k_t$ **then**
    Set $\alpha_{k,t+1} = \alpha_{k,t} + R_{k_t,t}$
    Set $\beta_{k,t+1} = \beta_{k,t} + 1 - R_{k_t,t}$
  else
    Set $\alpha_{k,t+1} = \alpha_{k,t}$
    Set $\beta_{k,t+1} = \beta_{k,t}$

The intuition for Thompson sampling is that the random samples of $p_{k,t}$ 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 [18]. 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

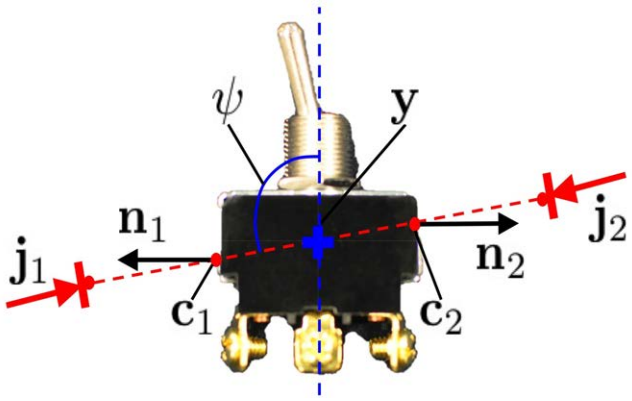

Fig. 2: Illustration of our grasping model for parallel jaw grippers on a mechanical switch. Jaw placements are illustrated by a red direction arrow and line. A grasp plan centered at $\mathbf{y}$ (plus symbol) at angle ψ consists of 2D locations for each of the parallel jaws $\mathbf{j}_1$ and $\mathbf{j}_2$. When following the grasp plan, the jaws contact the object at locations $\mathbf{c}_1$ and $\mathbf{c}_2$, and the object has outward pointing unit surface normals $\mathbf{n}_1$ and $\mathbf{n}_2$ at these locations. 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.

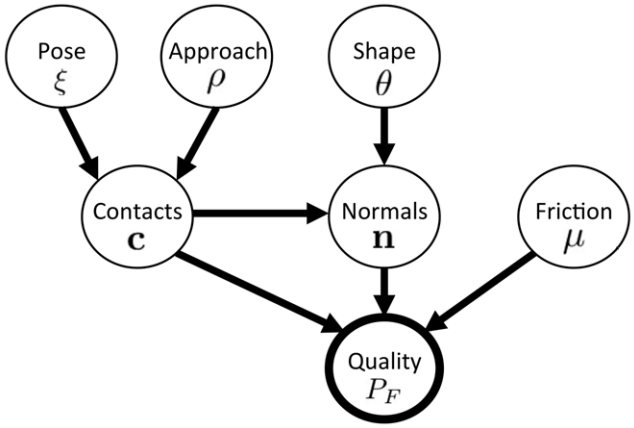

Fig. 3: A graphical model of the relationship between the uncertain parameters we consider. Uncertainty in object shape θ, object pose ξ, and grasp approach angle ρ affect the points of contact $\mathbf{c}$ with the object and the surface normals $\mathbf{n}$ at the contacts. Uncertainty in friction μ coefficient affects the forces and torques used to compute our quality measure, the probability of force closure P_F.

sampling that was sub-linear for k -arms and logarithmic in the case of 2 arms [2].

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 [24], [25]. We consider uncertainty in shape, pose, gripper approach, and friction coefficient. We assume that the distributions on these quantities are given and can be sampled from.

A. Candidate Grasp Model

The grasp model is illustrated in Fig. 2. We formulate the MAB problem for planar objects using parallel-jaw grippers as modeled in Fig. 2. Similar to [31], we parameterize a grasp using a *grasp axis*, the axis of approach for two jaws, with jaws of width $w_j \in \mathbb{R}$ and a maximum width $w_g \in \mathbb{R}$. The two location of the jaws can be specified as $\mathbf{j}_1, \mathbf{j}_2 \in \mathbb{R}^2$, where $||\mathbf{j}_1 - \mathbf{j}_2||_2 \leq w_g$. We define a grasp consisting of the tuple $\Gamma = \{\mathbf{j}_1, \mathbf{j}_2\}$.

Given a grasp and an object, we define the *contact points* as the spatial locations at which the jaws come into contact with the object when following along the 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 $\mathbf{n}_1, \mathbf{n}_2 \in \mathbb{R}^2$, the object center of mass as $\mathbf{z} \in \mathbb{R}^2$ and the friction coefficient as $\mu \in \mathbb{R}$.

B. Sources of Uncertainty

We consider uncertainty in object shape, object pose, grasp approach angle, and friction coefficient. Fig. 3 illustrates a graphical model of the relationship between these sources of uncertainty. In this section, we describe our model of each source of uncertainty.

1) Shape Uncertainty: Uncertainty in object shape results from sensor imprecision and missing sensor data, which can occur due to transparency, specularity, and occlusions [31]. Following [31], 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 over spatial locations $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ 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 $\mathcal{X} = \{\mathbf{x}_1, ...\mathbf{x}_n\}, \mathbf{x}_i \in \mathbb{R}^2$, $f(\mathbf{x}_i) \sim \mathcal{N}(\mu_f(\mathbf{x}_i), \Sigma_f(\mathbf{x}_i))$, where $\mu_f(\cdot)$ and $\Sigma_f(\cdot)$ are the mean and covariance functions of the GPIS [34]. See Mahler et al. for details on how to estimate a mean and covariance function and sample shapes from a GPIS [31]. For convenience, in later sections we will refer to the GPIS parameters as $\theta = \{\mu_f(\mathbf{x}), \Sigma_f(\mathbf{x})\}$.

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 $\xi = (\phi, \mathbf{t})^T \in \mathbb{R}^3$. We assume Gaussian uncertainty on the pose parameters $\xi \sim \mathcal{N}\left(\hat{\xi}, \Sigma_\xi\right)$, where $\hat{\xi}$ corresponds to the expected pose of the object.

3) Approach Uncertainty: In practice a robot may not be able to execute a desired grasp $\Gamma = \{\mathbf{j}_i, \mathbf{j}_2\}$ exactly due to errors in actuation or feedback measurements used for trajectory following [24]. We model approach uncertainty as Gaussian uncertainty around the angle of approach and centroid of a straight line grasp Γ. Formally, let $\hat{\mathbf{y}} = \frac{1}{2}(\mathbf{j}_1 + \mathbf{j}_2)$ denote the center of a planned grasp axis and $\hat{\psi}$ denote the clockwise angle that the planned axis $\mathbf{j}_1 - \mathbf{j}_2$ makes with the y-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 $\psi \sim \mathcal{N}(\hat{\psi}, \sigma^2_\psi)$. For shorthand in the remainder of this paper we will refer to the uncertain approach parameters as $\rho = \{\mathbf{y}, \psi\}$. In practice Σ^2_y and σ^2_ψ can be set from

Preprint submitted to 2015 IEEE International Conference on Automation Science and Engineering. Received March 20, 2015.

repeatability measurements for a robot [33].

4) Friction Uncertainty: As shown in [43], [20], 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 [24], [25], [28], [42] given a grasp Γ. Force closure is a binary-valued quantity F that is 1 if the grasp can resist wrenches in arbitrary directions and 0 otherwise. Let $\mathcal{W} \in \mathbb{R}^3$ denote the contact wrenches derived from contact locations $\mathbf{c}_1, \mathbf{c}_2$, normals $\mathbf{n}_1, \mathbf{n}_2$, 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 [30]. We rank grasps using the probability of force closure given uncertainty in shape, pose, robot approach, and friction coefficient [11], [25]:

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

To estimate $P_F(\Gamma_k)$, we first generate samples from the distributions on θ, ξ, ρ, and μ. Using the relationships defined by the graphical model in Fig. 3, we next compute the contact locations $\mathbf{c}_1, \mathbf{c}_2$ given a sampled SDF, pose, and grasp approach by ray tracing along the grasp axis defined by $\Gamma_k = \{\mathbf{j}_1, \mathbf{j}_2\}$ [31]. We then compute the surface normals $\mathbf{n}_1, \mathbf{n}_2$ at the contacts using the gradient of the sampled SDF at the contact locations. Finally, we use these quantities to compute the forces and torques that can be applied to form the contact wrench set $\mathcal{W}$ and evaluate the force closure condition [30].

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 candidate grasps $\mathcal{G} = \{\Gamma_1, ..., \Gamma_K\}$:

$$\Gamma^* = \underset{\Gamma_k \in \mathcal{G}}{\operatorname{argmax}} \; P_F\left(\Gamma_k\right) \tag{2}$$

One method to approximately solve Equation 2 is to exhaustively evaluate $P_F(\Gamma_k)$ for all grasp in $\mathcal{G}$ using Monte-Carlo integration and then sort the plans by this quality metric. We refer to this as a brute force approach. This method has been evaluated for shape uncertainty [11], [24] 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 work has considered adaptive sampling to discard grasps that are not likely to be optimal without fully evaluating their quality [25].

To try and reduce the number of samples needed, we instead maximize the sum of P_F values for each sampled grasp $\Gamma_{k,t}$ at time t up to a given time T_s:

$$\max_{\Gamma_{k,} \in \mathcal{G}} \sum_{t=1}^{T_s} P_F\left(\Gamma_{k,t}\right) \tag{3}$$

This attempts to perform as well as Equation 2 in as few samples as possible [38]. We then formulate problem as a MAB model and compare two different Bayesian MAB algorithms, Thompson sampling and Gittins indices.

IV. Grasp Planning as a Multi-Armed Bandit

We frame the grasp selection problem of Section III-D as a MAB problem. Each arm corresponds to a different grasp, Γ_k, 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 Γ_k has a Bernoulli reward distribution with probability of force closure, $P_F(\Gamma_k)$. In a MAB, we want to try and minimize cumulative regret which is an equivalent objective to the objective of Equation 3.

One can think of the proposed algorithm as an anytime algorithm. It can be stopped at anytime during its computation to return the current estimate of the best grasp or wait until a 95% confidence interval is smaller than some threshold ϵ. Using the quantile function of the beta distribution, B, we can measure the 95% confidence interval as:

$$B(0.025, \alpha_{k,t}, \beta_{k,t}) \leq P_F(\Gamma_{k,t})) \leq B(0.975, \alpha_{k,t}, \beta_{k,t}) \tag{4}$$

To summarize, the algorithm terminates and returns $\bar{k}$, or the grasp that has the highest estimated P_F when the following condition is met:

$$\{t \geq T_s \text{ OR } |B(0.025, \alpha_{\bar{k},t}, \beta_{\bar{k},t}) - B(0.975, \alpha_{\bar{k},t}, \beta_{\bar{k},t})| \leq \epsilon\}. \tag{5}$$

V. Simulation Experiments

We used the Brown Vision Lab 2D dataset [1], a database of 2D planar objects, as in [11]. GPIS models of 3 example objects are visualized using the GPIS-Blur method of [31] in Figure 5. We downsampled the silhouette 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 measurement noise were assigned uniformly at random to the SDF. We then construct a GPIS using the same method proposed in [31]. For illustrative purposes, the noise in approach, pose, and friction coefficient were set to the following variances $\sigma_\psi = 0.2$ rads, $\sigma_{\mathbf{y}} = 3$ grid cells, $\sigma_\mu = 0.4$, $\sigma_\phi = 0.3$ rads and $\sigma_{\mathbf{t}} = 3$ grid cells. We also performed experiments for the case of two hard contacts in 2-D. We drew random grasps 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. All experiments were run on machine with OS X with a 2.7 GHz Intel core i7 processor and 16 GB 1600 MHz memory in Matlab 2014a.

Preprint submitted to 2015 IEEE International Conference on Automation Science and Engineering. Received March 20, 2015.

A. Multi-Armed Bandit Experiments

For our experiments we look at selecting the best grasp out of a size of $|G| = 1000$. We draw samples from our graphical model using the technique described in Sec. III-C. We averaged over 100 randomly selected shapes in the Brown Vision Lab 2D dataset and for the grasps planned by Thompson sampling, Gittins indices, iterative pruning [25] and uniform allocation). Uniform allocation selects a grasp at random from the set to sample next, thus does not use any prior information. Iterative pruning prunes grasps every 1000 iterations based on lowest sample mean and removes 10% of the current grasp set. We set the discount factor $\gamma = 0.98$ for Gittins because that was the highest we could compute the indices for in a feasible amount of time.

In Fig. 1, we plot time t vs. $P(\Gamma_{\bar{k},t})/P(\Gamma^*)$, the normalized probability of force closure for the grasp returned by the algorithm $\Gamma_{\bar{k}}$. Non-MAB methods such as uniform sampling and iterative pruning (eliminating candidates that perform poorly on initial samples) eventually converge to within 3% of the optimal, requiring approximately 40,000 and 20,000 iterations. Gittins indices and Thompson sampling perform significantly better, converging after only 4000 iterations.For illustrative purposes in Fig. 5, we select a stopping time $T_s = 10,000$, which is 10 samples per grasp on average, and for each method visualize the grasp returned, $\Gamma_{\bar{k}}$, on 3 randomly selected shapes in the dataset.

The time per iteration is $t_i = t_a + t_p$, where t_a is the time to decide which arm to pull next and t_p is the time taken to draw a sample from the graphical model in Fig. 3. The time per iteration for Thompson sampling, Gittins indices, iterative pruning and uniform allocation is 31.6, 31.2, 30.4 and 30.2 ms. Most of t_i is dominated by sampling time, since $t_p \approx 30$ms.

The MAB algorithm can also be terminated when the 95% confidence interval around the returned grasp (see Equation 4) is below a set threshold ϵ. We plot the confidence intervals around the returned grasp $P_F(\Gamma_{\bar{k}})$ vs. the number of samples drawn in Fig. 4 for Gittins indices, Thompson sampling, iterative pruning[25] and uniform allocation. As illustrated, the confidence interval for Thompson sampling and Gttins indices converges at a faster rate than the other two methods.

B. Sensitivity Analysis

We also analyze the performance of Thompson sampling under variations in noise from friction coefficient uncertainty, shape uncertainty, rotational pose, and translation pose. 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 100 objects randomly selected with from the Brown dataset with $|\mathcal{G}| = 1000$, or 1000 grasps, Γ.

For friction coefficient we varied σ_μ 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 5000 evaluations.

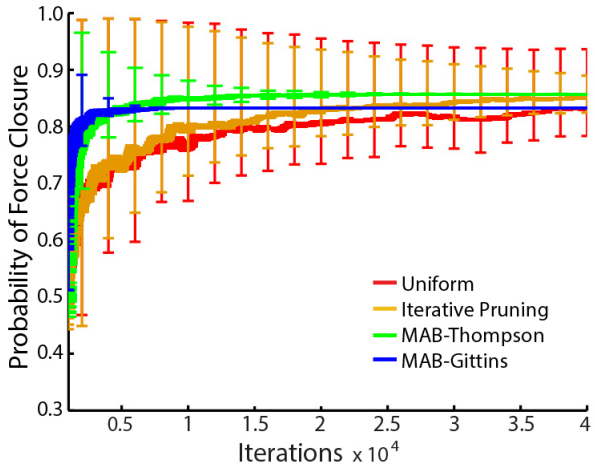

Fig. 4: Number of samples versus the 95% confidence intervals from Eq. 4 on the probability of force closure of the best estimated grasp after t samples using uniform allocation, iterative pruning, Gittins indices, and Thompson Sampling. The values are averaged over 100 independent runs on randomly selected objects from the Brown Vision 2D Dataset [1] with 1,000 candidate grasps for each object. An increasingly narrow confidence interval indicates that the algorithm allocated an increasing number of samples to its estimate of the best grasp.

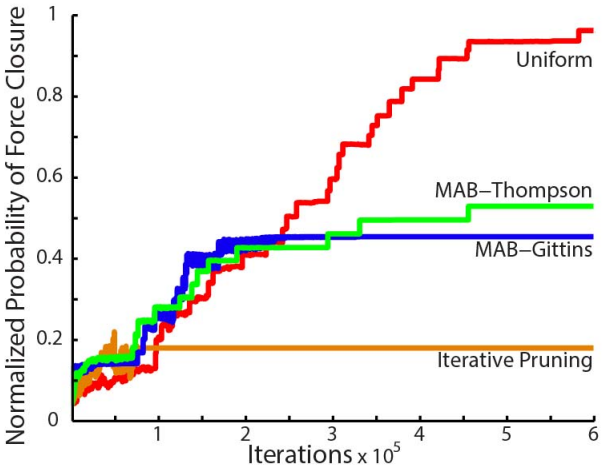

Fig. 6: Number of samples versus the probability of force closure of the best estimated grasp after t samples $P_F(\Gamma_{\bar{k},t})$ using uniform allocation, iterative pruning, Gittins indices, and Thompson sampling over 1,000 candidate grasps when samples are ordered such that many perturbations that do result in force closure are given to the algorithm initially. The normalized P_F is the ratio of the best estimated grasp at iteration t, $P_F(\Gamma_{\bar{k},t})$, to the highest P_F in the candidate grasp set $P_F(\Gamma^*)$ averaged over 100 independent runs on randomly selected objects from the Brown Vision 2D Dataset [1]. The highest quality grasp was determined by brute force search over all candidate grasps (which required 10x more time than any of these methods [25]). The results suggest that when samples are misleading the best policy is uniform allocation. However, Thompson sampling appears to continue making progress whereas iterative pruning and Gittins indices do not.

For rotational uncertainty in pose, we varied σ_ϕ 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° causes the convergence in simple regret to not be reached until around 6432 samples or an average of 6.4 samples per grasp.

For translational uncertainty in pose, we varied $\sigma_\mathbf{t}$ in the range of $\{3, 12, 24\}$ units (on a 40 x 40 unit workspace). Our

Preprint submitted to 2015 IEEE International Conference on Automation Science and Engineering. Received March 20, 2015.

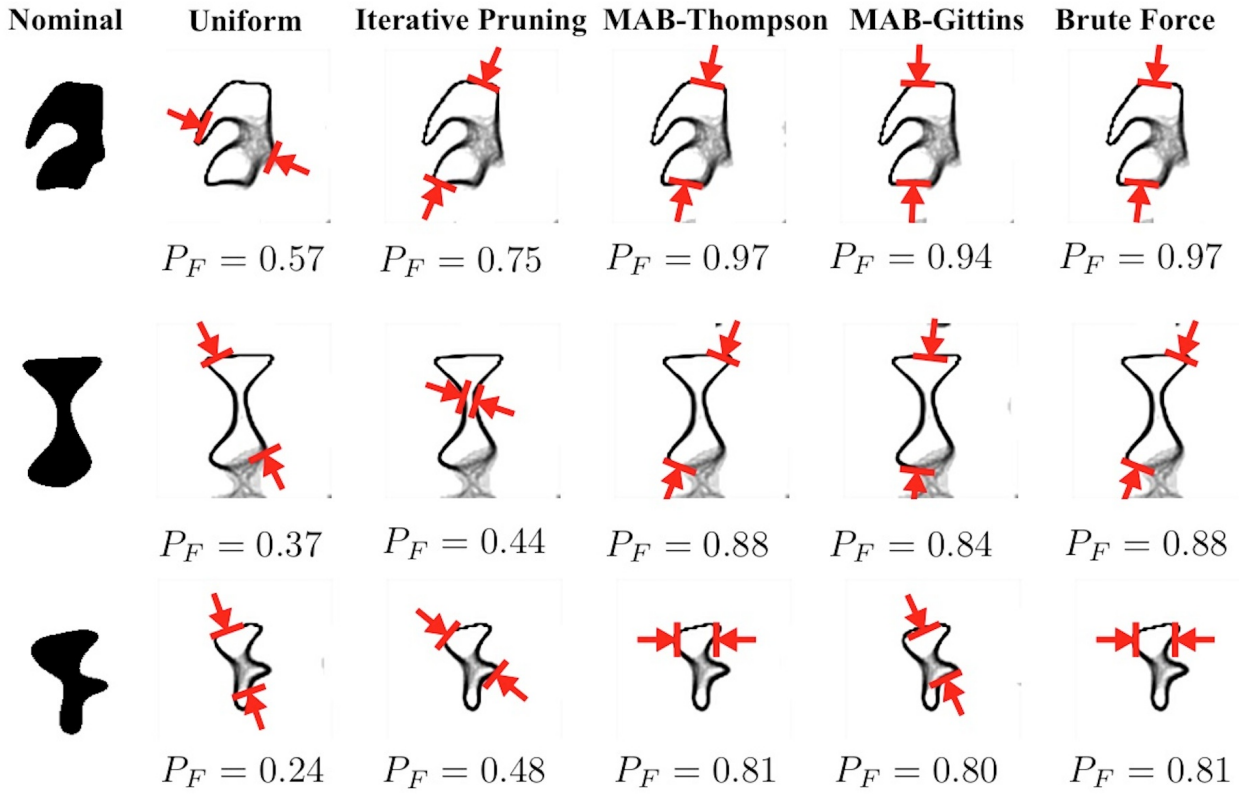

Fig. 5: Grasps with the highest estimated probabilty of force closure P_F after 10,000 samples using uniform allocation, iterative pruning, Gittins indices, and Thompson sampling on three objects from the Brown Vision 2D Dataset. For reference, we also show the grasp with highest P_F after brute force evaluation using 100,000 samples and the nominal shape. The candidate grasp set was of size $|\mathcal{G}| = 1000$. We visualize the GPIS representation of the object shape uncertainty using GPIS-Blur [31], which causes more uncertain areas to appear more blurry. On objects A and B Thompson sampling finds the best grasp after 10,000 samples, but for object C the grasp selected has 2% lower P_F.

results indicate that the performance of the bandit algorithms is affected by the change in rotation and an increased noise of $\sigma_{\mathbf{t}} = 24$ 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 order of best to worst. This provides misleading observations to the bandit algorithms. We demonstrate in Fig. 6 a case where the observations are misleading.

As illustrated in Fig. 6, all the methods are affected by worst case performance. The results indicate that when the observations are misleading the best thing to do is uniform allocation of grasp samples. Interestingly, Thompson sampling appears to to continue to improvement while Gittins indices and iterative pruning do not continue to make progress.

VI. Discussion and Future Work

In this work, we proposed a multi-armed bandit approach to efficiently identify high-quality grasps under uncertainty in shape, pose, friction coefficient and approach. A key insight from our work is that exhaustively sampling each grasp is inefficient, and we found that a MAB approach gives priority to promising grasps and can reduce computational time. Initial results have shown MAB algorithms to outperform the methods of prior work, uniform allocation and iterative pruning [25][24] in terms of finding a higher quality grasp faster. However, as shown in Fig. 6 there are some pathological cases that can mislead bandit algorithms to focus samples on the wrong grasps. Fortunately, the probability of many successive samples being misleading approaches zero at an exponential rate as the time horizon is increased.

In future work, we plan to scale our method to 3D objects. This could substantially increase the number of candidate

	# of Samples Until Convergence		
Uncertainty Type	**Low Uncertainty**	**Medium Uncertainty**	**High Uncertainty**
Orientation σ_ϕ	4230	5431	6432
Position $\sigma_{\mathbf{t}}$	4210	5207	8763
Friction σ_μ	4985	4456	4876

TABLE I: Number of iterations until convergence to within 3% of grasp with the highest estimated probability of force closure P_F for Thompson sampling under uncertainty in the object orientation $\sigma_\phi = \{0.03, 0.12, 0.24\}$ radians, uncertainty in the object position $\sigma_{\mathbf{t}} = \{3, 12, 24\}$ units, and uncertainty in friction coefficient $\sigma_\mu = \{0.05, 0.2, 0.4\}$ on a 40×40 grid averaged over 100 independent runs on random objects from the Brown Vision 2D Dataset. High variances in position and orientation uncertainty increases the amount of iterations needed for the bandit algorithm to converge.

grasps, further motivating the use of cloud computing. Glazebrook and Wilkinson showed that the Gittins index method could be parallelized by simply dividing the arms into M subsets, where M is the number of cores, and solving each MAB separately. [17]. A similar method could also be done for Thompson sampling. Another promising scheme for parrallelizing the MAB, is to have the algorithm not sample one arm at each algorithm but M. We will explore both of these approaches in future work.

VII. Acknowledgments

This work is supported in part by the U.S. National Science Foundation under Award IIS-1227536, NSF-Graduate Research Fellowship, and by grants from Google. We thank our colleagues who gave feedback and suggestions in particular Sanjay Krishnan, Peter Bartlett, Steve McKinley and Dylan Hadfield-Menell.

References

[1] "2d planar database," https://vision.lems.brown.edu/content/available-software-and-databasesl.

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

[3] P. Bachman and D. Precup, "Greedy confidence pursuit: A pragmatic approach to multi-bandit optimization," in *Machine Learning and Knowledge Discovery in Databases*. Springer, 2013, pp. 241–256.

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

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

[6] P. Brook, M. Ciocarlie, and K. Hsiao, "Collaborative grasp planning with multiple object representations," in *Proc. IEEE Int. Conf. Robotics and Automation (ICRA)*. IEEE, 2011, pp. 2851–2858.

[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] J.-S. Cheong, H. Kruger, and A. F. van der Stappen, "Output-sensitive computation of force-closure grasps of a semi-algebraic object," vol. 8, no. 3, pp. 495–505, 2011.

[11] 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.

[12] M. T. Ciocarlie and P. K. Allen, "Hand posture subspaces for dexterous robotic grasping," *Int. J. Robotics Research (IJRR)*, vol. 28, no. 7, pp. 851–867, 2009.

[13] 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.

[14] ——, "Uncertainty aware grasping and tactile exploration," in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 113–119.

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

[16] J. Gittins, K. Glazebrook, and R. Weber, *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

[17] K. Glazebrook and D. Wilkinson, "Index-based policies for discounted multi-armed bandits on parallel machines," *Annals of Applied Probability*, pp. 877–896, 2000.

[18] R. Goetschalckx, P. Poupart, and J. Hoey, "Continuous correlated beta processes." Citeseer.

[19] 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.

[20] K. Hang, F. T. Pokorny, and D. Kragic, "Friction coefficients and grasp synthesis," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Tokyo, Japan, 2013. [Online]. Available: http://www.csc.kth.se/~fpokorny/static/publications/hang2013a.pdf

[21] 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.

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

[23] 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.

[24] 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.

[25] ——, "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.

[26] B. Kehoe, S. Patil, P. Abbeel, and K. Goldberg, "A survey of research on cloud robotics and automation," 2015.

[27] F. Kelly *et al.*, "Multi-armed bandits with discount factor near one: The bernoulli case," *The Annals of Statistics*, vol. 9, no. 5, pp. 987–1001, 1981.

[28] 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.

[29] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.

[30] 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.

[31] 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."

[32] 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.

[33] 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.

[34] C. Rasmussen and C. Williams, *Gaussian processes for machine learning*. MIT Press, 2006.

[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] N. Srinivas, A. Krause, S. M. Kakade, and M. Seeger, "Gaussian process optimization in the bandit setting: No regret and experimental design," *arXiv preprint arXiv:0912.3995*, 2009.

[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] 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.