

Regret Bounds for Gaussian Process Bandit Problems

Steffen Grünewälder, Jean-Yves Audibert, Manfred Opper, John Shawe-Taylor

▶ **To cite this version:**

Steffen Grünewälder, Jean-Yves Audibert, Manfred Opper, John Shawe-Taylor. Regret Bounds for Gaussian Process Bandit Problems. AISTATS 2010 - Thirteenth International Conference on Artificial Intelligence and Statistics, May 2010, Chia Laguna Resort, Sardinia, Italy. pp.273-280. hal-00654517

HAL Id: hal-00654517
https://enpc.hal.science/hal-00654517

Submitted on 22 Dec 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Regret Bounds for Gaussian Process Bandit Problems

Steffen Grünewälder
University College London
steffen@cs.ucl.ac.uk

Jean-Yves Audibert
Université Paris-Est
& INRIA/ENS/CNRS
audibert@imagine.enpc.fr

Manfred Opper
TU-Berlin
opperm@cs.tu-berlin.de

John Shawe-Taylor
University College London
jst@cs.ucl.ac.uk

Abstract

Bandit algorithms are concerned with trading exploration with exploitation where a number of options are available but we can only learn their quality by experimenting with them. We consider the scenario in which the reward distribution for arms is modelled by a Gaussian process and there is no noise in the observed reward. Our main result is to bound the regret experienced by algorithms relative to the a posteriori optimal strategy of playing the best arm throughout based on benign assumptions about the covariance function defining the Gaussian process. We further complement these upper bounds with corresponding lower bounds for particular covariance functions demonstrating that in general there is at most a logarithmic looseness in our upper bounds.

1 INTRODUCTION

Bandit problems have become a topic of extensive research in recent years and many extensions of the classical framework (Robbins, 1952; Gittins and Jones, 1979; Berry and Fristedt, 1985; Auer et al., 2002; Cesa-Bianchi and Lugosi, 2006) have been developed (among many others, Abernethy et al., 2008; Slivkins and Upfal, 2008; Wang et al., 2008; Bubeck et al., 2008; Kleinberg et al., 2008b). A particularly important extension for real world applications is the use of a Gaussian process to model the reward distribution. It allows a canonical treatment of optimization problems with expensive-to-evaluate objective functions for which the expected similarity for different arguments (or parameter settings) are defined through kernels. Recently a number of different algorithms have been developed to tackle this sequential optimization problem with Gaussian process (GP) prior (see Section 2 of Ginsbourger et al. (2008) for an overview of them). The theoretical underpinnings in the form of regret bounds are, however, still missing.

To illustrate the setting, consider the problem of optimizing throughput of mobile phones. In this example, the arms of the bandit would correspond to system settings and the corresponding throughput to the reward. A kernel could be used to model the expected similarity between the rewards of two settings of the mobile. In the Bayesian framework, the kernel encodes our prior belief that the reward as a function of parameter settings is drawn from a GP defined by the given kernel (Figure 1 shows a draw from a GP with a Gaussian kernel). The approach will work well in practice if the GP reward functions are similar to that of the concrete optimisation problem at hand.

Due to the flexibility and power of GPs, the approach has sparked considerable interest (Schonlau, 1997; Jones et al., 1998; Jones, 2001; Ginsbourger and Riche, 2009; Osborne et al., 2009). The early works were done in the field of global optimisation. Using stochastic processes for global optimisation has a long tradition. The earliest works date back to Kushner (1964). The use of GPs is relatively new and can be found associated with the name "Kriging" and response surfaces (Schonlau, 1997; Jones et al., 1998; Jones, 2001). These works address exploration-exploitation tradeoffs to find the global optimum which is very similar to what Bandit algorithms do. Their motivation for introducing GPs to global optimisation comes from an engineering viewpoint, where the design space (also called the state space) can often be interpolated and extrapolated quite accurately and the GP allows them to "see" obvious trends. Furthermore, it allows them to provide what they call credible stopping rules. The suboptimality of the one-step lookahead policy has been recently exhibited and a multi-step optimal

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 6 of JMLR: W&CP 6. Copyright 2010 by the authors.

lookahead policy has been proposed (Osborne et al., 2009; Ginsbourger and Riche, 2009).

While a number of practical algorithms have been developed, there are no theoretical guarantees available on their performance in the form of regret bounds. One of the reason for this lies in the difficulty of deriving regret bounds for the GP setting. Bounding the regret requires to understand the behaviour of the supremum of the reward function for the *posterior* distribution, that is the distribution of the process knowing the past observations. Furthermore, each kernel induces various type of correlations between the rewards. Dealing with these problems in a generic way is also the target of this work. If we consider again Figure 1, we can ask where is the maximum? How high is it? How close to the maximum are rewards at arms that are ε close to the optimal arm? How consistent are these properties across multiple draws?

The key to dealing with these questions lies in a fundamental work from probability theory. One of the key tools is the celebrated chaining argument with, in particular, the Dudley integral (Dudley, 1967). This argument allows us to bound the expected supremum and, from concentration inequalities, with high probability, the supremum for a given GP sample. The interesting property of the Dudley integral is that it breaks the problem of handling the supremum down to understanding the *posterior* covariance function. The second important idea to make the approach general is the use of Hölder continuity assumptions on the kernel. These assumptions allows to control how the prior (and indirectly the posterior) Gaussian process when we move a short distance away from our current point. The assumption on the covariance is fulfilled by a majority of kernels (this is typically the case for the Brownian motion, the Brownian bridge, the OrnsteinUhlenbeck process, and the GP associated to the Gaussian kernel) and can be easily verified.

2 MAIN RESULTS

The first result is the upper bound on the regret for GPs with covariance functions that are Hölder continuous with a coefficient α and a constant L_k (details are given in Section 4). Assuming a zero mean Gaussian process prior, we know that after playing T arms in $[0,1]^D$ that the optimal strategy has a regret no bigger than

$$4\sqrt{\frac{L_k \log(2T)}{(2\tilde{T})^\alpha}} + 15\sqrt{\frac{(\alpha+3)DL_k}{\alpha(2\tilde{T})^\alpha}},$$

where $\tilde{T} := \lfloor T^{1/D} \rfloor$.

This upper bound is complemented by the following

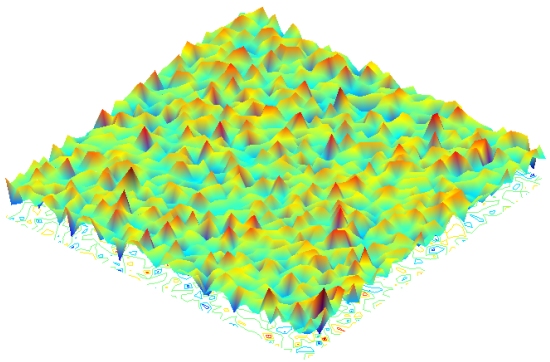


Figure 1: A draw from a Gaussian process on $[0,1]^2$ with a Gaussian kernel ($\sigma = 0.1$). A draw represents an optimization problem. The optimal arm is the one where the peak is highest.

lower bound on the optimal regret:

$$\kappa\sqrt{\frac{L_k}{2^\alpha (2T)^{\alpha/D} \log T}},$$

where $\kappa > 0$ is a universal constant. The lower bound is derived for a specific kernel that fulfills the α-Hölder continuity assumption and matches the upper bound up to a logarithmic factor.

During the reviewing process of this work, another work (Srinivas et al., 2009) addressing regret bounds in GP optimization has appeared. It provides a different approach linked to optimal empirical design and information gain and studies the cumulative regret. So, both the bounds presented here and theirs are of interest (and cannot really be compared).

3 BACKGROUND

We start with discussing the classical bandit setting. In its most basic formulation a finite number of arms (or actions) $x_1, \ldots, x_K$ is given. Each arm is associated to a reward distribution characterizing the arm. Successive plays of an arm x_i yields a sequence of rewards which are independent and identically distributed according to this unknown distribution (with unknown expectation). The goal is to define a strategy such that a functional called regret is minimised. Typically, the regret measures the expected loss for the strategy compared to playing the arm with the highest expected reward.

In our setting we do not have a finite number of arms but a continuum of arms. Different approaches to the continuum arm space setting have already been proposed Agrawal (1995); Kleinberg (2004); Auer et al. (2007); Wang et al. (2008); Bubeck et al. (2008); Klein-

berg et al. (2008a) but they do not integrate the Gaussian process assumption. The precise definition of our setting and regret is given in Section 4.

3.1 GAUSSIAN PROCESSES

We use a Gaussian process to model the reward distribution. A stochastic process $r(x)$ is Gaussian if and only if for every finite set of indices $x_1, \ldots, x_n$ the variable $(r(x_1), \ldots, r(x_n))$ is a multivariate Gaussian random variable (Rasmussen and Williams, 2006). A Gaussian process is completely specified by its mean function μ and its covariance function $k(x, y) = \mathbb{C}\mathrm{ov}[r(x), r(y)]$. The covariance function k is symmetric and positive semi-definite.

3.1.1 Metric Entropy Bound for the Expected Supremum

The Dudley integral (Dudley, 1967) gives a bound on the expected supremum of a Gaussian process. It is defined with respect to the canonical metric $d(x, y) = \sqrt{\langle x, x\rangle - 2\langle x, y\rangle + \langle y, y\rangle}$, where $\langle \cdot, \cdot \rangle$ denotes the covariance.

The bound on the expected supremum uses the size of the parameter space X. The size is measured with the packing number $N(\varepsilon, X)$, which is the maximal number of points that are all at least ε-distant from each other (in the $d(x, y)$ metric).

The bound on the expected supremum is given in the following theorem (Massart, 2003)[Th. 3.18, p. 74]:

Theorem 3.1 (Metric Entropy Bound) *Let $(f(x))_{x\in X}$ be some centered Gaussian process. Assume that (X, d) is totally bounded and denote by $N(\varepsilon, X)$ the ε-packing number of (X, d), for all positive ε. If $\sqrt{\log N(\varepsilon, X)}$ is integrable at 0, then $(f(x))_{x\in X}$ admits a version which is almost surely uniformly continuous on (X, d). Moreover, if $(f(x))_{x\in X}$ is almost surely continuous on (X, d), then*

$$\mathbb{E} \sup_{x\in X} f(x) \leq 12 \int_0^{\sigma_X} \sqrt{\log N(\varepsilon, X)} d\varepsilon,$$

where $\sigma_X = \sup_{x\in X} \mathbb{V}\mathrm{ar}\, f(x)$ is the supremum of the variance on X.

The important property of the bound is that it transforms the problem of understanding some complex stochastic process into understanding $d(x, y)$ which is a deterministic object that can be studied using analytic methods.

3.1.2 Inequalities

We will need two inequalities for Gaussian, respectively sub-Gaussian random variables. The first can be found in Massart (2003)[Prop. 3.19, p. 77]:

Proposition 3.2 *If $(f(x))_{x\in X}$ is some almost surely continuous Gaussian process on the totally bounded set (X, d) then for every $\lambda \in \mathbb{R}$*

$$\mathbb{E}[\exp(\lambda(\sup f(x) - \mathbb{E}[\sup f(x)]))] \leq \exp\left(\frac{\lambda^2 \sigma_X^2}{2}\right),$$

with $\sigma_X = \sup_{x\in X} \mathbb{V}\mathrm{ar}\, f(x)$.

The second is a statement about the supremum of sub-Gaussian random variables and can be found in Devroye and Lugosi (2001)[Lem. 2.2, p. 7]:

Lemma 3.3 *Let $\sigma > 0$, $n \geq 2$ and let $Y_1, \ldots, Y_n$ be real-valued random variables such that for all $s > 0$ and $1 \leq i \leq n, \mathbb{E}\exp(sY_i) \leq \exp\left(s^2\sigma^2/2\right)$. Then $\mathbb{E}\max_{i\leq n} Y_i \leq \sigma\sqrt{2\log n}$.*

3.1.3 The Posterior Process

The posterior process, i.e. the process conditional on observations, is again a Gaussian process. For a zero mean prior process f the mean μ_Z of the posterior process is given by

$$\mu_Z(x) = k(x, Z)' k(Z, Z)^{-1} f(Z),$$

where we use the short form $k(x, Z) := (k(x, Z_1), \ldots, k(x, Z_n))'$ and $k(Z, Z)$ for the kernel matrix of a vector $Z := (Z_1, \ldots, Z_n)$. The Z's denote the positions of the observations. And the values at these positions are $f(Z) = (f(Z_1), \ldots, f(Z_n))'$. The covariance of the posterior process is given by

$$\langle x, y\rangle := \langle x, y\rangle_Z := k(x, y) - k(x, Z)' k(Z, Z)^{-1} k(Z, y).$$

4 THE BANDIT GAME WITH A GAUSSIAN PROCESS PRIOR

Central to our approach are two Hölder continuity assumptions for the mean and the covariance. The covariance function k and the mean function μ are assumed to satisfy the following mild conditions:

(A1) For some $L_\mu \geq 0$, for any $x, y \in X$, $\left|\mu(x) - \mu(y)\right| \leq L_\mu \|x - y\|_\infty$.

(A2) For some $L_k \geq 0$ and some $\alpha > 0$, for any $x, y \in X$, $\left|k(x, x) - k(x, y)\right| \leq L_k \|x - y\|_\infty^\alpha$.

We use the supremum norms as it leads to simpler bounds (compared to the Euclidean norm for instance). However, as X is a finite dimensional space all norms are equivalent, meaning the results apply to arbitrary norms up to an extra constant.

Furthermore, we assume for simplicity that the space of arms is the D-dimensional unit cube $X = [0, 1]^D$. It is easy to extend the results to cubes of size R.

4.1 GAME DEFINITION

The player knows the (prior) distribution of the Gaussian process (he knows the mean function μ and the covariance function k). He also knows the number T of rounds to play. The game is the following.

For $t = 1, \dots, T$

the player chooses $\hat{x}_t \in X$.

the player observes $r(\hat{x}_t)$.

At the end of these T rounds, the player receives the reward

$$\max\big(r(\hat{x}_1), \dots, r(\hat{x}_T)\big).$$

A common alternative target encountered in the bandit literature is the cumulative reward $\sum_{t=1}^{T} r(\hat{x}_t)$. In applications, it is often the case that the 'max' reward is more interesting than the cumulative one. The two problems differ substantially in the sense that there is no cost of exploring in the 'max' reward setting (see Bubeck et al. (2009) for an interesting link between the two regret notions).

4.2 THE OPTIMAL STRATEGY

It is then natural to define the optimal strategy as the one having the highest expected reward:

$$\mathbb{E}\max\big(r(\hat{x}_1), \dots, r(\hat{x}_T)\big).$$

It occurs that the optimal strategy is the one obtained by dynamic programming or backward induction Bellman (1956), which in our finite horizon setting can be viewed as the T-steps lookahead strategy following the terminology of Ross (1970): at time t the player chooses

$$\begin{aligned}\hat{x}_t = \operatorname*{argmax}_{x_t \in X} \mathbb{E}\big[\max_{x_{t+1}\in X} \cdots \mathbb{E}\big[\max_{x_{T-1}\in X} \\ \mathbb{E}\big[\max_{x_T \in X} \mathbb{E}\big[\max\big(r(x_1), \dots, r(x_T)\big)\big|r(x_1), \dots, r(x_{T-1})\big] \\ \big|r(x_1), \dots, r(x_{T-2})\big] \cdots \big|r(x_1), \dots, r(x_{t-1})\big]. \qquad (1)\end{aligned}$$

The formula is a bit awful, but the idea is simple. If we are at time T, the optimal action $\hat{x}_T$ is

$$\operatorname*{argmax}_{x_T \in X} \mathbb{E}\big[\max\big(r(x_1), \dots, r(x_T)\big)\big|r(x_1), \dots, r(x_{T-1})\big],$$

which corresponds to a one-step lookahead. From this, we know that at time $T-1$, if we choose x_{T-1}, our expected reward knowing the past will be

$$\mathbb{E}\big[\max_{x_T \in X} \mathbb{E}\big[\max\big(r(x_1), \dots, r(x_T)\big)\big|r(x_1), \dots, r(x_{T-1})\big] \big|r(x_1), \dots, r(x_{T-2})\big].$$

This leads to the optimal choice for x_{T-1} defined in (1). Repeated this argument again, we obtain the T-steps lookahead strategy. This policy is, in spirit, the same as the one proposed in (Osborne et al., 2009, Section 3.2). Our main result is to provide a regret bound for it.

It is important to understand that the distribution of $r(x_t)$ knowing $r(x_1), \dots, r(x_{t-1})$ is known: it is the Gaussian posterior. The difficulty of proving results on the algorithm relies on the understanding of the behaviours of the mean and covariance of the Gaussian process conditional to the past observations.

4.3 GUARANTEE ON THE PERFORMANCE OF THE OPTIMAL STRATEGY

We call a guarantee on how the optimal strategy works a lower bound on the expected reward $\mathbb{E}\max\big(r(\hat{x}_1), \dots, r(\hat{x}_T)\big)$, or equivalently an upper bound on the expected regret $\mathbb{E}\big\{\sup_{x\in X} r(x) - \max\big(r(\hat{x}_1), \dots, r(\hat{x}_T)\big)\big\}$. We need a definition to simplify the notation: $\tilde{T} := \lfloor T^{1/D} \rfloor$. To establish the bound, we will use the following path.

- define a simpler strategy: we will consider the naive grid strategy. We partition the space of arms X into $\tilde{T}^D \leq T$ many cubes of side length $1/\tilde{T}$ and play the center of each of these cubes.
- lower bound the expected reward of this simpler strategy. This last point will rely on the use of concentration inequalities for the supremum of Gaussian and sub-Gaussian processes.

Theorem 4.1 *Under Assumptions (A1) and (A2), The optimal strategy satisfies*

$$\begin{aligned}&\mathbb{E}\Big\{\sup_{x\in X} r(x) - \max\big(r(\hat{x}_1), \dots, r(\hat{x}_T)\big)\Big\} \\ &\leq \mathbb{E}\Big\{\sup_{x\in X} r(x) - \max\big(r(x_1), \dots, r(x_T)\big)\Big\} \\ &\leq 4\sqrt{\frac{L_k \log(2T)}{(2\tilde{T})^\alpha}} + 15\sqrt{\frac{(\alpha+3)DL_k}{\alpha(2\tilde{T})^\alpha}} + \frac{L_\mu}{2\tilde{T}}.\end{aligned}$$

Proof. Let $\mathbf{i} = (\mathbf{i}_1, \dots, \mathbf{i}_D)$ be a D-dimensional index with $1 \leq \mathbf{i}_j \leq \tilde{T}$. Let $I_\mathbf{i} = \big[\frac{\mathbf{i}_1 - 1}{\tilde{T}}, \frac{\mathbf{i}_1}{\tilde{T}}\big] \times \ldots \times \big[\frac{\mathbf{i}_D - 1}{\tilde{T}}, \frac{\mathbf{i}_D}{\tilde{T}}\big]$ be the $\mathbf{i}$'s cube. We have $x_\mathbf{i} = (\frac{2\mathbf{i}_1 - 1}{2\tilde{T}}, \dots, \frac{2\mathbf{i}_D - 1}{2\tilde{T}}) \in I_\mathbf{i}$. Introduce the random variables

$$U_\mathbf{i} = \sup_{x\in I_\mathbf{i}} \Big\{r(x) - \mathbb{E}\big[r(x)\big|r(x_\mathbf{i})\big]\Big\}$$

$$V_\mathbf{i} = \sup_{x\in I_\mathbf{i}} \Big\{\mathbb{E}\big[r(x)\big|r(x_\mathbf{i})\big] - r(x_\mathbf{i})\Big\}$$

$$W_\mathbf{i} = \sup_{x\in I_\mathbf{i}} \big\{r(x) - r(x_\mathbf{i})\big\}.$$

We have

$$\begin{aligned}
&\mathbb{E}\Big\{ \sup_{x\in X} r(x) - \max\big(r(x_1),\dots,r(x_T)\big)\Big\} \\
&\leq \mathbb{E} \max_{\mathbf{i}\in\{1,\dots,\tilde{T}\}^D} W_{\mathbf{i}} \\
&\leq \mathbb{E} \max_{\mathbf{i}\in\{1,\dots,\tilde{T}\}^D} U_{\mathbf{i}} + \mathbb{E} \max_{\mathbf{i}\in\{1,\dots,\tilde{T}\}^D} V_{\mathbf{i}} \\
&\leq \mathbb{E} \max_{\mathbf{i}\in\{1,\dots,\tilde{T}\}^D} \big\{U_{\mathbf{i}} - \mathbb{E}[U_{\mathbf{i}}|r(x_{\mathbf{i}})]\big\} \\
&\quad + \mathbb{E} \max_{\mathbf{i}\in\{1,\dots,\tilde{T}\}^D} \mathbb{E}[U_{\mathbf{i}}|r(x_{\mathbf{i}})] + \mathbb{E} \max_{\mathbf{i}\in\{1,\dots,\tilde{T}\}^D} V_{\mathbf{i}}. \quad (2)
\end{aligned}$$

Before giving the propositions upper bounding the three terms of the last righthand-side, let us introduce a convenient notation. Define

$$\begin{aligned}
k(x,y|z) &= k(x,y) - \frac{k(x,z)k(y,z)}{k(z,z)}\mathbb{1}_{k(z,z)>0} \\
&= \mathbb{C}\mathrm{ov}(r(x), r(y)|r(z)),
\end{aligned}$$

i.e., $k(x,y|z)$ is the conditional covariance of $r(x)$ and $r(y)$ knowing the value of $r(z)$. Similarly, define $d(x,y|z)$ the posterior distance knowing the value of $r(z)$. We have

$$\begin{aligned}
d^2(x,y|z) &= k(x,x|z) + k(y,y|z) - 2k(x,y|z) \\
&= k(x,x) + k(y,y) - 2k(x,y) \\
&\quad - \frac{[k(x,z)-k(y,z)]^2}{k(z,z)}\mathbb{1}_{k(z,z)>0} \\
&\leq d^2(x,y), \quad (3)
\end{aligned}$$

where $d(x,y)$ denotes the prior distance and we used the characteristic function $\mathbb{1}_{k(z,z)>0}$ to account for cases where $k(z,z)=0$, i.e. for cases where at z no information can be gained as the value is known exactly. This inequality shows that the posterior distance is at most equal to the prior distance, which gives a rough but useful upper bound on the posterior distance.

The ε-covering number is the minimum number of balls of radius ε covering the space. The ε-packing number is known to be smaller than the $\varepsilon/2$-covering number. We will map the problem of covering X in the canonical $d(x,y)$ metric to the problem of covering X in the original metric from X. It is convenient to introduce a function for this. Let

$$\psi_{\mathbf{i}}(\beta) = \sup_{\substack{x,y\in I_{\mathbf{i}} \\ \|x-y\|_\infty\leq\beta}} d(x,y|x_{\mathbf{i}})$$

Define $\sigma_{\mathbf{i}} = \sup_{x\in I_{\mathbf{i}}} \sqrt{\mathbb{V}\mathrm{ar}\left\{r(x)\middle|r(x_{\mathbf{i}})\right\}}$. We have

$$\begin{aligned}
\sigma_{\mathbf{i}} &= \sup_{x\in I_{\mathbf{i}}} \sqrt{\mathbb{V}\mathrm{ar}\left\{r(x) - r(x_{\mathbf{i}})\middle|r(x_i)\right\}} \\
&= \sup_{x\in I_{\mathbf{i}}} d(x,x_{\mathbf{i}}|x_{\mathbf{i}}) \leq \sqrt{\frac{2L_k}{(2\tilde{T})^\alpha}},
\end{aligned}$$

where the last inequality uses (3), the decomposition $d^2(x,x_{\mathbf{i}}) = [k(x,x)-k(x,x_{\mathbf{i}})]+[k(x_{\mathbf{i}},x_{\mathbf{i}})-k(x,x_{\mathbf{i}})]$ and Assumption (A2). Introduce $\sigma = \max_{\mathbf{i}\in\{1,\dots,\tilde{T}\}^D}\sigma_{\mathbf{i}}$. We have

$$\sigma \leq \sqrt{\frac{2L_k}{(2\tilde{T})^\alpha}}. \quad (4)$$

Proposition 4.2 *For the first term we have that*

$$\mathbb{E} \max_{\mathbf{i}\in\{1,\dots,\tilde{T}\}^D} \big\{U_{\mathbf{i}} - \mathbb{E}[U_{\mathbf{i}}|r(x_{\mathbf{i}})]\big\} \leq 2\sqrt{\frac{L_k \log(T)}{(2\tilde{T})^\alpha}}.$$

Proof. Using Proposition 3.2 we have for any $\mathbf{i}\in\{1,\dots,\tilde{T}\}^D$ and $\lambda>0$ that

$$\begin{aligned}
&\mathbb{E}e^{\lambda\{U_{\mathbf{i}}-\mathbb{E}[U_{\mathbf{i}}|r(x_{\mathbf{i}})]\}} = \mathbb{E}_{r(x_{\mathbf{i}})}\mathbb{E}\big[e^{\lambda\{U_{\mathbf{i}}-\mathbb{E}[U_{\mathbf{i}}|r(x_{\mathbf{i}})]\}}\big|r(x_{\mathbf{i}})\big] \\
&\leq \mathbb{E}_{r(x_{\mathbf{i}})} e^{\lambda^2\sigma_{\mathbf{i}}^2/2} = e^{\lambda^2\sigma_{\mathbf{i}}^2/2} \leq e^{\lambda^2\sigma_X^2/2}.
\end{aligned}$$

We can use now apply Lemma 3.3 and we get that

$$\mathbb{E} \max_{\mathbf{i}=\{1,\dots,\tilde{T}\}^D} \big\{U_{\mathbf{i}} - \mathbb{E}[U_{\mathbf{i}}|r(x_{\mathbf{i}})]\big\} \leq \sigma\sqrt{2\log(\tilde{T}^D)},$$

and conclude by using (4). □

Proposition 4.3 *Under Assumption (A2) we have for the second term that*

$$\mathbb{E} \max_{\mathbf{i}\in\{1,\dots,\tilde{T}\}^D} \mathbb{E}[U_i|r(x_i)] \leq 24\sqrt{\frac{(\alpha+3)DL_k\log(2)}{2\alpha(2\tilde{T})^\alpha}}.$$

Proof. We use the metric entropy bound to control the $\mathbb{E}[U_i|r(x_i)]$ terms together with the Lipschitz assumption. Our bound will be the same for all $\mathbf{i}$ and will be a deterministic quantity. Therefore the outer expectation does not change anything. Let

$$\psi_{\mathbf{i}}^{-1}(\varepsilon) = \inf\big\{\beta>0 : \psi_{\mathbf{i}}(\beta) > \varepsilon\big\}.$$

Since $\psi_{\mathbf{i}}$ is continuous, we have $\psi_{\mathbf{i}}[\psi_{\mathbf{i}}^{-1}(\varepsilon/2)]$, and in particular, $\|x-y\|_\infty \leq \psi_{\mathbf{i}}^{-1}(\frac{\varepsilon}{2})$ implies $d(x,y|x_{\mathbf{i}}) \leq \frac{\varepsilon}{2}$. So a uniform grid of $I_{\mathbf{i}}$ of step $2\psi_{\mathbf{i}}^{-1}(\frac{\varepsilon}{2})$ allows to build a $\frac{\varepsilon}{2}$-covering net, implying that $N(\varepsilon, I_{\mathbf{i}}) \leq \lceil \frac{1}{2\tilde{T}\psi_{\mathbf{i}}^{-1}(\frac{\varepsilon}{2})}\rceil^D$. Using Theorem 3.1, we have

$$\mathbb{E}[U_{\mathbf{i}}|r(x_{\mathbf{i}})] \leq 12\sqrt{D}\int_0^\sigma \sqrt{\log\left(\frac{1}{2\tilde{T}\psi_{\mathbf{i}}^{-1}(\frac{\varepsilon}{2})}+1\right)}d\varepsilon.$$

Since $\psi_{\mathbf{i}}(\beta) \leq \sqrt{2L_k\beta^\alpha}$, we have $\psi_{\mathbf{i}}^{-1}(\frac{\varepsilon}{2}) \geq \big(\frac{\varepsilon}{2\sqrt{2L_k}}\big)^{\frac{2}{\alpha}}$. By using (4) and the integral computations in Appendix A with $a=\alpha/2$, $b=\frac{(2\sqrt{2L_k})^{2/\alpha}}{2\tilde{T}}$ and $c=\sqrt{\frac{2L_k}{(2\tilde{T})^\alpha}}$, we get

$$\begin{aligned}
\int_0^\sigma \sqrt{\log\left(\frac{1}{2\tilde{T}\psi_{\mathbf{i}}^{-1}(\frac{\varepsilon}{2})}+1\right)}d\varepsilon &\leq \int_0^c \sqrt{\log\big(1+b\varepsilon^{-1/a}\big)}d\varepsilon \\
&\leq 2\sqrt{\frac{L_k\log(2e2^{\alpha/2})}{\alpha(2\tilde{T})^\alpha}}.
\end{aligned}$$

□

Proposition 4.4 *Under Assumption (A1) and (A2) we have for the third term that*

$$\mathbb{E}\max_{\mathbf{i}\in\{1,\dots,\tilde{T}\}^D} V_{\mathbf{i}} \le \frac{L_\mu}{2\tilde{T}} + 2\sqrt{\frac{L_k\log(2T)}{(2\tilde{T})^\alpha}}.$$

PROOF. We have

$$V_{\mathbf{i}} = \sup_{x\in I_{\mathbf{i}}}\Big\{\mu(x)-\mu(x_{\mathbf{i}}) + \mathbb{1}_{k(x_{\mathbf{i}},x_{\mathbf{i}})>0}\Big[\frac{k(x,x_{\mathbf{i}})}{k(x_{\mathbf{i}},x_{\mathbf{i}})}-1\Big][r(x_{\mathbf{i}})-\mu(x_{\mathbf{i}})]\Big\}.$$

Let $a_{\mathbf{i}} = \sup_{x\in I_{\mathbf{i}}}\big|\frac{k(x,x_{\mathbf{i}})}{k(x_{\mathbf{i}},x_{\mathbf{i}})}-1\big|\mathbb{1}_{k(x_{\mathbf{i}},x_{\mathbf{i}})>0}$. We have

$$V_{\mathbf{i}} \le \sup_{x\in I_{\mathbf{i}}}\big\{\mu(x)-\mu(x_{\mathbf{i}})\big\} + \big|a_{\mathbf{i}}[\mu(x_{\mathbf{i}})-r(x_{\mathbf{i}})]\big|.$$

$a_{\mathbf{i}}[\mu(x_{\mathbf{i}})-r(x_{\mathbf{i}})]$ is a one dimensional Gaussian random variable with mean zero. Now using a property from the moment generating function of a one dimensional Gaussian we find that for any $\lambda\in\mathbb{R}$

$$\mathbb{E}e^{\lambda a_{\mathbf{i}}[\mu(x_{\mathbf{i}})-r(x_{\mathbf{i}})]} \le e^{\lambda^2 a_{\mathbf{i}}^2 k(x_{\mathbf{i}},x_{\mathbf{i}})/2}.$$

Using again Lemma 3.3, we have

$$\mathbb{E}\max_{\mathbf{i}=\{1,\dots,\tilde{T}\}^D} V_{\mathbf{i}} \le \max_{\mathbf{i}=\{1,\dots,\tilde{T}\}^D}\big\{\mu(x)-\mu(x_{\mathbf{i}})\big\} + \sqrt{2\log(2\tilde{T}^D)\max_{\mathbf{i}=\{1,\dots,\tilde{T}\}^D} a_{\mathbf{i}}^2 k(x_{\mathbf{i}},x_{\mathbf{i}})}.$$

The last term can be bounded in the following way

$$a_{\mathbf{i}}^2 k(x_{\mathbf{i}},x_{\mathbf{i}}) \le \sup_{x\in I_{\mathbf{i}}}\frac{[k(x,x_{\mathbf{i}})-k(x_{\mathbf{i}},x_{\mathbf{i}})]^2}{k(x_{\mathbf{i}},x_{\mathbf{i}})}\mathbb{1}_{k(x_{\mathbf{i}},x_{\mathbf{i}})>0} \le \sup_{x\in I_{\mathbf{i}}} d^2(x,x_{\mathbf{i}}) \underset{\mathbf{(A2)}}{\le} \frac{2L_k}{(2\tilde{T})^\alpha},$$

where the Cauchy-Schwarz inequality was used at the second inequality. □

Plugging the results of the three previous propositions into (2) leads to the announced result. □

4.4 A LOWER BOUND

In this section we present a lower bound to complement the upper bound from Theorem 4.1. On the positive side, this lower bound matches the leading term (in T) of the upper bound up to a logarithmic factor, and it holds for any dimension D. On the negative side, the bound is obtained by constructing a specific Gaussian process. Therefore, it does not apply to a specific kernel. Yet, it tells us that under our general assumptions a big improvement cannot be gained.

Theorem 4.5 *Let*

$$\kappa_T = \sqrt{\log T}\;\mathbb{E}\Big\{\max_{i=1,\dots,2T} Y_i - \max_{i=1,\dots,T} Y_i\Big\}$$

where $Y_1,\dots,Y_{2T}$ are independent standard normal random variables. For any $D\ge 1$, $L_k\ge 0$, $L_\mu\ge 0$, $0<\alpha\le 1$ and $T\in\big\{\frac{1}{2},\frac{2^D}{2},\frac{3^D}{2},\dots\big\}$, there exists a Gaussian process r defined on $[0,1]^D$ satisfying Assumptions (A1) and (A2) such that

$$\mathbb{E}\Big\{\sup_{x\in X} r(x) - \max\big(r(\hat{x}_1),\dots,r(\hat{x}_T)\big)\Big\} \ge \kappa_T\sqrt{\frac{L_k}{2^\alpha(2T)^{\alpha/D}\log T}} \ge \kappa\sqrt{\frac{L_k}{2^\alpha(2T)^{\alpha/D}\log T}},$$

for some universal constant $\kappa>0$.

PROOF. Let $m=(2T)^{1/D}$. By assumption on T, m is an integer. Define $h=\sqrt{L_k(2m)^\alpha}$. To prove the lower bound, we consider a Gaussian process $r(x)=\sum_{i=1}^{2T}\phi_i(x)Y_i$, where ϕ_i are real-valued functions with disjoint supports (up to the boundaries) and $Y_1,\dots,Y_{2T}$ are i.i.d. centered normal random variables with unit variance. Consider the regular grid on $[0,1]^D$ defined by

$$G=\Big\{\Big(\frac{2\mathbf{i}_1-1}{2m},\dots,\frac{2\mathbf{i}_D-1}{2m}\Big) : (\mathbf{i}_1,\dots,\mathbf{i}_D)\in\{1,\dots,m\}^D\Big\}.$$

This grid has $2T$ points. Let $s_1,\dots,s_{2T}$ denote these points (the ordering of these points has no importance). Define the function ϕ by

$$\phi(u)=\begin{cases} h\big[(2m)^{-\alpha}-\|u\|_\infty^\alpha\big] & \text{if } \|u\|_\infty\le(2m)^{-1}\\ 0 & \text{otherwise}\end{cases}$$

We consider the functions $\phi_i(x)=\phi(x-s_i)$ so that the Gaussian process is simply

$$r(x)=\sum_{i=1}^{2T}\phi(x-s_i)Y_i.$$

The function ϕ is continuous and "peaked around zero". Let us check that for any $s,t\in\mathbb{R}^D$, we have

$$\big|\phi(t)-\phi(s)\big|\le h\|t-s\|_\infty^\alpha. \tag{5}$$

First it is easy to see that we only need to prove this inequality for s and t in the cubic ball $B=\big\{u\in\mathbb{R}^D : \|u\|_\infty\le(2m)^{-1}\big\}$. Now, since for $0<\alpha\le 1$ and

any nonnegative numbers a and b we have $(a+b)^\alpha \leq a^\alpha + b^\alpha$, for any s and t in B, we obtain

$$\begin{aligned}\big|\phi(t)-\phi(s)\big| &\leq h\big|\|t\|_\infty^\alpha - \|s\|_\infty^\alpha\big| \\ &\leq h\big|\|t\|_\infty - \|s\|_\infty\big|^\alpha \leq h\|t-s\|_\infty^\alpha,\end{aligned}$$

so that (5) holds. We now use (5) to prove that Assumption (A2) is satisfied. We have

$$\begin{aligned}k(s,t) = \mathbb{E}[r(s)r(t)] &= \sum_{i=1}^{2T} \phi(s-s_i)\phi(t-s_i) \\ &= \phi(s-s_j)\phi(t-s_j),\end{aligned}$$

for j an integer such that $t - s_j \in B$. For such a j, from (5), we can also write

$$\begin{aligned}\big|k(t,t)-k(s,t)\big| &= \big|\phi(t-s_j)\big[\phi(t-s_j)-\phi(s-s_j)\big]\big| \\ &\leq h^2(2m)^{-\alpha}\|t-s\|_\infty^\alpha \\ &= L_k\|t-s\|_\infty^\alpha.\end{aligned}$$

So Assumptions (A1) and (A2) are satisfied (for any $L_\mu \geq 0$, since $\mu(x) = 0$ for any $x \in [0,1]^D$).

Whatever policy is used, after T observations of the Gaussian process, among the $2T$ random variables $Y_1, \ldots, Y_{2T}$, we know only the values of T of them, and we get absolutely no information on the T remaining ones[1]. Without loss of generality, let us consider that these T values correspond to $Y_1, \ldots, Y_T$. We thus have

$$\begin{aligned}&\mathbb{E}\left\{ \sup_{x\in X} r(x) - \max\big(r(\hat{x}_1),\ldots,r(\hat{x}_T)\big)\right\} \\ &\geq\ h(2m)^{-\alpha}\mathbb{E}\left\{ \max_{i=1,\ldots,2T} Y_i - \max_{i=1,\ldots,T} Y_i\right\} \\ &=\ \kappa_T\sqrt{\frac{L_k}{(2m)^\alpha \log T}}.\end{aligned}$$

This is the first inequality of the theorem. To obtain the second inequality, we will prove that there is an absolute constant $\kappa > 0$ such that $\kappa_T > \kappa$ for any $T \geq 1$. To do this, since κ_T is positive for any $T \geq 1$, it suffices to prove that there exists u_T such that $\kappa_T \geq u_T$ for any large enough T and u_T converges to a positive constant when T tends to infinity. Define

$$\begin{aligned}a_T &= 2\log\left(\frac{T}{\sqrt{\log(2T)}}\right), \\ M_1 &= \max_{i=1,\ldots,T} Y_i, \qquad M_2 = \max_{i=T+1,\ldots,2T} Y_i, \\ \Phi(t) &= \mathbb{P}(Y_1 > t), \text{ and} \\ u_T &= \sqrt{\log T}\,\big(\sqrt{a_T + 2\log 2} - \sqrt{a_T}\big) \\ &\quad \times \Big[1 - \mathbb{P}\big(M_1 \leq \sqrt{a_T + 2\log 2}\big)\Big]\mathbb{P}\big(M_1 \leq \sqrt{a_T}\big).\end{aligned}$$

[1] A different viewpoint is to compute explicitly the optimal policy in this setting. It is easy to see that the policy stems down to playing a different grid point at each time step.

For $T \geq 2$, we have

$$\begin{aligned}\kappa_T &= \sqrt{\log T}\,\mathbb{E}\max(0, M_2 - M_1) \\ &\geq \sqrt{\log T}\,\big(\sqrt{a_T + 2\log 2} - \sqrt{a_T}\big) \\ &\qquad \times \mathbb{P}\big(M_2 \geq \sqrt{a_T + 2\log 2}; M_1 \leq \sqrt{a_T}\big) \\ &= u_T.\end{aligned}$$

To obtain an equivalent on u_T when T goes to infinity, we use the well-known bound on tails of the normal distribution (e.g., see Pollard (2002, Appendix D)):

$$\frac{\exp(-\frac{t^2}{2})}{t\sqrt{2\pi}}\left(1 - \frac{1}{t^2}\right) \leq \Phi(t) \leq \frac{\exp(-\frac{t^2}{2})}{t\sqrt{2\pi}}.$$

Consequently, we obtain

$$\mathbb{P}\big(M_1 \leq \sqrt{a_T}\big) = [1 - \Phi(\sqrt{a_T})]^T \underset{T\to\infty}{\longrightarrow} \exp\left(-\frac{1}{\sqrt{4\pi}}\right)$$

and similarly

$$\mathbb{P}\big(M_1 \leq \sqrt{a_T + 2\log 2}\big) \underset{T\to\infty}{\longrightarrow} \exp\left(-\frac{1}{2\sqrt{4\pi}}\right).$$

Finally, elementary computations give

$$\sqrt{\log T}\,\big(\sqrt{a_T + 2\log 2} - \sqrt{a_T}\big) \underset{T\to\infty}{\longrightarrow} \frac{\log 2}{\sqrt{2}}.$$

Putting together the three last results, we obtain that u_T converges to a positive constant when T goes to infinity, hence there exists $\kappa > 0$ such that $\kappa_T \geq \kappa$ for any $T \geq 1$. □

5 CONCLUSIONS AND FUTURE WORK

The main contribution of the paper has been to our knowledge the first analysis of a new bandit scenario based on a Gaussian process model of the reward function. The results bound the regret (relative to always playing the optimal arm) of particular strategies in the case where there is no noise in the observed reward and the kernel satisfies benign continuity assumptions. We provide lower bounds that show our bounds are in general at most a logarithmic factor away from optimal.

We intend to extend this work to model of rewards incorporating a noise term so that the true reward would not then be directly observed. A more ambitious goal is to develop reward bounds for algorithms that merge the exploration and exploitation phases through for example selecting arms based on an upper confidence bound. It will also be interesting to see if more detail properties of the kernel (other than Hölder continuity) can be incorporated into the analysis to provide

tighter bounds for special types of kernel, such as for example specific spectral properties.

We believe that the line of analysis developed here will provide important groundwork for these further studies and will at the same time help to inform improvements in practical Gaussian process bandit algorithms.

A Dudley integral computations

To bound the Dudley integral, we use

$$\int_0^c \sqrt{\log(1+b\varepsilon^{-1/a})}d\varepsilon \leq c\sqrt{\frac{\log(e2^{a+1})}{a}}. \quad (6)$$

which holds for any a, b and c such that $b^a = 2c$. Indeed, letting $\xi = (1+2^{-1/a})^a$, we have

$$\int_0^c \sqrt{\log(1+b\varepsilon^{-1/a})}d\varepsilon \leq \int_0^c \sqrt{\log\left(\xi^a b\varepsilon^{-1/a}\right)}d\varepsilon$$
$$= \xi b^a \int_0^{\frac{1}{2\xi}} \sqrt{\log\left(u^{-1/a}\right)}du = \frac{\xi b^a}{\sqrt{a}} \int_0^{\frac{1}{2\xi}} \sqrt{-\log u}du$$
$$\leq \frac{\xi b^a}{\sqrt{2a\xi}} \sqrt{-\int_0^{\frac{1}{2\xi}} \log(u)\ du} \leq c\sqrt{\frac{\log(e2^{a+1})}{a}}.$$

Acknowledgements
We would like to thank the European Union for funding through PASCAL 2 and the ARAGORN project. The second author would like to acknowledge the French National Research Agency (ANR) through COSINUS program (project EXPLO-RA, ANR-08-COSI-004).

References

J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In Servedio and Zhang (2008), pages 263–274.

R. Agrawal. The continuum-armed bandit problem. *SIAM J. Control and Optimization*, 33:1926–1951, 1995.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *M. Lear.*, 2002.

P. Auer, R. Ortner, and C. Szepesvári. Improved rates for the stochastic continuum-armed bandit problem. *20th COLT, San Diego, CA, USA*, 2007.

R. Bellman. A problem in the sequential design of experiments. *Sankhyā: The Indian Journal of Statistics (1933-1960)*, 16(3):221–229, 1956.

D. Berry and B. Fristedt. *Bandit problems*. London: Chapman and Hall, 1985.

S. Bubeck, R. Munos, G. Stoltz, and C. Szepesvari. Online optimization in X-armed bandits. *NIPS*, 21, 2008.

S. Bubeck, R. Munos, and G. Stoltz. Pure exploration in multi-armed bandits problems. In *Proc. of the 20th International Conference on Algorithmic Learning Theory*, 2009.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

L. Devroye and G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer, 2001.

R. Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. *Journal of Functional Analysis*, 1967.

D. Ginsbourger and R. L. Riche. Towards gp-based optimization with finite time horizon. http://hal.archives-ouvertes.fr/hal-00424309/en/, 2009.

D. Ginsbourger, R. Le Riche, and L. Carraro. A Multi-points Criterion for Deterministic Parallel Global Optimization based on Gaussian Processes. c. DICE, 2008.

J. Gittins and D. Jones. A dynamic allocation index for the discounted multiarmed bandit problem. *Biometrika*, 66(3):561–565, 1979.

D. Jones. A taxonomy of global optimization methods based on response surfaces. *J. of Global Opt.*, 2001.

D. Jones, M. Schonlau, and W. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 1998.

R. Kleinberg. Nearly tight bounds for the continuum-armed bandit problem. In *NIPS*, 2004.

R. Kleinberg, A. Slivkins, and E. Upfal. Multi-armed bandit problems in metric spaces. In *Proceedings of the 40th ACM Symposium on Theory of Computing*, 2008a.

R. D. Kleinberg, A. Niculescu-Mizil, and Y. Sharma. Regret bounds for sleeping experts and bandits. In Servedio and Zhang (2008), pages 425–436.

H. Kushner. A new methods of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *Journal of Basic Engineering*, 1964.

P. Massart. *Concentration inequalities and model selection: Ecole d'été de Probabilités de Saint-Flour XXXIII - 2003*. Lecture Notes in Mathematics. Springer, 2003.

M. A. Osborne, R. Garnett, and S. J. Roberts. Gaussian processes for global optimization. In *LION3*, 2009.

D. Pollard. *A user's guide to measure theoretic probability*. Cambridge University Press, 2002.

C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006.

H. Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematics Society*, 58:527–535, 1952.

S. Ross. *Applied Probability Models with Optimization Applications*. Holden-Day, San Francisco, CA, 1970.

M. Schonlau. *Computer Experiments and Global Optimization*. PhD thesis, University of Waterloo, 1997.

R. A. Servedio and T. Zhang, editors. *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland, July 9-12, 2008*, 2008. Omnipress.

A. Slivkins and E. Upfal. Adapting to a changing environment: the brownian restless bandits. In Servedio and Zhang (2008), pages 343–354.

N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process bandits without regret: An experimental design approach. 2009. arXiv:0912.3995v3.

Y. Wang, J.-Y. Audibert, and R. Munos. Algorithms for infinitely many-armed bandits. *NIPS*, 21, 2008.