Regret Bounds for Gaussian Process Bandits Without Observation Noise

Nando de Freitas*
Department of Computer Science
University of British Columbia
Vancouver, BC
nando@cs.ubc.ca

Alex J. Smola
Yahoo! Research
Santa Clara, CA
alex@smola.org

Masrour Zoghi
Department of Computer Science
University of British Columbia
Vancouver, BC
mzoghi@cs.ubc.ca

Abstract

This paper presents some statistical refinements of the bandits approach presented in [1] in the situation where there is no observation noise. We give an improved bound on the cumulative regret of the samples chosen by an algorithm that is related (though not identical) to the UCB algorithm of [1] in a complementary setting. Given a function f on a domain $\mathcal{D}$, sampled from a Gaussian process with an anisotropic kernel that is four times differentiable at 0, and a lattice $\mathcal{L} \subseteq \mathcal{D}$, we show that if the points in $\mathcal{L}$ are chosen for sampling using our branch-and-bound algorithm, the regret asymptotically decreases according to $O\left(e^{-\sqrt[3]{T}}\right)$ with high probability.

1 Introduction

Let $f : \mathcal{D} \to \mathbb{R}$ be a function on a compact subset $\mathcal{D} \subseteq \mathbb{R}^d$. We would like to address the global optimization problem

$$x_M = \operatorname*{argmax}_{x \in \mathcal{D}} f(x).$$

Let us assume for the sake of simplicity that the objective function f has a unique global maximum (although it can have many local maxima). We will also assume in this paper that at each point the value of the function can be observed without noise.

Here, we will use Gaussian process bandits (aka Bayesian optimization) to approach this problem. The general idea is to use a surrogate function that provides a good upper bound for the objective function f and that is also easier to deal with: this idea is illustrated in Figure 1, where the red curve serves the role of the surrogate function.

2 Nonparametric bandits (aka Bayesian optimization)

2.1 Gaussian processes

As in [1], the reward function is distributed according to a Gaussian process prior:

$$f(x) \sim \mathrm{GP}(m(\cdot), \kappa(\cdot, \cdot)). \tag{1}$$

For convenience, and without loss of generality, we assume that the prior mean vanishes, i.e., $m(\cdot) = 0$.

*The names of the authors appear in alphabetical order.

There are many possible choices for the covariance kernel, e.g., we could use a linear kernel and recover the linear case. Since we assume a non-linear reward model, one of the obvious choices is an anisotropic kernel κ with a vector of automatic relevance determination (ARD) hyperparameters $\boldsymbol{\theta}$ (cf. [2], §4.2.1):

$$\kappa(x_i, x_j) \quad = \quad \widetilde{\kappa}\left(-(x_i - x_j)^\top \mathbf{D}(x_i - x_j)\right) \tag{2}$$

where $\widetilde{\kappa}$ is an isotropic kernel and $\mathbf{D} = \mathrm{diag}(\boldsymbol{\theta}^2)$ is a diagonal matrix with entries $\boldsymbol{\theta}^2$ along the diagonal and zeros elsewhere. Our results apply to the case when $\widetilde{\kappa}$ is either the squared exponential or the Matérn kernel with parameter $\nu > 2$. In this paper, we assume that the hyperparameters are fixed and known in advance.

We can sample the GP at t points by choosing points $\mathbf{x}_{1:t} := \{x_1, \ldots, x_t\}$ and sampling the values of the function at these points to produce the values $\mathbf{f}_{1:t} = [f(x_1) \cdots f(x_t)]^\top$. The function values are distributed according to a multivariate Gaussian distribution $\mathcal{N}(0, \mathbf{K})$, with covariance entries $\kappa(x_i, x_j)$. Assume that we already have the observations, say recorded data, and that we want to use Bayesian optimization to decide what action x_{t+1} should be considered next. Let us denote the value of the function at this arbitrary point as f_{t+1}. Then, by the properties of GPs, $\mathbf{f}_{1:t}$ and f_{t+1} are jointly Gaussian:

$$\begin{bmatrix} \mathbf{f}_{1:t} \\ f_{t+1} \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K} & \mathbf{k}^\top \\ \mathbf{k} & \kappa(x_{t+1}, x_{t+1}) \end{bmatrix}\right),$$

where $\mathbf{k} = [\kappa(x_{t+1}, x_1) \cdots \kappa(x_{t+1}, x_t)]^\top$. Using the Schur complement, one arrives at an expression for the posterior predictive distribution:

$$P(f_{t+1} | \mathbf{x}_{1:t+1}, \mathbf{f}_{1:t}) = \mathcal{N}(\mu_t(x_{t+1}), \sigma_t^2(x_{t+1})),$$

where

$$\begin{aligned} \mu_t(x_{t+1}) &= \mathbf{k}^\top \mathbf{K}^{-1} \mathbf{f}_{1:t}, \text{ where } \mathbf{f}_{1:t} = [f(x_1) \cdots f(x_t)]^\top \\ \sigma_t^2(x_{t+1}) &= \kappa(x_{t+1}, x_{t+1}) - \mathbf{k}^\top \mathbf{K}^{-1} \mathbf{k}. \end{aligned} \tag{P}$$

2.2 Bayesian optimization

Here, we will make more precise the ideas summarized in Section 1 in the context of a Gaussian Process prior. When it is assumed that our objective function f is sampled from a GP, then one can use a combination of the posterior

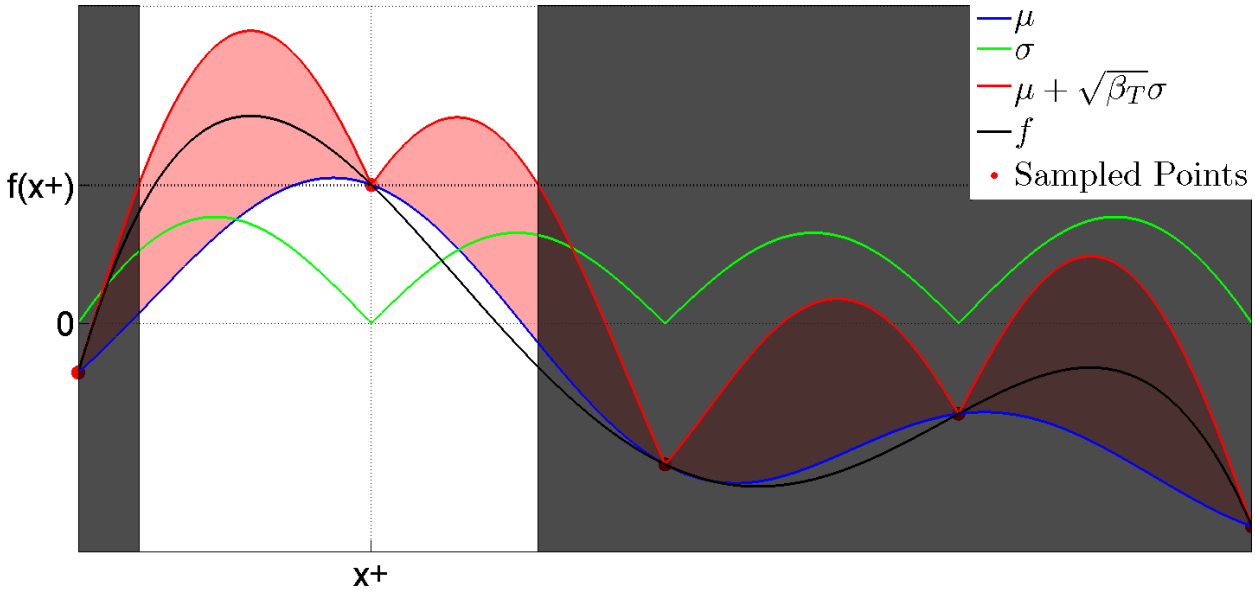

Figure 1: *An example of the branch and bound algorithm being applied to a function, and the shrinking of the search space: the algorithm does not have to sample in the shaded region anymore. As shown in the figure, the shaded region consist of the points x, for which $\mu(x) + \sqrt{\beta_T}\sigma(x) < f(x+)$, where $x+$ is the best sampled point so far. Moreover, Note that the UCB surrogate function (the red curve) bounds the objective function (the black curve) from above.*

predictive mean and variance given by equations (P) above to propose surrogate functions. The combination that we will utilize in this paper is the Upper Confidence Bound (UCB) given by the expression

$$\mu_T(x) + B_T\sigma_T(x),$$

where $\{B_T\}_{T=1}^{\infty}$ is a sequence of numbers that will be specified by the algorithm. The other choices are Probability of Improvement, Expected Improvement and Thompson sampling, for a description of which the reader is referred to [3], [4] and [5].

2.3 Our algorithm

The precise algorithm with which we work is given in Algorithm 1, the main idea of which is to tighten the bound on f given by the UCB surrogate function by sampling the search space more and more densely and shrinking the search space as more and more of the UCB surrogate function is "submerged" under the maximum observed value of the objective function, as depicted in Figure 1.

Let us point out here that the purpose of this algorithm is not to be of practical use in applications, but rather to provide a worst case scenario as far as performance is concerned given how manifestly wasteful it is when it comes to choosing the samples.

Algorithm 1 Branch and Bound

Given a compact subset $\mathcal{D} \subseteq \mathbb{R}^d$, a discrete lattice $\mathcal{L} \subseteq \mathcal{D}$ and a function $f : \mathcal{D} \to \mathbb{R}$, we execute the following algorithm:

$\mathcal{R} \leftarrow \mathcal{D}$
$\delta \leftarrow 1$
repeat
  Sample Twice as Densely:
    • $\delta \leftarrow \dfrac{\delta}{2}$
    • Sample f at enough points in $\mathcal{L}$ so that every point in $\mathcal{R}$ is contained in a simplex of size δ
  Shrink the Relevant Region:
    • Set

$$\widetilde{\mathcal{R}} := \left\{ x \in \mathcal{R} \middle| \mu_T(x) + \sqrt{\beta_T}\sigma_T(x) > \max_{t=1,\ldots,T-1} f(x_t) \right\},$$

      where T is the number points sampled so far and $\beta_T = 4\ln\left(\frac{|\mathcal{L}|T}{\alpha}\right)$ for a given $\alpha > 0$.
    • Solve the following constrained optimization problem:

$$(x_1^*, x_2^*) = \operatorname*{argmax}_{(x_1,x_2)\in\widetilde{\mathcal{R}}\times\widetilde{\mathcal{R}}} \|x_1 - x_2\|$$

    • $\mathcal{R} \leftarrow B\left(x_1^*, \|x_1^* - x_2^*\|\right) \cap B\left(x_2^*, \|x_1^* - x_2^*\|\right)$, where $B(p, r)$ is the ball of radius r centred around p.
until $\mathcal{R} \cap \mathcal{L} = \varnothing$

The idea of this algorithm was drawn from an observation made in [1] pointing out that the optimum point cannot lie outside our relevant set $\widetilde{\mathcal{R}}$.

3 Analysis

Let us recall that $\mathcal{D} \subseteq \mathbb{R}^d$ is assumed to be a non-empty compact subset and f a sample from the Gaussian Process $\mathrm{GP}\left(0, \kappa(\cdot,\cdot)\right)$ on $\mathcal{D}$. Moreover, in what follows we will use the notation $x_M := \mathrm{argmax}_{\mathcal{D}}\, f$. Also, by convention, for any set $\mathcal{S}$, we will denote its interior by $\mathcal{S}^\circ$ and its boundary by $\partial\mathcal{S}$.

Definition 1 *Given the above setup, the* **regret** *function is defined to be*

$$r(x) = \max_{\mathcal{D}} f - f(x).$$

Proposition 2 *The following holds true:*

$\forall\blacktriangleright$ *Suppose we are given the following setup:*

- $\alpha > 0$
- $\mathcal{D} \subseteq \mathbb{R}^d$ *a compact subset;*
- κ *an stationary kernel on* $\mathbb{R}^d$ *that is four times differentiable;*
- $f \sim \mathrm{GP}(0, \kappa)$ *a continuous sample on* $\mathcal{D}$ *that has a unique global maximum* x_M*, which satisfies one of the following two conditions:*
  - (†) $x_M \in \mathcal{D}^\circ$ *and* $f(x_M) - c_1\|x - x_M\|^2 < f(x) \leq f(x_M) - c_2\|x - x_M\|^2$ *for all* x *satisfying* $x \in B(x_M, \rho_0)$ *for some* $\rho_0 > 0$*;*
  - (‡) $x_M \in \partial\mathcal{D}$ *and both* f *and* $\partial\mathcal{D}$ *are smooth at* x_M*, with* $\nabla f(x_M) \neq 0$*;*

$\exists\blacktriangleright$ *Then, there exist positive numbers* A *and* τ *such that*

$\forall\blacktriangleright$ *for any lattice* $\mathcal{L} \subseteq \mathcal{D}$ *satisfying the following two conditions*

- (♣) $2\mathcal{L} \cap \mathrm{conv}(\mathcal{L}) \subseteq \mathcal{L}$*,*
- (♠) $2^{\lceil -\log_2 \frac{\rho_0}{\mathrm{diam}(\mathcal{D})} \rceil + 1}\mathcal{L} \cap \mathcal{L} \neq \varnothing$ *if* f *satisfies* (†)*,*

$\exists\blacktriangleright$ *there exists an interger* T *such that the points specified by the Branch and Bound algorithm,* $\{x_t\}$*, will satisfy the following asymptotic bound:*

$\forall\blacktriangleright$ *for all* $t > T$*, with probability* $1 - \alpha$ *we have*

$$r(x_t) < Ae^{-\tau\sqrt[3]{t}}.$$

Remark 3 *(1) Note that for a random sample* $f \sim \mathrm{GP}(0, \kappa)$ *one of conditions* (†) *and* (‡) *will be satisfied almost surely if* κ *is a Matérn kernel with* $\nu > 2$ *because the sample* f *is twice differentiable almost surely (by Theorem 1.4.2 of [6] and §2.6 of [7]) and the vanishing of at least one of the eigenvalues of the Hessian is a codimension 1 condition.*

(2) The two conditions (♣) *and* (♠) *simply require that the lattice be "divisible by 2" and that it be fine enough so that the algorithm can sample inside the ball* $B(x_M, \rho_0)$ *when the maximum of the function is located in the interior of the search space* $\mathcal{D}$*.*

References

[1] Niranjan Srinivas, Andreas Krause, Sham M Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *International Conference on Machine Learning*, 2010.

[2] Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

[3] Eric Brochu, Vlad M Cora, and Nando 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, arXiv:1012.2599v1, University of British Columbia, Department of Computer Science, 2009.

[4] Benedict May, Nathan Korda, Anthony Lee, and David Leslie. Optimistic Bayesian sampling in contextual-bandit problems. 2010.

[5] Matthew Hoffman, Eric Brochu, and Nando de Freitas. Portfolio allocation for Bayesian optimization. In *Uncertainty in Artificial Intelligence*, pages 327–336, 2011.

[6] Robert J. Adler and Jonathan E. Taylor. *Random Fields and Geometry*. Springer, 2007.

[7] Michael L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, 1999.