A Nonparametric Conjugate Prior Distribution for the Maximizing Argument of a Noisy Function

Pedro A. Ortega
Max Planck Institute for Intelligent Systems
Max Planck Institute for Biolog. Cybernetics
pedro.ortega@tuebingen.mpg.de

Jordi Grau-Moya
Max Planck Institute for Intelligent Systems
Max Planck Institute for Biolog. Cybernetics
jordi.grau@tuebingen.mpg.de

Tim Genewein
Max Planck Institute for Intelligent Systems
Max Planck Institute for Biolog. Cybernetics
tim.genewein@tuebingen.mpg.de

David Balduzzi
Max Planck Institute for Intelligent Systems
david.balduzzi@tuebingen.mpg.de

Daniel A. Braun
Max Planck Institute for Intelligent Systems
Max Planck Institute for Biolog. Cybernetics
daniel.braun@tuebingen.mpg.de

Abstract

We propose a novel Bayesian approach to solve stochastic optimization problems that involve finding extrema of noisy, nonlinear functions. Previous work has focused on representing possible functions explicitly, which leads to a two-step procedure of first, doing inference over the function space and second, finding the extrema of these functions. Here we skip the representation step and directly model the distribution over extrema. To this end, we devise a non-parametric conjugate prior based on a kernel regressor. The resulting posterior distribution directly captures the uncertainty over the maximum of the unknown function. Given t observations of the function, the posterior can be evaluated efficiently in time $\mathcal{O}(t^2)$ up to a multiplicative constant. Finally, we show how to apply our model to optimize a noisy, non-convex, high-dimensional objective function.

1 Introduction

Historically, the fields of statistical inference and stochastic optimization have often developed their own specific methods and approaches. Recently, however, there has been a growing interest in applying inference-based methods to optimization problems and vice versa [1–4]. Here we consider stochastic optimization problems where we observe noise-contaminated values from an unknown nonlinear function and we want to find the input that maximizes the expected value of this function.

The problem statement is as follows. Let $\mathcal{X}$ be a metric space. Consider a stochastic function $f : \mathcal{X} \rightsquigarrow \mathbb{R}$ mapping a test point $x \in \mathcal{X}$ to real values $y \in \mathbb{R}$ characterized by the conditional pdf $P(y|x)$. Consider the mean function

$$\bar{f}(x) := \mathbf{E}[y|x] = \int y P(y|x)\, dy. \tag{1}$$

The goal consists in modeling the optimal test point

$$x^* := \arg\max_x \{\bar{f}(x)\}. \tag{2}$$

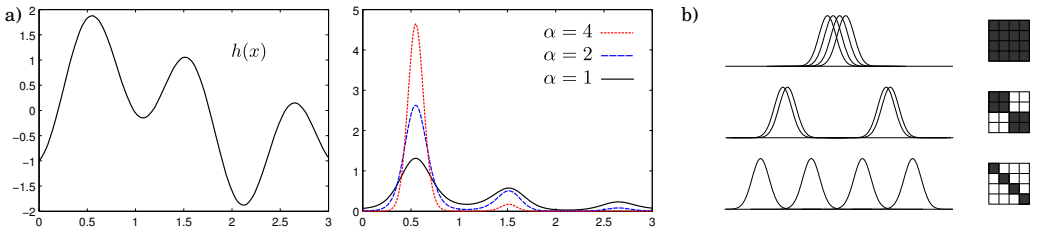

Figure 1: a) Given an estimate h of the mean function $\bar{f}$ (left), a simple probability density function over the location of the maximum x^* is obtained using the transformation $P(x^*) \propto \exp\{\alpha h(x^*)\}$, where $\alpha > 0$ plays the role of the precision (right). b) Illustration of the Gramian matrix for different test locations. Locations that are close to each other produce large off-diagonal entries.

Classic approaches to solve this problem are often based on stochastic approximation methods [5]. Within the context of statistical inference, Bayesian optimization methods have been developed where a prior distribution over the space of functions is assumed and uncertainty is tracked during the entire optimization process [6, 7]. In particular, non-parametric Bayesian approaches such as Gaussian Processes have been applied for derivative-free optimization [8, 9], also within the context of the continuum-armed bandit problem [10]. Typically, these Bayesian approaches aim to explicitly represent the unknown objective function of (1) by entertaining a posterior distribution over the space of objective functions. In contrast, we aim to model directly the distribution of the maximum of (2) conditioned on observations.

2 Brief Description

Our model is intuitively straightforward and easy to implement[1]. Let $h(x) : \mathcal{X} \to \mathbb{R}$ be an estimate of the mean $\bar{f}(x)$ constructed from data $\mathcal{D}_t := \{(x_i, y_i)\}_{i=1}^t$ (Figure 1a, left). This estimate can easily be converted into a posterior pdf over the location of the maximum by first multiplying it with a precision parameter $\alpha > 0$ and then taking the normalized exponential (Figure 1a, right)

$$P(x^*|\mathcal{D}_t) \propto \exp\{\alpha \cdot h(x^*)\}.$$

In this transformation, the precision parameter α controls the certainty we have over our estimate of the maximizing argument: $\alpha \approx 0$ expresses almost no certainty, while $\alpha \to \infty$ expresses certainty. The rationale for the precision is: the more *distinct* inputs we test, the higher the precision—testing the same (or similar) inputs only provides *local* information and therefore should not increase our knowledge about the *global* maximum. A simple and effective way of implementing this idea is given by

$$P(x^*|\mathcal{D}_t) \propto \exp\Bigg\{\rho \cdot \underbrace{\left(\xi + t \cdot \frac{\sum_i K(x_i, x_i)}{\sum_i \sum_j K(x_i, x_j)}\right)}_{\text{effective \# of locations}} \cdot \underbrace{\frac{\sum_i K(x_i, x^*) y_i + K_0(x^*) y_0(x^*)}{\sum_i K(x_i, x^*) + K_0(x^*)}}_{\text{estimate of } \bar{f}(x^*)}\Bigg\}, \quad (3)$$

where ρ, ξ, K, K_0 and y_0 are parameters of the estimator: $\rho > 0$ is the precision we gain for each new distinct observation; $\xi > 0$ is the number of prior points; $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ is a symmetric kernel function; $K_0 : \mathcal{X} \to \mathbb{R}^+$ is a prior precision function; and $y_0 : \mathcal{X} \to \mathbb{R}$ is a prior estimate of $\bar{f}$.

In (3), the mean function $\bar{f}$ is estimated with a kernel regressor [11] that combines the function observations with a prior estimate of the function, and the total effective number of locations is calculated as the sum of the prior locations ξ and the number of distinct locations in the data $\mathcal{D}_t$. The latter is estimated by multiplying the number of data points t with the coefficient

$$\frac{\sum_i K(x_i, x_i)}{\sum_i \sum_j K(x_i, x_j)} \in (0, 1], \quad (4)$$

[1]Implementations can be downloaded from http://www.adaptiveagents.org/argmaxprior

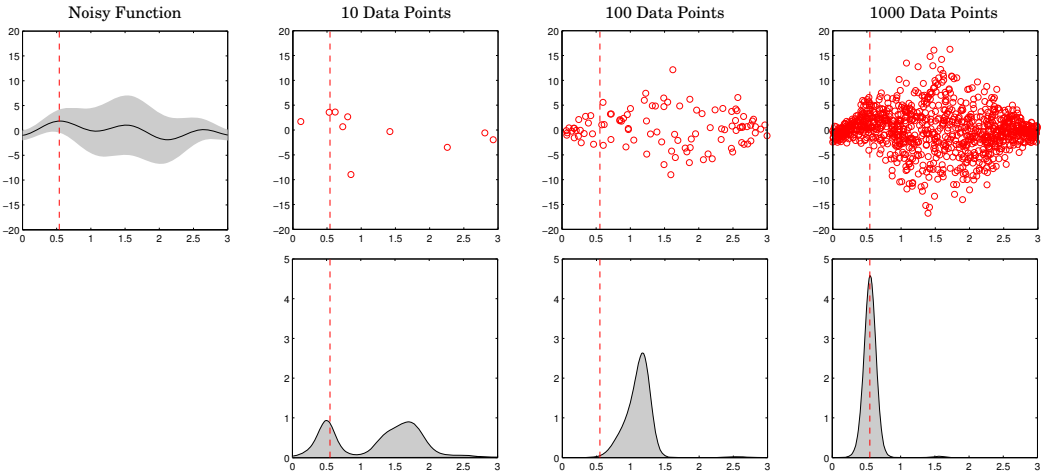

Figure 2: Illustration of the posterior distribution over the maximizing argument for 10, 100 and 1000 observations drawn from a function with varying noise. The top-left panel illustrates the function and the variance bounds (one standard deviation). The observations in the center region close to $x = 1.5$ are very noisy. It can be seen that the prior gets progressively washed out with more observations.

i.e. the ratio between the trace of the Gramian matrix $(K(x_i, x_j))_{i,j}$ and the sum of its entries. Inputs that are very close to each other will have overlapping kernels, resulting in large off-diagonal entries of the Gramian matrix—hence decreasing the number of distinct locations (Figure 1b). For example, if we have t observations from $n \ll t$ locations, and each location has t/n observations, then the coefficient (4) is equal to n/t and hence the number of distinct locations is exactly n, as expected.

Figure 2 illustrates the behavior of the posterior distribution. The expression for the posterior can be calculated up to a constant factor in $\mathcal{O}(t)$ time. The computation of the normalizing constant is in general intractable. Therefore, our proposed posterior can be easily combined with Markov chain Monte Carlo methods (MCMC) to implement stochastic optimizers as will be illustrated in Section 4.

3 Derivation

3.1 Function-Based, Indirect Model

Our first task is to derive an *indirect* Bayesian model for the optimal test point that builds its estimate via the underlying function space. Let $\mathcal{G}$ be the set of hypotheses, and assume that each hypothesis $g \in \mathcal{G}$ corresponds to a stochastic mapping $g : \mathcal{X} \rightsquigarrow \mathbb{R}$. Let $P(g)$ be the prior[2] over $\mathcal{G}$ and let the likelihood be $P(\{y_t\}|g, \{x_t\}) = \prod_t P(y_t|g, x_t)$. Then, the posterior of g is given by

$$P(g|\{y_t\}, \{x_t\}) = \frac{P(g)P(\{y_t\}|g, \{x_t\})}{P(\{y_t\}|\{x_t\})} = \frac{P(g)\prod_t P(y_t|g, x_t)}{P(\{y_t\}|\{x_t\})}. \tag{5}$$

For each $x^* \in \mathcal{X}$, let $\mathcal{G}(x^*) \subset \mathcal{G}$ be the subset of functions such that for all $g \in \mathcal{G}(x^*)$, $x^* = \arg\max_x \{\bar{g}(x)\}$[3]. Then, the posterior over the optimal test point x^* is given by

$$P(x^*|\{y_t\}, \{x_t\}) = \int_{\mathcal{G}(x^*)} P(g|\{y_t\}, \{x_t\})\, dg, \tag{6}$$

This model has two important drawbacks: (a) it relies on modeling the entire function space $\mathcal{G}$, which is potentially much more complex than necessary; (b) it requires calculating the integral (6), which is intractable for virtually all real-world problems.

[2]For the sake of simplicity, we neglect issues of measurability of $\mathcal{G}$.

[3]Note that we assume that the mean function $\bar{g}$ is bounded and that it has a unique maximizing test point.

3.2 Domain-Based, Direct Model

We want to arrive at a Bayesian model that bypasses the integration step suggested by (6) and directly models the location of optimal test point x^*. The following theorem explains how this *direct model* relates to the previous model.

Theorem 1. *The Bayesian model for the optimal test point x^* is given by*

$$P(x^*) = \int_{\mathcal{G}(x^*)} P(g)\, dg \qquad \textit{(prior)}$$

$$P(y_t|x^*, x_t, \mathcal{D}_{t-1}) = \frac{\int_{\mathcal{G}(x^*)} P(y_t|g, x_t) P(g) \prod_{k=1}^{t-1} P(y_k|g, x_k)\, dg}{\int_{\mathcal{G}(x^*)} P(g) \prod_{k=1}^{t-1} P(y_k|g, x_k)\, dg}, \qquad \textit{(likelihood)}$$

where $\mathcal{D}_t := \{(x_k, y_k)\}_{k=1}^t$ is the set of past tests.

Proof. Using Bayes' rule, the posterior distribution $P(x^*|\{y_t\}, \{x_t\})$ can be rewritten as

$$\frac{P(x^*) \prod_t P(y_t|x^*, x_t, \mathcal{D}_{t-1})}{P(\{y_t\}|\{x_t\})}. \tag{7}$$

Since this posterior is equal to (6), one concludes (using (5)) that

$$P(x^*) \prod_t P(y_t|x^*, x_t, \mathcal{D}_{t-1}) = \int_{\mathcal{G}(x^*)} P(g) \prod_t P(y_t|g, x_t)\, dg.$$

Note that this expression corresponds to the joint $P(x^*, \{y_t\}|\{x_t\})$. The prior $P(x^*)$ is obtained by setting $t = 0$. The likelihood is obtained as the fraction

$$P(y_t|x^*, x_t, \mathcal{D}_{t-1}) = \frac{P(x^*, \{y_k\}_{k=1}^t|\{x_k\}_{k=1}^t)}{P(x^*, \{y_k\}_{k=1}^{t-1}|\{x_k\}_{k=1}^{t-1})},$$

where it shall be noted that the denominator $P(x^*, \{y_k\}_{k=1}^{t-1}|\{x_k\}_{k=1}^{t-1})$ doesn't change if we add the condition x_t. □

From Theorem 1 it is seen that although the likelihood model $P(y_t|g, x_t)$ for the indirect model is i.i.d. at each test point, the likelihood model $P(y_t|x^*, x_t, \mathcal{D}_{t-1})$ for the direct model depends on the past tests $\mathcal{D}_{t-1}$, that is, *it is adaptive*. More critically though, the likelihood function's internal structure of the direct model corresponds to an integration over function space as well—thus inheriting all the difficulties of the indirect model.

3.3 Abstract Properties of the Likelihood Function

There is a way to bypass modeling the function space explicitly if we make a few additional assumptions. We assume that for any $g \in \mathcal{G}(x^*)$, the mean function $\bar{g}$ is continuous and has a unique maximum. Then, the crucial insight consists in realizing that the value of the mean function $\bar{g}$ inside a sufficiently small neighborhood of x^* is larger than the value outside of it (see Figure 3a).

We assume that, for any $\delta > 0$ and any $z \in \mathcal{X}$, let $B_\delta(z)$ denote the open δ-ball centered on z. The functions in $\mathcal{G}$ fulfill the following properties:

a. *Continuous:* Every function $g \in \mathcal{G}$ is such that its mean $\bar{g}$ is continuous and bounded.

b. *Maximum:* For any $x^* \in \mathcal{X}$, the functions $g \in \mathcal{G}(x^*)$ are such that for all $\delta > 0$ and all $z \notin B_\delta(x^*)$, $\bar{g}(x^*) > \bar{g}(z)$.

Furthermore, we impose a symmetry condition on the likelihood function. Let x_1^* and x_2^* be in $\mathcal{X}$, and consider their associated equivalence classes $\mathcal{G}(x_1^*)$ and $\mathcal{G}(x_2^*)$. There is no reason for them to be very different: in fact, they should virtually be indistinguishable outside of the neighborhoods of x_1^* and x_2^*. It is only inside of the neighborhood of x_1^* when $\mathcal{G}(x_1^*)$ becomes distinguishable from the other equivalence classes because the functions in $\mathcal{G}(x_1^*)$ systematically predict higher values

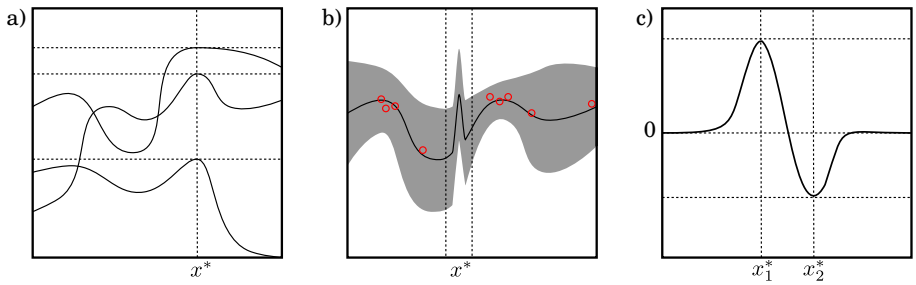

Figure 3: Illustration of assumptions. a) Three functions from $\mathcal{G}(x^*)$. They all have their maximum located at $x^* \in \mathcal{X}$. b) Schematic representation of the likelihood function of $x^* \in \mathcal{X}$ conditioned on a few observations. The curve corresponds to the mean and the shaded area to the confidence bounds. The density inside of the neighborhood is unique to the hypothesis x^*, while the density outside is shared amongst all the hypotheses. c) The log-likelihood ratio of the hypotheses x_1^* and x_2^* as a function of the test point x. The kernel used in the plot is Gaussian.

than the rest. This assumption is illustrated in Figure 3b. In fact, taking the log-likelihood ratio of two competing hypotheses

$$\log \frac{P(y_t|x_1^*, x_t, \mathcal{D}_{t-1})}{P(y_t|x_2^*, x_t, \mathcal{D}_{t-1})}$$

for a given test location x_t should give a value equal to zero unless x_t is inside of the vicinity of x_1^* or x_2^* (see Figure 3c). In other words, the amount of evidence a hypothesis gets when the test point is outside of its neighborhood is essentially zero (i.e. it is the same as the amount of evidence that most of the other hypotheses get).

3.4 Likelihood and Conjugate Prior

Following our previous discussion, we propose the following likelihood model. Given the previous data $\mathcal{D}_{t-1}$ and a test point $x_t \in \mathcal{X}$, the likelihood of the observation y_t is

$$P(y_t|x^*, x_t, \mathcal{D}_{t-1}) = \frac{1}{Z(x_t, \mathcal{D}_{t-1})} \lambda(y_t|x_t, \mathcal{D}_{t-1}) \exp\big\{\alpha_t \cdot h_t(x^*) - \alpha_{t-1} \cdot h_{t-1}(x^*)\big\}, \quad (8)$$

where: $Z(x_t, \mathcal{D}_{t-1})$ is a normalizing constant; $\lambda(y_t|x_t, \mathcal{D}_{t-1})$ is a posterior probability over y_t given x_t and the data $\mathcal{D}_{t-1}$; α_t is a precision measuring the knowledge we have about the whole function; and and h_t is an estimate of the mean function f. We have chosen the precision α_t as

$$\alpha_t := \rho \cdot \Big(\xi + \frac{\sum_i K(x_i, x_i)}{\sum_i \sum_j K(x_i, x_j)}\Big)$$

where $\rho > 0$ is a scaling parameter; $\xi > 0$ is a parameter representing the number prior locations tested; and $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$ is a symmetric kernel function[4]. For the estimate h_t, we have chosen a Naradaya-Watson kernel regressor [11]

$$h_t(x^*) := \frac{\sum_{i=1}^t K(x_i, x^*) y_i + K_0(x^*) y_0(x^*)}{\sum_{i=1}^t K(x_i, x^*) + K_0(x^*)}.$$

In the last expression, y_0 corresponds to a prior estimate of $\bar{f}$ with prior precision K_0. Inspecting (8), we see that the likelihood model favours positive changes to the estimated mean function *from new, unseen* test locations. The pdf $\lambda(y_t|x_t, \mathcal{D}_{t-1})$ does not need to be explicitly defined, as it will later drop out when computing the posterior. The only formal requirement is that it should be independent of the hypothesis x^*.

We propose the conjugate prior

$$P(x^*) = \frac{1}{Z_0} \exp\{\alpha_0 \cdot g_0(x^*)\} = \frac{1}{Z_0} \exp\{\xi \cdot y_0(x^*)\}. \quad (9)$$

[4]We refer the reader to the kernel regression literature for an analysis of the choice of kernel functions.

The conjugate prior just encodes a prior estimate of the mean function. In a practical optimization application, it serves the purpose of guiding the exploration of the domain, as locations x^* with high prior value $y_0(x^*)$ are more likely to contain the maximizing argument.

Given a set of data points $\mathcal{D}_t$, the prior (9) and the likelihood (8) lead to a posterior given by

$$\begin{aligned} P(x^*|\mathcal{D}_t) &= \frac{P(x^*)\prod_{k=1}^t P(y_k|x^*, x_k, \mathcal{D}_{k-1})}{\int_{\mathcal{X}} P(x')\prod_{k=1}^t P(y_k|x', x_k, \mathcal{D}_{k-1})\,dx'} \\ &= \frac{\exp\{\sum_{k=1}^t \alpha_k \cdot h_k(x^*) - \alpha_{k-1}\cdot h_{k-1}(x^*)\} Z_0^{-1}\prod_{k=1}^t Z(x_k, \mathcal{D}_{k-1})^{-1}}{\int_{\mathcal{X}} \exp\{\sum_{k=1}^t \alpha_k \cdot h_k(x') - \alpha_{k-1}\cdot h_{k-1}(x')\} Z_0^{-1}\prod_{k=1}^t Z(x_k, \mathcal{D}_{k-1})^{-1}\,dx'} \\ &= \frac{\exp\{\alpha_t \cdot h_t(x^*)\}}{\int_{\mathcal{X}} \exp\{\alpha_t \cdot h_t(x')\}\,dx'}. \end{aligned} \tag{10}$$

Thus, the particular choice of the likelihood function guarantees an analytically compact posterior expression. In general, the normalizing constant in (10) is intractable, which is why the expression is only practical for relative comparisons of test locations. Substituting the precision α_t and the mean function estimate h_t yields

$$P(x^*|\mathcal{D}_t) \propto \exp\left\{\rho \cdot \left(\xi + t \cdot \frac{\sum_i K(x_i, x_i)}{\sum_i \sum_j K(x_i, x_j)}\right) \cdot \frac{\sum_i K(x_i, x^*) y_i + K_0(x^*) y_0(x^*)}{\sum_i K(x_i, x^*) + K_0(x^*)}\right\}.$$

4 Experimental Results

4.1 Parameters.

We have investigated the influence of the parameters on the resulting posterior probability distribution. We have used the Gaussian kernel

$$K(x, x^*) = \exp\left\{-\frac{1}{2\sigma^2}(x - x^*)^2\right\}. \tag{11}$$

In this figure, 7 data points are shown, which were drawn as $y \sim N(f(x), 0.3)$, where the mean function is

$$f(x) = \cos(2x + \tfrac{3}{2}\pi) + \sin(6x + \tfrac{3}{2}\pi). \tag{12}$$

The prior precision K_0 and the prior estimate of the mean function y_0 were chosen as

$$K_0(x) = 1 \qquad \text{and} \qquad y_0(x) = -\frac{1}{2\sigma_0^2}(x - \mu_0)^2, \tag{13}$$

where the latter corresponds to the logarithm of a Gaussian with mean $\mu_0 = 1.5$ and variance $\sigma_0^2 = 5$. This prior favours the region close to μ.

Figure 4 shows how the choice of the precision scale ρ and the kernel width σ affect the shape of the posterior probability density. Here, it is seen that a larger kernel width σ increases the region of influence of a particular data point, and hence produce smoother posterior densities. The precision scale parameter ρ controls the precision per distinct data point: higher values for ρ lead to sharper updates of the posterior distribution.

4.2 Application to Optimization.

The main motivation behind our proposed model is its application to the optimization of noisy functions. Because of the noise, choosing new test locations requires carefully balancing explorative and exploitative tests—a problem well known in the *multiarmed bandits* literature. To overcome this, one can apply the Bayesian control rule/Thompson sampling [12, 13]: the next test location is chosen by *sampling* it from the posterior. We have carried out two experiments, described in the following.

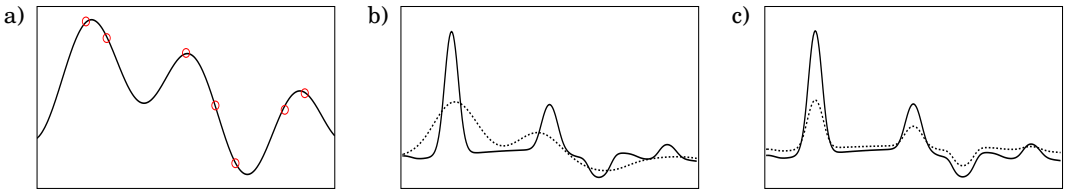

Figure 4: Effect of the change of parameters on the posterior density over the location of the maximizing test point. Panel (a) shows the 7 data points drawn from the noisy function (solid curve). Panel (b) shows the effect of increasing the width of the kernel (here, Gaussian). The solid and dotted curves correspond to $\sigma = 0.01$ and $\sigma = 0.1$ respectively. Panel (c) shows the effect of diminishing the precision on the posterior, where solid and shaded curves correspond to $\rho = 0.2$ and $\rho = 0.1$ respectively.

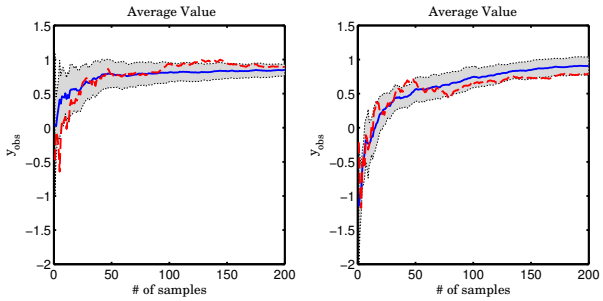

Figure 5: Observation values obtained by sampling from the posterior over the maximizing argument (left panel) and according to GP-UCB (right panel). The solid blue curve corresponds to the time-averaged function value, averaged over ten runs. The gray area corresponds to the error bounds (1 standard deviation), and the dashed curve in red shows the time-average of a single run.

Comparison to Gaussian Process UCB. We have used the model to optimize the same function (12) as in our preliminary tests but with higher additive noise equal to one. This is done by sampling the next test point x_t directly from the posterior density over the optimum location $P(x^*|\mathcal{D}_t)$, and then using the resulting pair (x_t, y_t) to recursively update the model. Essentially, this procedure corresponds to Bayesian control rule/Thompson sampling.

We compared our method against a Gaussian Process optimization method using an upper confidence bound (UCB) criterion [10]. The parameters for the GP-UCB were set to the following values: observation noise $\sigma_n = 0.3$ and length scale $\ell = 0.3$. For the constant that trades off exploration and exploitation we followed Theorem 1 in [10] which states $\beta_t = 2\log(|D|t^2\pi^2/6\delta)$ with $\delta = 0.5$. We have implemented our proposed method with a Gaussian kernel as in (11) with width $\sigma^2 = 0.05$. The prior sufficient statistics are exactly as in (13). The precision parameter was set to $\rho = 0.3$.

Simulation results over ten independent runs are summarized in Figure 5. We show the time-averaged observation values y of the noisy function evaluated at test locations sampled from the posterior. Qualitatively, both methods show very similar convergence (on average), however our method converges faster and with a slightly higher variance.

High-Dimensional Problem. To test our proposed method on a challenging problem, we have designed a non-convex, high-dimensional noisy function with multiple local optima. This *Noisy Ripples* function is defined as

$$f(x) = -\tfrac{1}{1000}\|x - \mu\|^2 + \cos(\tfrac{2}{3}\pi\|x - \mu\|)$$

where $\mu \in \mathcal{X}$ is the location of the global maximum, and where observations have additive Gaussian noise with zero mean and variance 0.1. The advantage of this function is that it generalizes well to any number of dimensions of the domain. Figure 6a illustrates the function for the 2-dimensional

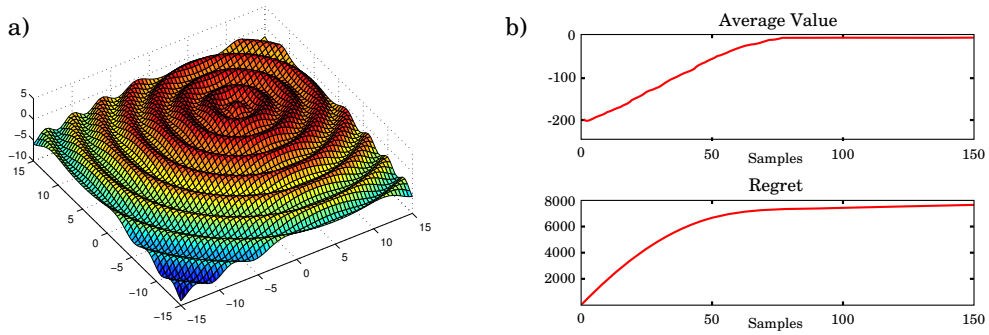

Figure 6: a) The *Noisy Ripples* objective function in 2 dimensions. b) The time-averaged value and the regret obtained by the optimization algorithm on a 50-dimensional version of the *Noisy Ripples* function.

input domain. This function is difficult to optimize because it requires averaging the noisy observations and smoothing the ridged landscape in order to detect the underlying quadratic form.

We optimized the 50-dimensional version of this function using a Metropolis-Hastings scheme to sample the next test locations from the posterior over the maximizing argument. The Markov chain was started at $[20, 20, \cdots, 20]^T$, executing 120 isotropic Gaussian steps of variance 0.07 before the point was used as an actual test location. For the arg-max prior, we used a Gaussian kernel with lengthscale $l = 2$, precision factor $\rho = 1.5$, prior precision $K_0(x^*) = 1$ and prior mean estimate $y_0(x^*) = -\frac{2}{1000}\|x + 5\|^2$. The goal μ was located at the origin.

The result of one run is presented in Figure 6b. It can be seen that the optimizer manages to quickly (≈ 100 samples) reach near-optimal performance, overcoming the difficulties associated with the high-dimensionality of the input space and the numerous local optima. Crucial for this success was the choice of a kernel that is wide enough to accurately estimate the mean function. The authors are not aware of any method capable of solving a problem of similar characteristics.

5 Conclusions

Our goal was to design a probabilistic model over the maximizing argument that is algorithmically efficient and statistically robust even for large, high-dimensional noisy functions. To this end, we have derived a Bayesian model that directly captures the uncertainty over the maximizing argument, thereby bypassing having to model the underlying function space—a much harder problem.

Our proposed model is computationally very efficient when compared to Gaussian process-based (which have cubic time complexity) or models based on upper confidence bounds (which require finding the input maximizing the bound—a generally intractable operation). In our model, evaluating the posterior up to a constant factor scales quadratically with the size of the data.

In practice, we have found that one of the main difficulties associated with our proposed method is the choice of the parameters. As in any kernel-based estimation method, choosing the appropriate kernel bandwidth can significantly change the estimate and affect the performance of optimizers that rely on the model. There is no clear rule on how to choose a good bandwidth.

In a future research, it will be interesting to investigate the theoretical properties of the proposed nonparametric model, such as the convergence speed of the estimator and its relation to the extensive literature on active learning and bandits.

References

[1] E. Brochu, V. Cora, and N. de Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. Technical Report TR-2009-023, University of British Columbia, Department of Computer Science, 2009.

[2] K. Rawlik, M. Toussaint, and S. Vijayakumar. Approximate inference and stochastic optimal control. *arXiv:1009.3958*, 2010.

[3] A. Shapiro. *Probabilistic Constrained Optimization: Methodology and Applications*, chapter Statistical Inference of Stochastic Optimization Problems, pages 282–304. Kluwer Academic Publishers, 2000.

[4] H.J. Kappen, V. Gómez, and M. Opper. Optimal control as a graphical model inference problem. *Machine Learning*, 87(2):159–182, 2012.

[5] H.J. Kushner and G.G. Yin. *Stochastic Approximation Algorithms and Applications*. Springer-Verlag, 1997.

[6] J. Mockus. Application of bayesian approach to numerical methods of global and stochastic optimization. *Journal of Global Optimization*, 4(4):347–365, 1994.

[7] D. Lizotte. *Practical Bayesian Optimization*. Phd thesis, University of Alberta, 2008.

[8] D.R. Jones, M. Schonlau, and W.J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13(4):455–492, 1998.

[9] M.A. Osborne, R. Garnett, and S.J. Roberts. Gaussian processes for global optimization. In *3rd International Conference on Learning and Intelligent Optimization (LION3)*, 2009.

[10] N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning*, 2010.

[11] T. Hastie, R. Tbshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, second edition, 2009.

[12] P.A. Ortega and D.A. Braun. A minimum relative entropy principle for learning and acting. *Journal of Artificial Intelligence Research*, 38:475–511, 2010.

[13] B.C. May and D.S. Leslie. Simulation studies in optimistic Bayesian sampling in contextual-bandit problems. Technical Report 11:02, Statistics Group, Department of Mathematics, University of Bristol, 2011.