On Kernelized Multi-armed Bandits

Sayak Ray Chowdhury [1] **Aditya Gopalan** [1]

Abstract

We consider the stochastic bandit problem with a continuous set of arms, with the expected reward function over the arms assumed to be fixed but unknown. We provide two new Gaussian process-based algorithms for continuous bandit optimization – Improved GP-UCB (IGP-UCB) and GP-Thomson sampling (GP-TS), and derive corresponding regret bounds. Specifically, the bounds hold when the expected reward function belongs to the reproducing kernel Hilbert space (RKHS) that naturally corresponds to a Gaussian process kernel used as input by the algorithms. Along the way, we derive a new self-normalized concentration inequality for vector-valued martingales of arbitrary, possibly infinite, dimension. Finally, experimental evaluation and comparisons to existing algorithms on synthetic and real-world environments are carried out that highlight the favorable gains of the proposed strategies in many cases.

1. Introduction

Optimization over large domains under uncertainty is an important subproblem arising in a variety of sequential decision making problems, such as dynamic pricing in economics (Besbes & Zeevi, 2009), reinforcement learning with continuous state/action spaces (Kaelbling et al., 1996; Smart & Kaelbling, 2000), and power control in wireless communication (Chiang et al., 2008). A typical feature of such problems is a large, or potentially infinite, domain of decision points or covariates (prices, actions, transmit powers), together with only partial and noisy observability of the associated outcomes (demand, state/reward, communication rate); reward/loss information is revealed only for decisions that are chosen. This often makes it hard to balance exploration and exploitation, as available knowledge must be transferred efficiently from a finite set of observations so far to estimates of the values of infinitely many decisions. A classic case in point is that of the canonical stochastic MAB with finitely many arms, where the effort to optimize scales with the total number of arms or decisions; the effect of this is catastrophic for large or infinite arm sets.

With suitable structure in the values or rewards of arms, however, the challenge of sequential optimization can be efficiently addressed. Parametric bandits, especially linearly parameterized bandits (Rusmevichientong & Tsitsiklis, 2010), represent a well-studied class of structured decision making settings. Here, every arm corresponds to a known, finite dimensional vector (its feature vector), and its expected reward is assumed to be an unknown linear function of its feature vector. This allows for a large, or even infinite, set of arms all lying in space of finite dimension, say d, and a rich line of work gives algorithms that attain sublinear regret with a polynomial dependence on the dimension, e.g., Confidence Ball (Dani et al., 2008), OFUL (Abbasi-Yadkori et al., 2011) (a strengthening of Confidence Ball) and Thompson sampling for linear bandits (Agrawal & Goyal, 2013)[1] The insight here is that even though the number of arms can be large, the number of unknown parameters (or degrees of freedom) in the problem is really only d, which makes it possible to learn about the values of many other arms by playing a single arm.

A different approach to modelling bandit problems with a continuum of arms is via the framework of Gaussian processes (GPs) (Rasmussen & Williams, 2006). GPs are a flexible class of nonparametric models for expressing uncertainty over functions on rather general domain sets, which generalize multivariate Gaussian random vectors. GPs allow tractable regression for estimating an unknown function given a set of (noisy) measurements of its values at chosen domain points. The fact that GPs, being distributions on functions, can also help quantify function uncertainty makes it attractive for basing decision making strategies on them. This has been exploited to great advantage to

[1] Department of Electrical Communication Engineering, Indian Institute of Science, Bengaluru, 560012, India. Correspondence to: Sayak Ray Chowdhury <srchowdhury@ece.iisc.ernet.in>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017. Copyright 2017 by the author(s).

[1] Roughly, for rewards bounded in $[-1, 1]$, these algorithms achieve optimal regret $\tilde{O}\left(d\sqrt{T}\right)$, where $\tilde{O}\left(\cdot\right)$ hides polylog(T) factors.

build nonparametric bandit algorithms, such as GP-UCB (Srinivas et al., 2009), GP-EI and GP-PI (Hoffman et al., 2011). In fact, GP models for bandit optimization, in terms of their kernel maps, can be viewed as the parametric linear bandit paradigm pushed to the extreme, where each feature vector associated to an arm can have infinite dimension [2].

Against this backdrop, our work revisits the problem of bandit optimization with stochastic rewards. Specifically, we consider stochastic multiarmed bandit (MAB) problems with a continuous arm set, and whose (unknown) expected reward function is assumed to lie in a reproducing kernel Hilbert space (RKHS), with bounded RKHS norm – this effectively enforces smoothness on the function[3]. We make the following contributions-

- We design a new algorithm – Improved Gaussian Process-Upper Confidence Bound (IGP-UCB) – for stochastic bandit optimization. The algorithm can be viewed as a variant of GP-UCB (Srinivas et al., 2009), but uses a significantly reduced confidence interval width resulting in an order-wise improvement in regret compared to GP-UCB. IGP-UCB also shows a markedly improved numerical performance over GP-UCB.

- We develop a nonparametric version of Thompson sampling, called Gaussian Process Thompson sampling (GP-TS), and show that enjoys a regret bound of $\tilde{O}\left(\gamma_T\sqrt{dT}\right)$. Here, T is the total time horizon and γ_T is a quantity depending on the RKHS containing the reward function. This is, to our knowledge, the first known regret bound for Thompson sampling in the agnostic setup with nonparametric structure.

- We prove a new self-normalized concentration inequality for infinite-dimensional vector-valued martingales, which is not only key to the design and analysis of the IGP-UCB and GP-TS algorithms, but also potentially of independent interest. The inequality generalizes a corresponding self-normalized bound for martingales in finite dimension proven by Abbasi-Yadkori et al. (2011).

- Empirical comparisons of the algorithms developed above, with other GP-based algorithms, are presented, over both synthetic and real-world setups, demonstrating performance improvements of the proposed algorithms, as well as their performance under misspecification.

2. Problem Statement

We consider the problem of sequentially maximizing a fixed but unknown reward function $f : D \rightarrow \mathbb{R}$ over a (potentially infinite) set of decisions $D \subset \mathbb{R}^d$, also called actions or arms. An algorithm for this problem chooses, at each round t, an arm $x_t \in D$, and subsequently observes a reward $y_t = f(x_t) + \varepsilon_t$, which is a noisy version of the function value at x_t. The arm x_t is chosen causally depending upon the arms played and rewards obtained upto round $t-1$, denoted by the history $\mathcal{H}_{t-1} = \{(x_s, y_s) : s = 1, \ldots, t-1\}$. We assume that the noise sequence $\{\varepsilon_t\}_{t=1}^{\infty}$ is conditionally R-sub-Gaussian for a fixed constant $R \geq 0$, i.e.,

$$\forall t \geq 0,\ \ \forall \lambda \in \mathbb{R},\ \ \mathbb{E}\left[e^{\lambda \varepsilon_t} \;\middle|\; \mathcal{F}_{t-1}\right] \leq \exp\left(\frac{\lambda^2 R^2}{2}\right), \tag{1}$$

where $\mathcal{F}_{t-1}$ is the σ-algebra generated by the random variables $\{x_s, \varepsilon_s\}_{s=1}^{t-1}$ and x_t.This is a mild assumption on the noise (it holds, for instance, for distributions bounded in $[-R, R]$) and is standard in the bandit literature (Abbasi-Yadkori et al., 2011; Agrawal & Goyal, 2013).

Regret. The goal of an algorithm is to maximize its cumulative reward or alternatively minimize its cumulative *regret* – the loss incurred due to not knowing f's maximum point beforehand. Let $x^\star \in \mathrm{argmax}_{x \in D} f(x)$ be a maximum point of f (assuming the maximum is attained). The instantaneous regret incurred at time t is $r_t = f(x^\star) - f(x_t)$, and the cumulative regret in a time horizon T (not necessarily known a priori) is defined to be $R_T = \sum_{t=1}^{T} r_t$. A sub-linear growth of R_T in T signifies that $R_T/T \rightarrow 0$ as $T \rightarrow \infty$, or vanishing per-round regret.

Regularity Assumptions. Attaining sub-linear regret is impossible in general for arbitrary reward functions f and domains D, and thus some regularity assumptions are in order. In what follows, we assume that D is compact. The smoothness assumption we make on the reward function f is motivated by Gaussian processes[4] and their associated reproducing kernel Hilbert spaces (RKHSs, see Schölkopf & Smola (2002)). Specifically, we assume that f has small norm in the RKHS of functions $D \rightarrow \mathbb{R}$, with positive semi-definite kernel function $k : D \times D \rightarrow \mathbb{R}$. This RKHS, denoted by $H_k(D)$, is completely specified by its kernel function $k(\cdot, \cdot)$ and vice-versa, with an inner product $\langle \cdot, \cdot \rangle_k$ obeying the reproducing property: $f(x) = \langle f, k(x, \cdot) \rangle_k$ for all $f \in H_k(D)$. In other words, the kernel plays the role of delta functions to represent the evaluation map at each point $x \in D$ via the RKHS inner product. The RKHS norm $\lVert f \rVert_k = \sqrt{\langle f, f \rangle_k}$ is a measure of the smoothness[5]

[2] The completion of the linear span of all feature vectors (images of the kernel map) is precisely the reproducing kernel Hilbert space (RKHS) that characterizes the GP.

[3] Kernels, and their associated RKHSs,

[4] Other work has also studied continuum-armed bandits with weaker smoothness assumptions such as Lipschitz continuity – see Related work for details and comparison.

[5] One way to see this is that for every element g in the RKHS, $|g(x) - g(y)| = |\langle g, k(x, \cdot) - k(y, \cdot) \rangle| \leq \lVert g \rVert_k \lVert k(x, \cdot) - k(y, \cdot) \rVert_k$ by Cauchy-Schwarz.

of f, with respect to the kernel function k, and satisfies: $f \in H_k(D)$ if and only if $\|f\|_k < \infty$.

We assume a known bound on the RKHS norm of the unknown target function[6]: $\|f\|_k \leq B$. Moreover, we assume bounded variance by restricting $k(x,x) \leq 1$, for all $x \in D$. Two common kernels that satisfy bounded variance property are *Squared Exponential* and *Matérn*, defined as

$$\begin{aligned} k_{SE}(x,x') &= \exp\Big(- s^2/2l^2\Big), \\ k_{Matérn}(x,x') &= \frac{2^{1-\nu}}{\Gamma(\nu)}\Big(\frac{s\sqrt{2\nu}}{l}\Big)^{\nu} B_{\nu}\Big(\frac{s\sqrt{2\nu}}{l}\Big), \end{aligned}$$

where $l > 0$ and $\nu > 0$ are hyperparameters, $s = \|x - x'\|_2$ encodes the similarity between two points $x, x' \in D$, and $B_\nu(\cdot)$ is the modified Bessel function. Generally the bounded variance property holds for any stationary kernel, i.e. kernels for which $k(x,x') = k(x-x')$ for all $x, x' \in \mathbb{R}^d$. These assumptions are required to make the regret bounds scale-free and are standard in the literature (Agrawal & Goyal, 2013). Instead if $k(x,x) \leq c$ or $\|f\|_k \leq cB$, then our regret bounds would increase by a factor of c.

3. Algorithms

Design philosophy. Both the algorithms we propose use Gaussian likelihood models for observations, and Gaussian process (GP) priors for uncertainty over reward functions. A Gaussian process over D, denoted by $GP_D(\mu(\cdot), k(\cdot,\cdot))$, is a collection of random variables $(f(x))_{x\in D}$, one for each $x \in D$, such that every finite sub-collection of random variables $(f(x_i))_{i=1}^m$ is jointly Gaussian with mean $\mathbb{E}[f(x_i)] = \mu(x_i)$ and covariance $\mathbb{E}[(f(x_i)-\mu(x_i))(f(x_j)-\mu(x_j))] = k(x_i,x_j)$, $1 \leq i,j \leq m$, $m \in \mathbb{N}$. The algorithms use $GP_D(0, v^2 k(\cdot,\cdot))$, $v > 0$, as an initial prior distribution for the unknown reward function f over D, where $k(\cdot,\cdot)$ is the kernel function associated with the RKHS $H_k(D)$ in which f is assumed to have 'small' norm at most B. The algorithms also assume that the noise variables $\varepsilon_t = y_t - f(x_t)$ are drawn independently, across t, from $\mathcal{N}(0, \lambda v^2)$, with $\lambda \geq 0$. Thus, the prior distribution for each $f(x)$, is assumed to be $\mathcal{N}(0, v^2 k(x,x))$, $x \in D$. Moreover, given a set of sampling points $A_t = (x_1, \ldots, x_t)$ within D, it follows under the assumption that the corresponding vector of observed rewards $y_{1:t} = [y_1, \ldots, y_t]^T$ has the multivariate Gaussian distribution $\mathcal{N}(0, v^2(K_t + \lambda I))$, where $K_t = [k(x,x')]_{x,x' \in A_t}$ is the kernel matrix at time t. Then, by the properties of GPs, we have that $y_{1:t}$ and $f(x)$ are jointly Gaussian given A_t:

$$\begin{bmatrix} f(x) \\ y_{1:t} \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} v^2 k(x,x) & v^2 k_t(x)^T \\ v^2 k_t(x) & v^2(K_t + \lambda I) \end{bmatrix}\right),$$

where $k_t(x) = [k(x_1,x), \ldots, k(x_t,x)]^T$. Therefore conditioned on the history $\mathcal{H}_t$, the posterior distribution over f is $GP_D(\mu_t(\cdot), v^2 k_t(\cdot,\cdot))$, where

$$\begin{aligned} \mu_t(x) &= k_t(x)^T (K_t + \lambda I)^{-1} y_{1:t}, & (2) \\ k_t(x,x') &= k(x,x') - k_t(x)^T (K_t + \lambda I)^{-1} k_t(x'),(3) \\ \sigma_t^2(x) &= k_t(x,x). & (4) \end{aligned}$$

Thus for every $x \in D$, the posterior distribution of $f(x)$, given $\mathcal{H}_t$, is $\mathcal{N}(\mu_t(x), v^2 \sigma_t^2(x))$.

Remark. Note that the GP prior and Gaussian likelihood model described above is only an aid to algorithm design, and has nothing to do with the actual reward distribution or noise model as in the problem statement (Section 2). The reward function f is a fixed, unknown, member of the RKHS $H_k(D)$, and the true sequence of noise variables ε_t is allowed to be a conditionally R-sub-Gaussian martingale difference sequence (Equation 1). In general, thus, this represents a misspecified prior and noise model, also termed the *agnostic* setting by Srinivas et al. (2009).

The proposed algorithms, to follow, assume the knowledge of only the sub-Gaussianity parameter R, kernel function k and upper bound B on the RKHS norm of f. Note that v, λ are free parameters (possibly time-dependent) that can be set specific to the algorithm.

3.1. Improved GP-UCB (IGP-UCB) Algorithm

We introduce the IGP-UCB algorithm (Algorithm 1), that uses a combination of the current posterior mean $\mu_{t-1}(x)$ and standard deviation $v\sigma_{t-1}(x)$ to (a) construct an upper confidence bound (UCB) envelope for the actual function f over D, and (b) choose an action to maximize it. Specifically it chooses, at each round t, the action

$$x_t = \operatorname*{argmax}_{x \in D} \mu_{t-1}(x) + \beta_t \sigma_{t-1}(x), \tag{5}$$

with the scale parameter v set to be 1. Such a rule trades off exploration (picking points with high uncertainty $\sigma_{t-1}(x)$) with exploitation (picking points with high reward $\mu_{t-1}(x)$), with $\beta_t = B + R\sqrt{2(\gamma_{t-1} + 1 + \ln(1/\delta))}$ being the parameter governing the tradeoff, which we later show is related to the width of the confidence interval for f at round t. $\delta \in (0,1)$ is a free *confidence* parameter used by the algorithm, and γ_t is the *maximum information gain* at time t, defined as:

$$\gamma_t := \max_{A \subset D : |A| = t} I(y_A; f_A).$$

Here, $I(y_A; f_A)$ denotes the *mutual information* between $f_A = [f(x)]_{x \in A}$ and $y_A = f_A + \varepsilon_A$, where $\varepsilon_A \sim \mathcal{N}(0, \lambda v^2 I)$ and quantifies the reduction in uncertainty about f after observing y_A at points $A \subset D$. γ_t is a problem dependent quantity and can be found given the knowledge of domain D and kernel function k. For

[6] This is analogous to the bound on the weight θ typically assumed in regret analyses of linear parametric bandits.

a compact subset D of $\mathbb{R}^d$, γ_T is $O((\ln T)^{d+1})$ and $O(T^{d(d+1)/(2\nu+d(d+1))}\ln T)$, respectively, for the Squared Exponential and Matérn kernels (Srinivas et al., 2009), depending only polylogarithmically on the time T.

Algorithm 1 Improved-GP-UCB (IGP-UCB)

```
  Input: Prior $GP(0,k)$, parameters $B$, $R$, $\lambda$, $\delta$.
  for t = 1, 2, 3 ... T do
    Set $\beta_t = B + R\sqrt{2(\gamma_{t-1} + 1 + \ln(1/\delta))}$.
    Choose $x_t = \operatorname*{argmax}_{x\in D} \mu_{t-1}(x) + \beta_t\sigma_{t-1}(x)$.
    Observe reward $y_t = f(x_t) + \varepsilon_t$.
    Perform update to get $\mu_t$ and $\sigma_t$ using 2, 3 and 4.
  end for
```

Discussion. Srinivas et al. (2009) have proposed the GP-UCB algorithm, and Valko et al. (2013) the KernelUCB algorithm, for sequentially optimizing reward functions lying in the RKHS $H_k(D)$. Both algorithms play an arm at time t using the rule: $x_t = \operatorname{argmax}_{x\in D}\mu_{t-1}(x) + \tilde{\beta}_t\sigma_{t-1}(x)$. GP-UCB uses the exploration parameter $\tilde{\beta}_t = \sqrt{2B^2 + 300\gamma_{t-1}\ln^3(t/\delta)}$, with λ set to σ^2, where σ is additionally assumed to be a known, uniform (i.e., almost-sure) upper bound on all noise variables ε_t (Srinivas et al., 2009, Theorem 3). Compared to GP-UCB, IGP-UCB (Algorithm 1) reduces the width of the confidence interval by a factor roughly $O(\ln^{3/2} t)$ at every round t, and, as we will see, this small but critical adjustment leads to much better theoretical and empirical performance compared to GP-UCB. In KernelUCB, $\tilde{\beta}_t$ is set as $\eta/\lambda^{1/2}$, where η is the exploration parameter and λ is the regularization constant. Thus IGP-UCB can be viewed as a special case of KernelUCB where $\eta = \beta_t$.

3.2. Gaussian Process Thompson Sampling (GP-TS)

Our second algorithm, GP-TS (Algorithm 2), inspired by the success of Thompson sampling for standard and parametric bandits (Agrawal & Goyal, 2012; Kaufmann et al., 2012; Gopalan et al., 2014; Agrawal & Goyal, 2013), uses the time-varying scale parameter $v_t = B + R\sqrt{2(\gamma_{t-1} + 1 + \ln(2/\delta))}$ and operates as follows. At each round t, GP-TS samples a random function $f_t(\cdot)$ from the GP with mean function $\mu_{t-1}(\cdot)$ and covariance function $v_t^2 k_{t-1}(\cdot,\cdot)$. Next, it chooses a decision set $D_t \subset D$, and plays the arm $x_t \in D_t$ that maximizes f_t[7]. We call it GP-Thompson-Sampling as it falls under the general framework of Thompson Sampling, i.e., (a) assume a prior on the underlying parameters of the reward distribution, (b) play the arm according to the prior probability that it is optimal, and (c) observe the outcome and update the prior. However, note that the prior is nonparametric in this case.

Algorithm 2 GP-Thompson-Sampling (GP-TS)

```
  Input: Prior $GP(0,k)$, parameters $B$, $R$, $\lambda$, $\delta$.
  for t = 1, 2, 3 ..., do
    Set $v_t = B + R\sqrt{2(\gamma_{t-1} + 1 + \ln(2/\delta))}$.
    Sample $f_t(\cdot)$ from $GP_D(\mu_{t-1}(\cdot), v_t^2 k_{t-1}(\cdot,\cdot))$.
    Choose the current decision set $D_t \subset D$.
    Choose $x_t = \operatorname*{argmax}_{x\in D_t} f_t(x)$.
    Observe reward $y_t = f(x_t) + \varepsilon_t$.
    Perform update to get $\mu_t$ and $k_t$ using 2 and 3.
  end for
```

4. Main Results

We begin by presenting two key concentration inequalities which are essential in bounding the regret of the proposed algorithms.

Theorem 1 *Let $\{x_t\}_{t=1}^{\infty}$ be an $\mathbb{R}^d$-valued discrete time stochastic process predictable with respect to the filtration $\{\mathcal{F}_t\}_{t=0}^{\infty}$, i.e., x_t is $\mathcal{F}_{t-1}$-measurable $\forall t \geq 1$. Let $\{\varepsilon_t\}_{t=1}^{\infty}$ be a real-valued stochastic process such that for some $R \geq 0$ and for all $t \geq 1$, ε_t is (a) $\mathcal{F}_t$-measurable, and (b) R-sub-Gaussian conditionally on $\mathcal{F}_{t-1}$. Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a symmetric, positive-semidefinite kernel, and let $0 < \delta \leq 1$. For a given $\eta > 0$, with probability at least $1-\delta$, the following holds simultaneously over all $t \geq 0$:*

$$\left\lVert \varepsilon_{1:t} \right\rVert^2_{((K_t+\eta I)^{-1}+I)^{-1}} \leq 2R^2 \ln \frac{\sqrt{\det((1+\eta)I + K_t)}}{\delta}. \tag{6}$$

(Here, K_t denotes the $t \times t$ matrix $K_t(i,j) = k(x_i, x_j)$, $1 \leq i,j \leq t$ and for any $x \in \mathbb{R}^t$ and $A \in \mathbb{R}^{t\times t}$, $\lVert x \rVert_A := \sqrt{x^T A x}$). Moreover, if K_t is positive definite $\forall t \geq 1$ with probability 1, then the conclusion above holds with $\eta = 0$.

Theorem 1 represents a self-normalized concentration inequality: the 'size' of the increasing-length sequence $\{\varepsilon_t\}_t$ of martingale differences is normalized by the growing quantity $((K_t + \eta I)^{-1} + I)^{-1}$ that explicitly depends on the sequence. The following lemma helps provide an alternative, abstract, view of the self-normalized process of Theorem 1, based on the feature space representation induced by a kernel.

Lemma 1 *Let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a symmetric, positive-semidefinite kernel, with associated feature map $\varphi : \mathbb{R}^d \to H_k$ and the reproducing kernel Hilbert space*[8] *(RKHS) H_k.*

[7] If $D_t = D$ for all t, then this is simply *exact* Thompson sampling. For technical reasons, however, our regret bound is valid when D_t is chosen as a suitable discretization of D, so we include D_t as an algorithmic parameter.

[8] Such a pair (φ, H_k) always exists, see e.g., Rasmussen & Williams (2006).

Letting $S_t = \sum_{s=1}^t \varepsilon_s \varphi(x_s)$ and the (possibly infinite dimensional) matrix[9] *$V_t = I + \sum_{s=1}^t \varphi(x_s)\varphi(x_s)^T$, we have, whenever K_t is positive definite, that*

$$\left\|\varepsilon_{1:t}\right\|_{\left(K_t^{-1}+I\right)^{-1}} = \left\|S_t\right\|_{V_t^{-1}},$$

where $\left\|S_t\right\|_{V_t^{-1}} := \left\|V_t^{-1/2} S_t\right\|_{H_k}$ denotes the norm of $V_t^{-1/2} S_t$ in the RKHS H_k.

Observe that S_t is $\mathcal{F}_t$-measurable and also $\mathbb{E}\left[S_t \mid \mathcal{F}_{t-1}\right] = S_{t-1}$. The process $\{S_t\}_{t\geq 0}$ is thus a martingale with values[10] in the RKHS H, which can possibly be infinite-dimensional, and moreover, whose deviation is measured by the norm weighted by V_t^{-1}, which is itself derived from S_t. Theorem 1 represents the kernelized generalization of the finite-dimensional result of Abbasi-Yadkori et al. (2011), and we recover their result under the special case of a linear kernel: $\varphi(x) = x$ for all $x \in \mathbb{R}^d$.

We remark that when φ is a mapping to a finite-dimensional Hilbert space, the argument of Abbasi-Yadkori et al. (2011, Theorem 1) can be lifted to establish Theorem 1, but it breaks down in the generalized, infinite-dimensional RKHS setting, as the self-normalized bound in their paper has an explicit, growing dependence on the feature dimension. Specifically, the method of mixtures (de la Pena et al., 2009) or Laplace method, as dubbed by Maillard (2016), fails to hold in infinite dimension. The primary reason for this is that the mixture distribution for finite dimensional spaces can be chosen independently of time, but in a nonparametric setup like ours, where the dimensionality of the self-normalizing factor $\left(K_t^{-1}+I\right)^{-1}$ itself grows with time, the use of (random) stopping times, precludes using time-dependent mixtures. We get around this difficulty by applying a novel 'double mixture' construction, in which a pair of mixtures on (a) the space of real-valued functions on $\mathbb{R}^d$, i.e., the support of a Gaussian process, and (b) on real sequences is simultaneously used to obtain a more general result, of potentially independent interest.

Our next result shows that how the posterior mean is concentrated around the unknown reward function f.

Theorem 2 *Under the same hypotheses as those of Theorem 1, let $D \subset \mathbb{R}^d$, and $f : D \to \mathbb{R}$ be a member of the RKHS of real-valued functions on D with kernel k, with RKHS norm bounded by B. Then, with probability at least $1-\delta$, the following holds for all $x \in D$ and $t \geq 1$:*

$$\left|\mu_{t-1}(x) - f(x)\right| \leq \left(B + R\sqrt{2(\gamma_{t-1} + 1 + \ln(1/\delta))}\right)\sigma_{t-1}(x),$$

where γ_{t-1} is the maximum information gain after $t-1$ rounds and $\mu_{t-1}(x)$, $\sigma_{t-1}^2(x)$ are mean and variance of posterior distribution defined as in Equation 2, 3, 4, with λ set to $1+\eta$ and $\eta = 2/T$.

Theorem 3.5 of Maillard (2016) states a similar result on the estimation of the unknown reward function from the RKHS. We improve upon it in the sense that the confidence bound in Theorem 2 is *simultaneous* over all $x \in D$, while the bound has been shown only for a single, fixed x in the Kernel Least-squares setting. We are able to achieve this result by virtue of Theorem 1.

4.1. Regret Bound of IGP-UCB

Theorem 3 *Let $\delta \in (0,1)$, $\|f\|_k \leq B$ and ε_t is conditionally R-sub-Gaussian. Running IGP-UCB for a function f lying in the RKHS $H_k(D)$, we obtain a regret bound of $O\left(\sqrt{T}(B\sqrt{\gamma_T} + \gamma_T)\right)$ with high probability. More precisely, with probability at least $1-\delta$, $R_T = O\Big(B\sqrt{T\gamma_T} + \sqrt{T\gamma_T(\gamma_T + \ln(1/\delta))}\Big)$.*

Improvement over GP-UCB. Srinivas et al. (2009), in the course of analyzing the GP-UCB algorithm, show that when the reward function lies in the RKHS $H_k(D)$, GP-UCB obtains regret $O\left(\sqrt{T}(B\sqrt{\gamma_T} + \gamma_T \ln^{3/2}(T))\right)$ with high probability (see Theorem 3 therein for the exact bound). Furthermore, they assume that the noise ε_t is *uniformly bounded* by σ, while our sub-Gaussianity assumption (see Equation 1) is slightly more general, and we are able to obtain a $O(\ln^{3/2} T)$ multiplicative factor improvement in the final regret bound thanks to the new self-normalized inequality (Theorem 1). Additionally, in our numerical experiments, we observe a significantly improved performance of IGP-UCB over GP-UCB, both on synthetically generated function, and on real-world sensor measurement data (see Section 6).

Comparison with KernelUCB. Valko et al. (2013) show that the cumulative regret of KernelUCB is $\tilde{O}(\sqrt{\tilde{d}T})$, where $\tilde{d}$, defined as the *effective dimension*, measures, in a sense, the number of principal directions over which the projection of the data in the RKHS is spread. They show that $\tilde{d}$ is at least as good as γ_T, precisely $\gamma_T \geq \Omega(\tilde{d}\ln\ln T)$ and thus the regret bound of KernelUCB is roughly $\tilde{O}(\sqrt{T\gamma_T})$, which is $\sqrt{\gamma_T}$ factor better than IGP-UCB. However, KernelUCB requires the number of actions to be *finite*, so the regret bound is not applicable for infinite or continuum action spaces.

4.2. Regret Bound of GP-TS

For technical reasons, we will analyze the following version of GP-TS. At each round t, the decision set used by GP-TS is restricted to be a unique discretization D_t

[9] More formally, $V_t : H_k \to H_k$ is the linear operator defined by $V_t(z) = z + \sum_{s=1}^t \varphi(x_s)\langle\varphi(x_s), z\rangle \; \forall z \in H_k$.

[10] We ignore issues of measurability here.

of D with the property that $|f(x) - f([x]_t)| \leq 1/t^2$ for all $x \in D$, where $[x]_t$ is the closest point to x in D_t. This can always be achieved by choosing a compact and convex domain $D \subset [0,r]^d$ and discretization D_t with size $|D_t| = (BLrdt^2)^d$ such that $\left\|x - [x]_t\right\|_1 \leq rd/BLrdt^2 = 1/BLt^2$ for all $x \in D$, where $L = \sup_{x \in D} \sup_{j \in [d]} \left(\frac{\partial^2 k(p,q)}{\partial p_j \partial q_j} |_{p=q=x} \right)^{1/2}$. This implies, for every $x \in D$,

$$|f(x) - f([x]_t)| \leq \|f\|_k L \left\|x - [x]_t\right\|_1 \leq 1/t^2, \qquad (7)$$

as any $f \in H_k(D)$ is Lipschitz continuous with constant $\|f\|_k L$ (De Freitas et al., 2012, Lemma 1).

Theorem 4 (Regret bound for GP-TS) *Let $\delta \in (0,1)$, $D \subset [0,r]^d$ be compact and convex, $\|f\|_k \leq B$ and $\{\varepsilon_t\}_t$ a conditionally R-sub-Gaussian sequence. Running GP-TS for a function f lying in the RKHS $H_k(D)$ and with decision sets D_t chosen as above, with probability at least $1 - \delta$, the regret of GP-TS satisfies $R_T = O\Big(\sqrt{(\gamma_T + \ln(2/\delta))d\ln(BdT)}\Big(\sqrt{T\gamma_T} + B\sqrt{T\ln(2/\delta)}\Big)\Big)$.*

Comparison with IGP-UCB. Observe that regret scaling of GP-TS is $\tilde{O}(\gamma_T\sqrt{dT})$ which is a multiplicative $\sqrt{d}$ factor away from the bound $\tilde{O}(\gamma_T\sqrt{T})$ obtained for IGP-UCB and similar behavior is reflected in our simulations on synthetic data. The additional multiplicative factor of $\sqrt{d\ln(BdT)}$ in the regret bound of GP-TS is essentially a consequence of discretization. How to remove this extra logarithmic dependency, and make the analysis discretization-independent, remains an open question.

Remark. The regret bound for GP-TS is inferior compared to IGP-UCB in terms of the dependency on dimension d, but to the best of our knowledge, Theorem 4 is the first (frequentist) regret guarantee of Thompson Sampling in the agnostic, non-parametric setting of infinite action spaces.

Linear Models and a Matching Lower Bound. If the mean rewards are perfectly linear, i.e. if there exists a $\theta \in \mathbb{R}^d$ such that $f(x) = \theta^T x$ for all $x \in D$, then we are in the parametric setup, and one way of casting this in the kernelized framework is by using the *linear kernel* $k(x,x') = x^T x'$. For this kernel, $\gamma_T = O(d\ln T)$, and the regret scaling of IGP-UCB is $\tilde{O}(d\sqrt{T})$ and that of GP-TS is $\tilde{O}(d^{3/2}\sqrt{T})$, which recovers the regret bounds of their linear, parametric analogues OFUL (Abbasi-Yadkori et al., 2011) and Linear Thompson sampling (Agrawal & Goyal, 2013), respectively. Moreover, in this case $\tilde{d} = d$, thus the regret of IGP-UCB is $\sqrt{d}$ factor away from that of KernelUCB. But the regret bound of KernelUCB also depends on the number of arms N, and if N is exponential in d, then it also suffers $\tilde{O}(d\sqrt{T})$ regret. We remark that a similar $O(\ln^{3/2} T)$ factor improvement, as obtained by IGP-UCB over GP-UCB, was achieved in the linear parametric setting by (Abbasi-Yadkori et al., 2011) in the OFUL algorithm, over its predecessor ConfidenceBall (Dani et al., 2008). Finally we see that the for linear bandit problem with infinitely many actions, IGP-UCB attains the information theoretic lower bound of $\Omega(d\sqrt{T})$ (see (Dani et al., 2008)), but GP-TS is a factor of $\sqrt{d}$ away from it.

5. Overview of Techniques

We briefly outline here the key arguments for all the theorems in Section 4. See Chowdhury & Gopalan (2017) for complete proofs.

Proof Sketch for Theorem 1. It is convenient to assume that K_t, the induced kernel matrix at time t, is invertible, since this is where the crux of the argument lies. First we show that for any function $g : D \to \mathbb{R}$ and for all $t \geq 0$, thanks to the sub-Gaussian property (1), the process $\left\{M_t^g := \exp(\varepsilon_{1:t}^T g_{1:t} - \frac{1}{2}\|g_{1:t}\|^2)\right\}_t$ is a non-negative super-martingale with respect to the filtration $\mathcal{F}_t$, where $g_{1:t} := [g(x_1), \ldots, g(x_t)]^T$ and in fact satisfies $\mathbb{E}\left[M_t^g\right] \leq 1$. The chief difficulty is to handle the behavior of M_t at a (random) stopping time, since the sizes of quantities such as $\varepsilon_{1:t}$ at the stopping time will be random.

We next construct a mixture martingale M_t by mixing M_t^g over g drawn from an independent $GP_D(0,k)$ Gaussian process, which is a measure over a large space of functions, i.e., the space $\mathbb{R}^D$. Then, by a change of measure argument, we show that this induces a mixture distribution which is essentially $\mathcal{N}(0, K_t)$ over *any* desired finite dimension t, thus obtaining $M_t = \frac{1}{\sqrt{\det(I+K_t)}} \exp\left(\frac{1}{2}\|\varepsilon_{1:t}\|^2_{(I+K_t^{-1})^{-1}}\right)$. Next from the fact that $\mathbb{E}\left[M_\tau\right] \leq 1$ and from Markov's inequality, for any $\delta \in (0,1)$, we obtain

$$\mathbb{P}\left[\|\varepsilon_{1:\tau}\|^2_{(K_\tau^{-1}+I)^{-1}} > 2\ln\left(\sqrt{\det(I+K_\tau)}/\delta\right)\right] \leq \delta.$$

Finally, we lift this bound for all t through a standard stopping time construction as in Abbasi-Yadkori et al. (2011).

Proof Sketch for Theorem 2. Here we sketch the special case of $\eta = 0$, i.e. $\lambda = 1$. Observe that $|\mu_t(x) - f(x)|$ is upper bounded by sum of two terms, $P := \left|k_t(x)^T(K_t + I)^{-1}\varepsilon_{1:t}\right|$ and $Q := \left|k_t(x)^T(K_t + I)^{-1} f_{1:t} - f(x)\right|$. Now we observe that $\sigma_t^2(x) = \varphi(x)^T(\Phi_t^T\Phi_t + I)^{-1}\varphi(x)$ and use this observation to show that $P = \left|\varphi(x)^T(\Phi_t^T\Phi_t + I)^{-1}\Phi_t^T\varepsilon_{1:t}\right|$ and $Q = \left|\varphi(x)^T(\Phi_t^T\Phi_t + I)^{-1} f\right|$, which are in turn upper bounded by the terms $\sigma_t(x)\|S_t\|_{V_t^{-1}}$ and $\|f\|_k \sigma_t(x)$ respectively. Then the result follows using Theorem 1, along with the assumption that $\|f\|_k \leq B$ and the fact that $\frac{1}{2}\ln(\det(I + K_t)) \leq \gamma_t$ a.s. when K_t is invertible.

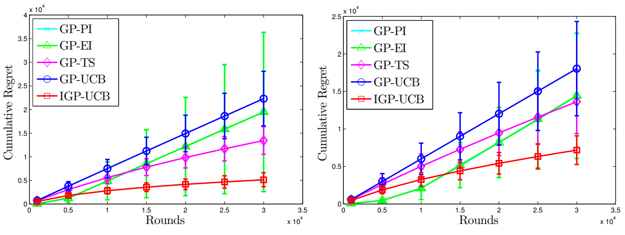
Proof Sketch for Theorem 3. First from Theorem 2 and the choice of x_t in Algorithm 1, we show that the instantaneous regret r_t at round t is upper bounded by $2\beta_t\sigma_{t-1}(x_t)$ with probability at least $1-\delta$. Then the result follows by bounding the term $\sum_{t=1}^{T}\sigma_{t-1}(x_t)$ by $O(\sqrt{T\gamma_T})$.

Proof Sketch for Theorem 4. We follow a similar approach given in Agrawal & Goyal (2013) to prove the regret bound of GP-TS. First observe that from our choice of discretization sets D_t, the instantaneous regret at round t is given by $r_t = f(x^\star) - f([x^\star]_t) + f([x^\star]_t) - f(x_t) \leq \frac{1}{t^2} + \Delta_t(x_t)$, where $\Delta_t(x) := f([x^\star]_t) - f(x)$ and $[x^\star]_t$ is the closest point to $x^\star$ in D_t. Now at each round t, after an action is chosen, our algorithm improves the confidence about true reward function f, via an update of $\mu_t(\cdot)$ and $k_t(\cdot,\cdot)$. However, if we play a suboptimal arm, the regret suffered can be much higher than the improvement of our knowledge. To overcome this difficulty, at any round t, we divide the arms (in the present discretization D_t) into two groups: *saturated arms*, S_t, defined as those with $\Delta_t(x) > c_t\sigma_{t-1}(x)$ and *unsaturated* otherwise, where c_t is an appropriate constant. The idea is to show that the probability of playing a saturated arm is small and then bound the regret of playing an unsaturated arm in terms of standard deviation. This is useful because the inequality $\sum_{t=1}^{T}\sigma_{t-1}(x_t) \leq O(\sqrt{T\gamma_T})$ allows us to bound the total regret due to unsaturated arms.

First we lower bound the probability of playing an unsaturated arm at round t. We define a filtration $\mathcal{F}'_{t-1}$ as the history $\mathcal{H}_{t-1}$ up to round $t-1$ and prove that for "most" (in a high probability sense) $\mathcal{F}'_{t-1}$, $\mathbb{P}\left[x_t \in D_t \setminus S_t \mid \mathcal{F}'_{t-1}\right] \geq p - 1/t^2$, where $p = 1/4e\sqrt{\pi}$. This observation, along with concentration bounds for $f_t(x)$ and $f(x)$ and "smoothness" of f, allow us to show that the expected regret at round t is upper bounded in terms of $\sigma_{t-1}(x_t)$, i.e. in terms of regret due to playing an unsaturated arm. More precisely, we show that for "most" $\mathcal{F}'_{t-1}$, $\mathbb{E}\left[r_t \mid \mathcal{F}'_{t-1}\right] \leq \frac{11c_t}{p}\mathbb{E}\left[\sigma_{t-1}(x_t) \mid \mathcal{F}'_{t-1}\right] + \frac{2B+1}{t^2}$, and use it to prove that $X_t \simeq r_t - \frac{11c_t}{p}\sigma_{t-1}(x_t) - \frac{2B+1}{t^2}; t \geq 1$ is a super-martingale difference sequence adapted to filtration $\{\mathcal{F}'_t\}_{t\geq 1}$. Now, using the Azuma-Hoeffding inequality, along with the bound on $\sum_{t=1}^{T}\sigma_{t-1}(x_t)$, we obtain the desired high-probability regret bound.

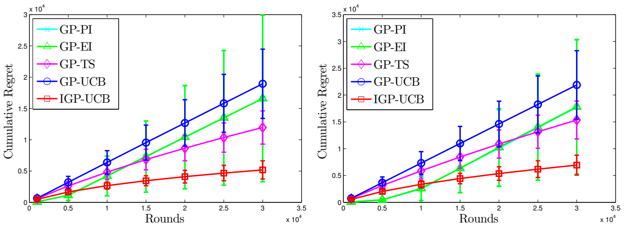
6. Experiments

In this section we provide numerical results on both synthetically generated test functions and functions from real-world data. We compare GP-UCB, IGP-UCB and GP-TS with GP-EI and GP-PI[11].



(a) (b)

Figure 1. Cumulative regret for functions lying in the RKHS corresponding to (a) Squared Exponential kernel and (b) Matérn kernel.



(a) (b)

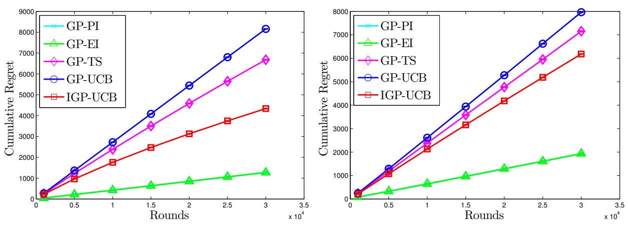
Figure 2. Cumulative regret for functions lying in the GP corresponding to (a) Squared Exponential kernel and (b) Matérn kernel.

Synthetic Test Functions. We use the following procedure to generate test functions from the RKHS. First we sample 100 points uniformly from the interval $[0,1]$ and use that as our decision set. Then we compute a kernel matrix K on those points and draw reward vector $y \sim \mathcal{N}(0,K)$. Finally, the mean of the resulting posterior distribution is used as the test function f. We set noise parameter R^2 to be 1% of function range and use $\lambda = R^2$. We used Squared Exponential kernel with lengthscale parameter $l = 0.2$ and Matérn kernel with parameters $\nu = 2.5, l = 0.2$. Parameters $\beta_t, \tilde{\beta}_t, v_t$ of IGP-UCB, GP-UCB and GP-TS are chosen as given in Section 3, with $\delta = 0.1, B^2 = f^T K f$ and γ_t set according to theoretical upper bounds for corresponding kernels. We run each algorithm for $T = 30000$ iterations, over 25 independent trials (samples from the RKHS) and plot the average cumulative regret along with standard deviations (Figure 1). We see a significant improvement in the performance of IGP-UCB over GP-UCB. In fact IGP-UCB performs the best in the pool of competitors, while GP-TS also fares reasonably well compared to GP-UCB and GP-EI/GP-PI.

We next sample 25 random functions from the $GP(0,K)$ and perform the same experiment (Figure 2) for both kernels with exactly same set of parameters. The relative performance of all methods is similar to that in the previous experiment, which is the arguably harder "agnostic" setting of a fixed, unknown target function.

Standard Test Functions. We consider 2 well-known

[11] GP-EI and PI perform similarly and thus are not separately distinguishable in the plots.



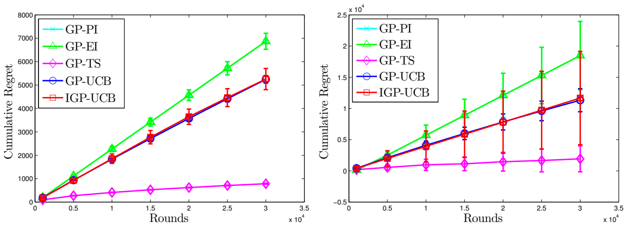
(a) (b)

Figure 3. Cumulative regret for (a) *Rosenbrock* and (b) *Hartman3* benchmark function.

synthetic benchmark functions for Bayesian Optimization: *Rosenbrock* and *Hartman*3 (see Azimi et al. (2012) for exact analytical expressions). We sample 100 d points uniformly from the domain of each benchmark function, d being the dimension of respective domain, as the decision set. We consider the Squared Exponential kernel with $l = 0.2$ and set all parameters exactly as in previous experiment. The cumulative regret for 25 independent trials on *Rosenbrock* and *Hartman3* benchmarks is shown in Figure 3. We see GP-EI/PI perform better than the rest, while IGP-UCB and GP-TS show competitive performance. Here no algorithm is aware of the underlying kernel function, hence we conjecture that the UCB- and TS- based algorithms are somewhat less robust on the choice of kernel than EI/PI.

Temperature Sensor Data. We use temperature data[12] collected from 54 sensors deployed in the Intel Berkeley Research lab between February 28th and April 5th, 2004 with samples collected at 30 second intervals. We tested all algorithms in the context of learning the maximum reading of the sensors collected between 8 am to 9 am. We take measurements of first 5 consecutive days (starting Feb. 28th 2004) to learn algorithm parameters. Following Srinivas et al. (2009), we calculate the empirical covariance matrix of the sensor measurements and use it as the kernel matrix in the algorithms. Here R^2 is set to be 5% of the average empirical variance of sensor readings and other algorithm parameters is set similarly as in the previous experiment with $\gamma_t = 1$ (found via cross-validation). The functions for testing consist of one set of measurements from all sensors in the two following days and the cumulative regret is plotted over all such test functions. From Figure 4, we see that IGP-UCB and GP-UCB performs the same, while GP-TS outperforms all its competitors.

Light Sensor Data. We take light sensor data collected in the CMU Intelligent Workplace in Nov 2005, which is available online as Matlab structure[13] and contains locations of 41 sensors, 601 train samples and 192 test samples.



(a) (b)

Figure 4. Cumulative regret plots for (a) temperature data and (b) light sensor data.

We compute the kernel matrix, estimate the noise and set other algorithm parameters exactly as in the previous experiment. Here also GP-TS is found to perform better than the others, with IGP-UCB performing better than GP-EI/PI (Figure 4).

Related work. An alternative line of work pertaining to $\mathcal{X}$-armed bandits (Kleinberg et al., 2008; Bubeck et al., 2011; Carpentier & Valko, 2015; Azar et al., 2014) studies continuum-armed bandits with smoothness structure. For instance, Bubeck et al. (2011) show that with a Lipschitzness assumption on the reward function, algorithms based on discretizing the domain yield nontrivial regret guarantees, of order $\Omega(T^{\frac{d+1}{d+2}})$ in $\mathbb{R}^d$. Other Bayesian approaches to function optimization are GP-EI (Močkus, 1975), GP-PI (Kushner, 1964), GP-EST (Wang et al., 2016) and GP-UCB, including the contextual (Krause & Ong, 2011), high-dimensional (Djolonga et al., 2013; Wang et al., 2013), time-varying (Bogunovic et al., 2016) safety-aware (Gotovos et al., 2015), budget-constraint (Hoffman et al., 2013) and noise-free (De Freitas et al., 2012) settings. Other relevant work focuses on best arm identification problem in the Bayesian setup considering pure exploration (Grünewälder et al., 2010). For Thompson sampling (TS), Russo & Van Roy (2014) analyze the Bayesian regret of TS, which includes the case where the target function is sampled from a GP prior. Our work obtains the first frequentist regret of TS for unknown, fixed functions from an RKHS.

7. Conclusion

For bandit optimization, we have improved upon the existing GP-UCB algorithm, and introduced a new GP-TS algorithm. The proposed algorithms perform well in practice both on synthetic and real-world data. An interesting case is when the kernel function is also not known to the algorithms a priori and needs to be learnt adaptively. Moreover, one can consider classes of time varying functions from the RKHS, and general reinforcement learning with GP techniques. There are also important questions on computational aspects of optimizing functions drawn from GPs.

[12] http://db.csail.mit.edu/labdata/labdata.html

[13] http://www.cs.cmu.edu/~guestrin/Class/10708-F08/projects/lightsensor.zip

Acknowledgement

Aditya Gopalan was supported by the DST INSPIRE faculty grant IFA13-ENG-69. The authors are grateful to O.-A. Maillard for helpful discussions, and to anonymous reviewers for providing useful comments.

References

Abbasi-Yadkori, Yasin, Pál, Dávid, and Szepesvári, Csaba. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

Agrawal, Shipra and Goyal, Navin. Analysis of thompson sampling for the multi-armed bandit problem. In *COLT*, pp. 39–1, 2012.

Agrawal, Shipra and Goyal, Navin. Thompson sampling for contextual bandits with linear payoffs. In *ICML*, pp. 127–135, 2013.

Azar, Mohammad Gheshlaghi, Lazaric, Alessandro, and Brunskill, Emma. Online stochastic optimization under correlated bandit feedback. In *ICML*, pp. 1557–1565, 2014.

Azimi, Javad, Jalali, Ali, and Fern, Xiaoli. Hybrid batch bayesian optimization. *arXiv preprint arXiv:1202.5597*, 2012.

Besbes, Omar and Zeevi, Assaf. Dynamic pricing without knowing the demand function: Risk bounds and near-optimal algorithms. *Operations Research*, 57(6):1407–1420, 2009.

Bogunovic, Ilija, Scarlett, Jonathan, and Cevher, Volkan. Time-varying gaussian process bandit optimization. *arXiv preprint arXiv:1601.06650*, 2016.

Bubeck, Sébastien, Munos, Rémi, Stoltz, Gilles, and Szepesvári, Csaba. X-armed bandits. *Journal of Machine Learning Research*, 12(May):1655–1695, 2011.

Carpentier, Alexandra and Valko, Michal. Simple regret for infinitely many armed bandits. In *ICML*, pp. 1133–1141, 2015.

Chiang, Mung, Hande, Prashanth, Lan, Tian, and Tan, Chee Wei. Power control in wireless cellular networks. *Foundations and Trends in Networking*, 2(4):381–533, 2008. ISSN 1554-057X. doi: 10.1561/1300000009.

Chowdhury, Sayak Ray and Gopalan, Aditya. On kernelized multi-armed bandits. *arXiv preprint arXiv:1704.00445*, 2017.

Dani, Varsha, Hayes, Thomas P, and Kakade, Sham M. Stochastic linear optimization under bandit feedback. In *COLT*, pp. 355–366, 2008.

De Freitas, Nando, Smola, Alex, and Zoghi, Masrour. Exponential regret bounds for gaussian process bandits with deterministic observations. *arXiv preprint arXiv:1206.6457*, 2012.

de la Pena, Victor H, Lai, Tze Leung, and Shao, Qi-Man. Self-normalized processes. probability and its applications, 2009.

Djolonga, Josip, Krause, Andreas, and Cevher, Volkan. High-dimensional gaussian process bandits. In *Advances in Neural Information Processing Systems*, pp. 1025–1033, 2013.

Gopalan, Aditya, Mannor, Shie, and Mansour, Yishay. Thompson sampling for complex online problems. In *ICML*, volume 14, pp. 100–108, 2014.

Gotovos, Alkis, CH, ETHZ, and Burdick, Joel W. Safe exploration for optimization with gaussian processes. 2015.

Grünewälder, Steffen, Audibert, Jean-Yves, Opper, Manfred, and Shawe-Taylor, John. Regret bounds for gaussian process bandit problems. In *AISTATS*, pp. 273–280, 2010.

Hoffman, Matthew D, Brochu, Eric, and de Freitas, Nando. Portfolio allocation for bayesian optimization. In *UAI*, pp. 327–336, 2011.

Hoffman, Matthew W, Shahriari, Bobak, and de Freitas, Nando. Exploiting correlation and budget constraints in bayesian multi-armed bandit optimization. *arXiv preprint arXiv:1303.6746*, 2013.

Kaelbling, Leslie Pack, Littman, Michael L, and Moore, Andrew W. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.

Kaufmann, Emilie, Korda, Nathaniel, and Munos, Rémi. Thompson sampling: An asymptotically optimal finite-time analysis. In *International Conference on Algorithmic Learning Theory*, pp. 199–213. Springer, 2012.

Kleinberg, Robert, Slivkins, Aleksandrs, and Upfal, Eli. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pp. 681–690. ACM, 2008.

Krause, Andreas and Ong, Cheng S. Contextual gaussian process bandit optimization. In *Advances in Neural Information Processing Systems*, pp. 2447–2455, 2011.

Kushner, Harold J. A new method of locating the maximum point of an arbitrary multipeak curve in the presence of noise. *Journal of Basic Engineering*, 86(1):97–106, 1964.

Maillard, Odalric-Ambrym. Self-normalization techniques for streaming confident regression. 2016.

Močkus, J. On bayesian methods for seeking the extremum. In *Optimization Techniques IFIP Technical Conference*, pp. 400–404. Springer, 1975.

Rasmussen, Carl Edward and Williams, Christopher KI. Gaussian processes for machine learning. 2006.

Rusmevichientong, Paat and Tsitsiklis, John N. Linearly parameterized bandits. *Math. Oper. Res.*, 35(2):395–411, May 2010.

Russo, Daniel and Van Roy, Benjamin. Learning to optimize via posterior sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

Schölkopf, Bernhard and Smola, Alexander J. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Smart, William D and Kaelbling, Leslie Pack. Practical reinforcement learning in continuous spaces. In *ICML*, pp. 903–910, 2000.

Srinivas, Niranjan, Krause, Andreas, Kakade, Sham M, and Seeger, Matthias. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*, 2009.

Valko, Michal, Korda, Nathaniel, Munos, Rémi, Flaounas, Ilias, and Cristianini, Nelo. Finite-time analysis of kernelised contextual bandits. *arXiv preprint arXiv:1309.6869*, 2013.

Wang, Zi, Zhou, Bolei, and Jegelka, Stefanie. Optimization as estimation with gaussian processes in bandit settings. In *International Conf. on Artificial and Statistics (AISTATS)*, 2016.

Wang, Ziyu, Zoghi, Masrour, Hutter, Frank, Matheson, David, Freitas, N, et al. Bayesian optimization in high dimensions via random embeddings. AAAI Press/International Joint Conferences on Artificial Intelligence, 2013.