On Information Gain and Regret Bounds in Gaussian Process Bandits

Sattar Vakili Kia
MediaTek Research, UK
sattar.vakili@mtkresearch.com
Victor Picheny
Secondmind, UK
victor@secondmind.ai
Khezeli.kia@mayo.edu

Abstract

Consider the sequential optimization of an expensive to evaluate and possibly non-convex objective function $f$ from noisy feedback, that can be regarded as a continuum-armed bandit problem. Upper bounds on the regret performance of several learning algorithms (GP-UCB, GP-TS, and their variants) are known under both a Bayesian (when $f$ is a sample from a Gaussian process (GP)) and a frequentist (when $f$ lives in a reproducing kernel Hilbert space) setting. The regret bounds often rely on the maximal information gain $\gamma_T$ between $T$ observations and the underlying GP (surrogate) model. We provide general bounds on $\gamma_T$ based on the decay rate of the eigenvalues of the GP kernel, whose specialisation for commonly used kernels improves the existing bounds on $\gamma_T$, and subsequently the regret bounds relying on $\gamma_T$ under numerous settings. For the Matérn family of kernels, where the lower bounds on $\gamma_T$, and regret under the frequentist setting, are known, our results close a huge polynomial in $T$ gap between the upper and lower bounds (up to logarithmic in $T$ factors).

1 Introduction

Bayesian optimization building on Gaussian Process (GP) models has been shown to efficiently address the exploration-exploitation trade-off in the sequential optimization of non-convex objective functions with bandit feedback. There have been significant recent advances in the analysis of GP-based Bayesian optimization algorithms, providing performance guarantees in terms of regret. Regret is defined as the cumulative loss in the value of the objective function $f$ at a sequence of observation points $\{x_t\}_{t=1}^T$, $T \in \mathbb{N}$, in comparison to its value at a global maximum $x^* \in \arg\max_{x \in X} f(x)$ over the search space $X \subseteq \mathbb{R}^d$ (see (1)). In their seminal paper, Srinivas et al. (2010) established performance guarantees for GP-UCB, an optimistic optimization algorithm which sequentially selects the $x_t$ that maximize an upper confidence bound score over the search space. They considered a fully Bayesian setting where $f$ is assumed to be a sample from a GP with a known kernel, as well as a frequentist setting (referred to as agnostic in Srinivas et al. (2010)) where $f$ is assumed to live in a reproducing kernel Hilbert space (RKHS) with a known kernel. They showed an $\tilde{O}(\sqrt{\gamma_T T})$ and an $O(\sqrt{\gamma_T T})$ regret bound for GP-UCB under the Bayesian and frequentist settings, respectively, where $\gamma_T$ is the maximal information gain between the observed sequence and the underlying model (see §2.5). Sublinearly scaling with $T$, $\gamma_T$ depends on the underlying kernel and is interpreted as a measure for the difficulty of the optimization task. Since the pioneering work of Srinivas et al., there have been several results on improving the bounds toward their optimal value. Chowdhury and Gopalan (2017) improved the regret bounds under the frequentist setting by multiplicative logarithmic in $T$ factors. A variant of GP-UCB (called SupKernelUCB), which builds on episodic independent batches of observations, was shown in Valko et al. (2013) to achieve the $O(\sqrt{\gamma_T T})$ regret bound under the frequentist setting. Furthermore, Chowdhury and Gopalan (2017) showed that $O(\sqrt{\gamma_T T})$ regret bounds, under the frequentist setting, also hold for GP-TS, a Bayesian optimization

---

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).
algorithm based on Thompson Sampling which sequentially draws $x_t$ from the posterior distribution of $x^*$. Under the Bayesian setting, Kandasamy et al. (2018) built on ideas from Russo and Van Roy (2014, 2016) to show that GP-TS achieves the same order of regret as GP-UCB.

The regret bounds mentioned above become complete only when $\gamma_T$ is properly bounded, which proves challenging. We first overview the existing upper and lower bounds on the maximal information gain in the literature. We then discuss our contribution and compare it against the state of the art.

1.1 Upper and Lower Bounds on $\gamma_T$

Srinivas et al. (2010) showed that $\gamma_T = \tilde{O}(T^{d(d+4)/(d+3)})$ for the Matérn-$\nu$ kernel (a Matérn kernel with smoothness parameter $\nu$; see §2.2 for the details), and $\gamma_T = O(\log^{d+1}(T))$ for the Squared Exponential (SE) kernel. Recently, Janz et al. (2020) introduced a GP-UCB based algorithm (specific to the Matérn family of kernels), that constructs a cover for the search space (as many hypercubes) and fits an independent GP to each cover element. This analysis elicited an improved bound for the Matérn-$\nu$ kernel; $\gamma_T = \tilde{O}(T^{d(d+1)/(d+2)})$. Plugging these bounds on $\gamma_T$ into the $O(\sqrt{\gamma_T}T)$ regret bounds mentioned above yields explicit upper bounds, in terms of $T$, which are in the forms of $O(T^{d/(2+\nu)}T)$ and $O(T^{\nu} \log^{\frac{d}{\nu}+1}(T))$, with Matérn-$\nu$ and SE, respectively.

Finding the order optimal regret bounds is a long standing open question. Under the frequentist setting, Scarlett et al. (2017) proved the $\tilde{O}(T^{d(d+4)/(d+3)})$ and $\Omega(T^{\frac{d}{d+3}} \log^{\frac{d}{d+3}}(T))$ lower bounds on the regret performance of any learning algorithm, with Matérn-$\nu$ and SE, respectively. From the results of Scarlett et al. (2017) and Valko et al. (2013), $\Omega(T^{\frac{d}{d+\nu}})$ and $O(\log^{d/\nu}(T))$ lower bounds on $\gamma_T$ can be concluded for Matérn-$\nu$ and SE, respectively (see also Janz et al., 2020), which facilitate the assessment of the upper bounds.

While the bounds are tight up to logarithmic factors for SE, a comparison between the lower and upper bounds for the practically useful Matérn family of kernels shows a drastic gap, which can surprisingly be as large as $O(\sqrt{T})$ in the case of regret, and as large as $O(T)$ in the case of $\gamma_T$, with particular configurations of parameters $\nu$ and $d$. Motivated by this huge gap in the literature,

we aim to provide tight bounds on $\gamma_T$ (reducing these polynomial in $T$ gaps to logarithmic ones), as outlined in the next section.

1.2 Contribution

Our contribution is in establishing novel bounds on $\gamma_T$, which directly translate to new regret bounds for Bayesian optimization algorithms. To achieve this, we use Mercer’s theorem to represent the GP kernel in terms of its eigenvalue-eigenfeature decomposition—an inner product in the corresponding reproducing kernel Hilbert space (RKHS)—which is infinite dimensional for typical kernels. To overcome the difficulty of working in infinite dimensional spaces, we use a projection on a finite $D$ dimensional space that allows us to bound the information gain in terms of $D$ and the spectral properties of the GP kernel. For a kernel with decreasing eigenvalues $\{\lambda_m\}_{m=1}^\infty$, we consider two cases of polynomial, $\lambda_m = O(m^{-\beta_p})$, $\beta_p > 1$, and exponential, $\lambda_m = O(\exp(-m^{\beta_e}))$, $\beta_e > 0$, decays. We prove $O(T^{\frac{d}{d+1}} \log^{1+ \frac{d}{\nu}}(T))$ and $O(T^{\frac{d}{d+\nu}} \log^{\frac{d}{\nu}+1}(T))$ upper bounds on $\gamma_T$ under these two cases, respectively. The application of our bounds on $\gamma_T$ to the regret bounds results in new upper bounds based on $\beta_p$ and $\beta_e$ which are summarized in Table 1. In comparison to the existing works, which rely on specific kernels (e.g. Matérn-$\nu$ and SE) for explicit regret bounds, our results provide general explicit regret bounds, providing the conditions on the decay rate of the eigenvalues of the GP kernel (referred to as eigendecay for brevity) are satisfied.

As an instance of polynomially decaying eigenvalues, our results apply to the Matérn-$\nu$ kernel (see §2.2) showing $\tilde{O}(T^{\frac{d}{d+\nu}})$ and $O(T^{\frac{d}{d+\nu}})$ bounds on $\gamma_T$ and regret, respectively. Our bounds on $\gamma_T$ and regret (under the frequentist setting) are tight, closing the gap with the respective $\Omega(T^{\frac{d}{d+\nu}})$ and $\Omega(T^{\frac{d}{d+\nu}})$ lower bounds reported in Scarlett et al. (2017); Janz et al. (2020), both up to logarithmic factors. As an instance of exponentially decaying eigenvalues, our results apply to the Squared Exponential (SE) kernel. A summary of the results is given in Table 1.

While we focus on the standard sequential optimization problem in this paper, it is worth noting that the and the upper bound on regret become arbitrarily close to $\Omega(\sqrt{T})$ and $O(T)$, respectively. Therefore, the worst case gap between them is in $O(\sqrt{T})$. The lower bound and the upper bound on $\gamma_T$ become arbitrarily close to $\Omega(1)$ and $O(T)$, respectively. Therefore, the worst case gap between them is in $O(T)$.
Table 1: The Upper bounds on the maximal information gain $\gamma_T$ and the regret of Bayesian optimization algorithms under general polynomial and exponential conditions on the eigendecay of the GP kernel (see Definition 1), as well as, with Matérn-$\nu$ and SE kernels (established in this paper). The lower bounds on regret under the frequentist setting (on the third column of the table) were reported in Scarlett et al. (2017). The gap between the upper and lower bounds is reduced to logarithmic factors.

<table>
<thead>
<tr>
<th>Kernel</th>
<th>Bound on $\gamma_T$</th>
<th>Regret Lower Bound</th>
<th>Regret Upper Bound ($\mathcal{O}(\sqrt{\gamma_T T})$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Polynomial eigendecay</td>
<td>$\mathcal{O}\left(T^{\frac{1}{\nu}} \log^{1-\frac{1}{2\nu}}(T)\right)$</td>
<td>-</td>
<td>$\widetilde{\mathcal{O}}\left(T^{\frac{\beta p}{2}}\right)$</td>
</tr>
<tr>
<td>Exponential eigendecay</td>
<td>$\mathcal{O}\left(\log^{1+\frac{1}{\nu}}(T)\right)$</td>
<td>-</td>
<td>$\mathcal{O}\left(T^{\frac{1}{2}} \log \frac{2}{\nu}(T)\right)$</td>
</tr>
<tr>
<td>Matérn-$\nu$</td>
<td>$\mathcal{O}\left(T^{\frac{4d}{\nu+\lambda}} \log^{\frac{2\nu}{\nu+\lambda}}(T)\right)$</td>
<td>$\Omega\left(T^{\frac{4d}{\nu+\lambda}}\right)$</td>
<td>$\mathcal{O}\left(T^{\frac{4d}{\nu+\lambda}}\right)$</td>
</tr>
<tr>
<td>SE</td>
<td>$\mathcal{O}\left(\log^{d+1}(T)\right)$</td>
<td>$\Omega\left(T^{\frac{1}{2}} \log^{\frac{1}{2}}(T)\right)$</td>
<td>$\mathcal{O}\left(T^{\frac{1}{2}} \log^{\frac{1}{2}}(T)\right)$</td>
</tr>
</tbody>
</table>

Most of existing work reports regret bounds in terms of $\gamma_T$. Our results directly apply to, and improve, the regret bounds in all of the works mentioned above, should our bounds on $\gamma_T$ replace the existing ones.

The interest in the bounds on $\gamma_T$ goes beyond the regret bounds. For example, the confidence bounds for the RKHS elements (see e.g., Chowdhury and Gopalan, 2017, Theorem 2) depend on $\gamma_T$. Another closely related quantity is the so called effective dimension $\tilde{D}_T$ of the problem that satisfies $\tilde{D}_T = \mathcal{O}(\gamma_T)$ (see Valko et al., 2013; Calandriello et al., 2019; Janz et al., 2020, and Remark 1). Calandriello et al. (2019) introduced a variation of GP-UCB which improves its computational cost. The improved computational cost depends on $\tilde{D}_T$. Our bounds on $\gamma_T$ (consequently on $\tilde{D}_T$) improve such bounds on the algorithmic properties of GP-based methods.

1.3 Other Related Work

Recent years have shown an increasing interest in Bayesian optimization based on GP models. Performance guarantees in terms of regret are studied under various settings including contextual information (Krause and Ong, 2011), high dimensional spaces (Djolonga et al., 2013; Mutny and Krause, 2018), safety constraints (Berkenkamp et al., 2016; Sui et al., 2018), parallelization (Kandasamy et al., 2018), multifidelity evaluations (Kandasamy et al., 2019), ordinal models (Picheny et al., 2019), online control (Kakade et al., 2020), and corruption tolerance (Bogunovic et al., 2020), to name a few. Javidi and Shekhar (2018) introduced an adaptive discretization of the search space improving the computational complexity of a GP-UCB based algorithm. Sparse approximation of GP posteriors are shown to preserve the regret orders while significantly improving the computational complexity of both GP-UCB (Mutny and Krause, 2018; Calandriello et al., 2019) and GP-TS (Vakili et al., 2020). Most of existing work reports regret bounds in terms of $\gamma_T$. Our results directly apply to, and improve, the regret bounds in all of the works mentioned above, should our bounds on $\gamma_T$ replace the existing ones.

Our analytical approach and conditions on the eigendecay of GP kernels bear similarity to Chatterji et al. (2019), where the authors studied the problem of online learning with kernel losses. The problems and their analysis, however, hold substantial differences. A more challenging adversarial setting was considered for the objective function in Chatterji et al. (2019). However, the objective function was restricted to the subspace of one dimensional functions in the RKHS, which is very limiting for our purposes (one of the main challenges in our analysis is the infinite-dimensionality of the RKHS). The algorithmic designs, based on exponential weights, under the adversarial setting, are also significantly different from GP-UCB and GP-TS, especially, in the sense that their analysis does not rely on the information gain.

Instead of Mercer’s Theorem, other decompositions of GP kernels may also be used in a similar way to our analysis. For instance, decompositions based on Fourier features were used in Mutny and Krause (2018) to implement computationally efficient variations of GP-TS and GP-UCB. They did not however consider the analysis of $\gamma_T$. 

Under the Bayesian setting, Scarlett (2018) proved tight \( \Omega(\sqrt{T}) \) (up to a logarithmic in \( T \) factor) lower bounds on regret when the search space is one-dimensional \((d = 1)\). To the best of our knowledge, lower bounds are unknown for the general case \((d > 1)\), under the Bayesian setting.

Both GP-UCB and GP-TS are rooted in the classic multi-armed bandit literature (see Auer et al., 2002; Russo and Van Roy, 2016; Slivkins, 2019; Zhao, 2019, and references therein). Our work strengthens the link between linear (Dani et al., 2008; Rusmevichientong and Tsitsiklis, 2010; Abbasi-Yadkori et al., 2011; Agrawal and Goyal, 2013; Abeille and Lazaric, 2017) and kernelized (GP-based) (Srinivas et al., 2010; Chowdhury and Gopalan, 2017) models for sequential optimization with bandit feedback, as we build our analysis based on a finite-dimensional projection that is equivalent to linear bandits.

The remainder of the paper is organized as follows. The problem formulation, the preliminaries on GPs, GP-UCB, GP-TS, and the background on the connection between the regret bounds and the information gain are presented in §2. The analysis of the bounds on \( \gamma_T \) is provided in §3. The explicit regret bounds (in terms of \( T \)) for Bayesian optimization algorithms are given in §4. The paper is concluded in §5.

2 Problem Formulation and Preliminaries

In this section, we provide background information on sequential optimization, GPs, and the connection between the information gain and the regret bounds for Bayesian optimization algorithms.

We use the following notations throughout the paper. For a square matrix \( M \in \mathbb{R}^{n \times n} \), the notations \( \det(M) \) and \( \text{tr}(M) \) denote the determinant and the trace of \( M \), respectively. The notation \( M^\top \) is used for the transpose of an arbitrary matrix \( M \). For a positive definite matrix \( P \), \( \log \det(P) \) denotes \( \log(\det(P)) \). The identity matrix of dimension \( n \) is denoted by \( I_n \). For a vector \( z \in \mathbb{R}^n \), the notation \( \|z\|_2 \) denotes its \( L^2 \) norm.

2.1 The Sequential Optimization Problem

Consider the sequential optimization of a fixed and unknown objective function \( f \) over a compact set \( \mathcal{X} \subset \mathbb{R}^d \). A learning algorithm \( \pi \) sequentially selects an observation point \( x_t \in \mathcal{X} \) at each discrete time instance \( t = 1, 2, \ldots, \) and receives the corresponding real-valued reward \( y_t = f(x_t) + \epsilon_t \), where \( \epsilon_t \) is the observation noise. Specifically, \( \pi = \{\pi_t\}_{t=1}^\infty \) is a sequence of mappings \( \pi_t : \mathcal{H}_{t-1} \to \mathcal{X} \) from the history of observations to a new observation point; \( \mathcal{H}_t = \{X_t, Y_t\} \), \( X_t = [x_1, x_2, \ldots, x_t]^\top \), \( Y_t = [y_1, y_2, \ldots, y_t]^\top \), \( x_s \in \mathcal{X} \), \( y_s \in \mathbb{R} \), for all \( s \geq 1 \). The regularity assumptions on \( f \) and \( \epsilon_t \) are specified in §2.4.

The goal is to minimize regret, defined as the cumulative loss compared to the maximum attainable objective, over a time horizon \( T \). Specifically,

\[
R(T; \pi) = \sum_{t=1}^T (f(x^*) - f(x_t)),
\]

where \( x^* = \arg\max_{x \in \mathcal{X}} f(x) \) is a global maximum of \( f \). To simplify the notation, the dependency on \( \pi \) in the notation of \( X_t \) has been omitted.

2.2 Gaussian Processes

The learning algorithms considered here build on GP (surrogate) models. A GP is a random process \( \{\hat{f}(x)\}_{x \in \mathcal{X}} \), in which all finite subsets follow multivariate Gaussian distributions (Rasmussen and Williams, 2006). The distribution of a GP can be specified by its mean function \( \mu(x) = \mathbb{E}[\hat{f}(x)] \) and a positive definite kernel (or covariance function) \( k(x, x') = \mathbb{E}[(\hat{f}(x) - \mu(x))\,(\hat{f}(x') - \mu(x'))] \). Without loss of generality, it is typically assumed that \( \forall x \in \mathcal{X}, \mu(x) = 0 \) for prior GP distributions.

Conditioning GPs on available observations provides us with powerful non-parametric Bayesian (surrogate) models over the space of functions. In particular, conditioned on \( \mathcal{H}_t \), the posterior of \( \hat{f} \) is a GP with mean function \( \mu_t(x) = \mathbb{E}[(\hat{f}(x)|\mathcal{H}_t)] \) and kernel function \( k_t(x, x') = \mathbb{E}[(\hat{f}(x) - \mu_t(x))(\hat{f}(x') - \mu_t(x'))|\mathcal{H}_t] \) specified as follows:

\[
\mu_t(x) = k_{x_t, x}^\top(K_{x_t, x} + \tau I)^{-1}y_t, \quad k_t(x, x') = k(x, x') - k_{x_t, x}^\top K_{x_t, x'} (K_{x_t, x'} + \tau I)^{-1}k_{x_t, x'},
\]

where \( k_{x_t, x} = [k(x_1, x), k(x_2, x), \ldots, k(x_t, x)]^\top \) and \( K_{x_t, x} \) is the \( t \times t \) positive definite covariance matrix, \( [k(x_i, x_j)]_{i,j=1}^t \). The posterior variance of \( \hat{f}(x) \) is denoted by \( \sigma_t^2(x) = k_t(x, x) \).

Matérn and squared exponential (SE) are perhaps the most popular kernels in practice for Bayesian optimiza-
tion (see e.g., Snoek et al., 2012; Shahriari et al., 2016),
\[
    k_{\text{Matérn}}(x, x') = \frac{1}{\Gamma(\nu)2^{\nu-1}l^\nu} \left( \frac{\sqrt{2r}}{l} \right)^\nu B_\nu \left( \frac{\sqrt{2r}}{l} \right),
\]
\[
    k_{SE}(x, x') = \exp \left( -\frac{r^2}{2
\right),
\]
where \( l > 0, r = \|x - x'\|_2 \) is the Euclidean distance between \( x \) and \( x' \), \( \nu > 0 \) is referred to as the smoothness parameter, \( \Gamma \) is the gamma function, and \( B_\nu \) is the modified Bessel function of the second kind. Variation over parameter \( \nu \) creates a rich family of kernels. The SE kernel can also be interpreted as a special case of Matérn family when \( \nu \to \infty \).

2.3 Bayesian Optimization Algorithms (GP-UCB and GP-TS)

GP-UCB relies on an optimistic upper confidence bound score to select the observation points. Specifically, at each time \( t \), \( x_t \) is selected as
\[
    x_t = \arg\max_{x \in X} \mu_{t-1}(x) + \alpha_t \sigma_{t-1}(x),
\]
where \( \mu_{t-1} \) and \( \sigma_{t-1} \) are the posterior mean and the standard deviation, based on previous observations defined in §2.2, and \( \alpha_t \) is a user-specified scaling parameter.

GP-TS selects the observation points by posterior sampling. Specifically, at each time \( t \), a sample \( \hat{f}_t(x) \) is drawn from a GP with mean \( \mu_{t-1} \) and kernel function \( \alpha_t^2 k_{t-1} \) where \( \mu_{t-1} \) and \( k_{t-1} \) are the posterior mean and the posterior kernel, based on previous observations defined in §2.2, and \( \alpha_t \) is a user-specified scaling parameter. Then, \( x_t \) is selected as
\[
    x_t = \arg\max_{x \in X} \hat{f}_t(x).
\]

The scaling parameters \( \alpha_t \) are designed to balance the trade-off between exploitation and exploration of the search space and increase with \( t \) (\( \alpha_t > \alpha_{t'} \) when \( t > t' \)). See, e.g., Srinivas et al. (2010); Chowdhury and Gopalan (2017) for the specifications of \( \alpha_t \).

2.4 Regularity Assumptions

The regret performance of the learning algorithms is analysed under two different settings, referred to as Bayesian and frequentist.

Under the Bayesian setting, \( f \) is assumed to be a sample from a prior GP with kernel \( k \). The observation noise \( \{\epsilon_t\}_{t=1}^T \) are assumed to be i.i.d. zero mean Gaussian random variables with variance \( \tau \).

Under the frequentist setting, \( f \) is assumed to live in the RKHS corresponding to \( k \). In particular, \( \| f \|_{\mathcal{H}_k} \leq B \), for some \( B > 0 \), where \( \| \cdot \|_{\mathcal{H}_k} \) denote the RKHS norm (see §3.1 for the definition of the RKHS norm). The observation noise are assumed to be i.i.d. sub-Gaussian random variables. Specifically, it is assumed that \( \forall h \in \mathbb{R}, \forall t \in \mathbb{N}, E[e^{h\epsilon_t}] \leq \exp(\frac{h^2R^2}{2}) \), for some \( R > 0 \). The sub-Gaussian assumption implies that \( E[\epsilon_t] = 0 \), for all \( t \).

2.5 The Information Gain and The Upper Bounds on Regret

The regret analysis of Bayesian optimization algorithms typically consists of two main components. One is a bound on the maximal information gain \( \gamma_T \), and the other is a confidence bound for random processes. The bound on \( \gamma_T \) is treated identically under both Bayesian and frequentist settings. Confidence bounds which are utilized under each setting are different, however. To motivate the analysis of \( \gamma_T \), we provide a sketch for the analysis of GP-UCB. Since our bounds on \( \gamma_T \) are independent of the learning algorithm, they are indiscriminately applicable to all settings where the regret bound is given in terms of \( \gamma_T \).

A classic approach to the sequential optimization problem is to construct a \( 1 - \delta \) upper confidence bound for \( f \), after \( t - 1 \) observations, in the form of
\[
    U_t(x) = \mu_{t-1}(x) + \beta_t(\delta)\sigma_{t-1}(x).
\]

GP modelling provides us with closed form expressions for \( \mu_{t-1}(x) \) and \( \sigma_{t-1}(x) \), while \( \beta_t(\delta) \) is a properly chosen confidence width multiplier that ensures \( U_t(x) \geq f(x) \), with probability at least \( 1 - \delta \). Following the standard analysis (see e.g., Srinivas et al., 2010), a probability union bound implies that, with probability at least \( 1 - \delta \), for all \( t \geq 1 \),
\[
    f(x^*) - f(x_t) \leq \mu_{t-1}(x^*) + \beta_t(3\delta/(\pi^2t^2))\sigma_{t-1}(x^*) - \mu_{t-1}(x_t) + \beta_t(3\delta/(\pi^2t^2))\sigma_{t-1}(x_t).
\]

The selection rule of GP-UCB with \( \alpha_t = \beta_t(3\delta/(\pi^2t^2)) \) indicates that \( \mu_{t-1}(x^*) + \alpha_t\sigma_{t-1}(x^*) \leq \mu_{t-1}(x_t) + \alpha_t\sigma_{t-1}(x_t) \). Therefore, for GP-UCB, with probability at least \( 1 - \delta \), \( f(x^*) - f(x_t) \leq 2\alpha_t\sigma_{t-1}(x_t) \), for all \( t \geq 1 \). Summing up both sides over \( t \) and applying
Cauchy-Schwarz inequality, we get, with probability at least \(1 - \delta\),
\[
R(T; \text{GP-UCB}) \leq 2\alpha_T \sqrt{D_T T},
\]
where \(D_T = \sum_{t=1}^{T} \sigma_{t-1}^2(x_t)\).

Srinivas et al. (2010) showed that, under the Bayesian setting with some mild regularity assumptions which hold for most typical kernels (e.g., SE and Matérn-\(\nu\) with \(\nu > 2\)), \(\beta_1(\delta) = \mathcal{O}(\sqrt{\log(t/\delta)})\), consequently \(\alpha_T = \mathcal{O}(\sqrt{T/\delta})\). Under the frequentist setting, Chowdhury and Gopalan (2017) established similar confidence bounds; albeit, with a much larger width multiplier \(\beta_1(\delta) = B + R \sqrt{2(\gamma_{t-1} + 1 + \log(1/\delta))}\) resulting in \(\alpha_T = \mathcal{O}(\sqrt{T})\), and consequently an \(\mathcal{O}(\sqrt{T})\) gap between the regret bounds, for GP-UCB, under the Bayesian and frequentist settings. It is unknown whether this confidence bound and the regret bounds for vanilla GP-UCB under the frequentist setting can be improved. The main challenge in establishing confidence bounds is the adaptivity of the observation sequence in the sequential optimization problem (in contrast to an offline setting with predetermined observation points). Of significant theoretical value, the SupKernelUCB algorithm (Valko et al., 2013) gets around this technicality and achieves \(\mathcal{O}(\sqrt{T} \log^2(T))\) regret through the use of an independent batch observation trick, which can be attributed back to Auer (2002). The original analysis of SupKernelUCB, which was given on a finite search space, can be extended to more general compact sets through a discretization argument, preserving \(\mathcal{O}(\sqrt{T})\) regret (see Cai and Scarlett, 2020, Appendix A.4).

It remains to bound the cumulative variance at the observation points, \(\mathcal{D}_T\). The standard approach for bounding \(\mathcal{D}_T\) is to use the information gain that refers to the mutual information \(I(y_t; \hat{f})\) (Cover and Thomas, 2012) between \(y_t\) and \(\hat{f}\). From the closed form expression of mutual information between two multivariate Gaussian distributions, we know that \(I(y_t; \hat{f}) = \frac{1}{2} \log \det(I + \frac{1}{\gamma} K_{X_tX_t})\). Using Jensen’s inequality, Srinivas et al. (2010) proved that \(\mathcal{D}_T \leq c_1 I(y_T; \hat{f})\) where \(c_1 = 2/\log(1 + 1/\gamma)\) is an absolute constant.

It is standard to proceed by defining a kernel-specific and \(X_T\)-independent maximal information gain,
\[
\gamma_T = \sup_{X_T \subseteq \mathcal{X}} I(y_T; \hat{f}).
\]

The regret bounds are then given in terms of \(\gamma_T\).

For specific kernels (Matérn and SE), Srinivas et al. (2010); Janz et al. (2020) proved upper bounds on \(\gamma_T\) which are commonly used to provide explicit regret bounds. The contribution of this paper is to derive novel bounds on \(\gamma_T\) (consequently, on \(\mathcal{D}_T\)) which immediately translate to improved regret bounds under various settings.

### 3 Upper Bounds on the Information Gain

Our bounds on the information gain are achieved through a finite dimensional projection of the GP model in the RKHS corresponding to \(k\). We start with outlining the details of the RKHS and the finite dimensional projection of the GP model. We then present the bounds on \(\gamma_T\).

#### 3.1 RKHS and Mercer’s Theorem

Consider a positive definite kernel \(k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}\) with respect to a finite Borel measure (e.g., the Lebesgue measure) supported on \(\mathcal{X}\). A Hilbert space \(H_k\) of functions on \(\mathcal{X}\) equipped with an inner product \(\langle \cdot, \cdot \rangle_{H_k}\) is called an RKHS with reproducing kernel \(k\) if the following are satisfied. For all \(x \in \mathcal{X}\), \(k(\cdot, x) \in H_k\), and for all \(x \in \mathcal{X}\) and \(f \in H_k\), \(\langle f, k(\cdot, x) \rangle_{H_k} = f(x)\) (reproducing property).

An RKHS is completely specified with its kernel function and vice-versa. The inner product induces the RKHS norm \(\|f\|_{H_k} = (f, f)_{H_k}\) that can be interpreted as a measure for the complexity of \(f\).

Mercer’s theorem provides an alternative representation for GP kernels as an inner product of infinite dimensional feature maps (see e.g., Kanagawa et al., 2018, Theorem 4.1).

**Theorem 1 (Mercer’s Theorem)**. Let \(k\) be a continuous kernel with respect to a finite Borel measure on \(\mathcal{X}\). There exists \(\{\lambda_m, \phi_m\}_{m=1}^{\infty}\) such that \(\lambda_m \in \mathbb{R}^+, \phi_m \in H_k\), for \(m \geq 1\), and
\[
k(x, x') = \sum_{m=1}^{\infty} \lambda_m \phi_m(x) \phi_m(x').
\]

The \(\{\lambda_m\}_{m=1}^{\infty}\) and the \(\{\phi_m\}_{m=1}^{\infty}\) are referred to as the eigenvalues and the eigenfeatures (or eigenfunctions) of \(k\), respectively. Throughout the paper, it is assumed that \(\{\lambda_m\}_{m=1}^{\infty}\) are in a decreasing order: \(\lambda_1 \geq \lambda_2 \geq \ldots\).
Our technical assumption on $k$, used in the analysis of $\gamma_T$, is specified next, that is the same as in Chat-terji et al. (2019), and holds for practically relevant kernels (cf. Riutort-Mayol et al., 2020).

**Assumption 1.** a) $k$ is a Mercer kernel (that is to satisfy the conditions of Mercer’s theorem). b) $\forall x, x' \in \mathcal{X}, |k(x, x')| \leq k$, for some $k > 0$. c) $\forall m \in \mathbb{N}, \forall x \in \mathcal{X}, |\phi_m(x)| \leq \psi$, for some $\psi > 0$.

As a result of Mercer’s theorem, we can express a GP sample $\hat{f}$ in terms of a weight vector in the feature space of $k$

$$\hat{f}() = \sum_{m=1}^{\infty} W_m \lambda_m^\frac{1}{2} \phi_m(),$$  \hspace{1cm} (5)

where the weights $W_m$ are i.i.d. random variables with standard normal distribution (see e.g., Kanagawa et al., 2018, Remark 4.4). It is straightforward to check that $\hat{f}$ given in (5) is a zero mean GP with kernel $k$. We refer to this representation as the feature space representation in contrast to the function space representation presented in §2.2.

The RKHS can also be represented in terms of $\{ (\lambda_m, \phi_m) \}_{m=1}^{\infty}$ using Mercer’s representation theorem (see e.g., Kanagawa et al., 2018, Theorem 4.2).

**Theorem 2** (Mercer’s Representation Theorem). Let $\{ (\lambda_m, \phi_m) \}_{m=1}^{\infty}$ be the same as in Theorem 1. Then, the RKHS of $k$ is given by

$$H_k = \left\{ f() = \sum_{m=1}^{\infty} w_m \lambda_m^\frac{1}{2} \phi_m() : \| f \|_{H_k} = \sum_{m=1}^{\infty} w_m^2 < \infty \right\}.$$  

Mercer’s representation theorem provides an explicit definition for the RKHS norm. It also indicates that $\{ \lambda_m^\frac{1}{2} \phi_m \}_{m=1}^{\infty}$ form an orthonormal basis for $H_k$.  

### 3.2 Projection onto a Finite Dimensional Space

The feature space representation of typical GP kernels is infinite dimensional. To overcome the difficulty of working in infinite dimensional spaces, we use a projection $P_D$ on a $D$ dimensional RKHS consisting of the first $D$ features (corresponding to the $D$ largest eigenvalues of the kernel). Specifically, consider the $D$-dimensional feature space $\phi_D() = [\phi_1(), \phi_2(), \ldots, \phi_D()]^T$, the $D$-dimensional column vector $W_D = [W_1, W_2, \ldots, W_D]^T$ and the diagonal matrix $\Lambda_D = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_D)$ with $[\lambda_1, \lambda_2, \ldots, \lambda_D]$ as the diagonal entries. The projection of $\hat{f}$ on the $D$-dimensional space is given by

$$P_D[\hat{f}()] = W_D^\frac{1}{2} \Lambda_D^\frac{1}{2} \phi_D().$$

Notice that $P_D[\hat{f}]$ is a zero mean GP with kernel $k_P(x, x') = \sum_{m=1}^{D} \lambda_m \phi_m(x) \phi_m(x')$. We used the subscript $P$ to signify the space resulted from the projection. In addition, let $P_D^\perp[\hat{f}] = f - P_D[\hat{f}]$ be the orthogonal part of $\hat{f}$ with respect to the projection. Notice that $P_D^\perp[\hat{f}]$ is also a GP, with kernel $k_O(x, x') = k(x, x') - k_P(x, x')$. We used the subscript $O$ to signify the orthogonal part.

We define the following quantity based on the tail mass of the eigenvalues of $k$

$$\delta_D = \sum_{m=D+1}^{\infty} \lambda_m \psi^2.$$  \hspace{1cm} (6)

If $\lambda_m$ diminishes at a sufficiently fast rate (see Definition 1), $\delta_D$ becomes arbitrarily small when $D$ is large enough. For all $x, x' \in \mathcal{X}$, we then have $k_O(x, x') \leq \delta_D$.

### 3.3 Analysis of the Information Gain

Here, we establish a novel upper bound on $\gamma_T$.

**Theorem 3** (Bounding $\gamma_T$). Consider a GP with a kernel $k$ satisfying Assumption 1. For $D \in \mathbb{N}$, let $\delta_D$ be as defined in (6). The following upper bound on $\gamma_T$, defined in (4), holds for all $D \in \mathbb{N}$.

$$\gamma_T \leq \frac{1}{2} D \log \left( 1 + \frac{k_T}{\tau D} \right) + \frac{1}{2} \delta_p T.$$  \hspace{1cm} (7)

The expression can be simplified as

$$\gamma_T = O\left(D \log(T) + \delta_D T\right).$$

In contrast to the existing results, Theorem 3 provides an upper bound in terms of the spectral properties of the GP kernel through $\delta_D$ that is applicable to all kernels based on their eigendecay. Specializing this bound for common kernels (Matérn and SE) results in tight upper bounds on $\gamma_T$ (up to a $\log(T)$ factor), and consequently improved bounds on regret, across various settings, compared to the existing ones.

**Proof Sketch.** Recall $I(y_t; \hat{f}) = \frac{1}{2} \log \det(I_t + \frac{1}{T} K_{X_t, X_t})$. The problem is thus bounding the log det of the covariance matrix $I_t + \frac{1}{T} K_{X_t, X_t}$ for an arbitrary sequence.
The following corollary is a consequence of Theorem 3.

**Corollary 1.** Consider \( \tau_T \) defined in (4). If \( k \) has a \((C_p, \beta_p) \) polynomial eigendecay, we have

\[
\tau_T \leq \left( \frac{C_p \psi^2 T}{\tau} \right)^{\frac{1}{\beta_p}} \log \left( \frac{1}{\tau} (1 + \frac{kT}{\tau}) + 1 \right) \log \left( 1 + \frac{kT}{\tau} \right).
\]

The expression can be simplified as \( \tau_T = O \left( T^{\frac{1}{\beta_p}} \log^{1+\frac{1}{\beta_p}} (T) \right) \).

If \( k \) has a \((C_{e,1}, C_{e,2}, \beta_e) \) exponential eigendecay, we have

\[
\tau_T \leq \left( \frac{2}{C_{e,2}} \log(T) + C_{\beta_e} \right)^{\frac{1}{\beta_e}} + 1 \right) \log \left( 1 + \frac{kT}{\tau} \right),
\]

where \( C_{\beta_e} = \log \left( \frac{C_{e,1} \psi^2}{C_{e,2}} \right) \) if \( \beta_e = 1 \), and \( C_{\beta_e} = \log \left( \frac{2C_{e,1} \psi^2}{\tau C_{e,2}} \right) + \left( \frac{1}{\beta_e} - 1 \right) \left( \log \left( \frac{1}{\tau} \right) - 1 \right) \), otherwise. The expression can be simplified as \( \tau_T = O \left( \log^{1+\frac{1}{\beta_e}} (T) \right) \).

Corollary 1 gives general bounds on \( \tau_T \) providing the polynomial and exponential conditions on the eigendecay of \( k \) are satisfied. A detailed proof is provided in the supplementary material.

**Remark 2.** It is known that, in the case of a Matérn kernel with smoothness parameter \( \nu > \frac{1}{2} \), \( \lambda_m = O \left( m^{-\frac{2+\nu}{2}} \right) \) (Santin and Schaback, 2016); and, in the case of SE kernel, \( \lambda_m = O \left( \exp(-m^\beta) \right) \) (Belkin, 2018). Also, see Rüttort-Mayol et al. (2020) which gave closed form expression of their eigenvalue-eigenfeature pairs on hypercubes. Thus, as special cases of polynomial and exponential eigendecays, we have

\[
\gamma_T = O \left( T^{\frac{\nu}{\nu+2}} \log^{\frac{\nu}{\nu+2}} (T) \right), \quad \text{for Matérn-\nu kernel,}
\]

\[
\gamma_T = O \left( \log^{d+1} (T) \right), \quad \text{for SE kernel,}
\]

which are tight up to \( \log(T) \) factors, based on the lower bounds reported in Scarlett et al. (2017).
We thank Jonathan Scarlett for insightful comments on an earlier version of this paper.

Acknowledgment

We thank Jonathan Scarlett for insightful comments on an earlier version of this paper.

References


S. R. Chowdhury and A. Gopalan. On kernelized multi-


