Budgeted Multi-Armed Bandit Models for Sample-Based Grasp Planning in the Presence of 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*— Sampling perturbations in shape, state, and control can facilitate grasp planning in the presence of uncertainty arising from noise, occlusions, and surface properties such as transparency and specularities. Monte-Carlo sampling is computationally demanding, even for planar models. We consider an alternative based on the multi-armed bandit (MAB) model for making sequential decisions, which can apply to a variety of uncertainty models. We formulate grasp planning as a "budgeted multi-armed bandit model" (BMAB) with finite stopping time to minimize "simple regret", the difference between the expected quality of the best grasp and the expected quality of the grasp evaluated at the stopping time. To evaluate MAB-based sampling, we compare it with Monte-Carlo sampling for grasping an uncertain planar object defined by a Gaussian process implicit surface (GPIS), but the method is applicable to other models of uncertainty. We derive distributions on contact points, surface normal, and center of mass and use these solve the associated MAB model, finding that it computes grasps of similar quality and can reduce computation time by an order of magnitude. This suggests a number of new research questions about how MAB can be applied to other models of uncertainty and how different MAB solution techniques can be applied to further reduce computation.**

I. Introduction

Consider a robot packing boxes in a shipping warehouse environment, where it may frequently encounter new consumer products and need to process them quickly. The robot may need to rapidly plan grasps for these objects without prior knowledge of their shape, pose and even material properties like friction coefficient or center of mass. Due to sensor noise and missing data due to partial visibility and object properties such as transparency, the robot may not be able to measure these quantities exactly. Grasp planners that assume exact prior knowledge of object geometry or an exact measurement of pose may fail in this environment.

Grasp quality metrics have been developed to determine if a grasp will be successful or not and how much force it needs to exert to resist an opposing force, however most of them evaluate a grasp assuming all the parameters are known [10]. This motivates using knowledge of uncertainty to select grasps, but most methods for evaluating the quality of a single grasp in the presence of uncertainty require use of an exhaustive sampling over the possible values of the uncertain quantity [17], [36]. To select a grasp with high quality this evaluation is often performed for a large set of potential grasps, which can be very time-consuming. However, when evaluating a set of grasps we may be able to determine the difference of quality between grasps with only a few samples and throw away grasps that are likely to be suboptimal [15]. Thus, we can adaptively concentrate grasp quality evaluation on the grasps that are most likely to have the highest quality based on the evaluation done so far.

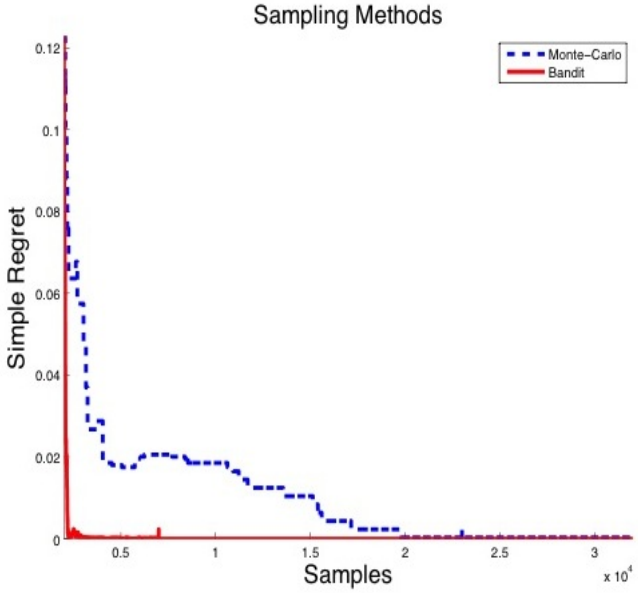

Fig. 1: Convergance in regret of the bandit sampling method (red), compared to the traditional Monte-Carlo method. The fast convergence of the bandit method is due to its ability to intelligently pick what grasp to sample next in a given set of proposed grasps on an object with shape uncertainty

The multi-armed bandit (MAB) model for sequential decision making problems [3], [19], [**?**] provides a way to reason about selecting the next grasp to evaluate and the grasps to discard from consideration. The goal in a MAB model is to make a sequence of decisions over a set of possible options such that a measure of the expected reward of such decisions is maximized. Solutions to the MAB model are particularly useful in applications where it is too expensive to fully evaluate a set of options; for example, in optimal design of clinical trials [31], market pricing [30], and choosing strategies for games [33]. The budgeted multi-armed bandit

[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

model [21] is a specialization of the MAB model with a finite stopping time where the objective is to maximize the expected reward of the decision made at the stopping time, or equivalently to minimize "simple regret", which is the difference between the true expected reward of an optimal arm and the true expected reward of the arm pulled at the stopping time.

Our main contribution in this paper is formulating the problem of planning grasps with a high expected quality in the presence of uncertainty as a budgeted multi-armed bandit model. We use this formulation to rank a set of potential grasps by expected Ferrari-Canny quality [10] under shape uncertainty. We use a budgeted multi-armed bandit model since we would like to execute only one grasp plan after evaluating the expected quality of many potential grasp plans. We choose the model of uncertainty to be a Gaussian process implicit surface (GPIS), a Bayesian representation of shape uncertainty that seen recent use in various robotic applications [9], [13], however the method applies to any model of uncertainty that can be sampled.

We also show how to estimate distributions on the contact points and surface normals and center of mass using a GPIS shape representation. We then leverage these distributions to reduce the computational complexity of sampling from the GPIS model from $O(n^6)$ to $O(n^3)$, where the workspace is discretized as a $n \times n$ grid. Our experiments demonstrate that using the MAB sampling method improves the time to rank a set of 1000 grasps by 10x, an order of magnitude improvement, over the baseline Monte-Carlo approach. These promising results suggest that our MAB approach could be used on other types of uncertainty, such as pose, friction coefficent or center of mass and provide significant speed improvements.

II. Related Work

Past work on grasping under uncertainty has considered state uncertainty [12], [34], uncertainty in contact locations with an object [38], uncertainty in object pose [8], [36], [17]. The effect of uncertainty in object geometry on grasp selection has been studied for spline representations of objects [8], extruded polygonal mesh models [15], [16], and point clouds [14].

Currently, the most common method of evaluating the expected grasp quality under uncertainty is to rank a set of random grasps on an object using samples on shapes, pose or parameters to evaluate a quality measure [8], [15], [16]. Monte-Carlo sampling involves drawing random samples from a distribution to approximate an expected value[6], which can be slow when the distribution is high-dimensional, such as for distributions on possible shapes. To address this, Kehoe et al. [15] demonstrated a procedure for finding a minimum bound on expected grasp quality given shape uncertainty, which reduced the number of terms needed in Monte-Carlo sampling in order to choose the highest quality grasps. The adaptive sampling pruned grasps using only the sample mean and did not utilize any estimates of how accurate the current sample mean is. Laaksonen et al. [18] used Markov Chain Monte-Carlo (MCMC) sampling to estimate grasp quality and object pose under shape and pose uncertainty. MCMC simplified sampling from a complicated joint distribution on pose and shape, but it can be slow to converge to the correct distribution due to burn in and mixing conditions [1].

We chose to study our MAB sampling method for shape uncertainty using a Gaussian process implicit surface representation. Our decision to use this uncertainty model is based on GPIS's ability to combine various modes of noise observations such as tactile, laser and visual [27], [37], [9] and its recent use in modeling uncertainty for a number of robotic applications. Hollinger et al. used GPIS as a model of uncertainty and performed active sensing on the hulls in underwater boats [13]. Dragiev et al. showed how GPIS can enable a grasp controller on the continuous signed distance function [9]. Mahler et al. used the GPIS representation to find locally optimal anti-podal grasps by framing grasp planning as an optimization problem [**?**].

III. Multi-Armed Bandit Model

The multi-armed bandit model, originally described by Robbins [29], is a statistical decision model of an agent trying to make correct decisions, while gathering information at the same time. The traditional setting of a multi-armed bandit model is a gambler that has K independent slot machine arms and decides what machines to play, how many times to play each one, what order to play them in. A successful gambler would want to exploit the machine that currently yields the highest reward and explore new arms to see if they give better rewards. Developing a policy that successfully trades between exploration and exploitation has been the focus of extensive research, since the problem formulation [5], [**?**], [4]. 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 [31], market pricing [30], and choosing strategies for games [33].

Traditional bandit algorithms minimize cumulative regret, which is the sum over each sub-optimal arm of the number of times that arm is pulled times the difference between the true expected reward of an optimal arm and the true expected reward of that sub-optimal arm. There are a number of algorithms for developing policies to balance exploration and exploitation. One algorithm is $\epsilon-$greedy, which is the idea of choosing the arm with the highest empirical expected reward with $1-\epsilon$ probability and choosing a random arm with probability ϵ [3]. A class of algorithms that have stronger theoretical guarantees are from the Upper Confidence Bound (UCB) family. UCB algorithms maintain the empirical expected reward based off of pulling each arm multiple times, while also estimating an upper bound on the true expected reward using assumptions on the probability distribution of rewards and number of times each arm has been sampled.

Different to the normal bandit problem, we do not actually care about cumulative regret because the exploration and exploitation stage is decoupled from each other. For example

in medical testing for patients it is important to always administer the best treatment for the given patient. However for the case of product testing in cosmetics it does not matter what products are administer in the testing phase, but only what final product is sold to the consumer. Thus, performance of Budgeted Multi-Armed Bandit models is measured in terms of the convergence of simple regret. Given a set of arms $\{1, ...K\}$ with respective mean rewards $\mu_1, ..., \mu_K$ and the optimal arm $\mu^* = \max_{k \in \{1,...,K\}} \mu_k$. The regret at time t is given by Eq. 1, where μ_t is the estimate of the best arm at time t from the previous observations.

$$r_t = \mu^* - \mu_t \tag{1}$$

Best arm identification has a wide variety of literature that largely falls into two camps: one where the algorithm terminates once a fixed confidence interval around the best arm is met and one where a set horizon is given and the algorithm has to choose a arm at that stopping time.

A. Fixed Confidence

In the fixed confidence setting the forecaster seeks to minimize the simple regret until a fixed confidence threshold is met at which point it terminates. Originally the problem was solved by racing algorithms, were methods used Hoeffding inequalities or the empirical Bernstein inequality to prune arms that were deemed not likely to succeed and used uniform allocation to explore the remaining set [22] [24]. These methods were later extended to include not only the best arm, but return the top m arms [11].

B. Fixed Budget

In the fixed budget setting the algorithm is given a stopping time and needs to return the best arm at that stopping time. Audibert el al. demonstrated an algorithm called Successive Rejects that given a budget divides up the time each arm is pull an can return the best arm with near-optimal probability depending on the hardness of the problem [2]. Other UCB like methods have been proposed as well that measure a confidence gap and then pull the arm with the highest confidence interval [11]. In [5], they showed a link between simple regret and cumulative regret that allowed for the analysis of the existing bandit algorithms like UCB1.

For the case of choosing the best grasp, we are working in the scenario where the robot has a fix amount of time to choose a grasp and thus the fixed budget setting is more appropriate.

IV. Bandit Algorithms For Best Arm Identification

A. Thompson Sampling

Thompson Sampling is a Bayesian method for the multi-armed bandit problem. The main idea behind it is to update the conjugate prior for the distribution of the arm recently pulled. Then sample from the prior distribution on all arms and pull the arm with the highest sample drawn. The full algorithm is shown in Algorithm 1. Empirically this method has been shown to outperform frequentist methods like UCB [7]

Algorithm 1: Thompson Sampling for Beta-Bernoulli Process

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

B. Upper Confidence Bound (UCB)

The UCB strategy was the first strategy shown to have asymptotic logarithmic regret [19] for distributions with bounded rewards. UCB estimates an empirical confidence bound for each arm based on the observations seen so far. Then it pulls the arm with the highest confidence bound. The confidence interval is derived from various concentration inequalities and assumes the reward function is bounded. The algorithm is in Algorithm 2 and assumes a Beta-Bernoulli Distribution, which is the case in our grasping.

Algorithm 2: UCB for Beta-Bernoulli Process

Result: Best Arm, g^*
for *t=1,2,...* **do**
  Pick arm: $k = \underset{k \in 1,...,K}{\operatorname{argmax}} \mu_k + \sqrt{\frac{6\log(t)}{T_k}}$
  Observe reward $X_{k_t,t} \in \{0, 1\}$
  Update
  $T_k = T_k + 1$
  $\mu_k = \frac{T_k - 1}{T_k} * \mu_k + \frac{1}{T_k} * X_{k,t}$

C. The Gittins Index Method

One possible solution to solve the MAB problem is to treat is as an Markov Decision Process (MDP) and use Markov Decision theory. This solution makes a lot of sense when the distribution is known because the all elements in the standard MDP tuple, $\{S, A, T, R, \gamma\}$, would be known and it is optimal with respect γ.

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

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

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

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

Result: Best Arm, g^*

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

V. Preliminaries and Problem Definition

Before we present the problem definition, we introduce a way to evaluate the quality of a grasp and our grasping model, the line of action.

A. Grasp Metric

In their work over two decades ago Ferrari and Canny [10], demonstrated a method to rank grasps by considering their contact points and surface normals. Importantly the magnitude of Q yields a measurement that allows one to rank grasps by their physical stability and evaluate the property of force-closure. Furthermore, it has wide spread use in grasp packages like GraspIT[23], OpenGrasp[20] and Simox [35], which motivates studying its effect with uncertainties.

The L^1 version of the metric works by taking as input the contact points $\mathbf{c}_1, ..., \mathbf{c}_m \in \mathcal{R}^2$, surface normals $\mathbf{n}_1, ..., \mathbf{n}_m \in \mathcal{R}^2$, center of mass $\mathbf{z}$ and friction coefficient μ. Then constructing a convex hull around the wrenches made up of those parameters and finding the radius of the largest unit ball centered at the origin in wrench space. A wrench is defined as concatenation of a force and torque vector. If the convex hull does not enclose the origin, the grasp is not in force-closure. Formally, not being in force closure means there exists a wrench for which the grasp is not able to resist against, which occurs if the Quality metric Q is less than zero. Thus a grasp can be parameterized by the following tuple $g = (\mathbf{c}_1, ..., \mathbf{c}_m, \mathbf{n}_1, ..., \mathbf{n}_m, \mu, \mathbf{z})$.

Since we are in an uncertain environment, we are interested in calculating the probability of achieving force closure, or $P(Q > 0)$, similar to [8][16] . $P(Q > 0)$ is computed by sampling from our distributions on pose, shape and material properties (friction coefficient and center of mass) and averaging the qualities that are computed. Reducing the time to find the best expected grasp from a set of a large number of grasps is the primary focus of the paper.

B. Line of action

In an uncertain environment one would not actually know the true g because the actual grasp that is executed is non-deterministic. Thus, we have to work with the trajectory of the gripper. Similar to the work of [8], we assume that each gripper finger approaches along a *line of action*, a 1D curve $\gamma(t)$ with endpoints a and b as seen in Fig. 3. A gripper finger starts at point a and moves towards b, we assume a is far enough away to be collision free of the object. Each gripper contact is defined by a line of action, so we assume the following tuple is provided $\Gamma = (\gamma_1(\cdot), ..., \gamma_m(\cdot))$, which designates a proposed *grasp plan*.

C. Types of Uncertainty

In this work we consider the following types of uncertainty shape, pose, movement and friction coefficient. The property of force closure is not effected, by the location of center of mass because it only measures if a wrench can be resisted not how large the force is need to resist. Thus we calculate the expected The probabilistic relationship of all these types of uncertainty can be seen in Fig. 2. We model each one of these with a corresponding distribution and demonstrate how MAB algorithms can handle all types of uncertainty.

1) Distribution on Shape: For shape we use a GPIS representation, which is explained in detail in Sec. XIII. Essenttialy, a GPIS is a Gaussian distribution on a signed distance function that describes the workspace. The GPIS can be sampled from vian $N(\mu(x), \Sigma(x))$, where $\mu(x)$ and $\Sigma(x)$ are the mean and covariance functions of the GPIS. For our application we set x to discretized points along each line of action γ_i and for each sample see where the sample is equal to zero. We will furthermore represent the $\mu(x)$ and $\Sigma(x)$ as the tuple $\theta = \}\mu(x), \Sigma(x)\{$.

2) Distribution on Pose: [TODO: JEFF CAN YOU WRITE ABOUT POSE]

3) Distribution on Movement: In practice a robot may not be able to execute a desired grasp plan Γ exactly due to errors in trajectory following or registration to the object [15]. To handle this uncertainity, we sample the angle of approach ρ of the proposed grasp plan Γ from a zero mean one-dimensional Guassian $\rho \sim N(0, \sigma^2)$. In practice σ^2 can be set from repeatibility measurements for a robot [**?**].

4) Distribution on Friction Coefficient: As shown in [38], uncertainty in friction coefficient can play a large role in grasp quality evaluation. The expected friction coefficient $E(\mu)$ can be derived by means of object classification and a look up table. However, because their could be material in between the objects surface and the robot gripper (i.e. dust,

water, moisture). It is important to assume noise. Thus, we purpose sampling the friction coefficient from a Gaussian around the expected friction coefficient $\mu \sim N(E(\mu), \sigma_\mu^2)$.

Thus to evaluate the $P(Q(\Gamma) > 0)$, we sample from the graphical model shown in Fig. 2. In the naive way sampling from the graphical model could be slow, however our goal is not to actually determine the probability of force closure but instead to determine the best grasp in a set G. Thus, our BMAB algorithms can efficiently sample from theses priors on the different types of uncertainty.

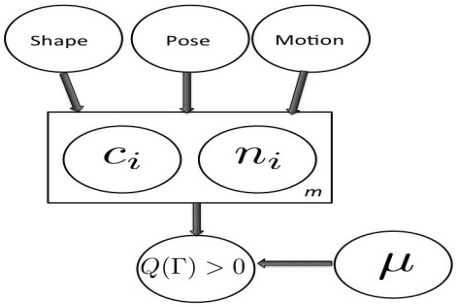

Fig. 2: A graphical model that illustrates the relationship between the different types of uncertainty in an object. Center of Mass uncertainty is dependent on the pose and shape of the object, however friction coefficient is independent of all other types.

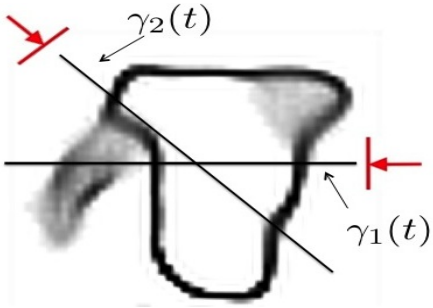

Fig. 3: Illustration of a grasp plan Γ composed of two lines of action, $\gamma_1(t)$ and $\gamma_2(t)$

D. Problem Definition

Given a 2-D workspace $\mathcal{W}$, with an unknown object represented as a trained GPIS model, which we describe in Section XIII-A and set of possible grasp plans G which are generated either randomly or with a heuristic as in [**?**]. We are interested in determining Eq. 3 with respect to a chosen grasp metric Q.

$$\Gamma^* \in \underset{\Gamma \in G}{\text{argmax}} P(Q(\Gamma) > 0) \tag{3}$$

VI. Distribution of Grasp Parameters

To sample from $p(Q(\Gamma) > 0)$, we need to sample from the distributions associated with a line of action $p(\mathbf{n}_i, \mathbf{c}_i | \gamma_i(t), \theta, \rho)$. Using Bayes rule and our graphical model we can rewrite this as

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

In section XIII-C, we look at how to sample from $p(\mathbf{c}_i | \gamma_i(t), \mu(x), \Sigma(x))$ from a GPIS model. Then in section VI-A, we look at how to sample from $p(\mathbf{n}_i | \mathbf{c}_i, \gamma_i(t), \theta, \rho)$ and present a novel visualization technique for the distribution on surface normals. Lastly in section VI-B, we show a way to calculate the expected center of mass assuming a uniform mass distribution. .

A. Distribution on Surface Normals

Using Eq. 9 and Eq. 10, we can compute the mean of the gradient $\mu_\nabla(x)$ and the covariance of the gradient $\Sigma_\nabla(x)$ respectively. Thus we can compute the distribution around the surface normal for a given point in $\mathcal{W}$. We can now write

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

$$p(\mathbf{n}_i) = \int_a^b p\big(\mathbf{n}_i = \mathbf{v} | \mathbf{c}_i = \gamma(t)\big) p\big(\mathbf{c}_i = \gamma(t)\big) dt \tag{4}$$

Grasp metrics such as Ferrari-Canny require $\mathbf{n}_i$ be normalized, or, equivalently, a member of the sphere $\mathcal{S}^{d-1}$ [10]. To account for this we densely sample from the distribution $p\big(\mathbf{n}_i)\big)$ and project onto $\mathcal{S}^{d-1}$. In Fig.9, we visualize the distribution on $\mathbf{n}_i$ calculated for a given GPIS and approach line of action.

B. Expected Center of Mass

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

$$\bar{z} = \frac{\int_{\mathcal{W}} x P(sd(x) < 0) dx}{\int_{\mathcal{W}} P(sd(x) < 0) dx} \tag{5}$$

which can be approximated by sampling $\mathcal{W}$ in a grid and approximating the spatial integral by a sum. Since this operation involves the entire SDF, one would want to use a low resolution grid for computational efficiency. We show the computed density and calculated expected center of mass for a marker in Fig. 4.

VII. Multi-Armed Bandits for Grasp Selection

While a standard approach to solving the problem in Eq. 3 would be to perform Monte-Carlo integration on each Γ_i and compute the expected grasp quality, we propose treating the problem as a multi-armed bandit model and forming a policy for selecting which grasp to sample. In our setting, we have a probabilistic shape representation and would like to evaluate many potential grasps on that shape model. Motivated by limited computational resources we are interested in how to intelligently allocate sampling resources to efficiently find

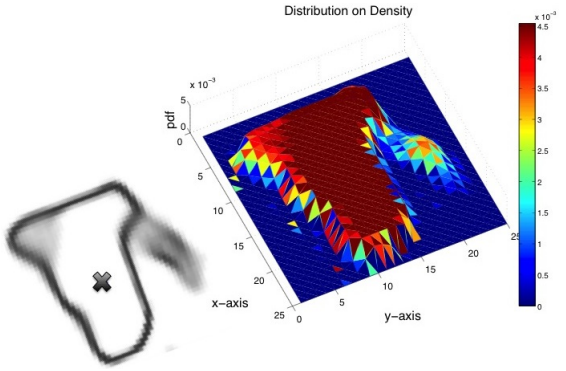

Fig. 4: Left: A surface with GPIS construction and expected center of mass (black X) Right: The distribution on the density of each point assuming uniform density

the best grasp plan Γ^*. Here each arm corresponds to a different grasp plan and pulling the arm is sampling from the graphical model in Fig. 2 and evaluating the arm's grasp plan on the sample. The reward for pulling an arm is 1 if the grasp is in force closure on the sample and 0 if the grasp is not in force closure.

We can now model the distribution on rewards as a Beta-Bernoulli process. Bernoulli distributions are formally the probability θ that an event has occurred and a Beta distribution is the conjugate prior on the probability of theta or $p(\theta)$. Beta distributions are specified by shape parameters α and β, to update the prior Beta distribution one simply adds the count of observed successes of the event to α and the count of the observed failures to β. The default $\alpha = 1$ and $\beta = 1$ corresponds to a uniform distribution on θ.

Since, we have framed the method as a Beta-Bernoulli problem we can use the methods in Sec.IV and obtain guarantees in the convergence properties of regret and overall performance of our policy for grasp allocation.

VIII. Calculating the Probability of Force Closure

Given a proposed grasp plan Γ , we draw samples from the shape distribution $P(S)$, the distribution on center of mass $P(z)$ and the distribution on friction coefficient $P(\mu)$. The distribution on force closure can then be estimated as Beta-Bernoulli Process with shape parameters α and β. Thus, we can write the probability of force closure as follows

$$P(Q(\Gamma|S,\mu,z) > 0) = \frac{\alpha}{\alpha+\beta} \tag{6}$$

Where $Q(\Gamma|S,\mu,z)$ is the grasp quality that is computed on a shape sample drawn from $p(S)$,$p(z)$,$p(\mu)$. To compute this we intersect the zero crossing of the level set with the propose grasp plan Γ and determine the parameters g, this has been the approach taken in previous work [15], [16], [8]. See Fig. 8 for an example of what samples drawn from $p(S)$ induced by GPIS look like. For computational reasons we approximate the integral via Monte-Carlo Integration. We use importance sampling to draw from the distribution induced by GPIS and count the number of sucesses of the grasp S and the number of failures F and calculate the following

$$P(Q(\Gamma|S,\mu,z) > 0) \approx \frac{(S+\alpha_0)}{(\alpha_0+S)+(\beta_0+F)} \tag{7}$$

Here α_0 and β_0 correspond to the initial shape parameters, we initialize them both to $\alpha_0 = \beta_0 = 1$ to reflect a uniform distribution on the probability of force closure .

IX. Experiments

For the experiments below we used common household objects. The objects used can be found at `http://rll.berkeley.edu/grasping/` We manually created a 25 x 25 grid, by tracing a point cloud of the object on a table taken with a Primesense Carmine depth sensor. To accompany the SDF, we created an 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. Our visualization technique follows the approach of [**?**] and consisted of drawing many shape samples from the distribution and blurring accordingly to a histogram equalization scheme.

We did experiments for the case of two hard contacts in 2-D, however our methods are not limited to this implementation. We drew random lines of actions $\gamma_1(t)$ and $\gamma_2(t)$ by sampling around a circle with radius $\sqrt{2}n$ and sampling the circles origin, then projecting onto the largest inscribing circle in the workspace.

A. Multi-Armed Bandit Experiments

We consider the problem of selecting the best grasp plan, Γ^* out of a set G. For our experiments we look at selecting the best grasp out of a size of $|G| = 1000$. In Fig. 5, we plotted the simple regret averaged over 100 of the shapes in our data set and compare the different methods (UCB, Thompson, Gittins and the naive random allocation). We initialize both the Monte-Carlo and bandit technique by sampling each grasp 1 time. We draw samples from our calculated distributions $p(g)$. Interestingly, Gittins and Thompson converge much faster than random and UCB. In Fig. 6, you can see that Gittins and Thompson allocate grasp samples to only the grasps of high quality, thus it is quickly able to ignore the low quality grasps. UCB takes a more conservative approach to sample allocation, which leads to poor performance in the best arm identification problem [5].

B. Sampling from Grasps vs. Shape

We first tested 1000 grasp plans and sampled each one 5000 times and measured the RMS error between converged expected grasp plan qualities for sampling shape Eq. 12 vs. grasps Eq. **??** was 0.004. After confirming the distributions converged close to the same value, we show the computational complexity in Fig. 7 of the two methods for evaluating 100 grasps on an $n \times n$ grid.

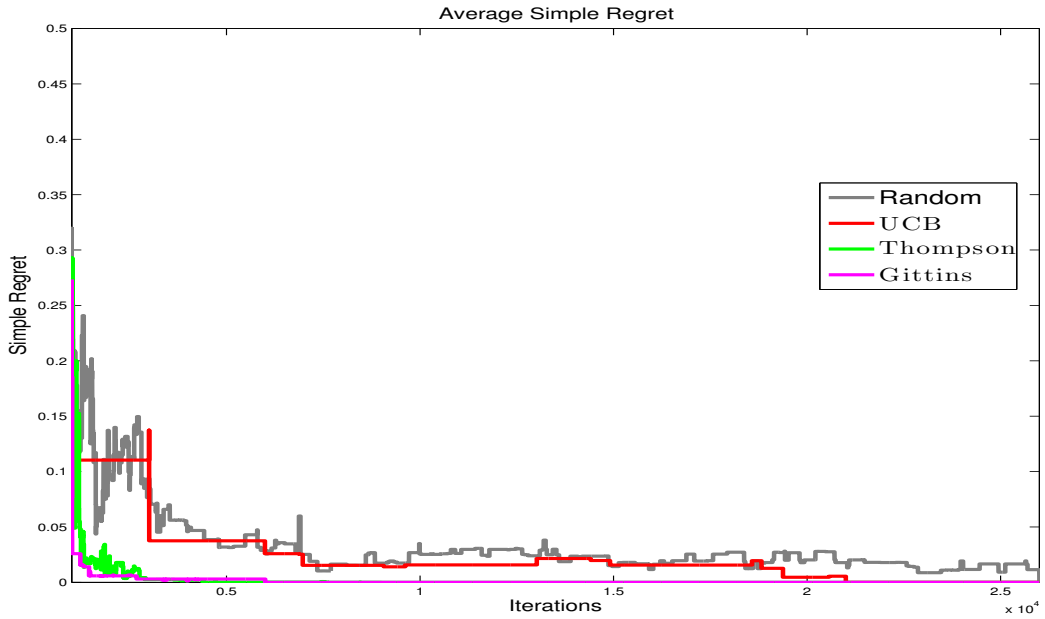

Fig. 5: Comparison of Simple Regret convergence for the four sequential decision methods (Random, UCB, Thompson, Gittins). Graph is averaged over 100 shapes from the Brown Sillohoute Dataset [**?**] with a set $|G| = 1000$ for each shape. As you can see the Thompson and Gittins method converge almost a magnitude faster than random allocation. UCB does poorly on simple regret which is expected [**?**]

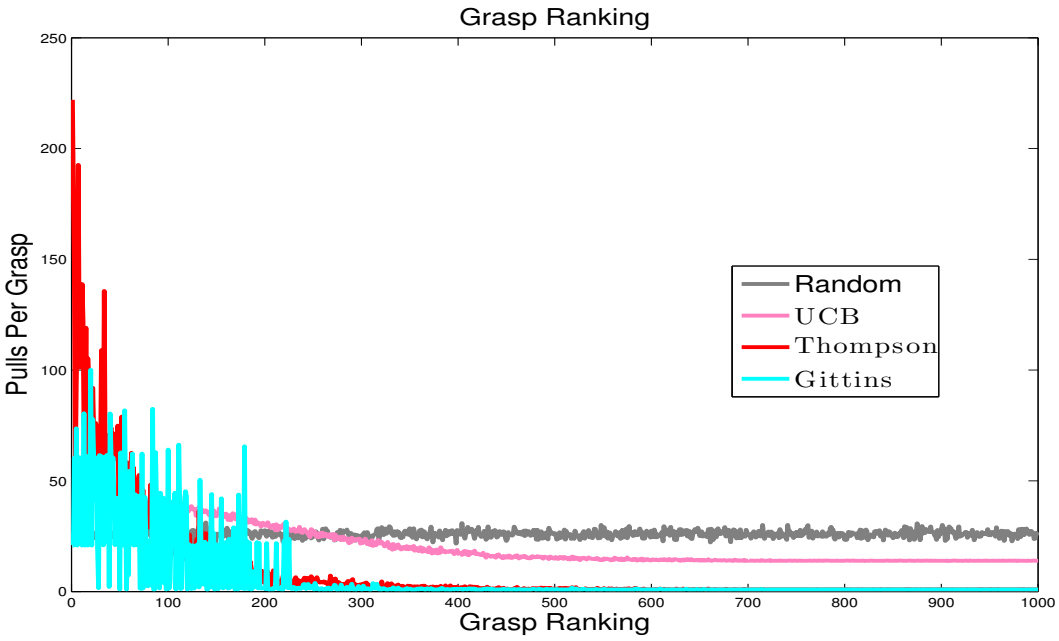

Fig. 6: Comparison of sample per grasp for the four sequential decision methods (Random, UCB, Thompson, Gittins). Graph is averaged over 100 shapes from the Brown Silhouette Dataset [**?**] with a set $|G| = 1000$ for each shape. The best grasps are ranked 1 and worst are 1000. As you can see the Thompson and Gittins method have a tendency to pull grasps with high ranking, while UCB has a more conservative policy [**?**]

X. Limitations

Our budgeted multi-armed bandit approach appears promising, but we still do not know how well it will perform on 3D shapes and large scale grids. Future work will be building an efficient construction of GPIS to scale to 3D and test the bandit method there. While we have seen the bandit method always converge to the correct value in our experiments, our method only has a statistical guarantee of doing so. A non-optimal grasp plan could be found, albeit with a low probability.

Sampling from our distribution $p(g)$ over $p(S)$ yields a reduction in computationally complexity, but only if the number of grasps one wants to evaluate remains small relative to n^3, techniques to ensure this could be to find locally optimal

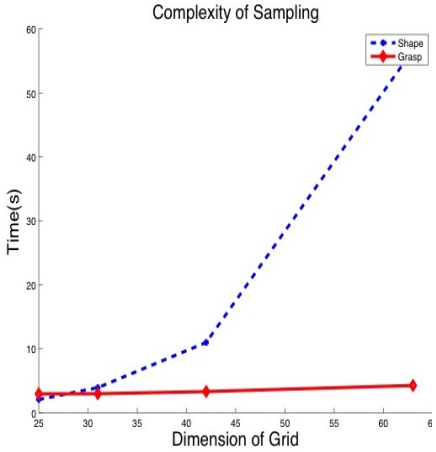

Fig. 7: Time it took to sample from 100 grasp distributions for a given resulution of the workspace. Blue line is sampling from $p(sd(\mathcal{R}))$ or shapes and Red is sampling from $p(g)$ or the calculated distribution on grasps. As you can see sampling from the calculated distributions scales much better.

potential grasps using optimization approaches [**?**].

An additional problem is that we only have an expected center of mass and not a distribution on the center of mass. This might prove to be to expensive to compute, however recent work by Panahi et al. showed a way to bound the center of mass for convex parts. Extension of his work to implicit surfaces could be of possible interest [25]. When using the shape sampling approach on the GPIS instead of sampling along the grasp plan trajectory, the center of mass is easy to compute on a sampled shape, so this is a limitation of our sampling method.

XI. Conclusion

Assessing grasp quality under shape 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 shape uncertainty. A key insight from our work is that uniformly allocating samples to grasps is inefficient, and we found that the Successive Elimination multi-armed bandit approach prioritizes evaluation of high-quality grasps while quickly pruning-out obviously poor grasps. A pre-requisite for applying a bandit approach is to formulate an efficient representation of how shape uncertainty affects grasp parameters and thus grasp quality. We modeled uncertainty with Gaussian process implicit surfaces (GPIS) and derived the distribution of grasp parameters when a nominal grasp is applied to the GPIS. As a result, we were able to more efficiently sample from a distribution of grasps executions rather than the shape; leading to a complexity improvement of n^3 in the resolution of the discretization n. We evaluated this theoretical model on a dataset of common objects and confirmed that: (1) the bandits approach always converged to the best grasp in the candidate set, (2) it converges on average an order of magnitude faster than a uniform sampling approach in our experiments, and (3) sampling from the grasp contacts is 12x faster than sampling from shapes for a 64x64 grid.

XII. Future Work

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

In principle, our method can be applied to other representations of shape uncertainty such as perturbations on polygonal vertices [15] or splines [8]. It can further be applied to other grasp quality metrics or simulation based evaluation methods [20].

Future work will also consider applying BMAB approach to grasp planners like GraspIt! [23] to see if our method can handle uncertainty while working under the time constraints needed for most real time applications. We currently have only tested one bandit algorithm, successive elimination. While our results are promising, it remains to be seen how well it deals with the increased complexity of 3D models over 2D models and larger scale experiments. However, the BMAB model has a large amount of literature to draw from as we encounter new and more challenging problems [4].

References

[1] 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.

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

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

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

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

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

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

[8] 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.

[9] 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.

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

[11] 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.

[12] 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.

[13] 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.

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

[15] 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.

[16] ——, "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.
[17] 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.
[18] 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.
[19] T. L. Lai and H. Robbins, "Asymptotically efficient adaptive allocation rules," *Advances in applied mathematics*, vol. 6, no. 1, pp. 4–22, 1985.
[20] 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.
[21] O. Madani, D. J. Lizotte, and R. Greiner, "The budgeted multi-armed bandit problem," in *Learning Theory*. Springer, 2004, pp. 643–645.
[22] O. Maron and A. W. Moore, "Hoeffding races: Accelerating model selection search for classification and function approximation," *Robotics Institute*, p. 263, 1993.
[23] 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.
[24] 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.
[25] F. Panahi and A. F. van der Stappen, "Bounding the locus of the center of mass for a part with shape variation," *Computational Geometry*, vol. 47, no. 8, pp. 847–855, 2014.
[26] K. B. Petersen, "The matrix cookbook."
[27] C. Rasmussen and C. Williams, *Gaussian processes for machine learning*. MIT Press, 2006.
[28] 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.
[29] H. Robbins, "Some aspects of the sequential design of experiments," in *Herbert Robbins Selected Papers*. Springer, 1985, pp. 169–177.
[30] M. Rothschild, "A two-armed bandit theory of market pricing," *Journal of Economic Theory*, vol. 9, no. 2, pp. 185–202, 1974.
[31] R. Simon, "Optimal two-stage designs for phase ii clinical trials," *Controlled clinical trials*, vol. 10, no. 1, pp. 1–10, 1989.
[32] 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.
[33] 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.
[34] 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.
[35] N. Vahrenkamp, T. Asfour, and R. Dillmann, "Simo: A simulation and motion planning toolbo for c+."
[36] 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.
[37] O. Williams and A. Fitzgibbon, "Gaussian process implicit surfaces," *Gaussian Proc. in Practice*, 2007.
[38] 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.

XIII. Appendix

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

A. Gaussian Process (GP) Background

We refer the reader to [**?**] for a more detailed explanation of the GP construction, which we summarize here. Given the training data $\mathcal{D} = \{\mathcal{X}, \mathbf{y}\}$ and covariance function $k(\cdot,\cdot)$, the posterior density $p(sd_*|\mathbf{x}_*, \mathcal{D})$, or the distribution on signed distance field, at a test point $\mathbf{x}_*$ is shown to be [28]:

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

where $K \in \mathbb{R}^{l \times l}$ is a matrix with entries $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ and $k(\mathcal{X}, \mathbf{x}_*) = [k(\mathbf{x}_1, \mathbf{x}_*), \ldots, k(\mathbf{x}_l, \mathbf{x}_*)]^\intercal$. This derivation can also be used to predict the mean and variance of the function gradient by extending the kernel matrices using the identities [32]:

$$\operatorname{cov}\left(sd(\mathbf{x}_i), sd(\mathbf{x}_j)\right) = k(\mathbf{x}_i, \mathbf{x}_j) \tag{8}$$

$$\operatorname{cov}\left(\frac{\partial sd(\mathbf{x}_i)}{\partial x_k}, sd(\mathbf{x}_j)\right) = \frac{\partial}{\partial x_k} k(\mathbf{x}_i, \mathbf{x}_j) \tag{9}$$

$$\operatorname{cov}\left(\frac{\partial sd(\mathbf{x}_i)}{\partial x_k}, \frac{\partial sd(\mathbf{x}_j)}{\partial x_l}\right) = \frac{\partial^2}{\partial x_k \partial x_l} k(\mathbf{x}_i, \mathbf{x}_j) \tag{10}$$

For our kernel choice we decided to use the square exponential kernel, similar to [9]. Other kernels relevant to GPIS are the thin-plate splines kernel and the Matern kernel [37].

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

B. Calculating the Probability of Force Closure

Given a proposed grasp plan Γ , we draw samples from the shape distribution $P(S)$, the distribution on center of mass $P(z)$ and the distribution on friction coefficient $P(\mu)$. The distribution on force closure can then be estimated as Beta-Bernoulli Process with shape parameters α and β. Thus, we can write the probability of force closure as follows

$$P(Q(\Gamma|S, \mu, z) > 0) = \frac{\alpha}{\alpha + \beta} \tag{11}$$

Where $Q(\Gamma|S, \mu, z)$ is the grasp quality that is computed on a shape sample drawn from $p(S)$,$p(z)$,$p(\mu)$. To compute this we intersect the zero crossing of the level set with the propose grasp plan Γ and determine the parameters g, this has been the approach taken in previous work [15], [16], [8]. See Fig. 8 for an example of what samples drawn from $p(S)$ induced by GPIS look like. For computational reasons we approximate the integral via Monte-Carlo Integration. We use importance sampling to draw from the distribution induced by GPIS and count the number of sucesses of the grasp S and the number of failures F and calculate the following

$$P(Q(\Gamma|S, \mu, z) > 0) \approx \frac{(S + \alpha_0)}{(\alpha_0 + S) + (\beta_0 + F)} \tag{12}$$

Here α_0 and β_0 correspond to the initial shape parameters, we initialize them both to $\alpha_0 = \beta_0 = 1$ to reflect a uniform distribution on the probability of force closure . To compute the above distribution we must draw samples from $p(S)$. In order to draw shape samples from a GPIS, one needs to sample from signed distance function, sd, over the joint on all points in the workspace $\mathcal{W}$ or $p(sd(\mathcal{W}))$. Since this is a GPIS, we know the following

$$p(S) = p(sd(\mathcal{W})) \sim N(\mu(\mathcal{W}), \Sigma(\mathcal{W})) \tag{13}$$

Thus if the workspace is an $n \times n$ grid, the joint distribution is an n^2 multi-variate Gaussian, due to $sd : \mathbb{R}^2 \rightarrow \mathbb{R}$. Sampling from a Gaussian involves inverting the covariance matrix and inversion is in the naive way scales with the cube of the number of input dimensions [26]. Thus the complexity of this operation is $O(n^6)$ in 2D and $O(n^9)$ in 3D.

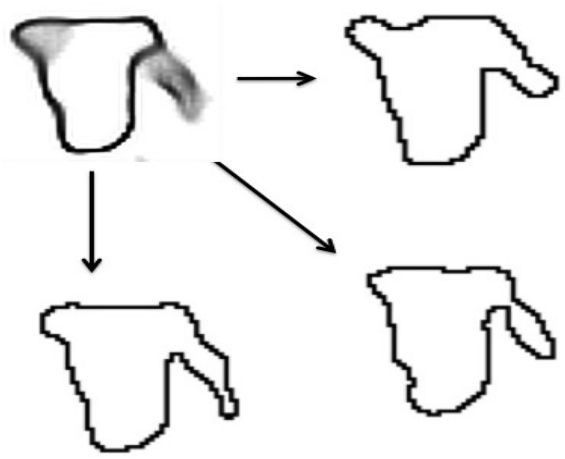

Fig. 8: Shape samples drawn from Eq. 13 on the object in the upper left corner. Given a shape sample we highlight the zero-crossing of the level set in black

To reduce complexity we propose sampling not from the shape distributions, but instead from the distributions on the grasps parameters themselves. We recall that a grasp according to our metric is defined as the tuple $g = (\mathbf{c}_1, ..., \mathbf{c}_m, \mathbf{n}_1, ..., \mathbf{n}_m, \mu, z)$. We are thus interested in calculating $p(g|\Gamma, \mu(x), \Sigma(x))$. The distribution on a grasp is defined then as:

$$p(g) = p\big(\mathbf{c}_1, ..., \mathbf{c}_m, \mathbf{n}_1, ..., \mathbf{n}_m | \Gamma, \mu(x), \Sigma(x)\big) \tag{14}$$

We note here that we currently use the friction coefficient μ and the expected center of mass $\bar{z}$ as deterministic values. For grippers that do not approach along the same line of action (i.e. non-parallel jaw grippers) we make the assumption that each contact and normal pair is independent, or

$$p(g) = \prod_{i=1}^{m} p\big(\mathbf{c}_i, \mathbf{n}_i | \gamma_i(t), \mu(x), \Sigma(x)\big) \tag{15}$$

We will now show how these distributions can be computed and how the computational complexity for sampling from a grasp plan is reduced from $O(n^6)$ to $O(n^3)$ for a GPIS model.

C. Distribution on Contact Points

We would like to find the distribution on contact point $\mathbf{c}_i$. A contact point in terms of the GPIS and line of action model can be defined as the point, t, when the signed distance function is zero and no points before said point along the line have touched the surface. We express this as the following conditions:

$$sd(\gamma(t)) = 0 \tag{16}$$

$$sd(\gamma(\tau)) > 0, \ \forall \tau \in [a, t) \tag{17}$$

We will now demonstrate how to efficiently sample from $p(\mathbf{c}_i | \mu x, \Sigma x, \gamma_i t)$

The probability distribution along the line $\gamma(t)$ is given by:

$$p\big(sd(\gamma(t)); \mu(t), \Sigma(t)\big) \ \forall t \in [a, b] \tag{18}$$

This gives the signed distance function distributions along the entire line of action in the workspace as a multivariate Gaussian. One could think of this as a marginalization of all other points in signed distance field except the line of action. To sample contact points, one can draw samples from Eq. 3 and iterate from a to b until they reach a point that satisfies Eq. 16 and Eq. 17.

D. Complexity Analysis on Sampling From Grasps Distributions

After deriving each distribution, we can now sample along each line of action $\gamma_i(t)$ from the joint distribution $p(\mathbf{c}_i, \mathbf{n}_i | \gamma_i(t))$, due to our independence assumption Eq. 15. This can be done by drawing samples from the GP for signed distance and normals simultaneously and using our projection technique for the normal distribution.

Having a distribution along a line of action model allows us to sample from those instead of the joint distribution $p(sd(\mathcal{W}))$. Assuming the number of discretization points along the line of action is n, sampling from this distribution for a single grasp is $O(n^3)$. However, each proposed grasp plan Γ requires the distribution to be computed, so if we have $T = |G|$ then the complexity is $O(Tn^3)$. In practice, this should be much smaller than $O(n^6)$.

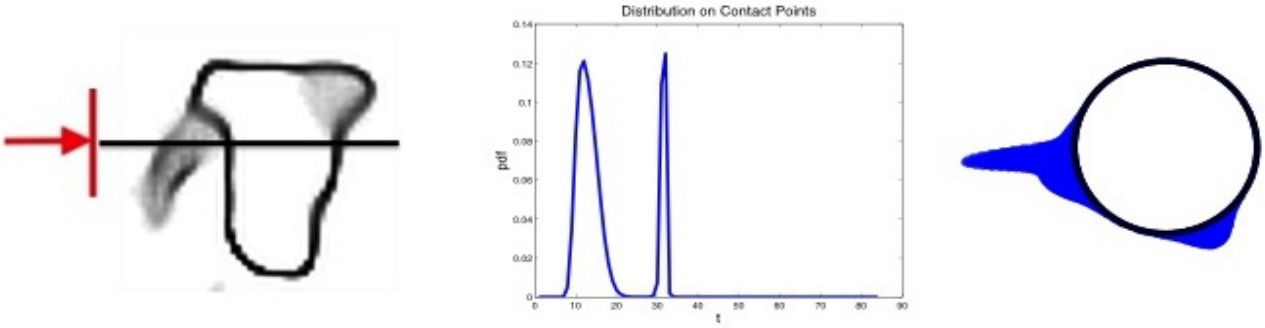

Fig. 9: (Left to Right): Line of action for a given gripper on an uncertain surface representing a measuring cup. Visualization technique comes from [**?**]. Distribution $p(c)$ as a function of t, the position along the line of action $\gamma(t)$. The two modes correspond to the different potential contact points, either the handle or the base of the cup. Lastly, the distribution on the surface normals (inward pointing) along $\gamma(t)$ described by equation 4. We plotted the probability mass along the unit circle, to reflect the normalization of the surface normals.