Thesis for The Degree of Licentiate of Engineering

Theoretical Understanding of Gaussian Process Bandits in Practical Applications

Jack Sandberg

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg
Gothenburg, Sweden, 2026

Theoretical Understanding of Gaussian Process Bandits in Practical Applications

Jack Sandberg

ISSN 1652-876X

Department of Computer Science and Engineering
Division of Data Science and AI
Chalmers University of Technology | University of Gothenburg
SE-412 96 Göteborg,
Sweden
Phone: +46(0)31 772 1000

Printed by Chalmers Digitaltryck,
Gothenburg, Sweden 2026.

Theoretical Understanding of Gaussian Process Bandits in Practical Applications

Jack Sandberg

Department of Computer Science and Engineering
Chalmers University of Technology | University of Gothenburg

Abstract

Bayesian optimization (BO) provides a principled framework for optimizing blackbox functions with noisy outputs, with applications ranging from aircraft design to hyperparameter tuning. BO algorithms can be theoretically analyzed through the lens of Gaussian process (GP) bandits, providing theoretical guarantees of their efficiency. This thesis provides theoretical analyses and experimental evaluation of GP bandit algorithms with relevance to practical applications.

The first part of the thesis is motivated by the need for navigation systems that prioritize energy-efficiency and adapt to collected data. As such, we develop a combinatorial GP bandit framework for online energy-efficient navigation of electric vehicles. We theoretically analyze three algorithms under this framework, providing bounds on their regret. The algorithms are evaluated on real-world road networks, demonstrating that they explore the road network more efficiently compared to previous work.

The second part of the thesis is focused on addressing a discrepancy between theory and practice. The theoretical literature commonly assumes that important characteristics (the prior) of the blackbox function being optimized is known before the optimization process starts. However in practice, the prior must often be inferred from the data, which invalidates any theoretical guarantees. To address this issue, we theoretically analyze two algorithms that simultaneously learn the prior and optimize the unknown function. We identify issues in previous theoretical analyses, correct and improve upon their results. Finally, we experimentally evaluate the algorithms on synthetic and real-world data and demonstrate their effectiveness at simultaneous optimization and prior identification.

Keywords

Multi-armed bandits, Gaussian processes, Bayesian optimization, machine learning, combinatorial bandits, energy-efficient navigation, Thompson sampling

List of Publications

Appended publications

This thesis is based on the following publications:

[**Paper I**] **J. Sandberg**, N. Åkerblom, M. Haghir Chehreghani, *Bayesian Analysis of Combinatorial Gaussian Process Bandits*
International Conference on Learning Representations (ICLR) 2025.

[**Paper II**] **J. Sandberg**, M. Haghir Chehreghani, *Comments on "Surrogate Modeling for Bayesian Optimization Beyond a Single Gaussian Process"*
Submitted, under review.

[**Paper III**] **J. Sandberg**, M. Haghir Chehreghani, *Adaptive Prior Selection in Gaussian Process Bandits with Thompson Sampling*
Submitted, under review.
An earlier version of this work was presented at the 18th European Workshop on Reinforcement Learning (EWRL), 2025, under the title "Efficient Prior Selection in Gaussian Process Bandits with Thompson Sampling".

For the publications listed above, **Jack Sandberg** derived the theoretical results, wrote and ran the code for the experiments, and wrote most parts of the publications. The coauthors provided input and guidance to the theoretical analysis, experimental design, writing and editing.

Other publications

The following publication was published during my PhD studies. However, my contributions to the publication were made prior to starting my PhD studies and the content of the publication is not related to the thesis.

[a] D. Spensieri, **J. Sandberg**, E. Åblad, J. Torstensson, J. Kressin, D. Strand, R. Lindqvist, *Automatic Planning and Optimization of a Laser Radar Inspection System*
International Journal of Computer Integrated Manufacturing 39(2) (April 2025), 303-317.

Acknowledgment

First, I would like to thank my supervisor Morteza Haghir Chehreghani. I am grateful for the support, supervision and encouragement you have provided through the journey thus far.

Second, I would like to thank all the colleagues at DSAI, past and present, for the great discussions and supportive atmosphere. I would also like to thank my office colleagues for the guidance and feedback you have provided, helping me become a better researcher.

Third, I would like to thank my friends and family for your support throughout the years.

Lastly, I want to thank the Wallenberg AI, Autonomous Systems and Software Program (WASP), funded by the Knut and Alice Wallenberg Foundation, for supporting the work in the thesis and the opportunities WASP has provided.

Contents

Paper III - Adaptive Prior Selection in Gaussian Process Bandits with Thompson Sampling

Part I

Summary

Chapter 1

Introduction

The problem of efficiently finding the best solution through sequential trial-and-error is widespread in modern applications. Examples include deciding what product should be at the top of a website, or identifying the frequency with minimal interference between a cell tower and a mobile phone. For such problems, naively trying all options is at best simply inefficient but at worst completely infeasible.

The multi-armed bandit literature provides a plethora of algorithms to solve sequential decision making problems with theoretical guarantees of their efficiency. While initially proposed for clinical trials (Thompson, 1933), modern applications include recommendation systems for videos, music, and podcasts (Yi et al., 2023; Feijer et al., 2025; Zhang et al., 2025), drug design (Svensson et al., 2022), and mobile health interventions for maternal and oral health (Dasgupta et al., 2025; Trella et al., 2025).

For these applications, additional structural assumptions must be made to ensure the best solution can be efficiently found within a reasonable time frame. Specifically, the Gaussian process (GP) bandit problem can be applied to problems where the available options (or actions) exhibit strong interdependencies and the number of actions is large or even infinite. GP bandits have been applied to hyperparameter optimization of machine learning algorithms (Snoek et al., 2012; Turner et al., 2021), portfolio optimization (Gonzalvez et al., 2019), aircraft design (Priem et al., 2020), among other problems.

A particular application that we focus on in this thesis is the problem of online energy-efficient navigation for electric vehicles. Rapid adoption of electric vehicles is an important step to reduce greenhouse gas emissions, making the mitigation of range anxiety an important issue to accelerate electric vehicle adoption (Alanazi, 2023). Software systems that provide energy-efficient routing are a cost-effective way to increase the effective range. However, the real-world energy consumption can be difficult to estimate in advance and as such online algorithms that learn to recommend routes through trial-and-error are desired (Åkerblom, 2024).

In **Paper I**, we extend a previous framework for online energy-efficient navigation of electric vehicles (Åkerblom et al., 2023) to use GP bandits and find

energy-efficient paths faster. We express the online energy-efficient navigation problem as an instance of a combinatorial and contextual multi-armed bandit problem and provide a theoretical analysis of three algorithms for this problem. The algorithms are experimentally evaluated using synthetic data on real-world road networks where a reduction in energy consumption is achieved compared to (Åkerblom et al., 2023).

As mentioned, GP bandits can be applied to problems where the actions exhibit dependence. Most of the theoretical literature assumes that the exact nature of this dependence is known to the practitioner beforehand. However, in real-world scenarios, this dependence must often be adaptively learned while simultaneously finding the best action. In **Papers II** and **III**, we narrow this gap by studying two Thompson sampling-based (Thompson, 1933) algorithms that can adaptively learn this dependence while maintaining theoretical guarantees. The first algorithm extends a previous algorithm (Ziomek et al., 2025), and the second has been studied before for GP bandits (Lu et al., 2023) and linear bandits (Hong et al., 2022). In **Paper II**, we identify issues with the theoretical analysis of the second algorithm for GP bandits (Lu et al., 2023). In **Paper III**, we analyze the regret of both algorithms, addressing issues with previous analyses, and experimentally evaluate them on synthetic and real-world data, observing the benefits of the Thompson sampling algorithms.

The remainder of the thesis is structured as follows. Chapter 2 introduces the concepts and problems studied in the thesis, Chapter 3 provides a summary of **Papers I** to **III**, and Chapter 4 provides a brief discussion of the findings in the papers and directions for future work.

Chapter 2

Background

In this chapter, we provide an introduction to the topics and problems considered in the appended papers.

2.1 Gaussian Processes

Gaussian processes (GPs) are a generalization of the multivariate Gaussian distribution. Whereas samples from the multivariate Gaussian distribution are vectors in finite dimensions and have a finite number of elements, samples from a GP can have an infinite (and even an uncountable) number of elements. As such, a sample from a GP can be viewed as a function on a continuous domain. The definition of GPs provides a principled method for probabilistically estimating non-linear functions from noisy data with only a few samples. In this section, we define Gaussian processes and discuss important properties of them.

2.1.1 Definition and kernels

Consider the set $\mathcal{X} \subseteq \mathbb{R}^d$, a Gaussian process $f(x) \sim \mathcal{GP}(\mu, k)$ with mean function $\mu : \mathcal{X} \mapsto \mathbb{R}$ and covariance (or kernel) function $k : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ is formally defined as a collection of random variables such that for any subset $\{x_1, \ldots, x_n\} \subseteq \mathcal{X}$, the vector $[f(x_1), \ldots, f(x_n)] \in \mathbb{R}^n$ has a multivariate Gaussian distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with mean vector $\boldsymbol{\mu} = [\mu(x_1), \ldots, \mu(x_n)]$ and covariance matrix $(\boldsymbol{\Sigma})_{ij} = k(x_i, x_j)$ for $i, j = 1, \ldots, n$. The mean function $\mu(x)$ equals the expected value of $f(x)$ and is often assumed to be a constant function or simply equal to zero for all $x \in \mathcal{X}$. The covariance function $k(x, x')$ is equal to the covariance between $f(x)$ and $f(x')$ but is more often referred to as the kernel function.

The kernel function determines many important properties of $f(x)$. In Fig. 2.1, we visualize sample functions $f(x) \sim \mathcal{GP}(0, k)$ for the following three

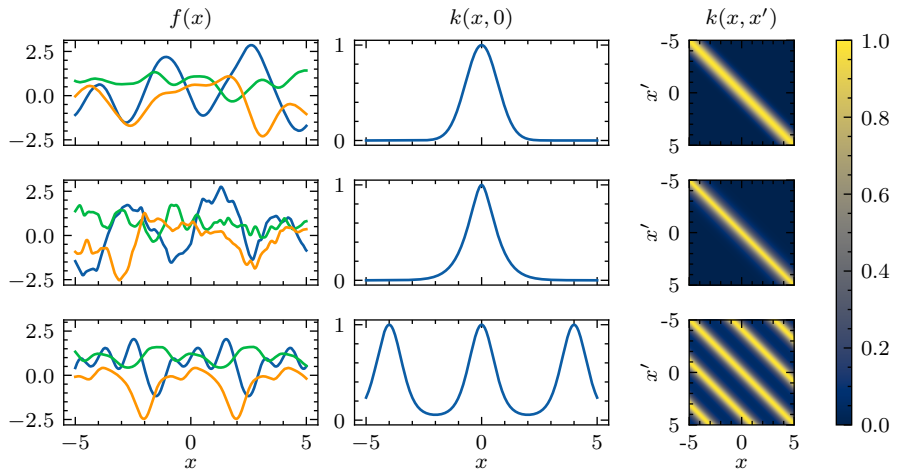


Figure 2.1: Samples from Gaussian processes (left) with the kernels visualized in one and two dimensions (middle and right). The top, middle and bottom row use the RBF, Matérn and periodic kernel respectively.

kernels:

$$k_{RBF}(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\ell^2}\right), \tag{2.1}$$

$$k_M(x, x') = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}\|x - x'\|}{\ell}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}\|x - x'\|}{\ell}\right), \tag{2.2}$$

$$k_P(x, x') = \exp\left(-2\sum_{i=1}^{d}\frac{1}{\ell}\sin^2\left(\frac{\pi}{p}(x_i - x_i')\right)\right). \tag{2.3}$$

The first kernel k_{RBF} is commonly called the radial basis function (RBF) kernel or the squared-exponential kernel. An important property that the RBF kernel imposes on the samples $f(x)$ is that they are infinitely continuously differentiable (i.e. $f \in C^\infty$). The second kernel k_M is called the Matérn kernel (Matérn, 1986) and it is defined in terms of the Gamma function $\Gamma(\nu)$ and the modified Bessel function of the second kind K_ν (Williams & Rasmussen, 2006). Unlike the RBF kernel, the Matérn kernel only guarantees that the samples $f(x)$ are k-times continuously differentiable ($f \in C^k$) for $k < \nu$ where ν is the smoothness parameter. However, in the limit $\nu \to \infty$, the Matérn kernel is equal to the RBF kernel. The third kernel k_P is called the periodic kernel and, as the name implies, it imposes that the samples $f(x)$ are periodic with period p (Mackay, 1998; Gardner et al., 2018). Inspecting the samples in Fig. 2.1, we can observe that the samples from the RBF kernel (top) are smoother than the jagged samples from the Matérn kernel with $\nu = 1.5$ (middle) and the samples from the periodic kernel exhibit periodicity (bottom).

For the kernels in Eqs. (2.1) to (2.3), the correlation between $f(x)$ and $f(x')$ depends on the distance $\|x - x'\|$. As $\|x - x'\|$ increases, $f(x)$ and $f(x')$ become

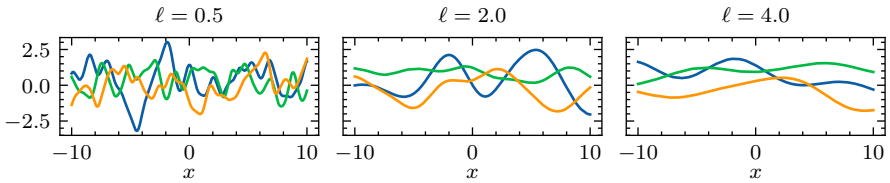


Figure 2.2: Sample functions $f(x)$ from $\mathcal{GP}(0, k_{\text{RBF}})$ with lengthscales $\ell \in \{0.5, 2, 4\}$.

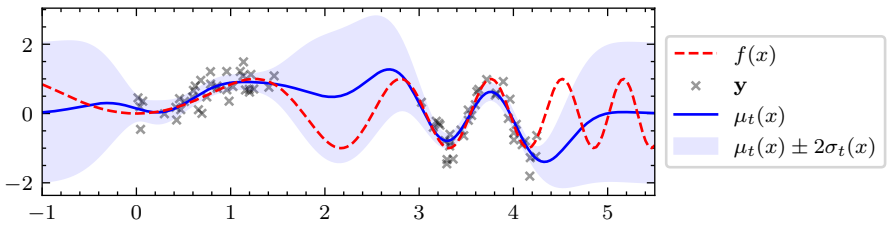


Figure 2.3: The conditional distribution of a Gaussian process after observing noisy data from the function $f(x) = \sin(x^2)$.

less correlated. The lengthscale parameter $\ell > 0$ regulates how quickly this correlation decreases. For large lengthscales, even distant points have a high correlation whereas for a small lengthscale the correlation between close points can be low. An alternative view is that the lengthscale parameter regulates how quickly the function $f(x)$ varies. Short (or long) lengthscales increase (or decrease) the magnitude of the derivative process $f'(x)$. We visualize sample functions with varying lengthscales in Fig. 2.2.

2.1.2 GP regression

A critical property of the multivariate Gaussian distribution that makes it such a versatile tool is that it is closed under many types of operations, i.e. if you apply an operation to one or more multivariate Gaussian distributions then the resulting distribution is still a multivariate Gaussian distribution (most often with different parameters). The specific operation that we are interested in when performing regression is conditioning on the observed data points. That is, if we have observed the values $(y_i)_{i=1}^{t-1}$ at locations $(x_i)_{i=1}^{t-1}$ where $y_i = f(x_i) + \epsilon_i$ for $i = 1, \ldots, t-1$ and the noise ϵ_i is identically and independently sampled from $\mathcal{N}(0, \sigma^2)$, then the distribution of $f(x)|(y_i)_{i=1}^{t-1}$ is a Gaussian process with mean and kernel function:

$$\mu_t(x) = \mu(x) - \mathbf{k}(x)^\top \left(\mathbf{K} + \sigma^2 I\right)^{-1} (\mathbf{y} - \boldsymbol{\mu}), \tag{2.4}$$

$$k_t(x, x) = k(x, x') - \mathbf{k}(x)^\top (\mathbf{K} + \sigma^2 I)^{-1} \mathbf{k}(x'). \tag{2.5}$$

Above, $\mathbf{k}(x), \mathbf{k}(x') \in \mathbb{R}^{t-1}$ are vectors with elements $(\mathbf{k}(x))_i = k(x_i, x)$ and $(\mathbf{k}(x')_i = k(x_i, x')$. Similarly, $\mathbf{y}, \boldsymbol{\mu} \in \mathbb{R}^{t-1}$ are vectors with elements $(\mathbf{y})_i = y_i$

and $(\boldsymbol{\mu})_i = \mu(x_i)$. The matrix $\mathbf{K} \in \mathbb{R}^{t-1 \times t-1}$ is the covariance matrix with elements $(K)_{ij} = k(x_i, x_j)$.

In Fig. 2.3, we perform GP regression with noisy data collected from the function $f(x) = \sin(x^2)$. The observations are clustered into two regions and inside each region the predicted distribution, or *the posterior distribution*, makes accurate predictions with tight confidence intervals. In between the two regions, the posterior mean $\mu_t(x)$ follows the overall shape of $f(x)$ but makes less accurate predictions. However, the uncertainty increases since no data has been collected nearby. Outside of the two regions with data, the posterior mean $\mu_t(x)$ reverts to the prior mean of zero and the uncertainty. This example highlights that GPs can be used to fit complex functions given sufficient data (even if perturbed by noise) and provides well-calibrated uncertainty quantification.

The major drawback of GP regression is that the computational requirements increase quickly as the number of samples increase. Computing the inverse of $\mathbf{K} + \sigma^2 I$ has a cubic complexity, $\mathcal{O}(t^3)$, with respect to the number of samples. Thus, GP regression can quickly become intractable as the number of samples increase.

To overcome this problem, Hensman et al. (2013) proposed Stochastic Variational Inference for GPs (SVGP). SVGP uses a sparse set of inducing points that summarizes the data. The location and values of the inducing points are optimized using small batches of the data. Therefore, if the data can be summarized by a small number of inducing points, SVGP regression can approximate the exact GP regression output at a much lower computational cost.

2.1.3 Bayesian Optimization

Another popular use case for Gaussian processes is optimizing an unknown function, so called Bayesian optimization (BO) (Garnett, 2023). BO provides a principled way to optimize functions that do not have a closed form, are costly to evaluate, and do not provide gradient information. The probabilistic nature of Gaussian processes ensures that BO can also handle outputs that have been corrupted by (Gaussian) noise.

BO is performed sequentially using an acquisition function $\alpha(x)$ with the goal of finding the optimizer $x^* = \arg\max f(x)$. The acquisition function $\alpha(x)$ describes the utility of evaluating $f(x)$ at an input location x based on the current posterior GP distribution. At time step t, the optimization procedure first selects an input location $x_t = \arg\max_x \alpha(x)$. Then, the function f is evaluated at input x_t, yielding an output y_t which is a (potentially) noisy evaluation of $f(x_t)$. The new observation (x_t, y_t) is used to update the posterior GP distribution and the procedure is repeated.

Examples of acquisition functions include probability of improvement, expected improvement and upper confidence bounds (Kushner, 1962, 1964; Mockus, 1975; Srinivas et al., 2012). The expected improvement is likely the most commonly used acquisition function (Garnett, 2023) and is defined as $\alpha(x) = \mathbb{E}[[f(x) - y^*]_+]$ where $[\cdot]_+ := \max(0, \cdot)$ and y^* is the incumbent, i.e. the believed optimal value. The incumbent is often set to the best observed

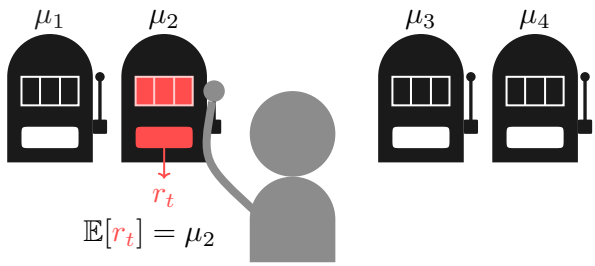


Figure 2.4: Agent pulling an arm in a multi-armed bandit problem.

value or the maximum of the posterior mean. As the name suggests, the input locations that maximize the expected improvement acquisition function are those that are expected to yield the highest improvement compared to the incumbent.

To analyze the theoretical performance of BO algorithms, the Bayesian optimization problem can be cast as a multi-armed bandit problem, which we will discuss next.

2.2 Multi-Armed Bandits

The multi-armed bandit (MAB) problem is a classical problem first considered by Thompson (1933) but reintroduced and modernized by Robbins (1952). In the MAB problem, we consider an agent that can select among K different actions (or *arms*) over a sequence of T rounds. Every round t, the agent must select an arm $a_t \in [K]$ where $[K] = \{1, \dots, K\}$. The agent receives a (stochastic) reward r_t from the distribution $\nu(a_t)$ with mean $\mu(a_t)$. The goal of the agent is to maximize the sum of rewards it collects over the horizon T. An efficient agent must balance trying new arms (exploration) against always selecting the best known arm (exploitation). If the agent selects a suboptimal arm, then it misses out on the extra bit of reward it could have obtained from the optimal arm and therefore, in expectation, suffers a regret. The goal of the agent can therefore be equivalently stated as to minimize the amount of regret it suffers. Formally, we define the cumulative regret as

$$R(T) = \sum_{t \in [T]} \mu(a^*) - \mu(a_t) \tag{2.6}$$

where $a^* \in \arg\max_{a \in [K]} \mu(a)$. In most settings in this thesis, we instead consider the expected regret $\mathrm{BR}(T) = \mathbb{E}[R(T)]$ where the expectation is taken with respect to all the randomness in the MAB problem and eventual randomness in the policy of the agent.

A relatively simple but effective algorithm for the standard MAB problem is ϵ-greedy. At every round t, the agent selects a random arm with probability ϵ_t or selects the arm with the highest estimated mean with probability $1 - \epsilon_t$. With an appropriate choice of the exploration schedule ϵ_t and assuming the rewards

$(r_t)_{t=1}^T$ are bounded, it can be shown that ϵ-greedy achieves *sublinear* expected regret, i.e. the regret grows slower than a linear function for sufficiently large T (Slivkins, 2019). The notion of sublinear regret is very important since it implies that the agent will eventually find the best arm.

Upper-confidence bound algorithms are a family of algorithms that build upon the idea of optimism in the face of uncertainty (Auer et al., 2002). These algorithms maintain both an estimated mean $\mu_t(a)$ and a confidence radius $\sigma_t(a)$ for each arm $a \in [K]$. Then at each iteration t, they select the arm that maximizes the upper confidence bound $U_t(a) = \mu_t(a) + \sigma_t(a)$. The idea is that $U_t(a)$ is large either if the estimated mean $\mu_t(a)$ is large (exploitation) or if the uncertainty $\sigma_t(a)$ is high (exploration).

The final algorithm that we discuss here is Thompson sampling (Thompson, 1933). Thompson sampling is a Bayesian algorithm that assumes that the mean of each arm $\mu(a)$ is sampled from a prior distribution. Then, at each time step t, the arm a_t is sampled from the posterior distribution of the optimal arm a^*. The idea of the algorithm is that Thompson sampling explores the different arms in the beginning because the uncertainty of the optimal arm a^* is high. But as more information is collected, the posterior distribution concentrates on fewer and fewer arms leading Thompson sampling to explore less and less. In many problems, computing the posterior distribution of a^* is intractable. However, Thompson sampling can be equivalently expressed in its more computationally tractable form; in every round t, sample the mean $\tilde{\mu}_t(a)$ from the posterior distribution for all arms $a \in [K]$ and select the arm $a_t = \arg\max_{a \in [K]} \tilde{\mu}_t(a)$.

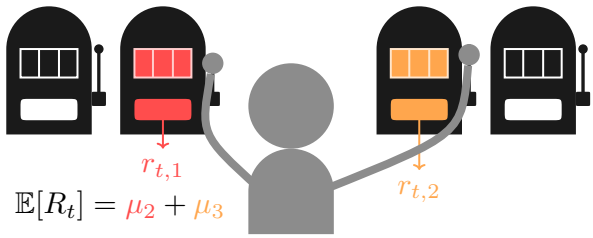
2.2.1 Extensions of the MAB problem

The simple formulation of the MAB problem allows many problems to be expressed as a MAB, such as news recommendation, molecule design, or optimal sensor placement. However, in many real-world applications, the agent has access to additional information about the structure of the problem that is either impossible to encode into the problem, or can be encoded but yields too many arms. For example, a news recommendation problem can involve ranking a set of news articles after their relevance. While it is possible to consider all permutations of the news articles as arms, trying all the arms in a reasonable time frame is not feasible.

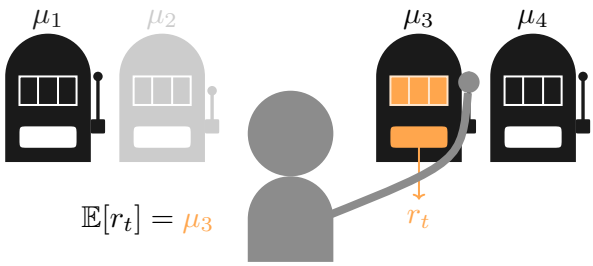
Therefore, a lot of extensions to the MAB problem have been proposed and analyzed. In particular, we will focus on *combinatorial, volatile and Gaussian process* extensions. We visualize the extensions in Fig. 2.5.

In a combinatorial semi-bandit problem, the agent selects a subset of the arms each round instead of a single arm, typically with some restrictions on which subsets can be chosen. The agent receives a base reward from each arm that is selected and a total reward which is the sum of the base rewards. The base rewards help the agent attribute which arms contribute most to the total reward and simplifies the learning process.

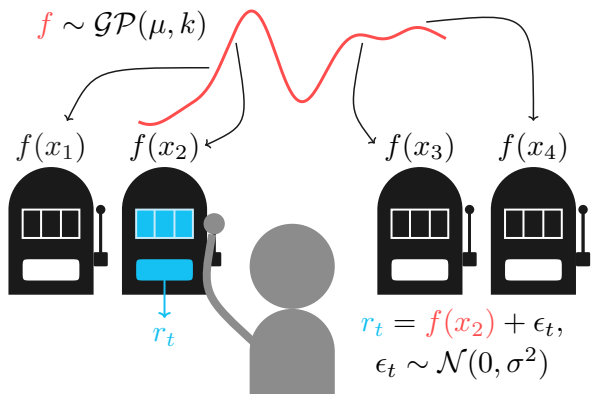
In a volatile multi armed bandit problem, the availability of each arm varies over time to model problems where an option might become unavailable. The



(a) A combinatorial semi-bandit problem.



(b) A volatile multi-armed bandit problem with the second arm unavailable.



(c) A Gaussian process bandit problem.

Figure 2.5: Visualizations of three extensions to the multi-armed bandit problem.

availability can either be fixed in advance or be stochastic.

In a Gaussian process bandit problem, the mean reward is a function $f : \mathcal{A} \mapsto \mathbb{R}$ of the arms $\mathcal{A} \subset \mathbb{R}^d$ and is assumed to be correlated and sampled from a Gaussian process $f \sim \mathcal{GP}(\mu, k)$. Since the arms are correlated, selecting one arm can reveal information about the other arms which simplifies the learning problem. Additionally, the GP structure permits us to consider an infinite number of arms that are distributed in a high-dimensional space.

Next, we will combine these three extensions and formulate the *combinatorial volatile Gaussian process* MAB (CV-GP-MAB) that we consider in **Paper I**. At

every time step t, a possibly random and finite subset of base arms $\mathcal{A}_t \subset \mathcal{A} \subset \mathbb{R}^d$ is available to the agent. The agent must select a feasible subset of base arms, which we call a super arm, $\mathbf{a}_t \in \mathcal{S}_t$ where $\mathcal{S}_t \subset 2^{\mathcal{A}_t}$ are the feasible and available super arms in round t. The agent observes the base rewards of the selected base arms $\boldsymbol{r}_t = \{r_{t,a} | a \in \mathbf{a}_t\}$ where $r_{t,a} = f(a) + \epsilon_{t,a}$ and $\epsilon_{t,a}$ is zero-mean Gaussian noise. The total reward is the sum of base arm rewards $R_t = \sum_{a \in \mathbf{a}_t} r_{t,a}$ and the goal of the agent is to minimize the Bayesian cumulative regret over a horizon T:

$$\mathrm{BR}(T) = \mathbb{E}\Big[\sum_{t \in [T]} f(\mathbf{a}_t^*) - f(\mathbf{a}_t) \Big] \tag{2.7}$$

where $\mathbf{a}_t^* = \arg\max_{\mathbf{a}} f(\mathbf{a})$ and $f(\mathbf{a}) = \sum_{a \in \mathbf{a}} f(a)$. The volatility appears both in the available base arms $\mathcal{A}_t$ and the feasible super arms $\mathcal{S}_t$.

2.2.2 Regret bounds and maximum information gain

As discussed in Section 2.2, the goal of the agent is to minimize the cumulative regret $R(T)$ or the Bayesian cumulative regret $\mathrm{BR}(T) = \mathbb{E}[R(T)]$. To theoretically evaluate algorithms, it is common practice to provide upper bounds of $R(T)$ or $\mathrm{BR}(T)$ with respect to the horizon T and other important parameters of the problem, such as the number of arms K. Upper bounds for the cumulative regret $R(T)$ provide a worst-case guarantee and it is common to only provide a bound that holds with high probability. In contrast, the Bayesian cumulative regret provides a guarantee on the average performance and these bounds hold with probability 1. An important criteria for the bounds is *sublinearity* with respect to T, i.e. that for any $c > 0$, $R(T)$ or $\mathrm{BR}(T)$ is smaller than cT for sufficiently large T.

For GP-bandits problems, the regret bounds are generally $\mathcal{O}(\sqrt{T\gamma_T \log(|\mathcal{A}|T)})$ where γ_T is the *maximum information gain* (MIG) (Srinivas et al., 2012).[1] The MIG is a measure of the complexity of learning a function $f \sim \mathcal{GP}(\mu, k)$. If learning the function f is difficult then γ_T is large and the agent will incur more regret. More formally, let $\mathbf{y}_A$ denote noisy observations of f at locations $A \subset \mathcal{A}$ then the MIG is defined as

$$\gamma_T := \sup_{A \subset \mathcal{A}, |A| \leq T} I(\mathbf{y}_A; f), \tag{2.8}$$

where $I(\mathbf{y}_A; f) = H(f) - H(f|\mathbf{y}_A)$ is the mutual information between $\mathbf{y}_A$ and f and $H(\cdot)$ is the entropy. The definition can be interpreted as the reduction in uncertainty of f after observing the T most informative data points. For the RBF and Matérn kernels, $\gamma_T = \mathcal{O}(\log^{d+1}(T))$ and $\gamma_T = \mathcal{O}(T^{\frac{d}{2\nu+d}} \log^{\frac{2\nu}{2\nu+d}}(T))$. We visualize upper bounds of γ_T for the three kernels in Eqs. (2.1) to (2.3) in Fig. 2.6.

[1] Note that we may assume that the set of arms is continuous ($|\mathcal{A}| = \infty$) in which case $|\mathcal{A}|$ is replaced by the number of discretization points used in the theoreotical analysis.

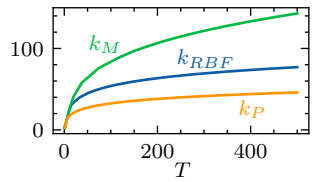


Figure 2.6: The maximum information gain γ_T for the Matérn kernel, the RBF kernel and the periodic kernel.

2.3 Online Energy-Efficient Navigation

To exemplify the online energy-efficient navigation problem, consider a driver commuting to work every day that wants to find a route that minimizes the energy consumption of their commute. Every morning, the driver tries a new route and measures the total energy consumed. However, the problem is complicated by traffic and weather that introduce an element of randomness to the measurements. In addition to the combinatorial amount of paths available, the driver must repeat the measurements to avoid being fooled by randomness. It seems that the drivers experiment will take forever. However, by formulating the online energy-efficient navigation problem as a CV-GP-MAB, we will see that the problem becomes manageable.

Let $\mathcal{G}(\mathcal{V}, \mathcal{E})$ be a directed graph where the vertices $\mathcal{V}$ represent intersections and edges $\mathcal{E}$ represent road segments. Each road segment e has a time-varying context vector $x_{t,e} \in \mathcal{X} \subset \mathbb{R}^d$ that includes features of the road segment that are indicative of the amount of energy the vehicle will consume such as length, incline and speed limit. We assume that there exists a start vertex $s \in \mathcal{V}$ and goal vertex $g \in \mathcal{V}$. Let $\mathcal{P}_t$ be the set of all simple and feasible paths at time t from s to g where a path is simple if no vertices are visited more than once.

At every round $t \in [T]$, the agent selects a path $\mathbf{p}_t \in \mathcal{P}_t$ and observes the energy consumed along each edge in the path. The goal of the agent is to minimize the total amount of energy consumed across the horizon T. To facilitate efficient learning, we assume the mean energy consumption along each road segment is sampled from a GP over the joint space $\mathcal{E} \times \mathcal{X}$: $f(e, x) \sim \mathcal{GP}(\mu, k)$. The energy consumption along a road segment is therefore dependent on both the local structure of the network and the features of the individual road segment.

2.4 Unknown prior GP-MAB

In the GP-bandit problems discussed so far, we have assumed that the mean μ and kernel function k of the prior $\mathcal{GP}(\mu, k)$ are known. When the prior is known, the GP extensions of UCB and Thompson sampling are known to have sublinear regret. In practical settings, is it rarely the case that the prior is known. Instead, it is common to parametrize the kernel function k_θ and estimate the parameter θ through maximum likelihood based on the data observed so far. However, since the data is collected adaptively there are no

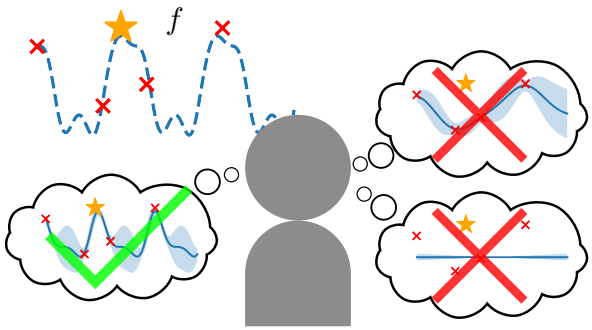


Figure 2.7: Posterior distributions of f for three different priors.

guarantees that the estimated parameter will converge to the true parameter and there are no known regret bounds for algorithms using maximum likelihood.

In the GP-MAB problem with an unknown prior considered in **Paper III**, we assume that there is an unknown prior $p^* \in P$ selected from a finite set of priors P. Each prior $p \in P$ has a corresponding prior mean μ_p and kernel function k_p and the mean reward function $f(x)$ is assumed to be sampled from $\mathcal{GP}(\mu_{p^*}, k_{p^*})$. We consider both the setting where the true prior p^* is arbitrarily selected from P and the setting where there exists a *hyperprior* distribution from which p^* is sampled. Since the prior determines many properties of the function f, an efficient agent must distinguish which prior the means are sampled from to ensure it explores efficiently. Believing f is sampled from the wrong prior can lead to poor predictions, as visualized in Fig. 2.7.

Chapter 3

Summary of Included Papers

In this chapter, we summarize the contributions of the papers appended to the thesis.

3.1 Paper I - Bayesian Analysis of Combinatorial Gaussian Process Bandits

In **Paper I**, we theoretically analyze three algorithms for the combinatorial volatile Gaussian process multi-armed bandit problem (CV-GP-MAB), formulate the online energy-efficient navigation problem as a combinatorial and contextual bandit problem, and experimentally evaluate the algorithms using synthetic data on the road networks of Luxembourg and Monaco, see Fig. 3.1.

The three algorithms analyzed are: GP-UCB, GP-TS and GP-BayesUCB (GP-BUCB). GP-UCB (Srinivas et al., 2012) and GP-TS (Russo & Roy, 2014) are natural extensions of the UCB and Thompson sampling algorithms to GP bandits. BayesUCB (Kaufmann et al., 2012) is a variant of UCB that selects the arm with the largest upper quantile and GP-BayesUCB is its GP extension (Nuara et al., 2018). The algorithms are extended to the combinatorial and volatile setting by providing an upper confidence bound (or posterior sample) to each available base arm. The super arm is selected by an oracle algorithm that outputs the optimal solution to the combinatorial problem given the upper confidence bounds (or posterior samples) as input.

The analysis for GP-UCB and GP-TS extends previous Bayesian regret bounds to the *infinite, volatile, and combinatorial* setting. The previous work of Nika et al. (2025) provided a *frequentist* analysis of GP-UCB for the CV-GP-MAB problem. However, previous analysis had not provided Bayesian regret bounds for GP-bandits with a set of arms that is both infinite and volatile. Additionally, we establish the first regret bound for GP-BayesUCB.

GP-BayesUCB can be seen as a variant of GP-UCB where the confidence parameter is defined in terms of the inverse error function $\text{erf}^{-1}(u)$. To bound the regret of GP-BayesUCB, we use a lower bound of $\text{erf}^{-1}(u)$ (Chang et al., 2011) and find a parametrization of the confidence parameter that can be used

Figure 3.1: Road networks of Luxembourg (left) and Monaco (right) with evaluation routes highlighted. Map data from OpenStreetMap: `www.openstreetmap.org/copyright`.

to tune the amount of exploration performed.

A standard technique for analyzing GP-bandits with infinite arms is to consider a discretization of the arm space that becomes finer over time. To establish the regret bound for infinite and volatile arms in a combinatorial, we propose a finer discretization that allows us to bound the discretization error of the upper confidence bound.

Åkerblom et al. (2023) introduced a combinatorial MAB framework based on Bayesian inference (BI) to learn the energy consumption on each road segment. A drawback of the framework of Åkerblom et al. (2023) is that the energy consumption of each road segment must be learned independent of all other road segments, necessitating additional exploration. **Paper I** extends the framework of Åkerblom et al. (2023) to a contextual GP setting and applies it to the real-world road networks in Fig. 3.1. We compare the GP-based algorithms against the respective edge-wise BI bandit algorithm. The correlations induced by the GP structure enables the agent to learn faster, reducing the exploration needed to find the most energy-efficient route.

3.2 Paper II - Comments on "Surrogate Modeling for Bayesian Optimization Beyond a Single Gaussian Process"

Lu et al. (2023) proposed Ensemble GP-TS (EGP-TS) for Bayesian optimization, an algorithm that adaptively chooses between a discrete set of priors using a bi-level Thompson sampling scheme. Lu et al. (2023) provided Bayesian cumulative regret bounds for the GP-bandit problem with an unknown prior, see Section 2.4, when the agent collects the data either sequentially or in parallel, similar to Kandasamy et al. (2018). The proof technique uses the idea of prior confidence sets introduced by Hong et al. (2022).

In **Paper II**, we show that the proof for the regret bounds of Lu et al. (2023)

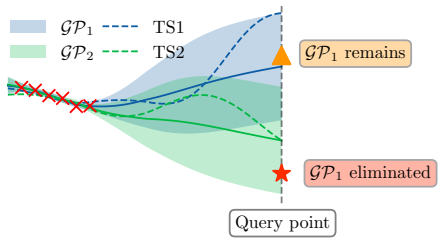


Figure 3.2: Elimination procedure of PE-GP-TS. The solid lines correspond to posterior means and the shaded regions are confidence intervals. The figure has been adapted from Ziomek et al. (2025). The dashed lines are samples from the posteriors.

do not hold due to four issues. First, it is stated that the maximum of $f(x)$ is bounded by a constant which cannot be true since $f(x)$ has infinite support. Second, it is claimed that the expectation of the maximum $f(x^*)$ is equal to the posterior mean $\mu_t(x^*)$, which does not hold since x^* is defined as the maximizer of $f(x)$ leading $f(x^*)$ to have a skew-Gaussian distribution (Arellano-Valle & Azzalini, 2022). Third, a result that bounds an expected value is used to bound a random variable absolutely. Fourth, the excess reward is bounded without accounting for the stopping time that changes its distribution.

3.3 Paper III - Adaptive Prior Selection in Gaussian Process Bandits with Thompson Sampling

In **Paper III**, we consider the GP-MAB problem with an unknown prior. Existing algorithms typically select a prior based on a specific rule and then maximize an acquisition function conditioned on the selected prior. The algorithms of Wang & de Freitas (2014); Berkenkamp et al. (2019) select the lengthscale according to a predetermined schedule whereas the Lengthscale Balancing GP-UCB (LB-GP-UCB) algorithm (Ziomek et al., 2024) selects the lengthscale according to a regret balancing scheme. The Prior Elimination GP-UCB (PE-GP-UCB) algorithm of Ziomek et al. (2025) jointly selects the combination of prior and arm that have the most optimistic upper confidence bound. Both LB-GP-UCB and PE-GP-UCB include an elimination criteria to remove priors whose predictions are inaccurate. As discussed in the previous section, Lu et al. (2023) proposed EGP-TS, an algorithm that samples the prior from the current hyperposterior and then generates a posterior sample based on the selected prior. However, as established in **Paper II**, the regret bounds for EGP-TS do not hold.

In **Paper III**, we investigate the use of Thompson sampling for solving GP-bandit problems with unknown priors and study two algorithms. The first algorithm we propose, Prior Elimination GP-TS (PE-GP-TS), is an extension of PE-GP-UCB that replaces the UCB acquisition function with Thompson sampling. In each iteration, PE-GP-TS jointly selects the arm and prior that

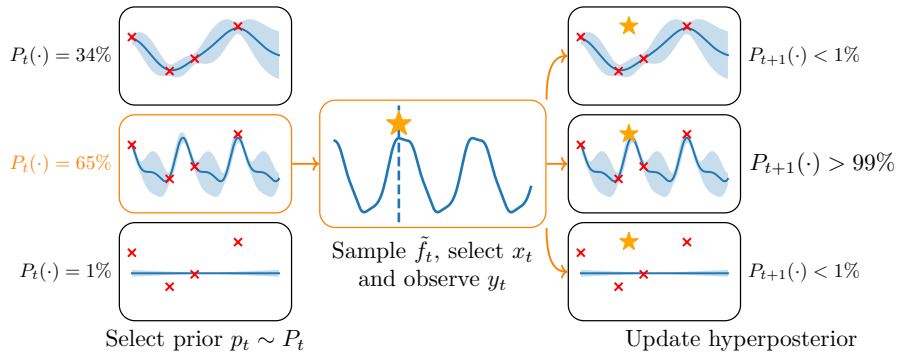


Figure 3.3: Overview of HP-GP-TS. The orange star represents the new observation y_t

maximizes the posterior samples drawn and eliminates priors whose predictive errors are too large, the elimination procedure is visualized in Fig. 3.2. We analyze the regret of PE-GP-TS and obtain a regret bound that matches that of PE-GP-TS plus a term (that is not shown to be sublinear) that depends on the uncertainty of the optimal arm under the correct prior.

The second algorithm we study is EGP-TS, which we refer to as HyperPrior GP-TS (HP-GP-TS), an overview of the algorithm is shown in Fig. 3.3. We prove a sublinear regret bound of HP-GP-TS that depends on the hyperprior weighted maximum information gain, instead of the worst-case maximum information gain. The proof uses the technique of prior confidence sets introduced by Hong et al. (2022). Hong et al. (2022) established regret bounds for MixTS, a general version of HP-GP-TS, for the standard and linear MAB. However, in **Paper III**, we identify technical issues with their proof in the linear setting that prevent directly extending their techniques to the GP setting. Our proof for HP-GP-TS thereby addresses the issues with the proof of Lu et al. (2023) and improves upon the regret bound with respect to the maximum information gain.

Finally, we experimentally evaluate the Thompson-sampling based algorithms on synthetic and real-world data. We observe that the regret of HP-GP-TS is close to the regret of an oracle GP-TS that knows the true prior, thus HP-GP-TS pays a small price for learning the prior. The optimistic bias of PE-GP-TS is less pronounced than that of PE-GP-TS, leading to less exploration and lower regret. We also observe that the regret of HP-GP-TS does not increase with the number of priors in our two scaling experiments whereas the regret of the prior-elimination methods increases in the subspace experiment.

Chapter 4

Discussion and Future Work

In this thesis, we have theoretically analyzed and evaluated GP bandit algorithms in two settings with practical relevance. The combinatorial and contextual GP bandit problem studied in **Paper I** was applied to online energy-efficient navigation of electric vehicles where the contextual information reduced exploration and energy consumption. A motivation for studying the GP bandit problem with an unknown prior in **Papers II** and **III** was the prevalence of maximum likelihood estimation of the kernel parameters in practice. Among the proposed algorithms in the literature, hyperposterior sampling is the most similar to maximum likelihood estimation and, by addressing the issues with previous analyses, our result is a step towards bridging the gap between theory and practice in Bayesian optimization.

Paper I provided a Bayesian framework for the combinatorial and volatile GP bandit problem, and presented regret bounds for three GP-bandit algorithms. The online energy-efficient navigation experiments showed that the GP-structure facilitated faster learning. To incorporate both network structure and road features, the kernel used in the experiments was carefully handcrafted but using the prior selection procedures from **Papers II** and **III** could allow for efficient kernel selection with an online and data-driven approach.

Our formulation of the CV-GP-MAB problem was motivated by the online energy-efficient navigation problem. In particular, the online energy-efficient navigation problem has the additive reward structure of the CV-GP-MAB problem. More broadly, many combinatorial problems do not have an additive reward structure. Nika et al. (2025) considers a Lipschitz continuous reward structure in a frequentist setting.

The navigation problem in **Paper I** is formulated from the perspective of a single vehicle exploring the road network. However, many routing problems involve a fleet of vehicles that can share data among themselves. Given that the energy-consumption can vary greatly depending on the characteristics of the vehicle and driver, all shared data may not be informative. A practically relevant future extension would be to first extend the work to homogeneous fleets and then to heterogeneous fleets where in-fleet correlations must be learned online.

Papers II and **III** consider two Thompson sampling-based algorithms for GP bandits where the prior is unknown but known to be in a finite set of options. Among the previous work, only Lu et al. (2023) had analyzed algorithms using posterior sampling as the acquisition function. In **Paper III**, we can compare the contribution of the acquisition function against the prior selection rule. We observed that the hyperposterior sampling of HP-GP-TS performed much closer to the equivalent oracle than the prior-elimination scheme of PE-GP-TS and more consistent performance than maximum likelihood estimation.

Hong et al. (2022); Lu et al. (2023) analyzed the equivalent HP-GP-TS for linear and GP-bandits respectively. In **Papers II** and **III**, we identify issues with the two proofs that we also address for the GP-setting. Our analysis yields a tighter result with respect to the maximum information gain compared to Lu et al. (2023) and results for other algorithms.

In practical settings, the set of priors is often continuous, i.e. representing the kernel lengthscale parameter. Unlike PE-GP-UCB and -TS, HP-GP-TS is practically extendable to a continuous setting using Markov chain Monte Carlo (MCMC) methods, as has been done for other algorithms. However, the question of extending the theoretical analysis is trickier. The analysis relies on scoring the selected prior based on how well it predicted the observed value and implicitly eliminating priors whose predictions do not fit the observations well. If the set of priors to choose from is infinite or uncountable, then no prior is likely to be selected more than once. Thus, extending the existing analysis to continuous priors would seem to require scoring all priors instead of only the selected priors.

Another common practice for handling uncertainty in the prior is to integrate out the uncertainty with MCMC. This is often more costly than maximum likelihood estimation but can yield higher quality predictions. Repeated hyperposterior sampling provides an approximation to integrating out the uncertainty and analyzing the regret of algorithms that integrate out the uncertainty in the prior is an interesting future direction of research.

References

Åkerblom, N. *Combinatorial Semi-Bandit Methods for Navigation of Electric Vehicles*. PhD thesis, Chalmers University of Technology, 2024. URL https://research.chalmers.se/en/publication/539812.

Åkerblom, N., Chen, Y., and Haghir Chehreghani, M. Online Learning of Energy Consumption for Navigation of Electric Vehicles. *Artificial Intelligence*, 317: 103879, April 2023.

Alanazi, F. Electric Vehicles: Benefits, Challenges, and Potential Solutions for Widespread Adaptation. *Applied Sciences*, 13(10):6016, January 2023. ISSN 2076-3417. doi: 10.3390/app13106016. URL https://www.mdpi.com/2076-3417/13/10/6016.

Arellano-Valle, R. B. and Azzalini, A. Some Properties of the Unified Skew-normal Distribution. *Statistical Papers*, 63(2):461–487, April 2022. ISSN 1613-9798. doi: 10.1007/s00362-021-01235-2. URL https://doi.org/10.1007/s00362-021-01235-2.

Auer, P., Cesa-Bianchi, N., and Fischer, P. Finite-time Analysis of the Multiarmed Bandit Problem. *Machine Learning*, 47(2):235–256, May 2002. ISSN 1573-0565. doi: 10.1023/A:1013689704352. URL https://doi.org/10.1023/A:1013689704352.

Berkenkamp, F., Schoellig, A. P., and Krause, A. No-Regret Bayesian Optimization with Unknown Hyperparameters. *Journal of Machine Learning Research*, 20(50):1–24, 2019. ISSN 1533-7928.

Chang, S.-H., Cosman, P. C., and Milstein, L. B. Chernoff-Type Bounds for the Gaussian Error Function. *IEEE Transactions on Communications*, 59 (11):2939–2944, 2011.

Dasgupta, A., Maniyar, M. P., Srivastava, A., Kumar, S., Mahale, A., Hegde, A., Suggala, A., Shanmugam, K., Tambe, M., and Taneja, A. Learning to Call: A Field Trial of a Collaborative Bandit Algorithm for Optimizing Call Timing in Mobile Maternal Health. In *Proceedings of the 10th Machine Learning for Healthcare Conference*. PMLR, October 2025. URL https://proceedings.mlr.press/v298/dasgupta25a.html.

Feijer, D., Abdollahpouri, H., Gupta, S., Clare, A., Wen, Y., Wasson, T., Dimakopoulou, M., Nazari, Z., Kretschman, K., and Lalmas, M. Calibrated Recommendations with Contextual Bandits, September 2025. URL http://arxiv.org/abs/2509.05460. arXiv:2509.05460 [cs].

Gardner, J. R., Pleiss, G., Bindel, D., Weinberger, K. Q., and Wilson, A. G. GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. In *Advances in Neural Information Processing Systems*, 2018.

Garnett, R. *Bayesian Optimization*. Cambridge University Press, 2023. ISBN 978-1-108-42578-0.

Gonzalvez, J., Lezmi, E., Roncalli, T., and Xu, J. Financial Applications of Gaussian Processes and Bayesian Optimization, 2019. URL https://arxiv.org/abs/1903.04841.

Hensman, J., Fusi, N., and Lawrence, N. D. Gaussian Processes for Big Data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, pp. 282–290, Arlington, Virginia, USA, 2013. AUAI Press.

Hong, J., Kveton, B., Zaheer, M., Ghavamzadeh, M., and Boutilier, C. Thompson Sampling with a Mixture Prior. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pp. 7565–7586. PMLR, May 2022. URL https://proceedings.mlr.press/v151/hong22b.html.

Kandasamy, K., Krishnamurthy, A., Schneider, J., and Poczos, B. Parallelised Bayesian Optimisation via Thompson Sampling. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pp. 133–142. PMLR, March 2018.

Kaufmann, E., Cappe, O., and Garivier, A. On Bayesian Upper Confidence Bounds for Bandit Problems. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, pp. 592–600. PMLR, March 2012.

Kushner, H. J. A Versatile Stochastic Model of a Function of Unknown and Time Varying Form. *Journal of Mathematical Analysis and Applications*, 5 (1):150–167, August 1962. ISSN 0022-247X. doi: 10.1016/0022-247X(62)90011-2. URL https://www.sciencedirect.com/science/article/pii/0022247X62900112.

Kushner, H. 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, March 1964. doi: 10.1115/1.3653121.

Lu, Q., Polyzos, K. D., Li, B., and Giannakis, G. B. Surrogate Modeling for Bayesian Optimization Beyond a Single Gaussian Process. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11283–11296, September 2023. ISSN 1939-3539. doi: 10.1109/TPAMI.2023.3264741. URL https://ieeexplore.ieee.org/abstract/document/10093035.

Mackay, D. J. C. Introduction to Gaussian Processes. In *NATO ASI Series. Series F : Computer and System Sciences*, pp. 133–165, 1998. ISBN 978-3-540-64928-1.

Matérn, B. Spatial Variation. In Brillinger, D., Fienberg, S., Gani, J., Hartigan, J., and Krickeberg, K. (eds.), *Spatial Variation*, volume 36 of *Lecture Notes in Statistics*. Springer, New York, 1986. ISBN 978-0-387-96365-5.

Mockus, J. On Bayesian Methods for Seeking the Extremum. In Marchuk, G. I. (ed.), *Optimization Techniques IFIP Technical Conference Novosibirsk, July 1–7, 1974*, pp. 400–404, Berlin, Heidelberg, 1975. Springer. ISBN 978-3-540-37497-8.

Nika, A., Elahi, S., and Tekin, C. Contextual Combinatorial Bandits With Changing Action Sets Via Gaussian Processes. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=2RgfAY3jnI.

Nuara, A., Trovò, F., Gatti, N., and Restelli, M. A Combinatorial-Bandit Algorithm for the Online Joint Bid/Budget Optimization of Pay-per-Click Advertising Campaigns. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018.

Priem, R., Gagnon, H., Chittick, I., Dufresne, S., Diouane, Y., and Bartoli, N. An Efficient Application of Bayesian Optimization to an Industrial MDO Framework for Aircraft Design. In *AIAA AVIATION 2020 FORUM*. American Institute of Aeronautics and Astronautics, 2020. doi: 10.2514/6.2020-3152. URL https://arc.aiaa.org/doi/abs/10.2514/6.2020-3152.

Robbins, H. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 58(5):527–535, 1952. ISSN 0002-9904, 1936-881X. doi: 10.1090/S0002-9904-1952-09620-8. URL https://www.ams.org/bull/1952-58-05/S0002-9904-1952-09620-8/.

Russo, D. and Roy, B. V. Learning to Optimize via Posterior Sampling. *Mathematics of Operations Research*, 39(4):1221–1243, 2014.

Slivkins, A. Introduction to Multi-Armed Bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, November 2019. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000068. URL https://www.nowpublishers.com/article/Details/MAL-068.

Snoek, J., Larochelle, H., and Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. W. Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, May 2012.

Svensson, H. G., Jannik Bjerrum, E., Tyrchan, C., Engkvist, O., and Chehreghani, M. H. Autonomous Drug Design with Multi-Armed Bandits. In *2022 IEEE International Conference on Big Data (Big Data)*, pp. 5584–5592, December 2022. doi: 10.1109/BigData55660.2022.10020357. URL https://ieeexplore.ieee.org/document/10020357.

Thompson, W. R. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, December 1933.

Trella, A. L., Zhang, K. W., Jajal, H., Nahum-Shani, I., Shetty, V., Doshi-Velez, F., and Murphy, S. A. A Deployed Online Reinforcement Learning Algorithm in an Oral Health Clinical Trial. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(28):28792–28800, April 2025. ISSN 2374-3468. doi: 10.1609/aaai.v39i28.35143. URL `https://ojs.aaai.org/index.php/AAAI/article/view/35143`.

Turner, R., Eriksson, D., McCourt, M., Kiili, J., Laaksonen, E., Xu, Z., and Guyon, I. Bayesian Optimization is Superior to Random Search for Machine Learning Hyperparameter Tuning: Analysis of the Black-Box Optimization Challenge 2020. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, pp. 3–26. PMLR, August 2021.

Wang, Z. and de Freitas, N. Theoretical Analysis of Bayesian Optimisation with Unknown Gaussian Process Hyper-Parameters, June 2014. URL `https://arxiv.org/abs/1406.7758`.

Williams, C. K. and Rasmussen, C. E. *Gaussian Processes for Machine Learning*, volume 2. MIT press Cambridge, MA, 2006.

Yi, X., Wang, S.-C., He, R., Chandrasekaran, H., Wu, C., Heldt, L., Hong, L., Chen, M., and Chi, E. H. Online Matching: A Real-time Bandit System for Large-scale Recommendations. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pp. 403–414, Singapore Singapore, September 2023. ACM. ISBN 979-8-4007-0241-9. doi: 10.1145/3604915.3608792. URL `https://dl.acm.org/doi/10.1145/3604915.3608792`.

Zhang, K. W., Baldwin-McDonald, T., Ciosek, K., Maystre, L., and Russo, D. Impatient bandits: Optimizing for the long-term without delay. *To appear in Journal of Machine Learning*, 2025.

Ziomek, J., Adachi, M., and Osborne, M. A. Bayesian Optimisation with Unknown Hyperparameters: Regret Bounds Logarithmically Closer to Optimal. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, November 2024.

Ziomek, J., Adachi, M., and Osborne, M. A. Time-varying Gaussian Process Bandits with Unknown Prior. In *The 28th International Conference on Artificial Intelligence and Statistics*, February 2025.

Part II

Appended Papers

Paper I

Bayesian Analysis of Combinatorial Gaussian Process Bandits

J. Sandberg, N. Åkerblom, M. Haghir Chehreghani

International Conference on Learning Representations (ICLR), 2025

The paper has been reformatted for uniformity.

Abstract

We consider the combinatorial volatile Gaussian process (GP) semi-bandit problem. Each round, an agent is provided a set of available base arms and must select a subset of them to maximize the long-term cumulative reward. We study the Bayesian setting and provide novel Bayesian cumulative regret bounds for three GP-based algorithms: GP-UCB, GP-BayesUCB and GP-TS. Our bounds extend previous results for GP-UCB and GP-TS to the *infinite*, *volatile* and *combinatorial* setting, and to the best of our knowledge, we provide the first regret bound for GP-BayesUCB. Volatile arms encompass other widely considered bandit problems such as contextual bandits. Furthermore, we employ our framework to address the challenging real-world problem of online energy-efficient navigation, where we demonstrate its effectiveness compared to the alternatives.

1 Introduction

The multi-armed bandit (MAB) problem is a classical problem in which an agent repeatedly has to choose between a number of available actions to perform (commonly called *arms*) and receives rewards depending on the selected action. The goal of the agent is to minimize its expected cumulative regret over a certain time horizon, either finite or infinite, where regret is defined as the expected difference in reward between the agent's selected arm and the best arm (Robbins, 1985). The MAB problem has applications in healthcare, advertising, telecommunications and more (Bouneffouf et al., 2020).

The combinatorial MAB (Gai et al., 2012; Cesa-Bianchi & Lugosi, 2012; Chen et al., 2013) considers a problem where the agent must select a subset of the available base arms, a *super arm*, in every round. The large number of super arms necessitates efficient exploration and may require solving difficult optimization problems.

The arms and environments in bandit applications often have some side-information (or context) that is correlated with the reward, e.g., the titles or user specifications in a news recommendation problem. In the contextual MAB (Li et al., 2010; Krause & Ong, 2011; Agarwal et al., 2014; Zhou, 2016), before selecting an arm, the agent is provided a context vector (for the entire environment or each individual arm) that may (randomly) vary over time. By utilizing the information in the context the agent can learn the expected rewards more efficiently.

When the set of arms or contexts is continuous (infinite), it is necessary to impose smoothness assumptions on the expected reward since the agent can only explore a finite number of arms. A common assumption is that the expected reward is a sample from a Gaussian process (GP) over the arm or context set. For a sufficiently smooth GP, this ensures that arms which lie close in the arm space have similar expected rewards. Integrating GPs into bandits can yield higher sample efficiency and improved learning.

In Table 1, we provide an overview and comparison of similar work in GP MABs. The seminal paper of Srinivas et al. (2012) first introduced the GP-UCB algorithm, which combines *upper confidence bounds* (UCB) with GPs for MABs with finite or infinite arm sets. Srinivas et al. provided frequentist regret bounds for GP-UCB on a MAB problem with a compact (infinite) arm space. Later work by Russo & Roy (2014) provided Bayesian regret bounds for GP-UCB and GP-TS, a similar algorithm based on Thompson sampling (Thompson, 1933), in the finite-arm setting with volatile arms. Volatile arms (often called *time-varying* or *sleeping* arms) means that not all arms are available to the agent in every round. This is a general formulation that encompasses other MAB extensions such as the contextual MAB.

For the infinite arm setting, Russo & Roy only hinted that the proof follows by discretization arguments. Using discretization, the recent work by Takeno et al. (2023, 2024) derives Bayesian regret bounds for GP-UCB and GP-TS in the infinite arm setting - but without volatile arms.

The combinatorial *and* contextual MAB with changing arm sets (C3-MAB) incorporates both extensions and has received much interest recently (Qin

Table 1: Comparison of similar work in GP MABs where T is the horizon, K is the maximum super arm size, and γ_T (γ_{TK}) is the maximum information gain from T (TK) base arms. Note that Takeno et al. (2023, 2024) obtain a regret bound of $\mathcal{O}(\sqrt{T\gamma_T})$ for IRGP-UCB, GP-TS and PIMS in the finite setting.

	Ours	(Nika, 2022)	(Takeno, 2023; 2024)	(Kandasamy, 2018)	(Russo, 2014)	(Srinivas, 2012)
Infinite/Finite	Infinite	Infinite	Infinite	Infinite	Finite	Infinite
Volatile/Static	Volatile	Volatile	Static	Static	Volatile	Static
Combinatorial	✓	✓	✗	✗	✗	✗
Bayesian/Frequentist	Bayesian	Frequentist	Bayesian	Bayesian	Bayesian	Frequentist
GP-UCB	✓	✓	✓	✗	✓	✓
GP-TS	✓	✗	✓	✓	✓	✗
GP-BUCB	✓	✗	✗	✗	✗	✗
Regret	$K\sqrt{T\gamma_{TK}\log T}$	$K\sqrt{T\gamma_{TK}\log T}$	$\sqrt{T\gamma_T\log T}$	$\sqrt{T\gamma_T\log T}$	$\sqrt{T\gamma_T\log T}$	$\sqrt{T\gamma_T\log T}$

et al., 2014; Chen et al., 2018; Nika et al., 2020, 2022; Elahi et al., 2023), with applications in online advertisement, epidemic control and base station assignment (Nuara et al., 2018; Lin & Bouneffouf, 2022; Shi et al., 2023). Recent work by Nika et al. (2022) considered the C3-MAB with base arm rewards sampled from a GP. Nika et al. (2022) provided high probability regret bounds for a combinatorial variant of GP-UCB with an approximation oracle.

In this work, we present novel Bayesian regret bounds for both GP-UCB and GP-TS in the combinatorial volatile Gaussian process semi-bandit problem that extend previous regret bounds for GP-UCB and GP-TS to the infinite, volatile and combinatorial setting. As discussed above, our results hold for the contextual setting as it is a special case of the volatile setting. Additionally, we present a Bayesian regret bound for a third bandit algorithm called GP-BayesUCB (GP-BUCB) which is based on the BayesUCB algorithm of Kaufmann et al. (2012). Whilst GP-BUCB was introduced by Nuara et al. (2018) for a combinatorial bandit problem, to the best of our knowledge there are no regret bounds for GP-BUCB - even in the non-combinatorial setting. We demonstrate that the parametrization of GP-BUCB is more flexible than GP-UCB and is less prone to over-exploration whilst retaining theoretical guarantees.

Furthermore, we demonstrate the applicability of combinatorial and contextual GP bandits to large scale problems with experiments on an online energy-efficient navigation problem for electric vehicles on real-world road networks. With the increasing emergence of electric vehicles, addressing this problem is crucial to mitigating the so-called *range anxiety*. Åkerblom et al. (2023) introduced a combinatorial MAB framework using Bayesian inference to learn the energy consumption on each road segment. In this paper, we extend the framework of Åkerblom et al. to a contextual setting and apply it to real-world road networks. The experimental results demonstrate that the contextual GP model achieves lower regret than the Bayesian inference model.

Our contributions can be summarized as follows.

- We extend previous Bayesian regret bounds for GP-UCB and GP-TS to the *infinite*, *volatile* (previous results only held for finite volatile arms) and *combinatorial* setting.

- To the best of our knowledge, we establish the first regret bound for GP-BayesUCB.
- We develop a combinatorial and contextual bandit framework for the important real-world application of online energy-efficient navigation.

2 Setup and Algorithms

In this section, we formulate our bandit problem, introduce Gaussian process bandit algorithms, and define the information gain, a quantity that is essential for GP bandits.

2.1 Problem formulation

We begin by formulating the combinatorial volatile Gaussian process semi-bandit problem. Let $\mathcal{A} \subset \mathbb{R}^d$ denote the set of base arms in a d-dimensional space and $\mathcal{S} = \{\mathbf{a} | \mathbf{a} \subset \mathcal{A}\} \subset 2^{\mathcal{A}}$ denote the set of feasible super arms. Note that $|\mathcal{A}|$ can either be finite or infinite, and $2^{\mathcal{A}}$ denotes the power set of $\mathcal{A}$. The expected reward for the base arms $f(a) \sim \mathcal{GP}(\mu(a), k(a, a'))$ is assumed to be a sample from a Gaussian process with mean function $\mu(a) : \mathcal{A} \to \mathbb{R}$ and covariance function $k(a, a') : \mathcal{A} \times \mathcal{A} \to [-1, 1]$.

At time t, a possibly random and finite[1] subset of base arms $\mathcal{A}_t \subseteq \mathcal{A}$ is available to the agent. In a combinatorial setting, the agent must select a feasible subset of base arms, a *super arm*, $\mathbf{a}_t \in \mathcal{S}_t$ where $\mathcal{S}_t \subset 2^{\mathcal{A}_t}$ is the set of feasible and available super arms. To facilitate a feasible combinatorial problem, the super arms have a maximum size K ($|\mathbf{a}| \leq K \ \forall \mathbf{a} \in \mathcal{S}_t$). The agent observes the rewards of the selected base arms (semi-bandit feedback) $\mathbf{r}_t = \{r_{t,a} | a \in \mathbf{a}_t\}$ where the base arm reward $r_{t,a} = f(a) + \epsilon_{t,a}$ is a sum of the expected reward and i.i.d. Gaussian noise with zero mean and variance ς^2. Motivated by the online energy-efficient navigation problem in Section 4.1, the total reward is assumed to be additive, and the agent also observes this reward at time t: $R_t = \sum_{a \in \mathbf{a}_t} r_{t,a}$. The total number of time steps, the horizon, is denoted by T. Let H_t denote the history $(\mathcal{A}_1, \mathcal{S}_1, \mathbf{a}_1, \mathbf{r}_1, \ldots, \mathcal{A}_{t-1}, \mathcal{S}_{t-1}, \mathbf{a}_{t-1}, \mathbf{r}_{t-1}, \mathcal{A}_t, \mathcal{S}_t)$ of past observations and the currently available arms at time t.

In this work, we are interested in minimizing the Bayesian cumulative regret which, with a horizon of T, is defined as

$$\mathrm{BR}(T) = \mathbb{E}\Big[\sum_{t \in [T]} f(\mathbf{a}_t^*) - f(\mathbf{a}_t) \Big], \tag{1}$$

where $[T] := \{1, ..., T\}$, $\mathbf{a}_t^* = \arg\max_{\mathbf{a} \in \mathcal{S}_t} f(\mathbf{a})$ and $f(\mathbf{a}) = \sum_{a \in \mathbf{a}} f(a)$. As discussed by Russo & Roy (2014), allowing stochastic arm sets permits us to consider broader sets of bandit problems, an example of particular interest to us will be contextual models. Even though $\mathcal{A}_t$ is finite, note that the infinite

[1] The restriction $|\mathcal{A}_t| < \infty$ prevents issues with limit points since the agent can only select the same base arm once. This limitation is not necessary in a non-combinatorial setting.

Algorithm 1 Framework for combinatorial volatile semi-bandit problem

Require: Prior agent parameters $\boldsymbol{\theta}_0$, base arm set $\mathcal{A}$, super arm set $\mathcal{S}$, horizon T.

1: **for** $t \leftarrow 1, \ldots, T$ **do**
2: $\quad \mathcal{A}_t, \mathcal{S}_t \leftarrow \text{ObserveAvailableArms}(\mathcal{A}, \mathcal{S})$
3: $\quad \mathbf{U}_t \leftarrow \text{GetBaseArmIndices}(t, \boldsymbol{\theta}_{t-1})$
4: $\quad \mathbf{a}_t \leftarrow \text{SelectOptimalSuperArm}(\mathcal{S}_t, \mathbf{U}_t)$
5: $\quad \mathbf{r}_t \leftarrow \text{ObserveRewards}(\boldsymbol{a}_t)$
6: $\quad \boldsymbol{\theta}_t \leftarrow \text{UpdateParameters}(\boldsymbol{a}_t, \boldsymbol{r}_t, \boldsymbol{\theta}_{t-1})$

case $|\mathcal{A}| = \infty$ is of great importance since it it is necessary for the context to be a continuous random variable.

Algorithm 1 provides a framework for the introduced bandit problem. In the framework, the agent parameters $\boldsymbol{\theta}_t$ are defined for a general agent and are not specified here. Similarly, $\mathbf{U}_t$ denotes the set of base arm indices which could be upper confidence bounds or a posterior sample, depending on the algorithm used.

2.2 Bayesian framework for combinatorial Gaussian process bandits

A Gaussian process $f(a) \sim \mathcal{GP}(\mu(a), k(a, a'))$ is a collection of random variables such that for any subset $\{a_1, \ldots, a_N\} \subset \mathcal{A}$ the vector $\mathbf{f} = [f(a_1), \ldots, f(a_N)]$ has a multivariate Gaussian distribution. We take a Bayesian view of the combinatorial problem and consider $\mathcal{GP}(\mu, k)$ as a prior over the base arm rewards. GPs are very useful for defining and solving bandit problems, due to their probabilistic nature and the flexibility they provide through the design of suitable kernels.

Let $N_{t-1} = \sum_{\tau=[t-1]} |\mathbf{a}_\tau|$ denote the total number of base arms selected up to time $t-1$ and let $a_1, \ldots, a_{N_{t-1}}$ denote the arms selected before time t. Additionally, let $\mathbf{y} \in \mathbb{R}^{N_{t-1}}$ denote the corresponding observed base arm rewards and $\boldsymbol{\mu} = [\mu(a_1), \ldots, \mu(a_{N_{t-1}})]^\top$ denote the corresponding prior expected base arm mean rewards. Then, for any $a \in \mathcal{A}$, the posterior GP distribution is given by:

$$\mu_{t-1}(a) = \mu(a) + \mathbf{k}(a)^\top \left(\mathbf{K} + \varsigma^2 I\right)^{-1} (\mathbf{y} - \boldsymbol{\mu})^\top \tag{2}$$

$$k_{t-1}(a, a') = k(a, a') - \mathbf{k}(a)^\top \left(\mathbf{K} + \varsigma^2 I\right)^{-1} \mathbf{k}(a'), \tag{3}$$

where $\mathbf{K} = \big(k(a_i, a_j)\big)_{i,j=1}^{N_{t-1}}$ is the covariance matrix of the previously selected arms and $\mathbf{k}(a) = \big[k(a, a_1), \ldots, k(a, a_{N_{t-1}})\big]^\top$ is the covariance between a and the previously selected arms. Let $\sigma_{t-1}(a)$ and $\sigma^2_{t-1}(a)$ denote the posterior standard deviation and variance respectively.

In 2012, Srinivas et al. introduced the GP-UCB algorithm, which selects the next arm based on an upper confidence bound. In our combinatorial setting, the GP-UCB algorithm selects the super arm $\mathbf{a}_t = \arg\max_{\mathbf{a} \in \mathcal{S}_t} U_t(\mathbf{a})$ where $U_t(\mathbf{a}) = \mu_{t-1}(\mathbf{a}) + \sqrt{\beta_t}\sigma_{t-1}(\mathbf{a})$, $\mu_{t-1}(\mathbf{a}) = \sum_{a \in \mathbf{a}} \mu_{t-1}(a)$, $\sigma_{t-1}(\mathbf{a}) = \sum_{a \in \mathbf{a}} \sigma_{t-1}(a)$ and

β_t is a confidence parameter, typically of order $\mathcal{O}(\log t)$. Kaufmann et al. (2012) introduced Bayes-UCB, which selects the arm with the largest $(1-\eta_t)$-quantile, where the quantile parameter η_t was of order $\mathcal{O}(1/t)$. Adapted to the combinatorial Gaussian process setting, we suggest the following selection rule for Bayes-UCB: $\mathbf{a}_t = \arg\max_{\mathbf{a}\in\mathcal{S}_t} \sum_{a\in\mathbf{a}} Q\left(1-\eta_t, \mathcal{N}\left(\mu_{t-1}(a), \sigma^2_{t-1}(a)\right)\right)$, where $Q(p,\lambda)$ denotes the p-quantile of the distribution λ. We refer to this adapted version as GP-BUCB. Note that for $\lambda = \mathcal{N}(\mu,\sigma^2)$, the p-quantile is given by $Q(p,\mathcal{N}(\mu,\sigma^2)) = \mu + \sigma\sqrt{2}\,\mathrm{erf}^{-1}(2p-1)$ where $\mathrm{erf}^{-1}(\cdot)$ is the inverse of the error function. Thus, GP-BUCB can be seen as a variant of GP-UCB where $\beta_t = 2\left(\mathrm{erf}^{-1}(1-2\eta_t)\right)^2$. GP-TS (Russo & Roy, 2014; Chowdhury & Gopalan, 2017) selects the next arm randomly by using posterior sampling. If $\hat{f}_t(a) \sim \mathcal{GP}(\mu_{t-1}, k_{t-1})$ is a sample from the posterior distribution, then, in the combinatorial setting, GP-TS selects the super arm $\mathbf{a}_t = \arg\max_{\mathbf{a}\in\mathcal{S}_t} \hat{f}_t(\mathbf{a})$, where $\hat{f}_t(\mathbf{a}) = \sum_{a\in\mathbf{a}} \hat{f}_t(a)$.

2.3 Information gain

The regret bounds of most GP bandit algorithms depend on a parameter called the maximal information gain γ_T (Srinivas et al., 2012; Vakili et al., 2021). The maximal information gain (MIG) describes how the uncertainty of f diminishes as the best set of sampling points $\mathbf{a} \subset \mathcal{A}, |\mathbf{a}| \leq T$ grows in size T. The MIG is defined using the mutual information between the observations $\mathbf{y_a}$ at locations $\mathbf{a}$ and expected reward function f:

$$\gamma_T := \sup_{\mathbf{a}\subset\mathcal{A}, |\mathbf{a}|\leq T} I(\mathbf{y_a}; f), \tag{4}$$

where $I(\mathbf{y_a}; f) = H(\mathbf{y_a}) - H(\mathbf{y_a}|f)$ and $H(\cdot)$ denotes the entropy. Both the true value of γ_T and most upper bounds depend on the kernel function k defining the GP from which f is sampled from. Srinivas et al. (2012) initially introduced bounds on γ_T for common kernels, such as the Matérn and RBF kernels. For the RBF kernel, Srinivas et al. showed that $\gamma_T = \mathcal{O}\left(\log^{d+1}(T)\right)$. Later, Vakili et al. (2021) presented a general method of bounding γ_T that utilizes the eigendecay of the kernel k. Using this method, Vakili et al. obtained improved bounds on the Matérn kernel with smoothness parameter ν: $\gamma_T = \mathcal{O}\left(T^{\frac{d}{2\nu+d}} \log^{\frac{2\nu}{2\nu+d}}(T)\right)$. To apply these bounds, we require that $\mathcal{A}$ is compact.

3 Regret Analysis

Whilst the work of Chen et al. (2013) can be seen as a standard combinatorial framework, we adopt the framework of (Russo & Roy, 2014) since it is better suited for Bayesian bandits with volatile and infinite arms. Russo & Roy (2014) first provided a Bayesian regret bound for GP-UCB in a volatile (but non-combinatorial) setting with a finite arm set. Recently, Takeno et al. (2023) presented explicit proof for the Bayesian regret of GP-UCB with a compact and static arm set. In this section, we present novel Bayesian regret bounds for

both GP-UCB and GP-TS in a combinatorial and volatile setting (including the contextual setting). Additionally, to the best of our knowledge, we present the first Bayesian regret bound for GP-BayesUCB. Similar to previous work, we first consider the finite arm case, $|\mathcal{A}| < \infty$, and then consider the infinite case, $|\mathcal{A}| = \infty$, by extending the finite arm results via a discretization.

3.1 Finite case

We start by highlighting our technical contributions for GP-BUCB. Following the proof framework of Russo & Roy (2014), we seek to bound two terms: $\mathbb{E}[f(\mathbf{a}_t^*) - U_t(\mathbf{a}_t^*)]$ and $\mathbb{E}[U_t(\mathbf{a}_t) - f(\mathbf{a}_t)]$. For GP-BUCB, establishing an upper bound for the second term requires us to work around the non-elementary function $\text{erf}^{-1}(u)$. Using Thm. 2 of Chang et al. (2011), we find that $\text{erf}^{-1}(u) \geq \sqrt{-\omega^{-1}\log((1-u)/\vartheta)}$ for $\omega > 1$ and $0 < \vartheta \leq \sqrt{2e/\pi}\sqrt{\omega-1}/\omega$, see Lemma A.13. The bound is tighter for larger values of the parameter ϑ (Chang et al., 2011), thus we set ϑ to its maximum value whilst ω is kept as a tunable parameter. Recall that the quantile parameter η_t determines how quickly the confidence bound grows and the order $\xi > 0$ (s.t. $\eta_t = \mathcal{O}(t^{-\xi})$) is another tunable parameter. As shown in the lemma below, these parameters influence the bound we get.

Lemma 3.1. *Let* $C_\omega = \left(\sqrt{\pi}\omega/\sqrt{2e(\omega-1)}\right)^{1/\omega}$, *then for GP-BUCB with confidence parameter* $\beta_t = 2\left(\text{erf}^{-1}(1-2\eta_t)\right)^2$ *and* $\eta_t = \frac{\sqrt{2\pi}^\omega}{2|\mathcal{A}|^\omega t^\xi}$, $\xi > 0$, $\omega > 1$,

$$\sum_{t\in[T]} \mathbb{E}[f(\mathbf{a}_t^*) - U_t(\mathbf{a}_t^*)] \leq C_\omega \cdot \begin{cases} \frac{\omega}{\omega-\xi} T^{1-\frac{\xi}{\omega}} & \text{if } \xi/\omega < 1, \\ 1 + \log T & \text{if } \xi/\omega = 1, \\ \frac{\xi}{\xi-\omega} & \text{if } \xi/\omega > 1. \end{cases}$$

Kaufmann et al. (2012) studied (non-GP and non-combinatorial) Bayes-UCB for a Bernoulli bandit with $\xi = 1$ whilst our analysis permit any $\xi > 0$. Lemma 3.1 shows that the ratio ξ/ω determines if the bound for the right term is sublinear, logarithmic or constant w.r.t T for GP-BUCB. The equivalent bounds for GP-UCB and GP-TS are both constant if $\beta_t = 2\log\left(|\mathcal{A}|t^2/\sqrt{2\pi}\right)$, see Lemma A.2, thus we assume $\xi/\omega > 1$ to simplify the regret bounds.

Srinivas et al. (2012) showed, in a non-combinatorial setting, that the sum of posterior variances can be bounded by the information gain between the sampled points and the expected reward function f. Lemma 3 in Nika et al. (2022) (adopted to our setting in Lemma A.12) generalizes this result to a combinatorial setting. The result depends on the maximum eigenvalue of all possible posterior covariance matrices of size at most K, which we denote as λ_K^*.

Then, we present the main theorems for GP-UCB, GP-BUCB and GP-TS in the finite case, see Section A.1 for the proofs.

Theorem 3.2 (Finite regret bounds)**.** *Let* $C_K := 2(\lambda_K^* + \varsigma^2)$. *When* $\mathcal{A}$ *is finite, the Bayesian regret of*

(i) GP-UCB with $\beta_t = 2\log(|\mathcal{A}|t^2/\sqrt{2\pi})$ *is bounded as* $BR(T) \leq \frac{\pi^2}{6} + \sqrt{C_K TK\beta_T\gamma_{TK}}$.

(ii) GP-BUCB with $\beta_t = 2\left(\text{erf}^{-1}(1-2\eta_t)\right)^2$ *for* $\eta_t = \frac{\sqrt{2\pi}^\omega}{2|\mathcal{A}|^\omega t^\xi}$, $\xi > \omega > 1$ *is bounded as* $BR(T) \leq \sqrt{C_K TK\beta_T\gamma_{TK}} + C_\omega \cdot \frac{\xi}{\xi-\omega}$ *where* $C_\omega = \left(\sqrt{\pi}\omega/\sqrt{2e(\omega-1)}\right)^{1/\omega}$.

(iii) GP-TS is bounded as $BR(T) \leq \frac{\pi^2}{3} + 2\sqrt{C_K TK\beta_T\gamma_{TK}}$ *where* $\beta_t = 2\log\left(|\mathcal{A}|t^2/\sqrt{2\pi}\right)$.

For all three algorithms (if $\xi/\omega > 1$ for GP-BUCB), we find that $\text{BR}(T) = \mathcal{O}(\sqrt{\lambda_K^* TK\beta_T\gamma_{TK}})$ where γ_{TK} is the MIG from TK base arms. Using the bounds of γ_T from Section 2.3, we get that the regret is sublinear in T for both the RBF and Matérn kernels. The closest work, by Nika et al. (2022), obtains a frequentist regret bound of the same order. Nika et al. (2022) noted that $\lambda_K^* = \mathcal{O}(K)$ which gives a linear dependence on K in the worst case (Zhan, 2005). For a linear kernel, the setting of Wen et al. (2015) is similar to our setting and they obtain $\mathcal{O}(K\sqrt{\log K})$ and $\mathcal{O}(K)$ dependencies on K whereas our dependency is $\mathcal{O}(K\sqrt{\log K})$. For combinatorial semi-bandits with linear reward functions (but independent arms), Merlis & Mannor (2020) obtain a $\Omega(\sqrt{K/\log K})$ lower bound which would suggest a gap of $\sqrt{K}\log K$ for the linear kernel. When $K = 1$, our results match the non-combinatorial results for GP-UCB. However, the improved random GP-UCB (IRGP-UCB) and GP-TS of Takeno et al. (2023, 2024) has a Bayesian regret of $\mathcal{O}(\sqrt{T\gamma_T})$ in the finite case, suggesting that a $\sqrt{\beta_T} = \mathcal{O}(\sqrt{\log T})$ improvement is possible. To our knowledge, there are no known lower bounds for the Bayesian regret of GP-bandit algorithms in general. However, for the SE-kernel the non-Bayesian regret satisfies $\Omega(\sqrt{T(\log T)^{d/2}})$ (Scarlett (2018); Cai & Scarlett (2021)). Taken at face value, this would imply that our bounds are tight up to logarithmic factors of T.

3.2 Infinite case

The infinite case, $|\mathcal{A}| = \infty$, is an important generalization since many decision problems have a continuum of actions to select from. Based on our framing of contextual MABs as a subset of volatile MAB, an infinite arm set permits contexts with support on infinite domains such as continuous time. This setting is often analytically more difficult and requires the following additional assumptions:

Assumption 3.3 (Regularity assumptions)**.** *Assume* $\mathcal{A} \subset [0, C_1]^d$ *is a compact and convex set for some* $C_1 > 0$*. Furthermore, assume that* μ *and* k *are both* L*-Lipschitz on* $\mathcal{A}$ *and* $\mathcal{A} \times \mathcal{A}$*, respectively, for some* $L > 0$*. In addition, for* $f \sim \mathcal{GP}(\mu, k)$ *assume that there exists constants* $C_2, C_3 > 0$ *such that:*

$$\mathbb{P}\left(\sup_{a\in\mathcal{A}}\left|\frac{\partial f}{\partial a^{(j)}}\right| > l\right) \leq C_2 \exp\left(-\frac{l^2}{C_3^2}\right), \tag{5}$$

for $j \in \{1, \ldots, d\}$ *and* $l > 0$ *where* $a^{(j)}$ *denotes the* j*-th element of* a*.*

Whilst the high probability bound on the derivatives of the sample paths is a common assumption in the literature (Srinivas et al., 2012; Kandasamy et al., 2018; Takeno et al., 2023), we additionally require that both μ and k are Lipschitz but this is not particularly restrictive, see Remark A.6.

Following Srinivas et al. (2012), proofs for the compact case tend to use a discretization $\mathcal{D}_t \subset \mathcal{A}$ where each dimension is divided into τ_t points such that $|\mathcal{D}_t| = \tau_t^d$. Let $[a]_{\mathcal{D}_t}$ denote the nearest point in $\mathcal{D}_t$ for $a \in \mathcal{A}$ and similarly let $[\mathbf{a}]_{\mathcal{D}_t} = \{[a]_{\mathcal{D}_t} | a \in \mathbf{a}\}$ for $\mathbf{a} \subset \mathcal{A}$. Due to the assumption of volatile arms, we require the following finer discretization (as compared to Takeno et al., 2023):

Assumption 3.4 (Discretization size)**.** *Let τ_t denote the number of discretization points per dimension and assume that*

$$\begin{cases} \tau_t \geq 2t^2 KLdC_1(1 + tK\varsigma^{-1}), & \text{(6a)} \\ \tau_t/\beta_t \geq 8t^4 K^2 LdC_1, & \text{(6b)} \\ \tau_t^2/\beta_t \geq 8t^5 K^3 L^2 d^2 C_1^2 \varsigma^{-2}, & \text{(6c)} \\ \tau_t \geq t^2 KdC_1 C_3(\sqrt{\log(C_2 d)} + \sqrt{\pi}/2) & \text{(6d)} \end{cases}$$

where the constants C_1, C_2, C_3 and L are given by Assumption 3.3 whilst the constants d, K and ς are defined by the bandit problem (Section 2.1).

We note that Eq. (6d) is equivalent to the discretization size used by Takeno et al. (2023) with an extra factor of K to account for the combinatorial setting whilst we introduce Eqs. (6a) to (6c) to bound $U_t([\mathbf{a}]_{\mathcal{D}_t}) - U_t(\mathbf{a})$. A key step to establish the regret bound of GP-UCB by Takeno et al. (2023) is to use the fact (for that setting) that $\mathbf{a}_t$ maximizes the upper confidence bound $U_t(\mathbf{a})$ and thus $U_t([\mathbf{a}_t^*]_{\mathcal{D}_t}) - U_t(\mathbf{a}_t) \leq 0$. Since we consider a setting with volatile arms, $[\mathbf{a}_t^*]_{\mathcal{D}_t}$ is not necessarily a feasible super arm and our technical contribution in the infinite setting is an analysis of the discretization error of $U_t([\mathbf{a}]_{\mathcal{D}_t}) - U_t(\mathbf{a})$.

Lemma 3.5. *If $U_t(\mathbf{a}) = \mu_{t-1}(\mathbf{a}) + \sqrt{\beta_t}\sigma_{t-1}(\mathbf{a})$, Assumption 3.3 holds and τ_t and β_t satisfy Eqs.* (6a) *to* (6c) *in Assumption 3.4, then for any sequence of super arms $\mathbf{a}_t \in \mathcal{S}_t$ $t \geq 1$:*

$$\sum_{t \in [T]} \mathbb{E}\left[U_t([\mathbf{a}_t]_{\mathcal{D}_t}) - U_t(\mathbf{a}_t)\right] \leq \frac{\pi^2}{6}. \tag{7}$$

To bound the difference in posterior mean, $\mu_{t-1}([\mathbf{a}]_{\mathcal{D}_t}) - \mu_{t-1}(\mathbf{a})$, we Cholesky decompose $\mathbf{K} + \varsigma^2 I = \mathbf{L}\mathbf{L}^\top$ and note that $||\mathbf{L}^{-1}(\mathbf{y} - \boldsymbol{\mu})||_2$ has a chi distribution with at most TK degrees of freedom. The difference in posterior standard deviation, $\sigma_{t-1}([\mathbf{a}]_{\mathcal{D}_t}) - \sigma_{t-1}(\mathbf{a})$, is bounded by using that k is Lipschitz, Assumption 3.4 and other smaller steps.

Next, we present our regret bounds for GP-UCB, GP-BUCB and GP-TS in the infinite setting:

Theorem 3.6 (Infinite regret bounds)**.** *If Assumption 3.3 holds and τ_t satisfies Assumption 3.4, then the Bayesian regret of*

(i) GP-UCB with $\beta_t = 2\log(\tau_t^d t^2/\sqrt{2\pi})$ *is bounded as* $BR(T) \leq \frac{\pi^2}{2} + \sqrt{C_K T K \beta_T \gamma_{TK}}$.

(ii) GP-BUCB with $\beta_t = 2\left(\text{erf}^{-1}(1-2\eta_t)\right)^2$ *for* $\eta_t = (2\pi)^{\omega/2}/\left(2\tau_t^{d\omega} t^{\xi}\right)$, $\xi > \omega > 1$ *is bounded as* $BR(T) \leq \frac{\pi^2}{3} + \sqrt{C_K T K \beta_T \gamma_{TK}} + C_\omega \cdot \frac{\xi}{\xi - \omega}$ *where* $C_\omega = \left(\sqrt{\pi}\omega/\sqrt{2e(\omega-1)}\right)^{1/\omega}$.

(iii) GP-TS is bounded as $BR(T) \leq \frac{2\pi^2}{3} + 2\sqrt{C_K T K \beta_T \gamma_{TK}}$.

The proofs are presented in Section A.2. Similar to Takeno et al. (2023), the regret is decomposed into multiple terms which are either bounded by the finite case or by using results such as Lemma 3.5. Because of the stochastic arm selection, the regret for GP-TS must be decomposed into more terms compared to GP-UCB, which increases the constants in the bound. As in the finite case, we get that $\text{BR}(T) = \mathcal{O}(\sqrt{\lambda_K^* T K \beta_T \gamma_{TK}})$ for all three algorithms which matches the non-combinatorial result of Takeno et al. (2023) for $K = 1$.

4 Experiments

In this section, we consider the important real-world application of online energy-efficient navigation for electric vehicles and formulate it as a combinatorial and contextual bandit problem. Previous work by Åkerblom et al. (2023) introduced a framework based on Bayesian inference to address the online navigation problem when no contextual information is available. Bayesian combinatorial bandits allow us to combine imperfect initial estimates with exploration to find efficient routes. In this work, we extend the framework to incorporate contextual information, enabling us to make use of correlations for even faster learning.

4.1 Bandit formulation of online energy efficient navigation problem

The online energy-efficient navigation problem Consider a directed graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ where the vertices $\mathcal{V}$ denote intersections of road segments and the edges $e = (u_1, u_2) \in \mathcal{E}$ denote the road segment from intersection u_1 to intersection u_2. Additionally, let $\mathcal{L}(\mathcal{G}) = \mathcal{G}(\mathcal{E}, \mathcal{C}_t)$ denote the directed line graph of $\mathcal{G}$ where the set of connections $\mathcal{C}_t \subseteq \{(e_1, e_2) | e_1 = (u, v) \in \mathcal{E}, e_2 = (v, w) \in \mathcal{E}\}$ determine which turns are legal in the road network at time t. Assume that we are given a start vertex $u_1 \in \mathcal{V}$ and a goal vertex $u_n \in \mathcal{V}$. Let $\mathbb{P}_t$ denote the set of simple feasible paths from u_1 to u_n at time t. A path $\mathbf{p} = \langle u_1, u_2, \ldots, u_n \rangle$ is legal if all the connections are legal, and $\mathbf{p}$ is simple if every vertex is visited at most once. At each time step t, we observe the set of available paths $\mathbb{P}_t$ and a context vector $x_{t,e} \in \mathbb{R}^d$ for each edge $e \in \mathcal{E}$. The context $x_{t,e}$ can include static features, such as the length of the road segment, and time-varying features, such as the congestion level. Based on the available connections and the context vector, we select a path $\mathbf{p}_t \in \mathbb{P}_t$ and observe the energy consumption associated

Algorithm 2 Compute Rectified Indices

procedure GETBASEARMINDICES($t, \mathcal{A}_t, \boldsymbol{\theta}_{t-1} = (\boldsymbol{\mu}_{t-1}, \boldsymbol{\sigma}_{t-1}, \boldsymbol{\varsigma}_{t-1})$)

for each edge $e \in \mathcal{A}_t$ **do**

  $\tilde{\mu}_e \leftarrow \mu_{t-1,e} - \sqrt{\beta_t}\sigma_{t-1,e}$ ▷ UCB

  $\tilde{\mu}_e \leftarrow Q(\frac{1}{t}, \mathcal{N}(\mu_{t-1,e}, \sigma^2_{t-1,e}))$ ▷ BUCB

  $\tilde{\mu}_e \leftarrow$ Sample from $\mathcal{N}(\mu_{t-1,e}, \sigma^2_{t-1,e})$ ▷ TS

  $U_{t,e} \leftarrow \mathbb{E}[z_e]$ where $z_e \sim \mathcal{N}^R(\tilde{\mu}_e, \varsigma^2_{t-1,e})$

return $\mathbf{U}_t$

with each edge in the path (negated reward): $R_t = \sum_{e \in \mathbf{p}_t} r_{t,e}$. The goal of online energy-efficient navigation is to minimize the total energy consumed over a horizon T. Note that the base arm set $\mathcal{A}_t$ corresponds to all edge-context tuples $(e, x_{t,e})$ and that the base arm space is defined as $\mathcal{A} = \mathcal{E} \times \mathcal{X}$ where $\mathcal{X} \subseteq [0, C_1]^d$ is a compact and convex set for some $C_1 > 0$. The super arm set $\mathcal{S}_t$ corresponds to sequences of edge-context tuples that form paths in $\mathbb{P}_t$.

Shortest paths with rectified Gaussians Using regenerative braking, the energy consumption of an electric vehicle can be negative along individual road segments which presents challenges when we wish to find the most energy-efficient path. The most common shortest path algorithm, Dijkstra's algorithm (Dijkstra, 1959), does not permit negative edge weights. Whilst alternative shortest path algorithms, such as Bellman-Ford (Shimbel, 1954; Bellman, 1958; Ford, 1956), allow negative edge weights, they are significantly slower and do not return a path if the graph has a reachable negative cycle. To avoid the complexity associated with negative weights, we use the rectified normal distribution to get non-negative energy consumption estimates $U_{t,e}$ as input for Dijkstra's algorithm. Upper confidence bound methods output optimistic estimates $\tilde{\mu}_e \in \mathbb{R}$ whereas Thompson sampling outputs posterior estimates $\tilde{\mu}_e \in \mathbb{R}$ by sampling from the posterior $\mathcal{N}(\mu_{t-1,e}, \sigma^2_{t-1,e})$. To ensure non-negative weights, the edge weight $U_{t,e}$ is set to $\mathbb{E}[z_e]$ where z_e is distributed as the rectified Gaussian $\mathcal{N}^R(\tilde{\mu}_e, \sigma^2_{t-1,e})$. A random variable $Y = \max(0, X)$ is said to have a rectified Gaussian distribution $\mathcal{N}^R(\mu, \sigma^2)$ if $X \sim \mathcal{N}(\mu, \sigma^2)$. In Algorithm 2, we show how to integrate UCB, BUCB and Thompson sampling with rectification within the framework of Algorithm 1. The notation $\mu_{t-1,e}$ and $\sigma^2_{t-1,e}$ refer respectively to the posterior mean and variance of the expected energy consumption for edge e whilst $\varsigma^2_{t-1,e}$ refers to the variance of the noise. Since the number of edges $|\mathcal{E}|$ may be large, each edge is sampled independently in TS, as by Nuara et al. (2018). Note that Algorithm 2 decouples the probabilistic regression model and Thompson sampling. In the next sections, we describe two probabilistic regression models for energy-efficient navigation.

GP regression for energy-efficient navigation To our knowledge, this study is the first combinatorial Gaussian process bandit solution for online energy-efficient navigation. The energy consumption depends on both the structure of the graph and the provided context. We use the graph Matérn kernel

Algorithm 3 SVGP Optimization Procedure

procedure UPDATEPARAMETERS($\mathbf{a}_t, \mathbf{r}_t, \boldsymbol{\theta}_{t-1}$)
1: Add $\mathbf{a}_t, \mathbf{r}_t$ to history.
2: Set inducing points $\mathbf{Z}_t$ to top M most visited edges.
3: **for** $i \in \{1, \dots, G\}$ **do**
4: $\quad \tilde{\mathbf{a}}, \tilde{\mathbf{r}} \leftarrow$ Subsample batch of size B from history.
5: $\quad$ Compute batch ELBO.
6: $\quad$ Optimize variational parameters with NGD.
7: Compute $\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t, \boldsymbol{\varsigma}_t$ using the optimized GP.
8: **return** $\boldsymbol{\mu}_t, \boldsymbol{\sigma}_t, \boldsymbol{\varsigma}_t$ and the optimized variational parameters.

$k_G : \mathcal{E} \times \mathcal{E} \to \mathbb{R}$ from Borovitskiy et al. (2021) to encode the structure of the line graph $\mathcal{L}(\mathcal{G})$ into the GP and an ordinary 5/2-Matérn kernel $k_f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ to encode the dependence on the context. The two kernels are combined $k_{G \cdot f + f} = k_G \cdot k_f + k'_f$ where the two feature kernels k_f and k'_f use separate sets of lengthscale and outputscale parameters. The cubic cost of exact GPs prohibits their application to large datasets. The sparse variational Gaussian processes (SVGP) (Titsias, 2009; Hensman et al., 2013) approximate the posterior distribution using a set of inducing points $\mathbf{Z}_t = \{z_1, \dots, z_M\}$ where $z_i \in \mathcal{A}$ and M is significantly smaller than the number of datapoints. By defining a prior distribution $q(\mathbf{u}_t) = \mathcal{N}(\mathbf{m}_t, \mathbf{S}_t)$ for the inducing variables $\mathbf{u}_t$, an approximate GP posterior can be obtained such that the complexity to perform N predictions is $\mathcal{O}(M^2 N)$, i.e. linear w.r.t. N. The variational paramaters $(\mathbf{m}_t, \mathbf{S}_t)$ are optimized by minimizing the evidence lower bound (ELBO) by performing G stochastic (natural) gradient descent steps using batch size B. Since the inducing points z_i lie in a mixed discrete and continuous space ($\mathcal{A} = \mathcal{E} \times \mathcal{X}$ for $\mathcal{X} \subset [0, C_1]^d$), we heuristically set z_i equal to the edge-context tuple of the i-th most visited edge at the start of the SVGP optimization. Then, the continuous dimensions of z_i are optimized together with $(\mathbf{m}_t, \mathbf{S}_t)$ using natural gradient descent (NGD) (Salimbeni et al., 2018). The procedure is described in Algorithm 3. Further details of the kernels and parameter values are provided in Sections B.1 and B.3.

Bayesian inference for energy-efficient navigation Åkerblom et al. (2023) introduced a framework for energy-efficient navigation using Bayesian inference to learn the distribution of the energy consumption in each road segment. The key assumption is that the energy consumption of an electric vehicle driving along a road segment is stochastic and follows a Gaussian distribution with unknown mean and known variance. Additionally, it is assumed that the energy consumption along different edges is independent. Using a Gaussian prior, the posterior distribution for edge e is computed using standard conjugate update rules.

Real-world road networks In our experiments we use the road networks of Luxembourg and Monaco (Codeca et al., 2017; Codeca & Härri, 2018, based on

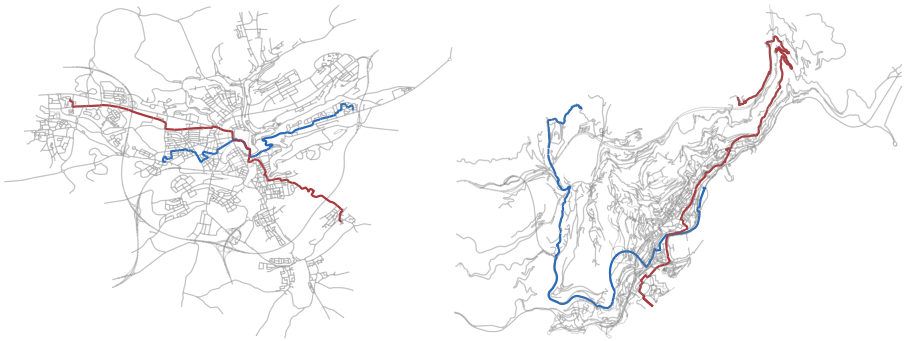


Figure 1: Road networks of Luxembourg (left) and Monaco (right) with evaluation routes A and B highlighted in blue and red.

data by OpenStreetMap contributors, 2017). Elevation data (Administration de la navigation aérienne, 2018) is added to the network using QGIS and the *netconvert* tool from SUMO. In Fig. 1, the two road networks are visualized along with two evaluation routes (A and B) per network. The evaluation routes span multiple regions of the network, allowing for many alternative paths. The context for each road segment consists of three fixed scalar properties: the length, the speed limit and the incline. Each property is standardized to have unit-variance. The prior expected energy consumption is computed by a deterministic model that assumes that the vehicle drives along an edge $e \in \mathcal{E}$, with length ℓ_e and inclination α_e, at constant speed v_e. The expended energy is then

$$E_e^{\text{det}} := (mg\ell_e \sin(\alpha_e) + mgC_r\ell_e \cos(\alpha_e) + 0.5C_dA\rho\ell_e v_e^2)/3600\eta. \tag{8}$$

The deterministic energy consumption E_e^{det} in Eq. (8) is given in Watt-hours and depends on the following vehicle-specific parameters: mass m, rolling resistance C_r, front surface area A, air drag coefficient C_d and powertrain efficiency η. The gravitational acceleration g and air density ρ also determine E_e^{det}. The parameter values are specified in Table 2 in Section B.3. Let $\overline{E^{\text{det}}} = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} E_e^{\text{det}}$ and $\sigma^2_{\text{det}} = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}}(E_e^{\text{det}} - \overline{E^{\text{det}}})^2$ denote the mean and variance of the deterministic energy consumption. The expected energy consumption is sampled from $\mathcal{GP}(E^{\text{det}}, k_{G\cdot f+f})$ where the outputscale of $k_{G\cdot f+f}$ (i.e. the variance $\boldsymbol{\sigma}_0^2$) is set to $0.25^2\sigma^2_{\text{det}}$. The noise variance $\boldsymbol{\varsigma}^2$ is set to $0.1^2\sigma^2_{\text{det}}$ for all edges and the kernel lengthscales are set to 1. See Section B for further details.

4.2 Results

Here, we demonstrate our experimental studies in different settings. We begin by comparing GP algorithms to Bayesian inference methods, then we compare the parametrizations of GP-UCB and GP-BUCB. Finally, we study the impact of the kernel lengthscale. Visualizations of the exploration are provided in Section C.2.

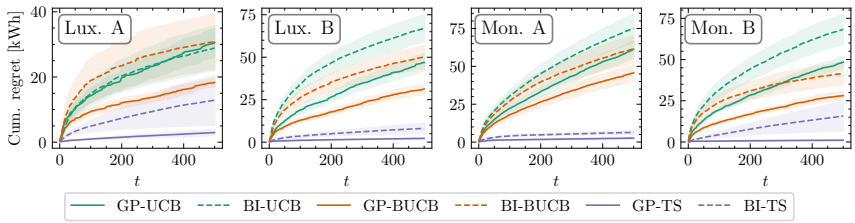


Figure 2: Cumulative regret for UCB, BUCB and TS using GP and Bayesian inference (BI) methods. The lines and regions correspond to the mean and ±1 standard error.

Investigation of different bandit algorithms In our first experiment, we compare the three algorithms GP-UCB, GP-BUCB and GP-TS. We use the Bayesian inference (BI) method of Åkerblom et al. (2023) with UCB, BUCB and TS as baselines. For UCB and BUCB (GP and BI), we use the β_t parametrization given by Theorem 3.2 with $\omega = 1$, $\xi = 1$. The six methods are evaluated 5 times each on the four routes in the Luxembourg and Monaco networks with a horizon of $T = 500$. The cumulative regret is shown in Fig. 2. The results show that the TS-based methods have significantly lower regret than both UCB and BUCB. Similarly, the GP-based methods generally have lower regret than their BI-based counterparts. Thereby, GP-TS yields the best results in terms of minimizing cumulative regret. Finally, we observe that GP-BUCB has lower regret than GP-UCB. In the next experiment, we investigate how the parametrization of these two algoritms affects the results.

BUCB parametrization As discussed in Section 2.2, GP-UCB and GP-BUCB differ mainly in their parametrization of the confidence parameter β_t. The confidence parameter determines the balance between exploration and exploitation. It is known that theoretical results tend to provide β_t values that overexplore (Russo & Roy, 2014). Using the parameters of β_t for GP-BUCB (ω and ξ), we can tune GP-BUCB towards more exploitation whilst retaining theoretical guarantees. We compare two theoretically valid choices of parametrizations for GP-BUCB ($\omega = 1$, $\xi = 1$ and $\omega = 1, \xi = 0.5$) against two parametrizations of GP-UCB where the first is theoretically valid and the second has scaled β_t by 0.5. The four parametrizations are evaluated 5 times each on the four routes in the Luxembourg and Monaco networks with a horizon of $T = 500$. The cumulative regret and the β_t values are shown in Fig. 3. The theoretically valid β_t values for GP-BUCB are smaller than for GP-UCB. By lowering ξ from 1 to 0.5, the quantile parameter η_t goes from $\mathcal{O}(t^{-1})$ to $\mathcal{O}(t^{-0.5})$ and using $\omega = 1$ the theoretial cumulative regret remains $\mathcal{O}(\sqrt{T})$.[2] The experimental results indicate that the parametrization with lower β_t generally has lower cumulative regret. Using GP-BUCB, we gain more

[2]Technically, one must use $\xi \leq 0.5 - \delta$ and $\omega \geq 1 + \delta$ for some $\delta > 0$. However, we could choose δ to be small enough such that GP-BUCB would select the exact same routes in all experiments.

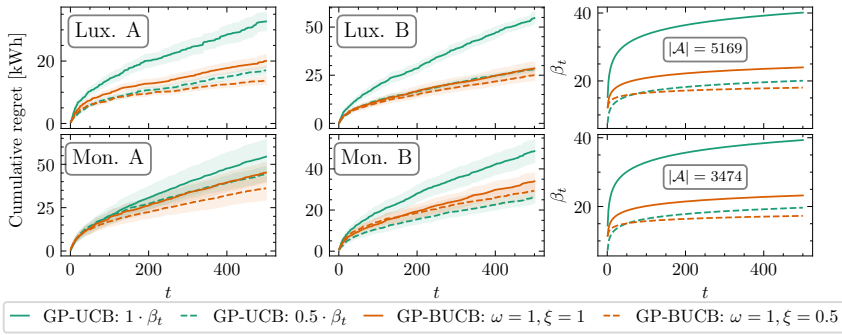


Figure 3: Cumulative regret of GP-UCB and GP-BUCB (left and middle column) for different parametrizations of β_t (right column).

control of β_t without sacrificing the theoretical guarantees.

Impact of lengthscale Finally, we investigate varying the kernel lengthscale to ensure our results are consistent and stable. A large lengthscale increases the correlation between edges, which should lower the regret of the GP-methods. Whilst a lower lengthscale decreases the correlation which should increase the regret of the GP-methods. We evaluate GP-BUCB and GP-TS against BI-BUCB and BI-TS with the kernel lengthscale varying between 0.1 and 2.0. Each combination of lengthscale and bandit-method is evaluated 5 times on all four routes with a horizon of $T = 500$. The final cumulative regret at $t = 500$ for the different lengthscales is shown in Figs. 4 and 5 in Section C.1. For GP-based methods, increasing the lengthscale increases the cumulative regret overall but for BI-based methods, there is no discernable pattern.

5 Conclusion

We presented novel Bayesian regret bounds for the combinatorial volatile Gaussian process semi-bandit for three GP-based bandit algorithms: GP-UCB, GP-BayesUCB and GP-TS. Additionally, we experimentally evaluated our contextual combinatorial GP method on the online energy-efficient navigation problem on real-world networks.

Acknowledgments

The work of Jack Sandberg and Morteza Haghir Chehreghani was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The work of Niklas Åkerblom was partially funded by the Strategic Vehicle Research and Innovation Programme (FFI) of Sweden, through the project EENE (reference number: 2018-01937). The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in

Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725. Map data from Openstreetmap and available from `www.openstreetmap.org/copyright`.

References

Administration de la navigation aérienne. Digital Terrain Model (high DEM resolution), 2018. URL `https://data.public.lu/en/datasets/digital-terrain-model-high-dem-resolution/`.

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the Monster: A Fast and Simple Algorithm for Contextual Bandits. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 1638–1646. PMLR, June 2014.

Niklas Åkerblom, Yuxin Chen, and Morteza Haghir Chehreghani. Online learning of energy consumption for navigation of electric vehicles. *Artificial Intelligence*, 317:103879, April 2023.

Richard Bellman. On a routing problem. *Quarterly of Applied Mathematics*, 16(1):87–90, 1958.

Viacheslav Borovitskiy, Iskander Azangulov, Alexander Terenin, Peter Mostowsky, Marc Deisenroth, and Nicolas Durrande. Matérn Gaussian Processes on Graphs. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2593–2601. PMLR, April 2021.

Djallel Bouneffouf, Irina Rish, and Charu Aggarwal. Survey on Applications of Multi-Armed and Contextual Bandits. In *2020 IEEE Congress on Evolutionary Computation (CEC)*, 2020. doi: 10.1109/CEC48606.2020.9185782.

Xu Cai and Jonathan Scarlett. On Lower Bounds for Standard and Robust Gaussian Process Bandit Optimization. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 1216–1226. PMLR, July 2021.

Nicolò Cesa-Bianchi and Gábor Lugosi. Combinatorial bandits. *Journal of Computer and System Sciences*, 78(5):1404–1422, September 2012.

Seok-Ho Chang, Pamela C. Cosman, and Laurence B. Milstein. Chernoff-Type Bounds for the Gaussian Error Function. *IEEE Transactions on Communications*, 59(11):2939–2944, 2011.

Lixing Chen, Jie Xu, and Zhuo Lu. Contextual Combinatorial Multi-armed Bandits with Volatile Arms and Submodular Reward. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial Multi-Armed Bandit: General Framework and Applications. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 151–159. PMLR, February 2013. ISSN: 1938-7228.

Sayak Ray Chowdhury and Aditya Gopalan. On Kernelized Multi-armed Bandits. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 844–853. PMLR, July 2017.

Lara Codeca and Jérôme Härri. Monaco SUMO Traffic (MoST) Scenario: A 3D Mobility Scenario for Cooperative ITS. In *SUMO 2018, SUMO User Conference, Simulating Autonomous and Intermodal Transport Systems, May 14-16, 2018, Berlin, Germany*, Berlin, GERMANY, May 2018.

Lara Codeca, Raphael Frank, Sebastien Faye, and Thomas Engel. Luxembourg SUMO Traffic (LuST) Scenario: Traffic Demand Evaluation. *IEEE Intelligent Transportation Systems Magazine*, 9(2):52–63, 2017.

E. W. Dijkstra. A Note on Two Problems in Connexion with Graphs. *Numerische Mathematik*, 1:269–271, 1959.

Sepehr Elahi, Baran Atalar, Sevda Öugüt, and Cem Tekin. Contextual Combinatorial Multi-output GP Bandits with Group Constraints. *Transactions on Machine Learning Research*, January 2023.

L. R. Ford. Network Flow Theory. Technical report, RAND Corporation, January 1956.

Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Combinatorial Network Optimization With Unknown Variables: Multi-Armed Bandits With Linear Rewards and Individual Observations. *IEEE/ACM Transactions on Networking*, 20(5):1466–1478, October 2012.

Jacob R Gardner, Geoff Pleiss, David Bindel, Kilian Q Weinberger, and Andrew Gordon Wilson. GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. In *Advances in Neural Information Processing Systems*, 2018.

Subhashis Ghosal and Anindya Roy. Posterior consistency of Gaussian process prior for nonparametric binary regression. *The Annals of Statistics*, 34(5), October 2006.

James Hensman, Nicolò Fusi, and Neil D. Lawrence. Gaussian Processes for Big Data. In *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence*, UAI'13, pp. 282–290, Arlington, Virginia, USA, 2013. AUAI Press.

Kirthevasan Kandasamy, Akshay Krishnamurthy, Jeff Schneider, and Barnabas Poczos. Parallelised Bayesian Optimisation via Thompson Sampling. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pp. 133–142. PMLR, March 2018.

Emilie Kaufmann, Olivier Cappe, and Aurelien Garivier. On Bayesian Upper Confidence Bounds for Bandit Problems. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, pp. 592–600. PMLR, March 2012.

Andreas Krause and Cheng Ong. Contextual Gaussian Process Bandit Optimization. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pp. 661–670. Association for Computing Machinery, April 2010.

Baihan Lin and Djallel Bouneffouf. Optimal Epidemic Control as a Contextual Combinatorial Bandit with Budget. In *2022 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–8, July 2022.

Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wiessner. Microscopic Traffic Simulation using SUMO. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pp. 2575–2582, November 2018.

Nadav Merlis and Shie Mannor. Tight Lower Bounds for Combinatorial Multi-Armed Bandits. In *Proceedings of Thirty Third Conference on Learning Theory*, pp. 2830–2857. PMLR, 2020. URL `https://proceedings.mlr.press/v125/merlis20a.html`.

Andi Nika, Sepehr Elahi, and Cem Tekin. Contextual Combinatorial Volatile Multi-armed Bandit with Adaptive Discretization. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 1486–1496. PMLR, June 2020.

Andi Nika, Sepehr Elahi, and Cem Tekin. Contextual Combinatorial Bandits with Changing Action Sets via Gaussian Processes, October 2022. arXiv:2110.02248.

Alessandro Nuara, Francesco Trovò, Nicola Gatti, and Marcello Restelli. A Combinatorial-Bandit Algorithm for the Online Joint Bid/Budget Optimization of Pay-per-Click Advertising Campaigns. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018.

OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . `https://www.openstreetmap.org`, 2017.

K. B. Petersen and M. S. Pedersen. The Matrix Cookbook, November 2012. URL `http://www2.compute.dtu.dk/pubdb/pubs/3274-full.html`.

QGIS Development Team. *QGIS Geographic Information System*. QGIS Association, 2023. URL `https://www.qgis.org`.

Lijing Qin, Shouyuan Chen, and Xiaoyan Zhu. Contextual Combinatorial Bandit and its Application on Diversified Online Recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining (SDM)*, Proceedings, pp. 461–469. Society for Industrial and Applied Mathematics, April 2014.

Herbert Robbins. Some aspects of the sequential design of experiments. *Herbert Robbins Selected Papers*, pp. 169–177, 1985.

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

Hugh Salimbeni, Stefanos Eleftheriadis, and James Hensman. Natural Gradients in Practice: Non-Conjugate Variational Inference in Gaussian Process Models. In Amos Storkey and Fernando Perez-Cruz (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 689–697. PMLR, April 2018.

Jonathan Scarlett. Tight Regret Bounds for Bayesian Optimization in One Dimension. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 4500–4508. PMLR, July 2018.

Fang Shi, Weiwei Lin, Lisheng Fan, Xiazhi Lai, and Xiumin Wang. Efficient Client Selection Based on Contextual Combinatorial Multi-Arm Bandits. *IEEE Transactions on Wireless Communications*, 22(8):5265–5277, August 2023.

Alfonso Shimbel. Structure in communication nets. *Proceedings of the Symposium on Information Networks*, pp. 119–203, 1954.

Niranjan Srinivas, Andreas Krause, Sham M. Kakade, and Matthias W. Seeger. Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting. *IEEE Transactions on Information Theory*, 58(5): 3250–3265, May 2012.

Michael L. Stein. *Interpolation of Spatial Data*. Springer Series in Statistics. Springer, New York, NY, 1999.

Shion Takeno, Yu Inatsu, and Masayuki Karasuyama. Randomized Gaussian Process Upper Confidence Bound with Tighter Bayesian Regret Bounds. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 33490–33515. PMLR, July 2023.

Shion Takeno, Yu Inatsu, Masayuki Karasuyama, and Ichiro Takeuchi. Posterior Sampling-Based Bayesian Optimization with Tighter Bayesian Regret Bounds. In *Proceedings of the 41st International Conference on Machine Learning*. PMLR, 2024.

William R Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, December 1933.

Michalis Titsias. Variational Learning of Inducing Variables in Sparse Gaussian Processes. In *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, pp. 567–574. PMLR, April 2009.

Sattar Vakili, Kia Khezeli, and Victor Picheny. On Information Gain and Regret Bounds in Gaussian Process Bandits. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pp. 82–90. PMLR, March 2021.

Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient Learning in Large-Scale Combinatorial Semi-Bandits. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1113–1122. PMLR, June 2015.

Xingzhi Zhan. Extremal Eigenvalues of Real Symmetric Matrices with Entries in an Interval. *SIAM Journal on Matrix Analysis and Applications*, 27(3): 851–860, 2005. ISSN 0895-4798. doi: 10.1137/050627812.

Li Zhou. A Survey on Contextual Multi-armed Bandits, February 2016. arXiv:1508.03326.

A Proofs

A.1 Finite case

In this section, we state and prove the regret bounds in the finite case for the three bandit algorithms GP-UCB, GP-BUCB and GP-TS. To begin, we establish lemmas that demonstrate the general procedure for the proofs and later we combine the lemmas to get the desired regret bounds.

In the following lemma, the Bayesian regret is separated into two terms.

Lemma A.1. *For GP-TS or any GP-UCB method the following upper bound on the Bayesian regret holds (with equality for GP-TS):*

$$BR(T) \leq \sum_{t\in[T]} \mathbb{E}[f(\mathbf{a}_t^*) - U_t(\mathbf{a}_t^*)] + \mathbb{E}[U_t(\mathbf{a}_t) - f(\mathbf{a}_t)]. \tag{9}$$

Proof. The proof follows the procedure of Prop. 1 by Russo & Roy (2014) for GP-TS and Thm. B.1. by Takeno et al. (2023) for GP-UCB. For GP-TS,

$$\begin{aligned}
\mathrm{BR}(T) &= \sum_{t\in[T]} \mathbb{E}\left[f(\mathbf{a}_t^*) - f(\mathbf{a}_t)\right] && (10)\\
&= \sum_{t\in[T]} \mathbb{E}_{H_t}\left[\mathbb{E}_t\left[f(\mathbf{a}_t^*) - U_t(\mathbf{a}_t^*) + U_t(\mathbf{a}_t) - f(\mathbf{a}_t)|H_t\right]\right] \\
&\qquad\qquad \left(\mathbf{a}_t^*|H_t \overset{d}{=} \mathbf{a}_t|H_t\right) && (11)\\
&= \sum_{t\in[T]} \mathbb{E}[f(\mathbf{a}_t^*) - U_t(\mathbf{a}_t^*)] + \sum_{t\in[T]} \mathbb{E}[U_t(\mathbf{a}_t) - f(\mathbf{a}_t)]. && (12)
\end{aligned}$$

Similarly, for any GP-UCB method,

$$\begin{aligned}
\mathrm{BR}(T) &= \sum_{t\in[T]} \mathbb{E}\left[f(\mathbf{a}_t^*) - f(\mathbf{a}_t)\right] && (13)\\
&= \sum_{t\in[T]} \mathbb{E}\left[f(\mathbf{a}_t^*) - U_t(\mathbf{a}_t^*) + U_t(\mathbf{a}_t^*) - U_t(\mathbf{a}_t) + U_t(\mathbf{a}_t) - f(\mathbf{a}_t)\right] && (14)\\
&\leq \sum_{t\in[T]} \mathbb{E}\left[f(\mathbf{a}_t^*) - U_t(\mathbf{a}_t^*) + U_t(\mathbf{a}_t) - f(\mathbf{a}_t)\right] && (15)
\end{aligned}$$

where the final step uses that $U_t(\mathbf{a}_t^*) - U_t(\mathbf{a}_t) \leq 0$ since $\mathbf{a}_t = \arg\max_{\mathbf{a}\in\mathcal{S}_t} U_t(\mathbf{a})$. □

Whilst Lemma A.1 applies to all the considered bandit algorithms, the two terms in the decomposition requires knowing the specific bandit algorithm. Bounding the left term requires knowledge of the confidence parameter β_t. Therefore we present a lemma that applies to GP-UCB and GP-TS, and another lemma that applies to GP-BUCB.

Lemma A.2. *If $|\mathcal{A}| < \infty$, then*

$$\sum_{t\in[T]} \mathbb{E}[f(\mathbf{a}_t^*) - U_t(\mathbf{a}_t^*)] \leq \frac{\pi^2}{6} \tag{16}$$

holds for GP-UCB and GP-TS with $\beta_t = 2\log\left(|\mathcal{A}|t^2/\sqrt{2\pi}\right)$.

Proof. The proof closely follows the proof of Thm. B.1 by Takeno et al. (2023). Let $R_1 = \sum_{t\in[T]} \mathbb{E}[f(\mathbf{a}_t^*) - U_t(\mathbf{a}_t^*)]$, then

$$\begin{aligned}
R_1 &= \sum_{t\in[T]} \mathbb{E}_{H_t}\left[\mathbb{E}_t\left[f(\mathbf{a}_t^*) - U_t(\mathbf{a}_t^*)|H_t\right]\right] && (17)\\
&= \sum_{t\in[T]} \mathbb{E}_{H_t}\left[\mathbb{E}_t\left[\sum_{a\in\mathbf{a}_t^*} f(a) - U_t(a)\middle| H_t\right]\right] && (18)\\
&\leq \sum_{t\in[T]} \mathbb{E}_{H_t}\left[\mathbb{E}_t\left[\sum_{a\in\mathbf{a}_t^*} \left(f(a) - U_t(a)\right)_+ \middle| H_t\right]\right] \qquad \left(\begin{matrix}(x)_+ := \\ \max(0,x) \geq x\end{matrix}\right) && (19)\\
&\leq \sum_{t\in[T]} \mathbb{E}_{H_t}\left[\sum_{a\in\mathcal{A}} \mathbb{E}_t\left[\left(f(a) - U_t(a)\right)_+ \middle| H_t\right]\right]. \qquad (\mathbf{a}_t^* \subseteq \mathcal{A}) && (20)
\end{aligned}$$

Note that $f(a) - U_t(a)|H_t \sim \mathcal{N}(-\sqrt{\beta_t}\sigma_{t-1}(a), \sigma_{t-1}^2(a))$. As Russo & Roy (2014), by using that if $X \sim \mathcal{N}(\mu, \sigma^2)$ for $\mu \leq 0$, then $\mathbb{E}[(X)_+] \leq \frac{\sigma}{\sqrt{2\pi}} \exp\left(\frac{-\mu^2}{2\sigma^2}\right)$, we get the following for R_1:

$$\begin{aligned}
R_1 &\leq \sum_{t\in[T]} \mathbb{E}_{H_t}\left[\sum_{a\in\mathcal{A}} \mathbb{E}_t\left[\frac{\sigma_{t-1}(a)}{\sqrt{2\pi}} \exp\left(\frac{-\beta_t}{2}\right)\middle| H_t\right]\right] && (21)\\
&\leq \sum_{t\in[T]} \frac{|\mathcal{A}|}{\sqrt{2\pi}} \exp\left(\frac{-\beta_t}{2}\right) && \left(\sigma_{t-1}^2(a) \leq k(a,a) \leq 1\right) && (22)\\
&\leq \sum_{t\in[T]} \frac{1}{t^2} && \left(\beta_t = 2\log\left(\frac{|\mathcal{A}|t^2}{\sqrt{2\pi}}\right)\right) && (23)\\
&\leq \frac{\pi^2}{6}. && \left(\sum_{t=1}^{\infty} \frac{1}{t^2} = \frac{\pi^2}{6}\right) && (24)
\end{aligned}$$

□

Lemma 3.1. *Let $C_\omega = \left(\sqrt{\pi}\omega/\sqrt{2e(\omega-1)}\right)^{1/\omega}$, then for GP-BUCB with confidence parameter $\beta_t = 2\left(\mathrm{erf}^{-1}(1-2\eta_t)\right)^2$ and $\eta_t = \frac{\sqrt{2\pi}^\omega}{2|\mathcal{A}|^\omega t^\xi}$, $\xi > 0$, $\omega > 1$,*

$$\sum_{t\in[T]} \mathbb{E}[f(\mathbf{a}_t^*) - U_t(\mathbf{a}_t^*)] \leq C_\omega \cdot \begin{cases} \frac{\omega}{\omega-\xi} T^{1-\frac{\xi}{\omega}} & \text{if } \xi/\omega < 1, \\ 1 + \log T & \text{if } \xi/\omega = 1, \\ \frac{\xi}{\xi-\omega} & \text{if } \xi/\omega > 1. \end{cases}$$

Proof. Following the proof of Lemma A.2, we get that

$$\sum_{t\in[T]} \mathbb{E}\left[f(\mathbf{a}_t^*) - U_t(\mathbf{a}_t^*)\right] \leq \sum_{t\in[T]} \frac{|\mathcal{A}|}{\sqrt{2\pi}} \exp\left(-\frac{\beta_t}{2}\right). \tag{25}$$

Note that, according to Lemma A.13, $\mathrm{erf}^{-1}(u) \geq \sqrt{-\omega^{-1}\log((1-u)/\vartheta)}$ for $\omega > 1$ and $\vartheta = \sqrt{2e/\pi}\sqrt{\omega-1}/\omega$. We use the largest value of ϑ permitted by Lemma A.13 since it yields the tightest bound. Then,

$$\sum_{t\in[T]} \frac{|\mathcal{A}|}{\sqrt{2\pi}} \exp\left(-\frac{\beta_t}{2}\right) = \sum_{t\in[T]} \frac{|\mathcal{A}|}{\sqrt{2\pi}} \exp\left(-\left(\mathrm{erf}^{-1}(1-2\eta_t)\right)^2\right) \tag{26}$$

$$\leq \sum_{t\in[T]} \frac{|\mathcal{A}|}{\sqrt{2\pi}} \exp\left(\omega^{-1}\log\left(\frac{1-(1-2\eta_t)}{\vartheta}\right)\right) \qquad \text{(Lemma A.13)} \tag{27}$$

$$= \sum_{t\in[T]} \frac{|\mathcal{A}|}{\sqrt{2\pi}} \left(\frac{2\eta_t}{\vartheta}\right)^{\frac{1}{\omega}} \tag{28}$$

$$= \sum_{t\in[T]} \vartheta^{-\frac{1}{\omega}} t^{-\frac{\xi}{\omega}} \qquad \text{(Def. of } \eta_t) \tag{29}$$

$$= \left(\frac{\sqrt{\pi}\omega}{\sqrt{2e(\omega-1)}}\right)^{\frac{1}{\omega}} \sum_{t\in[T]} t^{-\frac{\xi}{\omega}}. \qquad \text{(Def. of } \vartheta) \tag{30}$$

The behaviour of $\sum_{t\in[T]} t^{-\frac{\xi}{\omega}}$ critically depends on the ratio ξ/ω. First, if $\xi/\omega < 1$, then

$$\sum_{t\in[T]} t^{-\frac{\xi}{\omega}} \leq \int_0^T t^{-\frac{\xi}{\omega}} \mathrm{d}t = T^{1-\frac{\xi}{\omega}} \frac{1}{1-\frac{\xi}{\omega}}. \tag{31}$$

Second, if $\xi/\omega = 1$, then

$$\sum_{t\in[T]} t^{-1} \leq 1 + \int_1^T t^{-1} dt = 1 + \log T. \tag{32}$$

Finally, if $\xi/\omega > 1$, then

$$\sum_{t\in[T]} t^{-\frac{\xi}{\omega}} \leq 1 + \int_1^\infty t^{-\frac{\xi}{\omega}} \mathrm{d}t = 1 + \left[\frac{1}{1-\frac{\xi}{\omega}} t^{1-\frac{\xi}{\omega}}\right]_1^\infty = 1 - \frac{1}{1-\frac{\xi}{\omega}} = \frac{\xi}{\xi-\omega}. \tag{33}$$

□

Before we bound the right term in Lemma A.1, we introduce a lemma for the confidence radius that applies to all the bandit algorithms considered.

Lemma A.3.

$$\sum_{t\in[T]} \mathbb{E}\left[\sum_{a\in\mathbf{a}_t} \sqrt{\beta_t}\sigma_{t-1}(a)\right] \leq \sqrt{2(\lambda_K^* + \varsigma^2)TK\beta_T\gamma_{TK}} \tag{34}$$

for GP-TS or any GP-UCB method with increasing confidence parameter β_t.

Proof.

$$\sum_{t\in[T]} \mathbb{E}\left[\sum_{a\in\mathbf{a}_t} \sqrt{\beta_t}\sigma_{t-1}(a)\right] \tag{35}$$

$$= \mathbb{E}\left[\sum_{t\in[T]}\sum_{a\in\mathbf{a}_t} \sqrt{\beta_t}\sigma_{t-1}(a)\right] \tag{36}$$

$$\leq \mathbb{E}\left[\sqrt{\sum_{t\in[T]}\sum_{a\in\mathbf{a}_t} \beta_t}\sqrt{\sum_{t\in[T]}\sum_{a\in\mathbf{a}_t} \sigma_{t-1}^2(a)}\right] \qquad \begin{pmatrix}\text{Cauchy-Schwarz}\\ \text{inequality}\end{pmatrix} \tag{37}$$

$$\leq \mathbb{E}\left[\sqrt{TK\beta_T}\sqrt{\sum_{t\in[T]}\sum_{a\in\mathbf{a}_t} \sigma_{t-1}^2(a)}\right] \qquad \left(|\mathbf{a}_t| \leq K, \max_{t\in[T]} \beta_t = \beta_T\right) \tag{38}$$

$$= \sqrt{TK\beta_T}\mathbb{E}\left[\sqrt{\sum_{t\in[T]}\sum_{a\in\mathbf{a}_t} \sigma_{t-1}^2(a)}\right] \tag{39}$$

$$\leq \sqrt{TK\beta_T}\mathbb{E}\left[\sqrt{2(\lambda_K^* + \varsigma^2)\gamma_{TK}}\right] \qquad \text{(Lemma A.12)} \tag{40}$$

$$\leq \sqrt{2(\lambda_K^* + \varsigma^2)TK\beta_T\gamma_{TK}}. \tag{41}$$

□

Next, we show how the right term in Lemma A.1 can be rewritten in terms of the confidence radius for any GP-UCB method.

Lemma A.4.

$$\sum_{t\in[T]} \mathbb{E}\left[U_t(\mathbf{a}_t) - f(\mathbf{a}_t)\right] = \sum_{t\in[T]} \mathbb{E}\left[\sqrt{\beta_t}\sigma_{t-1}(\mathbf{a}_t)\right] \tag{42}$$

for any GP-UCB method with confidence parameter β_t.

Proof. Note that given the history H_t, $\mathbf{a}_t := \arg\max_{\mathbf{a}\in\mathcal{S}_t} U_t(\mathbf{a})$ is deterministic. Thus,

$$\sum_{t\in[T]} \mathbb{E}\left[U_t(\mathbf{a}_t) - f(\mathbf{a}_t)\right] = \sum_{t\in[T]} \mathbb{E}_{H_t}\left[\mathbb{E}_t\left[U_t(\mathbf{a}_t) - f(\mathbf{a}_t)|H_t\right]\right] \tag{43}$$

$$= \sum_{t\in[T]} \mathbb{E}_{H_t}\left[\mathbb{E}_t\left[\mu_{t-1}(\mathbf{a}_t) + \sqrt{\beta_t}\sigma_{t-1}(\mathbf{a}_t) - f(\mathbf{a}_t)\middle| H_t\right]\right] \tag{44}$$

$$= \sum_{t\in[T]} \mathbb{E}_{H_t}\left[\mathbb{E}_t\left[\mu_{t-1}(\mathbf{a}_t) + \sqrt{\beta_t}\sigma_{t-1}(\mathbf{a}_t) - \mu_{t-1}(\mathbf{a}_t)\middle| H_t\right]\right] \tag{45}$$

$$= \sum_{t\in[T]} \mathbb{E}\left[\sqrt{\beta_t}\sigma_{t-1}(\mathbf{a}_t)\right]. \tag{46}$$

□

For the final lemma in the finite case, we bound the right term in Lemma A.1 for Thompson sampling using the previous results.

Lemma A.5.

$$\sum_{t\in[T]} \mathbb{E}\left[U_t(\mathbf{a}_t) - f(\mathbf{a}_t)\right] \leq 2\sqrt{C_K T K \beta_T \gamma_{TK}} + \frac{\pi^2}{6} \tag{47}$$

holds for GP-TS with $\beta_t = 2\log\left(|\mathcal{A}|t^2/\sqrt{2\pi}\right)$.

Proof. By adding and subtracting the lower bound $L(\mathbf{a}_t)$, we obtain

$$\sum_{t\in[T]} \mathbb{E}\left[U_t(\mathbf{a}_t) - f(\mathbf{a}_t)\right] = \sum_{t\in[T]} \mathbb{E}\left[U_t(\mathbf{a}_t) - f(\mathbf{a}_t) + L(\mathbf{a}_t) - L(\mathbf{a}_t)\right] \tag{48}$$

$$= \sum_{t\in[T]} \mathbb{E}\left[U_t(\mathbf{a}_t) - L(\mathbf{a}_t)\right] + \sum_{t\in[T]} \mathbb{E}\left[L(\mathbf{a}_t) - f(\mathbf{a}_t)\right] \tag{49}$$

$$= \underbrace{2\sum_{t\in[T]} \mathbb{E}\left[\sqrt{\beta_t}\sigma_{t-1}(\mathbf{a}_t)\right]}_{(1)} + \underbrace{\sum_{t\in[T]} \mathbb{E}\left[L(\mathbf{a}_t) - f(\mathbf{a}_t)\right]}_{(2)}. \tag{50}$$

By Lemma A.3, $(1) \leq \sqrt{2(\lambda_K^* + \varsigma^2)TK\beta_T\gamma_{TK}}$. The bound $(2) \leq \frac{\pi^2}{6}$ is obtained using the same steps as in Lemma A.2 due to the symmetry of $L(\mathbf{a}_t) - f(\mathbf{a}_t)$ and $f(\mathbf{a}_t) - U(\mathbf{a}_t)$. □

Finally, we present and prove the regret bounds for GP-UCB, GP-BUCB and GP-TS using the established lemmas.

Theorem 3.2 (Finite regret bounds)**.** *Let* $C_K := 2(\lambda_K^* + \varsigma^2)$. *When* $\mathcal{A}$ *is finite, the Bayesian regret of*

(i) GP-UCB with $\beta_t = 2\log(|\mathcal{A}|t^2/\sqrt{2\pi})$ *is bounded as* $BR(T) \leq \frac{\pi^2}{6} + \sqrt{C_K T K \beta_T \gamma_{TK}}$.

(ii) GP-BUCB with $\beta_t = 2\left(\mathrm{erf}^{-1}(1-2\eta_t)\right)^2$ *for* $\eta_t = \frac{\sqrt{2\pi}^\omega}{2|\mathcal{A}|^\omega t^\xi}$, $\xi > \omega > 1$ *is bounded as* $BR(T) \leq \sqrt{C_K T K \beta_T \gamma_{TK}} + C_\omega \cdot \frac{\xi}{\xi-\omega}$ *where* $C_\omega = \left(\sqrt{\pi}\omega/\sqrt{2e(\omega-1)}\right)^{1/\omega}$.

(iii) GP-TS is bounded as $\mathrm{BR}(T) \leq \frac{\pi^2}{3} + 2\sqrt{C_K T K \beta_T \gamma_{TK}}$ *where* $\beta_t = 2\log\left(|\mathcal{A}|t^2/\sqrt{2\pi}\right)$.

Proof. (i) The regret bound for GP-UCB is obtained as follows:

$$\mathrm{BR}(T) \leq \sum_{t\in[T]} \mathbb{E}\left[f(\mathbf{a}_t^*) - U_t(\mathbf{a}_t^*)\right] + \sum_{t\in[T]} \mathbb{E}\left[U_t(\mathbf{a}_t) - f(\mathbf{a}_t)\right] \quad \text{(Lemma A.1)} \tag{51}$$

$$\leq \frac{\pi^2}{6} + \sum_{t\in[T]} \mathbb{E}\left[\sqrt{\beta_t}\sigma_{t-1}(\mathbf{a}_t)\right] \quad \text{(Lemmas A.2 and A.4)} \tag{52}$$

$$\leq \frac{\pi^2}{6} + \sqrt{2(\lambda_K^* + \varsigma^2)TK\beta_T\gamma_{TK}}. \quad \text{(Lemma A.3)} \tag{53}$$

(ii) The regret of GP-BUCB can be decomposed as follows:

$$\mathrm{BR}(T) \leq \sum_{t\in[T]} \mathbb{E}\left[U_t(\mathbf{a}_t) - f(\mathbf{a}_t)\right] + \sum_{t\in[T]} \mathbb{E}\left[f(\mathbf{a}_t^*) - U_t(\mathbf{a}_t^*)\right] \quad \text{(Lemma A.1)} \tag{54}$$

$$\leq \sum_{t\in[T]} \mathbb{E}\left[\sqrt{\beta_t}\sigma_{t-1}(\mathbf{a}_t)\right] \quad \text{(Lemma A.4)} \tag{55}$$

$$+ \left(\frac{\sqrt{\pi}\omega}{\sqrt{2e(\omega-1)}}\right)^{1/\omega} \cdot \begin{cases} \frac{\omega}{\omega-\xi} T^{1-\frac{\xi}{\omega}} & \text{if } \xi/\omega < 1, \\ \frac{\xi}{\xi-\omega} & \text{if } \xi/\omega > 1. \end{cases} \quad \text{(Lemma 3.1)} \tag{56}$$

From Lemma A.3, $\sum_{t\in[T]} \mathbb{E}\left[\sqrt{\beta_t}\sigma_{t-1}(\mathbf{a}_t)\right] \leq \sqrt{2(\lambda_K^* + \varsigma^2)TK\beta_T\gamma_{TK}}$ and we obtain the desired result.

(iii) The regret of GP-TS is obtained as follows:

$$\mathrm{BR}(T) = \sum_{t\in[T]} \mathbb{E}\left[f(\mathbf{a}_t^*) - U_t(\mathbf{a}_t^*)\right] + \sum_{t\in[T]} \mathbb{E}\left[U_t(\mathbf{a}_t) - f(\mathbf{a}_t)\right] \quad \text{(Lemma A.1)} \tag{57}$$

$$\leq \frac{\pi^2}{6} + \frac{\pi^2}{6} + 2\sqrt{C_K T K \beta_T \gamma_{TK}}. \quad \text{(Lemmas A.2 and A.5)} \tag{58}$$

□

A.2 Infinite case

Similar to the finite case, we establish lemmas that hold for all bandit algorithms and finally state and prove the regret bounds.

Before stating the first lemma, we restate the assumptions for convenience:

Assumption 3.3 (Regularity assumptions)**.** *Assume* $\mathcal{A} \subset [0, C_1]^d$ *is a compact and convex set for some* $C_1 > 0$*. Furthermore, assume that* μ *and* k *are both* L*-Lipschitz on* $\mathcal{A}$ *and* $\mathcal{A} \times \mathcal{A}$*, respectively, for some* $L > 0$*. In addition, for* $f \sim \mathcal{GP}(\mu, k)$ *assume that there exists constants* $C_2, C_3 > 0$ *such that:*

$$\mathbb{P}\left(\sup_{a\in\mathcal{A}} \left|\frac{\partial f}{\partial a^{(j)}}\right| > l\right) \leq C_2 \exp\left(-\frac{l^2}{C_3^2}\right), \tag{5}$$

for $j \in \{1, \ldots, d\}$ *and* $l > 0$ *where* $a^{(j)}$ *denotes the* j*-th element of* a*.*

Remark A.6. *By Thm. 5 of Ghosal & Roy (2006), the high probability bound holds if μ is continuously differentiable and k is 4 times differentiable, which would also imply the Lipschitzness of μ and k. As discussed by Srinivas et al. (2012), this holds for the Matérn kernel if $\nu \geq 2$ by a result of Stein (1999) and holds trivially for the squared exponential kernel. Thus, the Lipschitz assumption of μ and k is not particularly restrictive.*

Assumption 3.4 (Discretization size)**.** *Let τ_t denote the number of discretization points per dimension and assume that*

$$\begin{cases} \tau_t \geq 2t^2 KLdC_1(1 + tK\varsigma^{-1}), & \text{(6a)} \\ \tau_t/\beta_t \geq 8t^4 K^2 LdC_1, & \text{(6b)} \\ \tau_t^2/\beta_t \geq 8t^5 K^3 L^2 d^2 C_1^2 \varsigma^{-2}, & \text{(6c)} \\ \tau_t \geq t^2 KdC_1 C_3(\sqrt{\log(C_2 d)} + \sqrt{\pi}/2) & \text{(6d)} \end{cases}$$

where the constants C_1, C_2, C_3 and L are given by Assumption 3.3 whilst the constants d, K and ς are defined by the bandit problem (Section 2.1).

Remark A.7. *For the theorems to be relevant, the assumptions imposed on τ_t must be satisfiable for some τ_t. If $\beta_t = 2\log\left(\frac{\tau_t^d t^2}{\sqrt{2\pi}}\right)$, then Assumption 3.4 is satisfied by*

$$\tau_t = \max \begin{cases} 2KLdC_1(1 + tK\varsigma^{-1})t^2, \\ \left(\left(16t^4 K^2 LdC_1\right)\left(d + \log\left(\frac{t^2}{\sqrt{2\pi}}\right)\right)\right)^{\frac{1}{1-1/e}}, \\ \left(\left(16t^5 K^3 L^2 d^2 C_1^2 \varsigma^{-2}\right)\left(d + \log\left(\frac{t^2}{\sqrt{2\pi}}\right)\right)\right)^{\frac{1}{2-1/e}}, \\ t^2 KdC_1 C_3\left(\sqrt{\log(C_2 d)} + \frac{\sqrt{\pi}}{2}\right). \end{cases} \tag{59}$$

This can be shown by noting that $\log \tau_t \leq \sqrt[e]{\tau_t}$ and $1 \leq \sqrt[e]{\tau_t}$ and then deriving that $\frac{1}{\beta_t} \geq \frac{1}{\tau_t^{1/e}(d + \log(t^2/\sqrt{2\pi}))}$.

Similarly, if $\beta_t = 2\left(\mathrm{erf}^{-1}(1 - 2\eta_t)\right)^2$ and $\eta_t = \frac{\sqrt{2\pi}^\omega}{2\tau_t^{d\omega} t^\xi}$, $\omega > 1$, then Assumption 3.4 is satisfied by

$$\tau_t = \max \begin{cases} 2t^2 KLdC_1(1 + tK\varsigma^{-1}), \\ \left(\left(16t^4 K^2 LdC_1\right)\left(d\omega + \log\left(\frac{t^\xi}{2\sqrt{2\pi}^\omega}\right)\right)\right)^{\frac{1}{1-1/e}}, \\ \left(\left(16t^5 K^3 L^2 d^2 C_1^2 \varsigma^{-2}\right)\left(d\omega + \log\left(\frac{t^\xi}{2\sqrt{2\pi}^\omega}\right)\right)\right)^{\frac{1}{2-1/e}}, \\ t^2 KdC_1 C_3\left(\sqrt{\log(C_2 d)} + \frac{\sqrt{\pi}}{2}\right). \end{cases} \tag{60}$$

This is shown similarly as before but using Lemma A.14 to upper bound $\mathrm{erf}^{-1}(1 - 2\eta_t)$ in β_t.

Next, we present a lemma that bounds the discretization error of the expected reward of optimal super arm.

Lemma A.8. *Let $\mathcal{D}_t \subset \mathcal{A}$ be a finite discretization with each dimension equally divided into $\tau_t = t^2 K d C_1 C_3 \left(\sqrt{\log(C_2 d)} + \sqrt{\pi}/2\right)$ such that $|\mathcal{D}_t| = \tau_t^d$. Then,*

$$\sum_{t\in[T]} \mathbb{E}\left[f(\mathbf{a}_t^*) - f([\mathbf{a}_t^*]_{\mathcal{D}_t})\right] \leq \frac{\pi^2}{6}. \tag{61}$$

Proof.

$$\sum_{t\in[T]} \mathbb{E}[f(\mathbf{a}_t^*) - f([\mathbf{a}_t^*]_{\mathcal{D}_t})] = \sum_{t\in[T]} \mathbb{E}\left[\sum_{a\in\mathbf{a}_t^*} f(a) - f([a]_{\mathcal{D}_t})\right] \tag{62}$$

$$\leq K \sum_{t\in[T]} \mathbb{E}\left[\sup_{a\in\mathcal{A}} f(a) - f([a]_{\mathcal{D}_t})\right] \qquad (|\mathbf{a}_t^*| \leq K) \tag{63}$$

$$\leq K \sum_{t\in[T]} \frac{1}{Kt^2} \qquad \left(\begin{array}{l}\text{Lemma H.2 of Takeno et al.}\\ \text{(2023) with } u_t = Kt^2\end{array}\right) \tag{64}$$

$$\leq \frac{\pi^2}{6} \qquad \left(\sum_{t=1}^{\infty} \frac{1}{t^2} = \frac{\pi^2}{6}\right) \tag{65}$$

□

In the following lemma, we bound the discretization error of the posterior mean and standard deviation in terms of the regularity parameters, the discretization size and number of arms selected.

Lemma A.9. *Let μ_{t-1} and σ_{t-1} denote the posterior mean and standard deviation of $\mathcal{GP}(\mu, k)$ after sampling N_{t-1} base arms. If $a \in \mathcal{A}$, then*

$$\mu_{t-1}([a]_{\mathcal{D}_t}) - \mu_{t-1}(a) \leq L\frac{dC_1}{\tau_t} + L\frac{dC_1}{\tau_t}\sqrt{N_{t-1}}\varsigma^{-1}\sqrt{\|\mathbf{L}^{-1}(\mathbf{y}-\boldsymbol{\mu})\|_2^2} \tag{66}$$

and

$$\sigma_{t-1}([a]_{\mathcal{D}_t}) - \sigma_{t-1}(a) \leq \sqrt{L\frac{dC_1}{\tau_t} + N_{t-1}L^2\left(\frac{dC_1}{\tau_t}\right)^2\varsigma^{-2}} \tag{67}$$

for $L-$Lipschitz μ and k where $\mathbf{L}$ is the Cholesky decomposition of $\mathbf{K} + \varsigma^2 I$ and $\|\mathbf{L}^{-1}(\mathbf{y}-\boldsymbol{\mu})\|_2^2 \sim \chi^2$ with N_{t-1} degrees of freedom.

Proof. Consider first the difference in posterior mean:

$$\mu_{t-1}([a]_{\mathcal{D}_t}) - \mu_{t-1}(a) \tag{68}$$

$$= \mu([a]_{\mathcal{D}_t}) - \mu(a) + \left(\mathbf{k}([a]_{\mathcal{D}_t}) - \mathbf{k}(a)\right)^\top \left(\mathbf{K} + \varsigma^2 I\right)^{-1}(\mathbf{y}-\boldsymbol{\mu}) \tag{69}$$

$$\leq L \sup_{a\in\mathcal{A}} \|a - [a]_{\mathcal{D}_t}\|_1 + \left\|\left(\mathbf{k}([a]_{\mathcal{D}_t}) - \mathbf{k}(a)\right)^\top (\mathbf{K} + \varsigma^2 I)^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\|_2 \qquad (\mu \text{ } L\text{-Lipschitz}) \tag{70}$$

$$\leq L\frac{dC_1}{\tau_t} + \left\|\left(\mathbf{k}([a]_{\mathcal{D}_t}) - \mathbf{k}(a)\right)^\top (\mathbf{K} + \varsigma^2 I)^{-1}(\mathbf{y}-\boldsymbol{\mu})\right\|_2 \tag{71}$$

where the last step uses that $\sup_{a\in\mathcal{A}} \|a - [a]_{\mathcal{D}_t}\|_1 \leq \frac{dC_1}{\tau_t}$.

Next, we will appropriately split the norm into a product of norms and bound the individual factors. Let $\mathbf{K} + \varsigma^2 I = \mathbf{L}\mathbf{L}^\top$ denote the Cholesky decomposition. Note that $\mathbf{y} - \boldsymbol{\mu} \sim \mathcal{N}(0, \mathbf{K} + \varsigma^2 I)$. Then, $\mathbf{L}^{-1}(\mathbf{y} - \boldsymbol{\mu}) \sim \mathcal{N}(0, \mathbf{L}^{-1}\mathbf{L}\mathbf{L}^\top(\mathbf{L}^{-1})^\top) = \mathcal{N}(0, I)$ and thus $\|\mathbf{L}^{-1}(\mathbf{y} - \boldsymbol{\mu})\|_2$ has a chi distribution with N_{t-1} degrees of freedom.

Let $\mathrm{eig}(A)$ denote the set of eigenvalues of the square matrix A. The matrix norm of the inverted Cholesky decomposition $\mathbf{L}^{-1}$ can be bounded as:

$$\|\mathbf{L}^{-1}\|_2 = \sqrt{\max \mathrm{eig}\left((\mathbf{L}^{-1})^\top \mathbf{L}^{-1}\right)} \qquad \begin{pmatrix}\text{Eq. (538) of Petersen} \\ \text{\& Pedersen (2012)}\end{pmatrix} \tag{72}$$

$$= \sqrt{\max \mathrm{eig}\left((\mathbf{K} + \varsigma^2 I)^{-1}\right)} \tag{73}$$

$$= \sqrt{\max \frac{1}{\mathrm{eig}\left(\mathbf{K} + \varsigma^2 I\right)}} \tag{74}$$

$$= \sqrt{\max \frac{1}{\mathrm{eig}\left(\mathbf{K}\right) + \varsigma^2}} \leq \sqrt{\frac{1}{\varsigma^2}} \leq \frac{1}{\varsigma}. \qquad (\mathbf{K} \text{ p.s.d.}, \varsigma > 0) \tag{75}$$

Similarly, we also get that

$$\|(\mathbf{K} + \varsigma^2 I)^{-1}\|_2 \leq \varsigma^{-2}. \tag{76}$$

The kernel difference can be bounded as follows:

$$\|\mathbf{k}([a]_{\mathcal{D}_t}) - \mathbf{k}(a)\|_2 = \sqrt{\sum_{i=1}^{N_{t-1}} \left(k([a]_{\mathcal{D}_t}, x_i) - k(a, x_i)\right)^2} \tag{77}$$

$$\leq \sqrt{\sum_{i=1}^{N_{t-1}} L^2 \left(\frac{dC_1}{\tau_t}\right)^2} \leq L\frac{dC_1}{\tau_t}\sqrt{N_{t-1}} \tag{78}$$

where we use the fact that k is L-Lipschitz. Applying Cauchy-Schwarz and the obtained bounds, we find that

$$\mu_{t-1}([a]_{\mathcal{D}_t}) - \mu_{t-1}(a) \leq L\frac{dC_1}{\tau_t} + L\frac{dC_1}{\tau_t}\sqrt{N_{t-1}}\varsigma^{-1}\|\mathbf{L}^{-1}(\mathbf{y} - \boldsymbol{\mu})\|_2. \tag{79}$$

The posterior standard deviation is bounded similarly:

$$\sigma_{t-1}([a]_{\mathcal{D}_t}) - \sigma_{t-1}(a) \leq \sqrt{\left|\sigma^2_{t-1}([a]_{\mathcal{D}_t}) - \sigma^2_{t-1}(a)\right|}. \tag{80}$$

Continuing,

$$\left|\sigma^2_{t-1}([a]_{\mathcal{D}_t}) - \sigma^2_{t-1}(a)\right| \tag{81}$$

$$= \Big|k([a]_{\mathcal{D}_t}, [a]_{\mathcal{D}_t}) - k(a, a)$$
$$+ \left(\mathbf{k}([a]_{\mathcal{D}_t}) - \mathbf{k}(a)\right)^\top \left(\mathbf{K} + \varsigma^2 I)\right)^{-1} \left(\mathbf{k}([a]_{\mathcal{D}_t}) - \mathbf{k}(a)\right)\Big| \tag{82}$$

$$\leq \left|k([a]_{\mathcal{D}_t}, [a]_{\mathcal{D}_t}) - k(a, a)\right|$$
$$+ \left|\left(\mathbf{k}([a]_{\mathcal{D}_t}) - \mathbf{k}(a)\right)^\top \left(\mathbf{K} + \varsigma^2 I)\right)^{-1} \left(\mathbf{k}([a]_{\mathcal{D}_t}) - \mathbf{k}(a)\right)\right| \tag{83}$$

$$\leq L\frac{dC_1}{\tau_t} + \|\mathbf{k}([a]_{\mathcal{D}_t}) - \mathbf{k}(a)\|_2^2 \left\|(\mathbf{K}+\varsigma^2 I)^{-1}\right\|_2 \tag{84}$$

$$\leq L\frac{dC_1}{\tau_t} + \left(L\frac{dC_1}{\tau_t}\sqrt{N_{t-1}}\right)^2 \varsigma^{-2}. \qquad \text{(Eqs. (76) and (78))} \tag{85}$$

Combining the above, the final bound is:

$$\sigma_{t-1}([a]_{\mathcal{D}_t}) - \sigma_{t-1}(a) \leq \sqrt{L\frac{dC_1}{\tau_t} + N_{t-1}\left(L\frac{dC_2}{\tau_t}\right)^2 \varsigma^{-2}}. \tag{86}$$

□

Using Lemma A.9, we are ready to construct a constant bound for the expected discretization error of the posterior mean:

Lemma A.10. *If Assumption 3.3 holds and τ_t satisfies Eq.* (6a) *in Assumption 3.4, then for any sequence of super arms $\mathbf{a}_t \in \mathcal{S}_t$ $t \geq 1$, the posterior mean $\mu_{t-1}(\mathbf{a})$ satisfies*

$$\sum_{t\in[T]} \mathbb{E}\left[\mu_{t-1}([\mathbf{a}_t]_{\mathcal{D}_t}) - \mu_{t-1}(\mathbf{a}_t)\right] \leq \frac{\pi^2}{12}. \tag{87}$$

Proof. Note that the assumption on τ_t is equivalent to $KL\frac{dC_1}{\tau_t}(1+tK\varsigma^{-1}) \leq \frac{1}{2t^2}$. Then, we can bound the discretization error of the posterior mean as follows:

$$\sum_{t\in[T]} \mathbb{E}\left[\sum_{a\in\mathbf{a}_t} \mu_{t-1}([a]_{\mathcal{D}_t}) - \mu_{t-1}(a)\right] \tag{88}$$

$$\leq \sum_{t\in[T]} \mathbb{E}\left[K \sup_{a\in\mathcal{A}} \left[\mu_{t-1}([a]_{\mathcal{D}_t}) - \mu_{t-1}(a)\right]\right] \qquad (|\mathbf{a}_t| \leq K) \tag{89}$$

$$\leq \sum_{t\in[T]} \mathbb{E}\left[K \sup_{a\in\mathcal{A}} L\frac{dC_1}{\tau_t}\left(1 + \sqrt{tK}\varsigma^{-1}\sqrt{\|\mathbf{L}^{-1}(\mathbf{y}-\boldsymbol{\mu})\|_2^2}\right)\right]$$

$$\text{(Lemma A.9 and } N_{t-1} < tK) \tag{90}$$

$$= \sum_{t\in[T]} \mathbb{E}\left[KL\frac{dC_1}{\tau_t}\left(1 + \sqrt{tK}\varsigma^{-1}\sqrt{\|\mathbf{L}^{-1}(\mathbf{y}-\boldsymbol{\mu})\|_2^2}\right)\right]$$

$$\left(\mathbf{L}^{-1}, \mathbf{y}, \boldsymbol{\mu} \text{ independent of } a\right) \tag{91}$$

$$= \sum_{t\in[T]} KL\frac{dC_1}{\tau_t}\left(1 + \sqrt{tK}\varsigma^{-1}\mathbb{E}\left[\sqrt{\|\mathbf{L}^{-1}(\mathbf{y}-\boldsymbol{\mu})\|_2^2}\right]\right) \tag{92}$$

$$\leq \sum_{t\in[T]} KL\frac{dC_1}{\tau_t}\left(1 + \sqrt{tK}\varsigma^{-1}\sqrt{\mathbb{E}\left[\|\mathbf{L}^{-1}(\mathbf{y}-\boldsymbol{\mu})\|_2^2\right]}\right)$$

$$\text{(Concave Jensen's inequality)} \tag{93}$$

$$= \sum_{t\in[T]} KL\frac{dC_1}{\tau_t}\left(1+tK\varsigma^{-1}\right) \qquad \left(\begin{matrix}\|\mathbf{L}^{-1}(\mathbf{y}-\boldsymbol{\mu})\|_2^2 \sim \chi^2 \text{ with at most}\\ (t-1)K \text{ d.o.f.}\end{matrix}\right) \tag{94}$$

$$\leq \sum_{t\in[T]} \frac{1}{2t^2} \qquad \text{(Assumption on } \tau_t\text{)} \tag{95}$$

$$\leq \frac{\pi^2}{12}. \qquad \left(\sum_{t=1}^{\infty}\frac{1}{t^2}=\frac{\pi^2}{6}\right) \tag{96}$$

See the proof of Lemma A.9 for the motivation that $\|\mathbf{L}^{-1}(\mathbf{y}-\boldsymbol{\mu})\|_2^2 \sim \chi^2$. □

Similar to Lemma A.10, we establish a constant bound for the discretization error of the posterior standard deviation:

Lemma A.11. *If Assumption 3.3 holds; τ_t and β_t satisfy Eqs.* (6b) *and* (6c) *in Assumption 3.4 then, for any sequence of super arms $\mathbf{a}_t \in \mathcal{S}_t$ $t \geq 1$, the posterior standard deviation $\sigma_{t-1}(\mathbf{a})$ satisfies*

$$\sum_{t\in[T]} \mathbb{E}\left[\sqrt{\beta_t}\left(\sigma_{t-1}([\mathbf{a}_t]_{\mathcal{D}_t}) - \sigma_{t-1}(\mathbf{a}_t)\right)\right] \leq \frac{\pi^2}{12}. \tag{97}$$

Proof. Note that Eqs. (6b) and (6c) are equivalent to

$$\beta_t K^2 L\frac{dC_1}{\tau_t} \leq \frac{1}{8t^4} \text{ and } \beta_t t K^3 L^2 \frac{d^2C_1^2}{\tau_t^2}\varsigma^{-2} \leq \frac{1}{8t^4}. \tag{98}$$

Then,

$$\sum_{t\in[T]} \mathbb{E}\left[\sqrt{\beta_t}\left(\sigma_{t-1}([\mathbf{a}]_{\mathcal{D}_t}) - \sigma_{t-1}(\mathbf{a})\right)\right] \tag{99}$$

$$= \sum_{t\in[T]} \mathbb{E}\left[\sum_{a\in\mathbf{a}} \sqrt{\beta_t}\left(\sigma_{t-1}([a]_{\mathcal{D}_t}) - \sigma_{t-1}(a)\right)\right] \tag{100}$$

$$\leq \sum_{t\in[T]} \mathbb{E}\left[\sum_{a\in\mathbf{a}} \sqrt{\beta_t}\sqrt{L\frac{dC_1}{\tau_t} + tKL^2\frac{d^2C_1^2}{\tau_t^2}\varsigma^{-2}}\right] \qquad \text{(Lemma A.9)} \tag{101}$$

$$\leq \sum_{t\in[T]} K\sqrt{\beta_t}\sqrt{L\frac{dC_1}{\tau_t} + tKL^2\frac{d^2C_1^2}{\tau_t^2}\varsigma^{-2}} \qquad (|\mathbf{a}| \leq K) \tag{102}$$

$$= \sum_{t\in[T]} \sqrt{\beta_t K^2 L\frac{dC_1}{\tau_t} + \beta_t t K^3 L^2\frac{d^2C_1^2}{\tau_t^2}\varsigma^{-2}} \tag{103}$$

$$\leq \sum_{t\in[T]} \sqrt{\frac{1}{8t^4}+\frac{1}{8t^4}} \qquad \text{(Eq. (98))} \tag{104}$$

$$\leq \sum_{t\in[T]} \frac{1}{2t^2} \leq \frac{\pi^2}{12}. \qquad \left(\sum_{t=1}^{\infty}\frac{1}{t^2}=\frac{\pi^2}{6}\right) \tag{105}$$

□

Lemma 3.5. *If $U_t(\mathbf{a}) = \mu_{t-1}(\mathbf{a}) + \sqrt{\beta_t}\sigma_{t-1}(\mathbf{a})$, Assumption 3.3 holds and τ_t and β_t satisfy Eqs.* (6a) *to* (6c) *in Assumption 3.4, then for any sequence of super arms $\mathbf{a}_t \in \mathcal{S}_t$ $t \geq 1$:*

$$\sum_{t\in[T]} \mathbb{E}\left[U_t([\mathbf{a}_t]_{\mathcal{D}_t}) - U_t(\mathbf{a}_t)\right] \leq \frac{\pi^2}{6}. \tag{7}$$

Proof. Follows by combining Lemmas A.10 and A.11. □

Finally, we are ready to prove the regret bounds for the infinite case:

Theorem 3.6 (Infinite regret bounds)**.** *If Assumption 3.3 holds and τ_t satisfies Assumption 3.4, then the Bayesian regret of*

(i) GP-UCB with $\beta_t = 2\log(\tau_t^d t^2/\sqrt{2\pi})$ is bounded as $BR(T) \leq \frac{\pi^2}{2} + \sqrt{C_K TK\beta_T\gamma_{TK}}$.

(ii) GP-BUCB with $\beta_t = 2\left(\mathrm{erf}^{-1}(1-2\eta_t)\right)^2$ for $\eta_t = (2\pi)^{\omega/2}/\left(2\tau_t^{d\omega}t^{\xi}\right)$, $\xi > \omega > 1$ is bounded as $BR(T) \leq \frac{\pi^2}{3} + \sqrt{C_K TK\beta_T\gamma_{TK}} + C_\omega \cdot \frac{\xi}{\xi-\omega}$ where $C_\omega = \left(\sqrt{\pi}\omega/\sqrt{2e(\omega-1)}\right)^{1/\omega}$.

(iii) GP-TS is bounded as $BR(T) \leq \frac{2\pi^2}{3} + 2\sqrt{C_K TK\beta_T\gamma_{TK}}$.

Proof. (i) Similar to Takeno et al. (2023); Srinivas et al. (2012), we use a fixed discretization $\mathcal{D}_t \subset \mathcal{A}$ for $t \geq 1$. Let $\mathcal{D}_t \subset \mathcal{A}$ be a finite set with $|\mathcal{D}_t| = \tau_t^d$ and each dimension equally divided into τ_t points with τ_t satisfying Assumption 3.3. Let $[a]_{\mathcal{D}_t}$ denote the nearest point in $\mathcal{D}_t$ for $a \in \mathcal{A}$ and similarly let $[\mathbf{a}]_{\mathcal{D}_t} = \{[a]_{\mathcal{D}_t} | a \in \mathbf{a}\}$ for $\mathbf{a} \subset \mathcal{A}$.

As Takeno et al. (2023), we decompose the Bayesian regret into several parts:

$$\mathrm{BR}(T) = \sum_{t\in[T]} \mathbb{E}\big[\underbrace{f(\mathbf{a}_t^*) - f([\mathbf{a}_t^*]_{\mathcal{D}_t})}_{(1)} + \underbrace{f([\mathbf{a}_t^*]_{\mathcal{D}_t}) - U_t([\mathbf{a}_t^*]_{\mathcal{D}_t})}_{(2)} \tag{106}$$

$$+ \underbrace{U_t([\mathbf{a}_t^*]_{\mathcal{D}_t}) - U_t(\mathbf{a}_t^*)}_{(3)} + \underbrace{U_t(\mathbf{a}_t^*) - U_t(\mathbf{a}_t)}_{(4)} \tag{107}$$

$$+ \underbrace{U_t(\mathbf{a}_t) - f(\mathbf{a}_t)}_{(5)}\big] \tag{108}$$

Term (1) can be bounded using Lemma A.8: $\sum_{t\in[T]} \mathbb{E}[f(\mathbf{a}_t^*) - f([\mathbf{a}_t^*]_{\mathcal{D}_t})] \leq \frac{\pi^2}{6}$. Terms (2) and (5) can be bounded using the finite case with $\beta_t = 2\log(|\mathcal{D}_t|t^2/\sqrt{2\pi})$. Then, by Lemmas A.2 to A.4

$$\sum_{t\in[T]} \mathbb{E}[f([\mathbf{a}_t^*]_{\mathcal{D}_t}) - U_t([\mathbf{a}_t^*]_{\mathcal{D}_t}) + U_t(\mathbf{a}_t) - f(\mathbf{a}_t)]$$

$$\leq \frac{\pi^2}{6} + \sqrt{2(\lambda_K^* + \varsigma^2)TK\beta_T\gamma_{TK}}. \tag{109}$$

Takeno et al. (2023) consider the term $U_t([\mathbf{a}_t^*]_{\mathcal{D}_t}) - U_t(\mathbf{a}_t^*)$ and argue that it is non-positive since $\mathbf{a}_t = \arg\max_{\mathbf{a}\in\mathcal{S}_t} U_t(\mathbf{a})$. Unlike Takeno et al., we do not assume that all arms are available at time t and thus $[\mathbf{a}_t^*]_{\mathcal{D}_t} \in \mathcal{S}_t$ does not necessarily hold. By further decomposing this term into (3) and (4), the same argument can be applied to term (4): $U_t(\mathbf{a}_t^*) - U_t(\mathbf{a}_t) \leq 0$. Then, term (3) can be bounded using Lemma 3.5: $\sum_{t\in[T]} \mathbb{E}[U_t([\mathbf{a}_t^*]_{\mathcal{D}_t}) - U_t(\mathbf{a}_t^*)] \leq \pi^2/6$.

Finally, by combining the bounds for all terms we get that

$$\mathrm{BR}(T) \leq \frac{\pi^2}{2} + \sqrt{C_K T K \beta_T \gamma_{TK}}. \tag{110}$$

(ii) The proof for GP-BUCB is shown by following the steps of GP-UCB and using the finite case for Bayes-GP-UCB (Theorem 3.2 (ii)).

(iii) As in the proof for GP-UCB, assume that we have a discretization $\mathcal{D}_t$ and decompose the Bayesian regret into 4 terms:

$$\mathrm{BR}(T) = \sum_{t\in[T]} \mathbb{E}\big[\underbrace{f(\mathbf{a}_t^*) - f([\mathbf{a}_t^*]_{\mathcal{D}_t})}_{(1)} + \underbrace{f([\mathbf{a}_t^*]_{\mathcal{D}_t}) - U_t([\mathbf{a}_t^*]_{\mathcal{D}_t})}_{(2)} \tag{111}$$

$$+ \underbrace{U_t([\mathbf{a}_t^*]_{\mathcal{D}_t}) - U_t(\mathbf{a}_t)}_{(3)} + \underbrace{U_t(\mathbf{a}_t) - f(\mathbf{a}_t)}_{(4)} \big]. \tag{112}$$

As in the proof for GP-UCB, term (1) is dealt with using Lemma A.8 and term (2) and (4) are handled as in the finite case (Theorem 3.2 (iii)):

$$\sum_{t\in[T]} \mathbb{E}[(1) + (2) + (3)] \leq \frac{\pi^2}{6} + \frac{\pi^2}{3} + 2\sqrt{C_K T K \beta_T \gamma_{TK}}. \tag{113}$$

To bound term (3), we start by utilizing that $\mathbf{a}_t^* | H_t \overset{d}{=} \mathbf{a}_t | H_t$ and $U_t([\cdot]_{\mathcal{D}_t}) | H_t$ is deterministic and thus:

$$\sum_{t\in[T]} \mathbb{E}[(3)] = \sum_{t\in[T]} \mathbb{E}_{H_t}\left[\mathbb{E}_t\left[U_t([\mathbf{a}_t^*]_{\mathcal{D}_t}) - U_t(\mathbf{a}_t) | H_t\right]\right] \tag{114}$$

$$= \sum_{t\in[T]} \mathbb{E}_{H_t}\left[\mathbb{E}_t\left[U_t([\mathbf{a}_t]_{\mathcal{D}_t}) - U_t(\mathbf{a}_t) | H_t\right]\right] \tag{115}$$

$$\leq \frac{\pi^2}{6} \qquad \text{(Lemma 3.5)} \tag{116}$$

Put together, we have that

$$\mathrm{BR}(T) \leq \frac{2\pi^2}{3} + 2\sqrt{C_K T K \beta_T \gamma_{TK}}. \tag{117}$$

□

A.3 Additional lemmas

Lemma A.12. *For any sequence of superarms* $\mathbf{a}_1, \ldots, \mathbf{a}_T$*,*

$$\sum_{t=1}^{T} \sigma_{t-1}^2(\mathbf{a}_t) \leq 2(\lambda_K^* + \varsigma^2)\gamma_{TK}. \tag{118}$$

where λ_K^* *is the largest eigenvalue of all possible posterior covariance matrices of size at most* K*.*

Proof. This proof follows the proof of Lemma 3 of (Nika et al., 2022). Let $K_t = |\mathbf{a}_t|$ denote the number of base arms selected at time t. Similarly, let $N_T = \sum_{t\in[T]} K_t$ denote the number of base arms selected *up to* time T. Note that the information gain can be decomposed into two entropy terms: $I(\mathbf{r}_{[T]}; f) = H(\mathbf{r}_{[T]}) - H(\mathbf{r}_{[T]}|f)$.

Since $\mathbf{r}_{[T]}|\mathbf{f}_{[T]} \sim \mathcal{N}(\mathbf{f}_{[T]}, \varsigma^2 I_{K_t}$, $H(\mathbf{r}_{[T]}|\mathbf{f}_{[T]}) = \frac{1}{2}\log|2\pi e\varsigma^2 I_{N_T}|$. The first term can be analyzed by using the chain rule of entropy on the superarms:

$$\begin{aligned} H(\mathbf{r}_{[T]}) &= H(\mathbf{r}_T|\mathbf{r}_{[T-1]}) + H(\mathbf{r}_{[T-1]}) && (119)\\ &= \sum_{t=1}^{T} H(\mathbf{r}_t|\mathbf{r}_{[t-1]}). && (120) \end{aligned}$$

Then, $\mathbf{r}_t|\mathbf{r}_{[t-1]} \sim \mathcal{N}(\boldsymbol{\mu}_{t-1}, \boldsymbol{\Sigma}_{t-1} + \varsigma^2 I_{K_t})$ where $\boldsymbol{\mu}_{t-1} = [\mu_{t-1}(a)]_{a\in\mathbf{a}_t}$ is the posterior mean vector and $\boldsymbol{\Sigma}_{t-1} = (k_{t-1}(a,a'))_{a,a'\in\mathbf{a}_t\times\mathbf{a}_t}$ is the posterior covariance matrix for superarm $\mathbf{a}_t$ after observing $(\mathbf{a}_1, \boldsymbol{r}_1), \ldots, (\mathbf{a}_{t-1}, \boldsymbol{r}_{t-1})$. Let $\lambda_{t,k}$ denote the smallest kth eigenvalue of $\boldsymbol{\Sigma}_{t-1}$. Then,

$$\begin{aligned} H(\mathbf{r}_t|\mathbf{r}_{[t-1]}) &= \frac{1}{2}\log\left|2\pi e(\boldsymbol{\Sigma}_{t-1} + \varsigma^2 I_{K_t})\right| && (121)\\ &= \frac{1}{2}\log\left|2\pi e\varsigma^2(\varsigma^{-2}\boldsymbol{\Sigma}_{t-1} + I_{K_t})\right| && (122)\\ &= \frac{1}{2}\log\left|2\pi e\varsigma^2 I_{K_t}\right| + \frac{1}{2}\log\left|\varsigma^{-2}\boldsymbol{\Sigma}_{t-1} + I_{K_t}\right|. && (123) \end{aligned}$$

Let $\lambda_{t,k}$ denote the smallest kth eigenvalue of $\boldsymbol{\Sigma}_{t-1}$. Let $\mathcal{M} = \{\boldsymbol{\Sigma}_{t-1} | \forall t \in [T], \forall \mathbf{a}_1, \ldots, \mathbf{a}_t \in \mathcal{S}\}$ be the set of all possible posterior covariance matrices and let $\lambda_K^* = \sup_{\boldsymbol{\Sigma}\in\mathcal{M}} \max \mathrm{eig}(\boldsymbol{\Sigma})$ be the largest eigenvalue of all eigenvalues of the matrices in $\mathcal{M}$. Recall that $|A + I_n| = \prod_{k\leq n}(\lambda_k + 1)$ for any real and

symmetric matrix $A \in \mathbb{R}^{n\times n}$ with eigenvalues $\lambda_1, ..., \lambda_n$. Then,

$$\frac{1}{2}\log\left|\varsigma^{-2}\mathbf{\Sigma}_{t-1} + I_{K_t}\right| \tag{124}$$

$$= \frac{1}{2}\log\left(\prod_{k=1}^{K_t}\left(\varsigma^{-2}\lambda_{t,k}+1\right)\right) \tag{125}$$

$$= \frac{1}{2}\sum_{k=1}^{K_t}\log\left(\varsigma^{-2}\lambda_{t,k}+1\right) \tag{126}$$

$$\geq \frac{1}{2}\sum_{k=1}^{K_t}\frac{\varsigma^{-2}\lambda_{t,k}}{\varsigma^{-2}\lambda_{t,k}+1} \qquad (\log(x+1)\geq x/(x+1), \forall x>1) \tag{127}$$

$$\geq \frac{\varsigma^{-2}}{2(\varsigma^{-2}\lambda^*+1)}\sum_{k=1}^{K_t}\lambda_{t,k} \tag{128}$$

$$= \frac{\varsigma^{-2}}{2(\varsigma^{-2}\lambda^*+1)}\sum_{a\in\mathbf{a}_t}\sigma_{t-1}^2(a). \qquad \left(\mathrm{Tr}(A)=\sum_{\lambda\in\mathrm{eig}(A)}\lambda\right) \tag{129}$$

Put together, we get that $\sum_{t=1}^{T}\sigma_{t-1}^2(\mathbf{a}_t) \leq 2(\lambda^*+\varsigma^2)I(\mathbf{r}_{[T]};f)$. Since the maximum information γ_T is increasing w.r.t. T and $|\mathbf{a}_t|\leq K$, we get that $\sum_{t=1}^{T}\sigma_{t-1}^2(\mathbf{a}_t) \leq 2(\lambda^*+\varsigma^2)\gamma_{TK}$. □

Lemma A.13. *The inverse error function is lower bounded by*

$$\mathrm{erf}^{-1}(u) \geq \sqrt{-\omega^{-1}\log\left(\frac{1-u}{\vartheta}\right)} \tag{130}$$

for $u\in[0,1)$, $\omega>1$ *and* $0<\vartheta\leq\sqrt{\frac{2e}{\pi}}\frac{\sqrt{\omega-1}}{\omega}$.

Proof. According to Theorem 2 of Chang et al. (2011), $\mathrm{erfc}(u)\geq\vartheta\exp(-\omega u^2)$ for $\omega>1$ and $0<\vartheta\leq\sqrt{\frac{2e}{\pi}}\frac{\sqrt{\omega-1}}{\omega}$. Since $\mathrm{erf}(u)=1-\mathrm{erfc}(u)$, it follows that $\mathrm{erf}(u)\leq 1-\vartheta\exp(-\omega u^2) =: h(u)$.

In general, if $f(x)\leq g(x)$ then $f^{-1}(x)\geq g^{-1}(x)$. Thus, $\mathrm{erf}^{-1}(u)\geq h^{-1}(u)=\sqrt{-\omega^{-1}\log\left((1-u)/\vartheta\right)}$. □

Lemma A.14. *The inverse error function is upper bounded by*

$$\mathrm{erf}^{-1}(u) \leq \sqrt{-\omega^{-1}\log\left(\frac{1-u}{\vartheta}\right)} \tag{131}$$

for $u\in[0,1)$, $\vartheta\geq 1$ *and* $0<\omega\leq 1$.

Proof. The same arguments as in Lemma A.13 but using Theorem 1 of Chang et al. (2011). □

B Additional experimental details

B.1 Kernel details

Here, we provide further details on the graph kernel used in the experiments. The original graph Matérn GP of Borovitskiy et al. (2021) defines a GP on the vertices of a weighted and undirected graph. We extend the graph Matérn GP from Borovitskiy et al. (2021) to the edges of a directed graph by considering the incidence graph Laplacian of the line graph $\mathcal{L}(\mathcal{G})$.

Let $\mathbf{W}_{\mathcal{L}} \in \mathbb{R}^{|\mathcal{E}|\times|\mathcal{C}|}$ denote the weight matrix of $\mathcal{L}(\mathcal{G}) = (\mathcal{E}, \mathcal{C})$ where $\mathcal{E}$ is the set of edges and $\mathcal{C}$ is the set of all connections in the network. The weight $W_{\mathcal{L},e_1,e_2}$ is set to $\bar{\ell}/\ell_{e_1}$ where $\bar{\ell}$ is the average length of all edges and ℓ_{e_1} is the length of edge e_1. We replace the ordinary graph Laplacian used by Borovitskiy et al. (2021) with the incidence Laplacian:

$$\boldsymbol{\Delta}_I = \mathbf{B}\mathbf{B}^\top, \tag{132}$$

where the incidence matrix $\mathbf{B} \in \mathbb{R}^{|\mathcal{E}|\times|\mathcal{C}|}$ has entries

$$B_{e,c} = \begin{cases} -W_{\mathcal{L},e_1,e_2} & \text{if } e = e_1, \\ W_{\mathcal{L},e_1,e_2} & \text{if } e = e_2, \\ 0 & \text{otherwise} \end{cases} \quad \forall e \in \mathcal{E}, c = (e_1, e_2) \in \mathcal{C}. \tag{133}$$

Let $\boldsymbol{\Delta}_I = \mathbf{U}_I \boldsymbol{\Lambda}_I \mathbf{U}_I^\top$ denote the eigendecomposition of $\boldsymbol{\Delta}_I$, then the graph Matérn GP of the edges is given by

$$\mathbf{f} \sim \mathcal{N}\left(0, \mathbf{U}_I \left(\frac{2\nu_G}{\kappa_G^2}\mathbf{I} + \boldsymbol{\Lambda}_I\right)^{-\nu} \mathbf{U}_I^\top\right). \tag{134}$$

Recall that $k_f : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ denotes a feature kernel which measures the similarity between the contexts of the edges. The feature kernel is an ordinary Matérn kernel with fixed $\nu = 5/2$ but tunable outputscale σ_f and lengthscales $\boldsymbol{\ell}_f \in \mathbb{R}_+^d$ for each dimension:

$$k_f(x_e, x_{e'}) := \sigma_f \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} D\right)^\nu K_\nu \left(\sqrt{2\nu} D\right), \tag{135}$$

where x_e denotes the feature vector of edge e and the feature distance D between edge e and e' is given by

$$D = \sqrt{(x_e - x_{e'})^\top \operatorname{diag}(\boldsymbol{\ell}_f)^{-2} (x_e - x_{e'})}. \tag{136}$$

The kernels, the SVGP model and Algorithm 3 was implemented using GPyTorch (Gardner et al., 2018).

B.2 Road network

The set of available paths was restricted to edges within the largest strongly connected component. This mainly removed road segments in inaccessible

Table 2: Vehicle and environmental parameters for the energy model.

Variable	Value	Unit
Mass m	1830	kg
Rolling resistance coefficient C_r	0.01	
Front surface area A	2.6	m^2
Air drag coefficient C_d	0.35	
Power train efficiency η^+	0.98	
Recuperation efficiency η^-	0.96	
Gravitational acceleration g	9.82	m/s^2
Air density ρ	1.2	kg/m^3

areas and does not affect the navigational challenge. The route Luxembourg A starts in edge `-31118#2` and ends in edge `--32646#1`. The route Luxembourg B starts in edge `-30436#5` and ends in edge `-30946#0`. Similarly, the route Monaco A starts in edge `-30558` and ends in edge `-32888#0` whilst Monaco B starts in edge `-32166#0` and ends in edges `-32940#0`. For simplicity, the start and end points are edges since the shortest path was computed using the line graph $\mathcal{L}(\mathcal{G})$.

B.3 Detailed parameter values

In this section, we further specify the vehicle, environmental and algorithmic parameters used. We use the default parameters for electric vehicles provided by SUMO (Lopez et al., 2018), see Table 2.

The graph kernel is initialized with parameters $\nu_G = 2$, $\kappa_G = 1$ and σ_G set according to the prior. The natural gradient descent learning rate is set to 0.1 whilst the Adam learning rate is set to 0.01. The GP model uses a batch size B of 2500 and 1 gradient step per optimization procedure. The number of inducing points is set to 1000.

C Additional experimental results

C.1 Impact of lengthscale

In this section, we provide the full results for the lengthscale experiments in Section 4.2. The cumulative regret over time is visualized in Fig. 4 and the final cumulative regret as a function on the lengthscale ℓ is visualized in Fig. 5.

C.2 Visualization of exploration

In this section, we provide visualization of the routes selected by the algorithms. See Figs. 6 to 9 for visualization on Lux. A, Lux B, Mon. A and Mon. B, respectively. According to the results, the TS variants are able to find

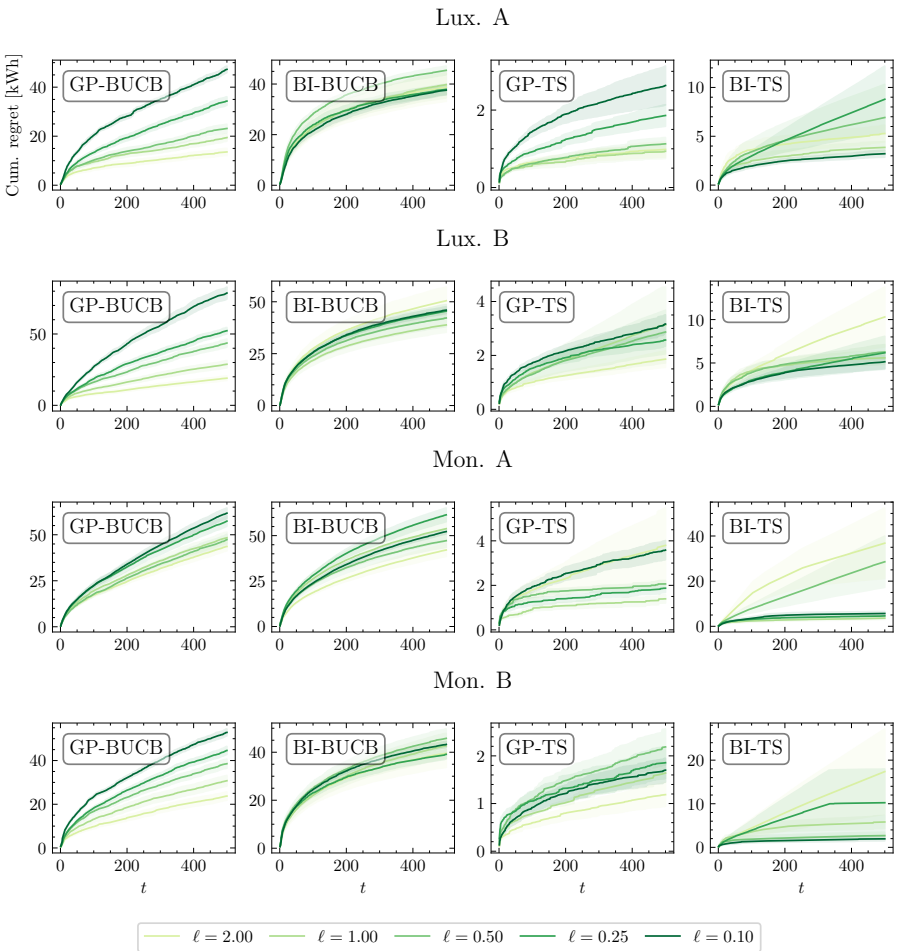


Figure 4: Cumulative regret of GP-BUCB, BI-BUCB, GP-TS and BI-TS for varying prior lengthscale values ℓ.

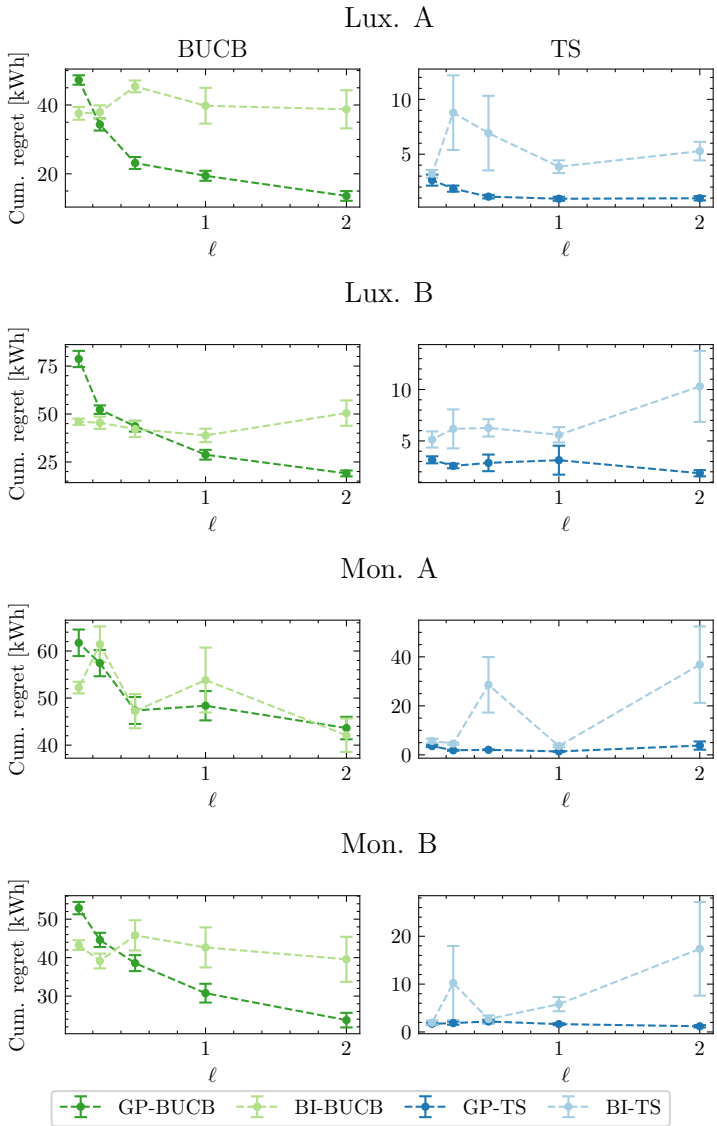
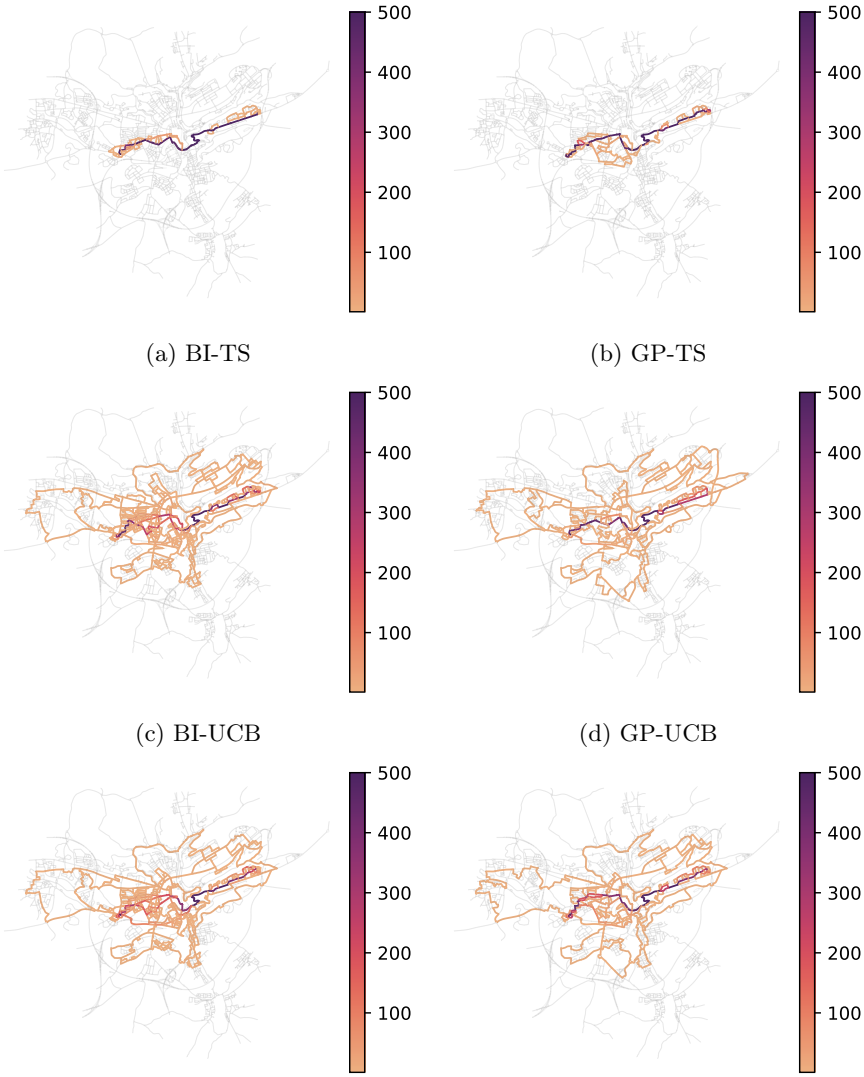


Figure 5: Cumulative regret at $t = 500$ for varying prior lengthscale values. Errorbars correspond to ± 1 standard error.

sophisticated paths with significantly less exploration compared to BUCB and UCB. This observation implies the sample efficiency of TS methods.



(a) BI-TS

(b) GP-TS

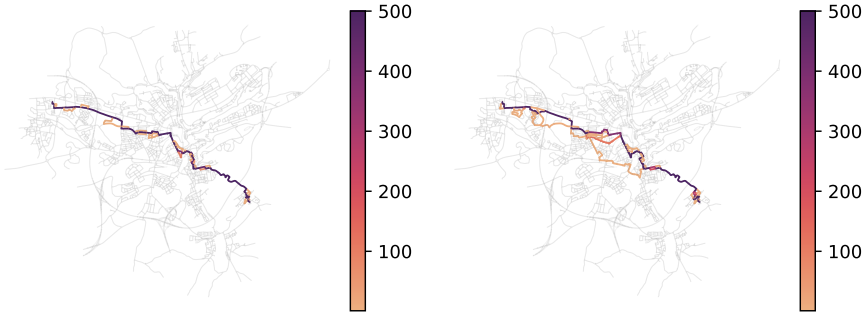
(c) BI-UCB

(d) GP-UCB

(e) BI-BUCB

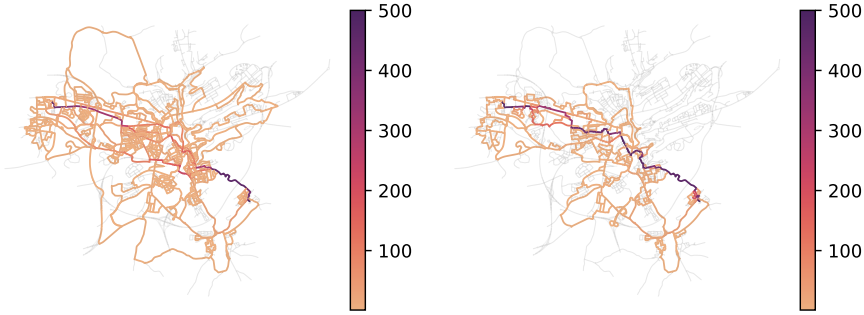
(f) GP-BUCB

Figure 6: Exploration of Luxembourg A.



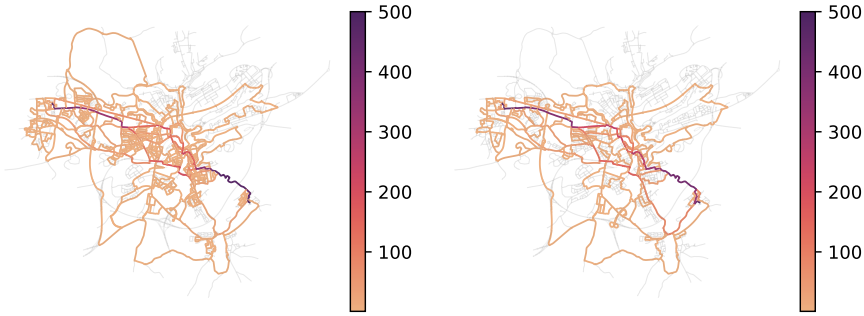
(a) BI-TS

(b) GP-TS



(c) BI-UCB

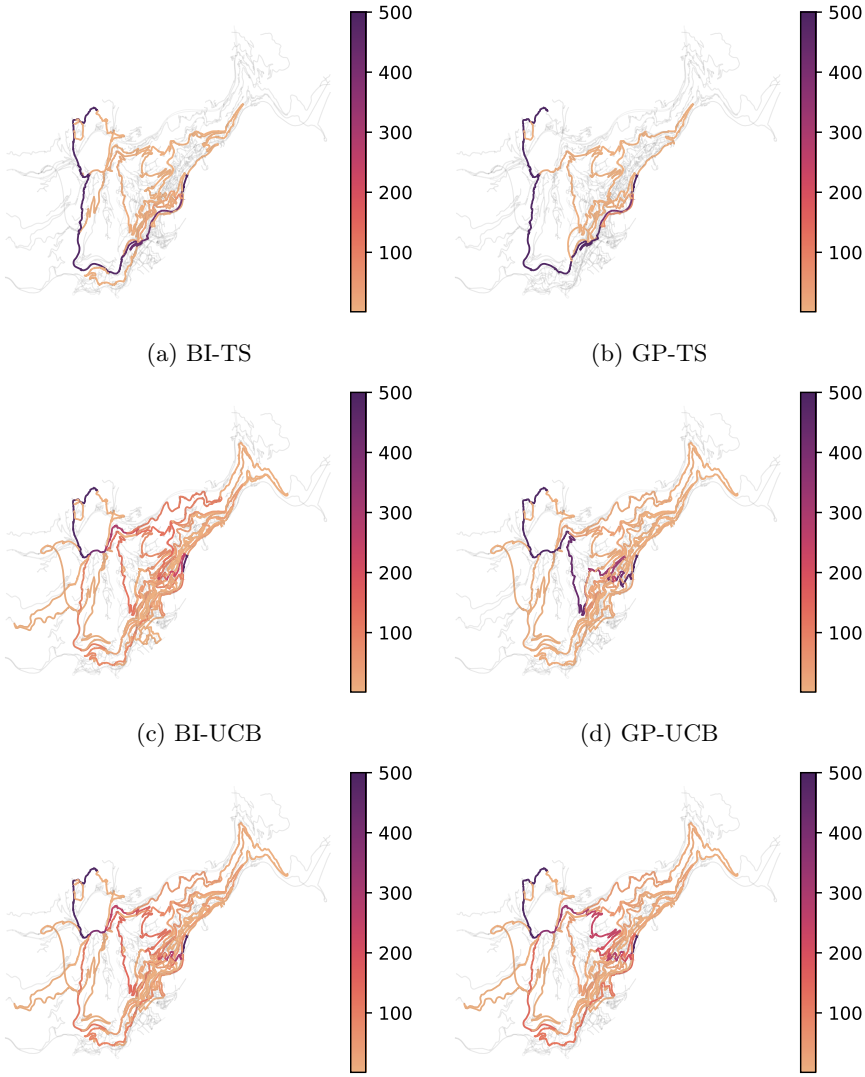
(d) GP-UCB



(e) BI-BUCB

(f) GP-BUCB

Figure 7: Exploration of Luxembourg B.



(a) BI-TS

(b) GP-TS

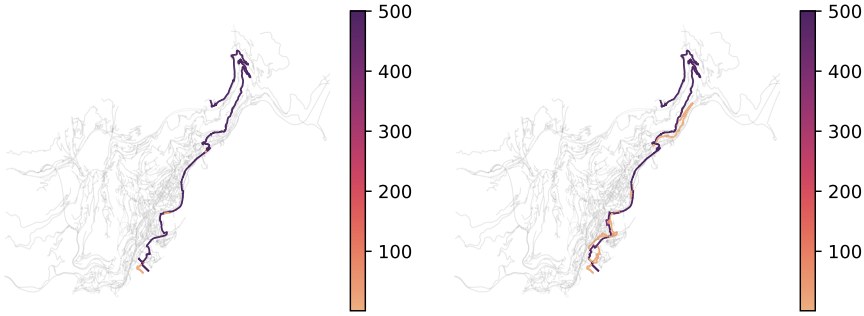
(c) BI-UCB

(d) GP-UCB

(e) BI-BUCB

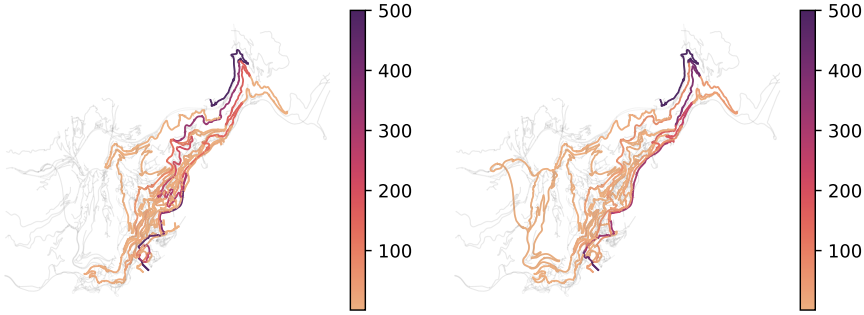
(f) GP-BUCB

Figure 8: Exploration of Monaco A.



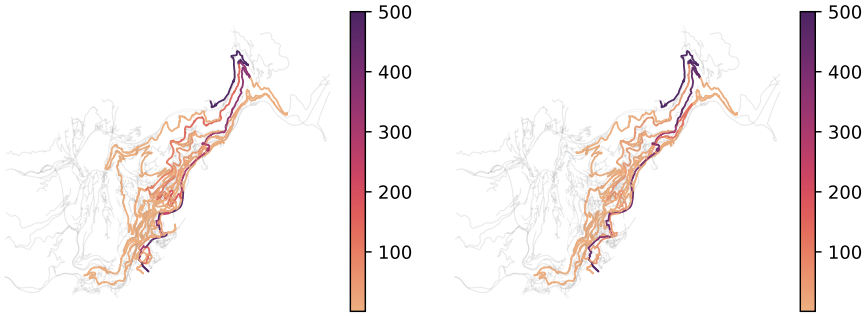
(a) BI-TS

(b) GP-TS



(c) BI-UCB

(d) GP-UCB



(e) BI-BUCB

(f) GP-BUCB

Figure 9: Exploration of Monaco B.

Paper II

Comments on "Surrogate Modeling for Bayesian Optimization Beyond a Single Gaussian Process"

J. Sandberg, M. Haghir Chehreghani

Submitted, under review

The paper has been reformatted for uniformity.

Abstract

Lu et al. (2023) proposed Ensemble Gaussian Process Thompson Sampling (EGP-TS) for Bayesian Optimization and derived bounds on the Bayesian regret in three settings: a standard sequential setting, an asynchronously parallel setting, and a synchronously parallel setting. In this paper, we show that the proof for the standard sequential setting does not hold due to multiple issues. Consequently, the proofs for the parallel settings are also invalidated.

1 Introduction

Bayesian optimization (BO) provides a flexible framework for optimizing unknown functions where (potentially noisy) observations of the function are costly to obtain Garnett (2023). In BO, the unknown function being optimized is implicitly or explicitly assumed to be sampled from a Gaussian process with a specific mean and kernel function. The choice of kernel function further imposes assumptions on the properties of the unknown function. E.g. the smoothness parameter of the Matérn kernel Matérn (1986) determines how many times the unknown function is continuously differentiable Da Costa et al. (2025) which can significantly impact the difficulty of the optimization problem.

Theoretical analysis of Bayesian optimization algorithms commonly assume that the mean and kernel function are known. In practice, parameters of the kernel function are often estimated using maximum likelihood allowing the kernel to adapt to the collected data. However, such adaptive methods lack theoretical performance guarantees. Lu et al. (2023) proposed Ensemble GP-TS (EGP-TS), a BO algorithm that adaptively selects probable kernel functions, and provided analytical performance guarantees for EGP-TS. EGP-TS combines Thompson sampling with a GP mixture prior. In each iteration, EGP-TS first samples a kernel according to its posterior probability of generating the data and then selects a location to query the unknown function based on the posterior probability of it being the location of the optima given the sampled kernel.

Lu et al. (2023) analyzed the Bayesian regret of EGP-TS and presented three theorems. Theorem 1 states that EGP-TS has sublinear regret in the standard setting without parallelism whereas Theorem 2 and 3 extend this result to the asynchronously and synchronously parallel setting respectively. The proofs rely on the technique of prior confidence sets introduced by Hong et al. (2022). In Hong et al. (2022), the authors introduced MixTS, a bi-level Thompson sampling algorithm for bandit problems with mixture priors for which EGP-TS can be seen as an instantiation of. Recently, Sandberg and Haghir Chehreghani (2026) identified issues with the proof of the regret bound for MixTS in the linear setting Hong et al. (2022).

In this paper, we highlight four issues with the proof of Theorem 1 and compare them with the issues of Hong et al. (2022). The first issue appears to be a typo. However, the three remaining issues are related to the novel part of the proof that bounds $\mathcal{BR}_2(T)$ (term A_4 in the proof in the appendix) using the techniques of Hong et al. (2022). The issues of Lu et al. (2023) and Hong et al. (2022) arise at mostly the same parts of the proofs but with some differences due to variations in the proof techniques. Consequently, the results of all three theorems appear to be invalid and a sublinear regret bound of EGP-TS has not been established.

2 Setup and notation

We briefly introduce the setup and necessary notation from Lu et al. (2023). We consider the optimization problem $\mathbf{x}_* = \arg\max_{\mathbf{x}\in\mathbf{X}} f(\mathbf{x})$ in a setting where f is sequentially queried at selected inputs $\mathbf{x}_t$ with noisy observations

$y_t = f(\mathbf{x}_t) + \epsilon_t$, $\epsilon_t \sim N(0, \sigma_n^2)$ for $t = 1, \ldots, T$. The prior model m_* is sampled from a categorical distribution $m_* \sim \mathcal{C}AT(\mathcal{M}, \boldsymbol{w}_0)$ of size $M = |\mathcal{M}|$ and with prior weights $\boldsymbol{w}_0$. The function f is sampled from a Gaussian process with a kernel function determined by m_*: $f \sim \mathcal{GP}(0, \kappa^{m_*})$. The Bayesian cumulative regret is defined as $\mathcal{BR}(T) = \mathbb{E}\left[\sum_{t=1}^{T} f(\mathbf{x}_*) - f(\mathbf{x}_t)\right]$ where the expectation is taken w.r.t. all of the randomness in the problem, including m_*, f, $(x_i, y_i)_{i=1}^T$ and internal randomness in the algorithm.

During the optimization process, the EGP-TS algorithm first samples a model from the posterior categorical distribution, $m_t \sim \mathcal{C}AT(\mathcal{M}, \boldsymbol{w}_t)$, then samples a function from the posterior GP distribution of m_t, $\tilde{f}_t(\mathbf{x}) \sim \mathcal{GP}(\mu_{t-1}^{m_t}(\mathbf{x}), \kappa_{t-1}^{m_t}(\mathbf{x}, \mathbf{x}'))$ where $\mu_{t-1}^{m_t}(\mathbf{x})$ and $\kappa_{t-1}^{m_t}(\mathbf{x}, \mathbf{x}'))$ is the posterior mean and kernel function of model m_t, and finally queries the black-box function at the input that maximizes the sampled function $\mathbf{x}_t := \arg\max_{\mathbf{x} \in \mathbf{X}} \tilde{f}_t(\mathbf{x})$.

The confidence set $\mathcal{C}_t \subset 2^{\mathcal{M}}$, introduced by Hong et al. (2022), is a tool in the regret analysis that tracks which prior models $m \in \mathcal{M}$ have lead to observations $\{y_t : m_t = m\}$ that are not too surprising assuming that $m_* = m$. It is defined as

$$\mathcal{C}_t = \{m \in \mathcal{M} : G_t^m \leq 2\sigma_n \sqrt{N_{t-1}^m \log T}\} \tag{1}$$

where the number of draws of m is $N_{t-1}^m = \sum_{\tau=1}^{t-1} \mathbf{I}(m_\tau = m)$, and the excess reward of m is $G_t^m = \sum_{\tau=1}^{t-1} \mathbf{I}(m_\tau = m)(L_\tau^m(\mathbf{x}_\tau) - y_\tau)$. The lower confidence bound for model m is $L_t^m(\mathbf{x}) = \mu_{t-1}^m(\mathbf{x}) - \eta\sigma_{t-1}^m(\mathbf{x})$ with $\eta = 2\sqrt{\log T}$.

3 Issues in the proofs of Lu et al. (2023)

In this section, we present the identified issues in the regret bounds of EGP-TS in the standard BO setting and discuss how they affect the parallel setting.

3.1 Lemma 3 statement

In Lemma 3, it is stated that $\max_{\mathbf{x} \in \mathbf{X}} |f(\mathbf{x}_*) - f(\mathbf{x})| \leq 2B$ where $B = \mathbb{E}[\sup_{\mathbf{x} \in \mathbf{X}} |f(x)|]$ for $f \sim \mathcal{GP}(0, \kappa^m)$. However, since $f(\mathbf{x})$ has a Gaussian distribution it has support on $\mathbb{R}$ $\forall \mathbf{x} \in \mathbf{X}$, and $\max_{\mathbf{x} \in \mathbf{X}} |f(\mathbf{x}_*) - f(\mathbf{x})|$ cannot be bounded by a constant w.p. 1 and thus the statement of the lemma is false. The lemma would be correct if stated with an expectation around $\max_{\mathbf{x} \in \mathbf{X}} |f(\mathbf{x}_*) - f(\mathbf{x})|$.

3.2 Proof of Lemma 5

Lemma 5 claims that $\mathbb{E}[\mu_{t-1}^{m_t}(\mathbf{x}_t) - f(\mathbf{x}_t)] \leq 2B$. In the proof of Lemma 5, the identity $\mathbb{E}_{t-1}[\mu_{t-1}^{m_*}(\mathbf{x}_*)] = \mathbb{E}_{t-1}[f(\mathbf{x}_*)]$ is used without motivation – which if incorrect would invalidate the lemma. Lemma 5 is used to bound term $A_{4,3}$ which is the next issue we will discuss. The function $\mu_{t-1}^{m_*}(\cdot)$ is the posterior mean of a Gaussian process with prior model m^* whereas $f(\mathbf{x}_*) = \sup_{\mathbf{x} \in \mathbf{X}} f(\mathbf{x})$. For any fixed m^* and $\mathbf{x}$, note that $f(\mathbf{x})|\mathbf{x}_* = \mathbf{x}$ has a skew-Gaussian distribution

Arellano-Valle and Azzalini (2022), not a Gaussian distribution, and therefore $f(\mathbf{x}_*)$ does not have the mean $\mathbb{E}[\mu_{t-1}^{m_*}(\mathbf{x}_*)]$ unless the skewness parameter equals 0. We illustrate this with a counterexample at time $t = 1$, a known prior (i.e. $M = 1$), and two input locations ($|\mathbf{X}| = 2$). Then, $\mathbb{E}_{t-1}[\mu_{t-1}^{m_*}(\mathbf{x}_*)]$ corresponds to the known prior mean of the optimal input $\mathbf{x}_*$ and $\mathbb{E}_{t-1}[f(\mathbf{x}_*)]$ is the prior mean of $\max_{\mathbf{x}\in\mathbf{X}} f(\mathbf{x})$.

Consider the bivariate Gaussian distribution $f \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$ for $\rho \in (-1, 1)$. Let f_1 and f_2 denote the first and second entry of f. By definition, $\mathbb{E}[f_1] = \mathbb{E}[f_2] = 0$ and therefore $\mathbb{E}_{t-1}[\mu_{t-1}^{m^*}(\mathbf{x}^*)] = \mathbb{E}_{t-1}[0] = 0$. However, we will show that $\mathbb{E}_{t-1}[f(\mathbf{x}_*)] = \mathbb{E}[\max(f_1, f_2)] > 0$ by direct calculations, contradicting the claimed identity. By the law of total expectation and symmetry of f_1 and f_2,

$$\begin{aligned} \mathbb{E}[\max(f_1, f_2)] &= \mathbb{E}[f_1 | f_1 > f_2]\mathbb{P}(f_1 > f_2) + \mathbb{E}[f_1 | f_1 \leq f_2]\mathbb{P}(f_1 \leq f_2) && (2) \\ &= \mathbb{E}[f_1 | f_1 > f_2]. && (3) \end{aligned}$$

By Bayes' rule, symmetry of f_1 and f_2, and conditional Gaussian update rules, the pdf $p(f_1 = t | f_1 > f_2) = \frac{\mathbb{P}(f_2 < t | f_1 = t)p(f_1 = t)}{\mathbb{P}(f_1 < f_2)} = 2\phi(t)\Phi(t\alpha)$ where $\alpha = \sqrt{(1-\rho)/(1+\rho)}$, $\phi(\cdot)$ and $\Phi(\cdot)$ is the pdf and cdf of the unit Gaussian distribution. Using integration by parts, it follows that $\mathbb{E}[\max(f_1, f_2)] = \sqrt{\frac{1-\rho}{\pi}} > 0$:

$$\begin{aligned} \mathbb{E}[f_1 | f_1 > f_2] &= \int_{\mathbb{R}} 2t\phi(t)\Phi(t\alpha)dt && (4) \\ &= [-2\phi(t)\Phi(t\alpha)]_{-\infty}^{\infty} + \int_{\mathbb{R}} 2\alpha\phi(t)\phi(t\alpha)dt && (5) \\ &= 0 + \frac{2\alpha}{\sqrt{2\pi}} \int_{\mathbb{R}} \phi\left(t\sqrt{1+\alpha^2}\right) dt && (6) \\ &= \sqrt{\frac{2}{\pi}} \frac{\alpha}{\sqrt{1+\alpha^2}} = \sqrt{\frac{1-\rho}{\pi}}. && (7) \end{aligned}$$

3.3 Bound of term $A_{4,3}$ in Eq. (23)

The term $A_{4,3} = \sum_{t=1}^{T} \mathbb{E}[2B\mathbf{I}(m_t \notin \mathcal{C}_t)]$ is derived in Eq. (23) by applying Lemma 5 inside the expectation, i.e. it is claimed that

$$\mathbb{E}[(\mu_{t-1}^{m_t}(\mathbf{x}_t) - f(\mathbf{x}_t))\mathbf{I}(m_t \notin \mathcal{C}_t)] \leq \mathbb{E}[2B\mathbf{I}(m_t \notin \mathcal{C}_t)]$$

due to Lemma 5. However, applying Lemma 5 is not sufficient to prove this since $\mu_{t-1}^{m_t}(\mathbf{x}_t) - f(\mathbf{x}_t)$ and $\mathbf{I}(m_t \notin \mathcal{C}_t)$ have not been demonstrated to be independent. Given that both terms depend on the selected model m_t (and the history H_{t-1}), one would assume the terms to be dependent unless proven otherwise.

3.4 Step (a) in the bound of term $A_{4,2}$

In step (a) in the bound of term $A_{4,2} = \sum_{t=1}^{T} \mathbb{E}[(\mu_{t-1}(\mathbf{x}_t) - f(\mathbf{x}_t))\mathbf{I}(m_t \in \mathcal{C}_t)]$, it is claimed that $L^m_{t^m_{\max}}(\mathbf{x}) - y_{t^m_{\max}}$ is bounded by $2B$ where $t^m_{\max}$ is defined as "being the last slot that m is selected". We assume that the authors intended to define $t^m_{\max} = \max\{t \in \{1, \dots, T\} : m_t = m, m \in \mathcal{C}_t\}$ and claim that

$$\mathbb{E}[L^m_{t^m_{\max}}(\mathbf{x}_{t^m_{\max}}) - y_{t^m_{\max}}] \leq 2B. \tag{8}$$

If Lemma 5 was true, it would be reasonable to assume this claim holds since $L^m_t(\mathbf{x}) \leq \mu^m_t(\mathbf{x})$ and $\mathbb{E}[y_t] = \mathbb{E}[f(\mathbf{x}_t)]$ for any fixed timestep t. However, since $t^m_{\max}$ is a random timestep that depends on the entire history H_T Lemma 5 cannot directly be applied. To see this, note that $t^m_{\max} = s$ implies that $G^m_s \leq 2\sigma_n\sqrt{N^m_{s-1}\log T}$ and $G^m_{s+1} > 2\sigma_n\sqrt{N^m_s \log T}$ where $G^m_{s+1} = \sum_{\tau=1}^{s} \mathbf{I}(m_\tau = m)(L^m_\tau(\mathbf{x}_\tau) - y_\tau)$. Therefore, $t^m_{\max} = s$ implies that the final term, $L^m_s(\mathbf{x}) - y_s$, made the sum in G^m_{s+1} exceed the threshold which changes the distribution of $L^m_s(\mathbf{x}) - y_s$.

3.5 Implications for Theorem 2 and 3

The proof of Theorem 2 states that "the term A_4 can be bounded as in (23)", referring to the step in the proof of Theorem 1 where terms $A_{4,2}$ and $A_{4,3}$ are bounded. In the proof of Theorem 3, the term C_5 is the equivalent of A_4. Since no motivation for the bound of C_5 is provided, we assume it follows from the same reasoning as in Theorem 1. As the proofs of Theorems 2 and 3 rely on the issues discussed for the bounds of $A_{4,2}$ and $A_{4,3}$, the proofs for both theorems appear to be invalid.

4 Comparison to linear setting of Hong et al. (2022)

Here, we compare the issues identified with the regret bound of EGP-TS with the issues identified by Sandberg and Haghir Chehreghani (2026) for MixTS in the linear setting Hong et al. (2022). The proof by Hong et al. (2022) introduces an event E_0 such that the linear parameter vector is close to its prior mean. In the GP setting, it corresponds to bounding the Mahalanobis distance of f to the true prior distribution $\mathcal{GP}(0, \kappa^{m_*})$. Hong et al. (2022) improperly applies the probability-matching property $(x_t|H_{t-1} \overset{d}{=} x^*|H_{t-1}$ where the history $H_{t-1} = (\mathbf{x}_i, y_i)_{i=1}^{t-1})$ when the event E_0 is conditioned on. The proof of Lu et al. (2023) does not introduce use the equivalent of E_0 and therefore avoids this mistake.

The event E_0 allows Hong et al. (2022) to bound the regret by a constant M. However, the use of M leads to several incorrect steps, such as bounding the equivalent of $\mu^{m_t}_t(x_t) - f(x_t)$ by M. Lu et al. (2023) instead (incorrectly) bounds the expectation of $\mu^{m_t}_t(x_t) - f(x_t)$ in Lemma 5 and applies it inside an expectation, as discussed in Sections 3.2 and 3.3.

To bound the equivalent of term $A_{4,2}$, Hong et al. (2022) surrounds $\mu_t^{m_t}(x_t) - f(x_t)$ with $\min(\cdot, M)$ (since E_0 has been conditioned on) and subtracts the equivalent confidence radius $\eta\sigma_{t,p_t}(x_t)$ and noise ϵ_t inside the minimum. This allows for $L_{t^m_{\max}}^m(\mathbf{x}) - y_{t^m_{\max}}$ to be bounded by M. However, adding the noise ϵ_t inside the minimum decreases the expectation rather than increase it. Therefore, both Lu et al. (2023) and Hong et al. (2022) incorrectly bound this term.

Acknowledgments

The work of Jack Sandberg and Morteza Haghir Chehreghani was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation.

References

Q. Lu, K. D. Polyzos, B. Li, and G. B. Giannakis, "Surrogate Modeling for Bayesian Optimization Beyond a Single Gaussian Process," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 9, pp. 11 283–11 296, Sep. 2023. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10093035

R. Garnett, *Bayesian Optimization*. Cambridge University Press, 2023.

B. Matérn, *Spatial Variation*, ser. Lecture Notes in Statistics, D. Brillinger, S. Fienberg, J. Gani, J. Hartigan, and K. Krickeberg, Eds. New York, NY: Springer, 1986, vol. 36. [Online]. Available: http://link.springer.com/10.1007/978-1-4615-7892-5

N. Da Costa, M. Pförtner, L. Da Costa, and P. Hennig, "Sample path regularity of Gaussian processes from the covariance kernel," *Analysis and Applications*, pp. 1–29, Dec. 2025. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/S0219530526500235

J. Hong, B. Kveton, M. Zaheer, M. Ghavamzadeh, and C. Boutilier, "Thompson Sampling with a Mixture Prior," in *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. PMLR, May 2022, pp. 7565–7586. [Online]. Available: https://proceedings.mlr.press/v151/hong22b.html

J. Sandberg and M. Haghir Chehreghani, "Adaptive Prior Selection in Gaussian Process Bandits with Thompson Sampling," Mar. 2026, arXiv:2502.01226 [cs]. [Online]. Available: http://arxiv.org/abs/2502.01226

R. B. Arellano-Valle and A. Azzalini, "Some properties of the unified skew-normal distribution," *Statistical Papers*, vol. 63, no. 2, pp. 461–487, Apr. 2022. [Online]. Available: https://doi.org/10.1007/s00362-021-01235-2

Paper III

Adaptive Prior Selection in Gaussian Process Bandits with Thompson Sampling

J. Sandberg, M. Haghir Chehreghani

Submitted, under review

The paper has been reformatted for uniformity.

Abstract

Gaussian process (GP) bandits provide a powerful framework for performing blackbox optimization of unknown functions. The characteristics of the unknown function depend heavily on the assumed GP prior. Most work in the literature assume that this prior is known but in practice this seldom holds. Instead, practitioners often rely on maximum likelihood estimation to select the hyperparameters of the prior - which lacks theoretical guarantees. In this work, we study two algorithms for joint prior selection and regret minimization in GP bandits based on GP Thompson sampling (GP-TS): Prior-Elimination GP-TS (PE-GP-TS) that disqualifies priors with poor predictive performance, and HyperPrior GP-TS (HP-GP-TS) that utilizes a bi-level Thompson sampling scheme. We theoretically analyze the algorithms and establish a sublinear regret bound for HP-GP-TS. In addition, we demonstrate the effectiveness of these algorithms compared to the alternatives through extensive experiments with synthetic and real-world data.

1 Introduction

The Gaussian process bandit problem is a variant of the multi-armed bandit problem where the arms are correlated and their expected reward is sampled from a Gaussian process (GP). The flexibility of GPs have made GP bandits applicable in a wide range of areas that need to optimize blackbox functions with noisy estimates, including hyperparameter tuning (Turner et al., 2021), online advertising (Nuara et al., 2018), and portfolio optimization (Gonzalvez et al., 2019). Most of the theoretical results in the literature assume that the GP prior is known but this is seldom the case in practical applications. Even with expert domain knowledge, selecting the exact prior to use can be a difficult task. Most practitioners tend to utilize maximum likelihood estimation (MLE) to identify suitable prior parameters. However, in a sequential decision making problem MLE is not guaranteed to recover the correct parameters which can hurt the performance.

As summarized in Table 1, previous works by Wang & de Freitas (2014); Berkenkamp et al. (2019) propose algorithms that use a decreasing sequence of lengthscales according to a fixed schedule. A drawback of these schedules is that they cannot adapt to the data and may therefore explore excessively. The Lengthscale Balancing GP-UCB algorithm of Ziomek et al. (2024) selects lengthscales such that each selected lengthscale incurs a similar amount of regret. However, this scheme relies on knowing the regret bounds, which can be impractical. Ziomek et al. (2025); Lu et al. (2023) propose algorithms that support unknown priors of (finite) arbitrary type. Prior-Elimination GP-UCB (PE-GP-UCB) (Ziomek et al., 2025) selects the prior and arm that maximize a joint upper confidence bound and eliminates priors with poor predictive performance. The joint upper confidence bound induces a double optimism in PE-GP-UCB that can lead to extra exploration. EGP-TS (Lu et al., 2023) uses bi-level Thompson sampling to select both a prior and an arm according to their posterior probabilities of being the true prior and optimal arm, respectively. Among these methods, posterior sampling is the only data-adaptive prior selection rule, and provides the closest analog to MLE.

EGP-TS is an instantiation of the more general MixTS algorithm (Hong et al., 2022b), whose regret was analyzed in the standard bandit and linear setting. However, the theoretical analyses for both algorithms are flawed. The technical issues in the regret analysis of EGP-TS were recently demonstrated by Sandberg & Haghir Chehreghani (2026) and, as we show in this work, the analysis of MixTS in the linear setting contains separate technical issues that invalidate the regret bound of Hong et al. (2022b).

Motivated by the excessive exploration of double optimism, alongside the flawed theoretical guarantees of existing Thompson sampling approaches, we investigate two distinct TS-based algorithms for GP-bandits with unknown priors. The first algorithm, Prior-Elimination GP-TS (PE-GP-TS), is an extension of PE-GP-UCB that replaces the doubly optimistic selection rule with posterior sampling and one less layer of optimism. We analyze the regret of PE-GP-TS and obtain a regret bound of order $\mathcal{O}(\sqrt{T|P|\hat{\gamma}_T \log T})$ (which matches that of PE-GP-UCB) plus a term (left unbounded) depending on the

Table 1: Comparison of similar work in GP bandits with an unknown prior.

Work	Algorithm	Prior selection	MIG dependence	Supports unknown
(Wang & de Freitas, 2014)	BOHO (EI)	Schedule	$\hat{\gamma}_T^{3/2}$	Lengthscale
(Berkenkamp et al., 2019)	A-GP-UCB	Schedule	γ_{T,p_T} †	Lengthscale and RKHS norm
(Lu et al., 2023)	EGP-TS	Posterior sampling	$\sqrt{\lvert P\rvert\hat{\gamma}_T}$ (invalid)	Arbitrary mean and kernel
(Ziomek et al., 2024)	LB-GP-UCB	Regret balancing	$\gamma_{T,\bar{p}}$ †	Lengthscale and RKHS norm
(Ziomek et al., 2025)	PE-GP-UCB	Optimistic	$\sqrt{\lvert P\rvert\hat{\gamma}_T}$	Arbitrary mean and kernel
This work	**PE-GP-TS**	Optimistic	$\sqrt{\lvert P\rvert\hat{\gamma}_T}$	Arbitrary mean and kernel
This work	**HP-GP-TS**‡	Posterior sampling	$\sqrt{\bar{\gamma}_T(P_1)}$	Arbitrary mean and kernel

† p_T is the final prior selected by A-GP-UCB, and $\bar{p}$ is the prior that minimizes the frequentist regret of GP-UCB.
‡ Equivalent to EGP-TS (Lu et al., 2023), we refer to it as HP-GP-TS.

uncertainty of the optimal arm under the correct prior. Here, T is the horizon, $|P|$ is the number of priors and $\hat{\gamma}_T$ is the worst-case maximum information gain. The second algorithm we study is EGP-TS, which we refer to as HyperPrior GP-TS (HP-GP-TS) to emphasize the use of a hyperprior, and it removes both levels of optimism. Our analysis of HP-GP-TS addresses the issues in the previous work and yields a regret bound of order $\mathcal{O}(\sqrt{T\bar{\gamma}_T(P_1)\log T})$ where $\bar{\gamma}_T(P_1)$ is a sum of maximum information gains with cardinality equal to the horizon T times the hyperprior probability $P_1(\cdot)$ s.t. $\bar{\gamma}_T(P_1) < |P|\hat{\gamma}_T$ generally holds.

We evaluate our methods on three sets of synthetic experiments and three experiments with real-world data. Across the experiments, the Thompson sampling based methods outperform PE-GP-UCB. Additionally, we find that the regret of HP-GP-TS does not increase with $|P|$ in our scaling experiments. Finally, we analyze the priors selected by the algorithms and observe that HP-GP-TS selects the correct prior more often than the other algorithms.

The contributions of this work can be summarized as:

- We propose a Thompson sampling based algorithm for GP-bandits with unknown prior, PE-GP-TS, and theoretically analyze its regret.
- We provide a sublinear regret bound for HP-GP-TS (Lu et al., 2023, EGP-TS) that depends on $\bar{\gamma}_T(P_1)$, correcting and improving upon the bound of Lu et al. (2023).
- We identify technical issues with the proof of the regret bound for MixTS (Hong et al., 2022b) in the linear setting, preventing its direct extension to the GP-setting.
- We experimentally evaluate the TS-based algorithms on both synthetic and real-world data, demonstrating that they achieve competitive performance and that the regret of HP-GP-TS does not empirically increase with $|P|$.

2 Background and problem statement

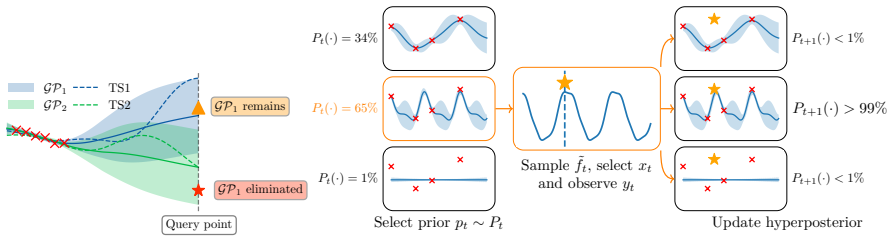
Problem statement We consider a sequential decision making problem where an agent repeatedly selects among a set of arms and receives a random

reward whose mean depends on the selected arm and is unknown to the agent. The goal of the agent is to maximize the cumulative sum of rewards over a finite time horizon. We assume that the distribution of the means, the *prior*, is sampled from a set of priors, the *hyperprior*. An effective agent must distinguish which prior the means are sampled from to ensure it explores efficiently.

Now, let us formally state the problem. Let $\mathcal{X} \subseteq [0, r]^d$ denote the finite set of arms and P a finite set of priors with associated prior mean and kernel functions $\mu_{1,p} : \mathcal{X} \mapsto \mathbb{R}$ and $k_{1,p} : \mathcal{X} \times \mathcal{X} \mapsto [-1, 1]$, $\forall p \in P$. Let $p^* \in P$ denote the true prior and assume the expected reward function $f : \mathcal{X} \mapsto \mathbb{R} \sim \mathcal{GP}(\mu_{1,p^*}, k_{1,p^*})$ is a sample from a Gaussian process with prior p^*. Both the function f and the true prior p^* are considered unknown. We will consider two settings: In the frequentist selection setting, the prior $p^* \in P$ is picked arbitrarily. In the Bayesian selection setting, the prior is sampled from a known hyperprior $p^* \sim P_1$. Then, for time step $t = 1, 2, \ldots, T$ where T is the horizon, the agent selects an arm $x_t \in \mathcal{X}$ and observes the reward $y_t = f(x_t) + \epsilon_t$ where $\{\epsilon_t\}_{t=1}^T$ are i.i.d. zero-mean Gaussian noise with variance σ^2. The goal of the agent is to select a sequence of arms $\{x_t\}_{t=1}^T$ that minimizes the regret $R(T) = \sum_{t \in [T]} f(x^*) - f(x_t)$ where $[T] = \{1, \ldots, T\}$ and $x^* = \arg\max_{x \in \mathcal{X}} f(x)$. In the Bayesian selection setting, we evaluate the agent based on the Bayesian regret $\mathrm{BR}(T) = \mathbb{E}[R(T)]$ where the expectation is taken over the prior p^*, the expected reward function f, the noise $\{\epsilon_t\}_{t=1}^T$ and the (potentially) stochastic selection of arms.

Gaussian processes A Gaussian process $f(x) \sim \mathcal{GP}(\mu, k)$ is a collection of random variables such that for any subset $\{x_1, \ldots, x_n\} \subset \mathcal{X}$, the vector $[f(x_1), \ldots, f(x_n)] \in \mathbb{R}^n$ has a multivariate Gaussian distribution. The probabilistic nature of GPs make them very useful for defining and solving bandit problems where the arms are correlated. Given the history $H_t = \{(x_i, y_i)\}_{i=1}^{t-1}$, the posterior mean and kernel functions of a Gaussian process $\mathcal{GP}(\mu, k)$ are given by $\mu_t(x) = \mu(x) + \mathbf{k}^\top \left(\mathbf{K} + \sigma^2 I\right)^{-1} (\mathbf{y} - \boldsymbol{\mu})$, and $k_t(x, \tilde{x}) = k(x, \tilde{x}) - \mathbf{k}^\top \left(\mathbf{K} + \sigma^2 I\right)^{-1} \tilde{\mathbf{k}}$. Above, $\mathbf{k}, \tilde{\mathbf{k}} \in \mathbb{R}^{t-1}$ are vectors such that $(\mathbf{k})_i = k(x_i, x)$ and $(\tilde{\mathbf{k}})_i = k(x_i, \tilde{x})$. Additionally, $\mathbf{y}, \boldsymbol{\mu} \in \mathbb{R}^{t-1}$ are also vectors such that $(\mathbf{y})_i = y_i$ and $(\boldsymbol{\mu})_i = \mu(x_i)$. The gram matrix is denoted by $\mathbf{K} \in \mathbb{R}^{t-1 \times t-1}$ where $(\mathbf{K})_{i,j} = k(x_i, x_j)$. Let $\mu_{t,p}$ and $k_{t,p}$ denote the posterior mean and kernel for a Gaussian process with prior $p \in P$ at time t and let $\sigma^2_{t,p}(x) = k_{t,p}(x, x)$ denote the posterior variance at time t. The kernel k determines important characteristics of the functions f, see Appendix D for more details and examples.

Information gain The maximal information gain (MIG) is a measure of the reduction in uncertainty of f after observing the most informative data points up to a specified size. The MIG commonly occurs in regret bounds for GP bandit algorithms (Srinivas et al., 2012; Vakili et al., 2021) and its growth rate is strongly determined by the prior kernel of the GP. Hence, we will define the MIG for any fixed GP prior $p \in P$. Let $\mathbf{y}_A$ denote noisy observations of f at the locations $A \subset \mathcal{X}$. Then, the MIG given prior $p \in P$ is defined as



(a) (b)

Figure 1: a) Elimination procedure of PE-GP-TS. The solid lines correspond to posterior means and the shaded regions are confidence intervals. The figure has been adapted from Ziomek et al. (2025). The dashed lines are samples from the posteriors. b) Overview of HP-GP-TS.

$\gamma_{T,p} := \sup_{A \subset \mathcal{X}, |A| \leq T} I_p(\mathbf{y}_A; f)$, where $I_p(\mathbf{y}_A; f) = H(\mathbf{y}_A|p) - H(\mathbf{y}_A|f,p)$ is the mutual information between $\mathbf{y}_A$ and f given p, and $H(\cdot)$ denotes the entropy. To aid our analysis later, we also define the worst-case MIG as $\hat{\gamma}_T := \max_{p \in P} \gamma_{T,p}$ and the hyperprior-weighted MIG as $\bar{\gamma}_T(P_1) := \sum_{p \in P} \Gamma_p(TP_1(p))$ for concave $\Gamma_p(\cdot)$ s.t. $\Gamma_p(t) \geq \gamma_{t,p}$ for all $t, p \in [T] \times P$. For the RBF and Matérn kernels, $\gamma_{T,p} = \mathcal{O}(\log^{d+1}(T))$ and $\gamma_{T,p} = \mathcal{O}(T^{\frac{d}{2\nu+d}} \log^{\frac{2\nu}{2\nu+d}}(T))$ (Srinivas et al., 2012; Vakili et al., 2021).

3 Algorithms

As discussed by Russo & Van Roy (2014), TS can offer advantages over UCB algorithms for problems where constructing tight confidence bounds is difficult. In addition, Thompson sampling is often observed to perform better than UCB in practice (Chapelle & Li, 2011; Wen et al., 2015; Kandasamy et al., 2018; Åkerblom et al., 2023b,a). Motivated by this, we present two algorithms for adaptive prior selection based on TS.

3.1 Prior-Elimination with Thompson sampling

Our first algorithm is an extension of PE-GP-UCB (Ziomek et al., 2025) to be employed with Thompson sampling – instead of UCB. The key difference is that instead of maximizing the upper confidence bound $U_t(x,p) = \mu_{t,p}(x) + \sqrt{\beta_t}\sigma_{t,p}(x)$ over $\mathcal{X} \times P_t$, we instead sample $\tilde{f}_{t,p}$ from the posterior $\mathcal{GP}(\mu_{t,p}, k_{t,p})$ for all priors $p \in P_t$ where P_t is the set of active priors. Then, we select the arm and prior x_t, p_t such that $x_t, p_t = \arg\max_{x,p \in \mathcal{X} \times P_t} \tilde{f}_{t,p}(x)$. Whilst PE-GP-UCB has two layers of optimism, the upper confidence bound and joint maximization of x and p, PE-GP-TS has only a single layer of optimism - which should alleviate potential overexploration issues.

The elimination procedure of PE-GP-TS is illustrated in Fig. 1(a). Samples $\tilde{f}_{t,p}$ are drawn from the active prior $p \in P_t$. Then, the unknown function f is queried at the selected arm x_t. If the observed value differs too much from

Algorithm 4 Prior Elimination GP-TS (PE-GP-TS)

procedure Horizon T, prior functions $\{\mu_{1,p}, k_{1,p}\}_{p\in P}$, confidence parameters $\{\beta_t\}_{t=1}^T$ and $\{\xi_t\}_{t=1}^T$.

1: $P_1 = P$, $S_{0,p} = \varnothing\ \forall p \in P$
2: **for** $t = 1, 2 \ldots, T$ **do**
3: Sample $\tilde{f}_{t,p} \sim \mathcal{GP}(\mu_{t,p}, k_{t,p})\ \forall p \in P_t$
4: Set $x_t, p_t = \arg\max_{x,p\in\mathcal{X}\times P_t} \tilde{f}_{t,p}(x)$
5: $S_{t,p_t} = S_{t-1,p_t} \cup \{t\}$ and $S_{t,p} = S_{t-1,p}$ for $p \in P \setminus \{p_t\}$
6: Observe $y_t = f(x_t) + \epsilon_t$
7: Set $\eta_t = y_t - \mu_{t,p_t}(x_t)$
8: Set $V_t = \sqrt{\xi_t |S_{t,p_t}|} + \sum_{i\in S_{t,p_t}} \sqrt{\beta_i}\sigma_{i,p_t}(x_i)$
9: **if** $\left|\sum_{i\in S_{t,p_t}} \eta_i\right| > V_t$ and $|P_t| > 1$ **then**
10: $P_{t+1} = P_t \setminus \{p_t\}$
11: **else**
12: $P_{t+1} = P_t$

Algorithm 5 HyperPrior GP-TS (HP-GP-TS)

procedure Horizon T, prior functions $\{\mu_{1,p}, k_{1,p}\}_{p\in P}$, hyperprior P_1.

1: **for** $t = 1, 2 \ldots, T$ **do**
2: Sample $p_t \sim P_t$
3: Sample $\tilde{f}_t \sim \mathcal{GP}(\mu_{t,p_t}, k_{t,p_t})$
4: Set $x_t = \arg\max_{x\in\mathcal{X}} \tilde{f}_t(x)$
5: Observe $y_t = f(x_t) + \epsilon_t$
6: Set $P_{t+1}(p) \propto \mathbb{P}(y_t | x_t, \{x_i, y_i\}_{i=1}^{t-1}, p) \cdot P_t(p)$ ▷ Update hyperposterior

the prediction made by the selected prior, then the selected prior is eliminated. Otherwise, it remains active.

The PE-GP-TS algorithm is presented in Algorithm 4. Similar to PE-GP-UCB, the set $S_{t,p}$ is used to store the time steps where prior p was selected up to and including time t. When prior p_t is selected, the prediction error $\eta_t = y_t - \mu_{t,p_t}(x_t)$ between the observed and predicted value made by the prior p_t is computed. If the sum of prediction errors made by the prior p_t exceeds the threshold value V_t, then p_t is eliminated from the active priors P_t, see line 9. Note that at time step t, only the selected prior p_t can be eliminated. As such, if a prior is very pessimistic it may never be selected and therefore will never be eliminated. Thus, the final set of active priors P_T should be viewed as non-eliminated priors rather than necessarily being reasonable priors.

3.2 HyperPrior Thompson sampling

In our first algorithm, we removed one layer of optimism. The second algorithm we study is a fully Bayesian algorithm that uses a hyperposterior sampling scheme where both the prior and the mean function are sampled from their respective posteriors. By shedding the optimism over the selected prior p_t,

HP-GP-TS should be able to avoid costly exploration by selecting likely priors instead of optimistic ones.

The algorithm is visualized in Fig. 1(b) and presented in detail in Algorithm 5. In the first step, the current prior p_t is sampled from the hyperposterior P_t. Then, a single sample $\tilde{f}_t$ is taken from the selected posterior $\mathcal{GP}(\mu_{t,p_t}, k_{t,p_t})$ and is used to select the current arm: $x_t = \arg\max_{x\in\mathcal{X}} \tilde{f}_t(x)$. After observing y_t, the hyperposterior is updated by computing the likelihood of y_t under the different priors. Note that since the set of priors P is finite, computing the posterior is tractable albeit computationally costly for large t with a complexity of $\mathcal{O}(t^3|P|)$. The algorithm can be extended to continuous priors P using MCMC sampling. In comparison to SCoreBO (Hvarfner et al., 2023) and other fully Bayesian algorithms that compute expected values over the hyperposterior through sampling, HP-GP-TS requires only one sample from the posterior and hyperposterior – potentially reducing the computational cost significantly. The likelihood $\mathbb{P}(y_t|x_t, \{x_i, y_i\}_{i=1}^{t-1}, p) = \mathcal{N}(y_t; \mu_{t,p}(x_t), \sigma^2_{t,p}(x_t) + \sigma^2)$ is simply the Gaussian likelihood of the posterior at x_t with added Gaussian noise with variance σ^2.

4 Regret analysis

In this section, we analyze the regret for the proposed algorithms. Recall from the problem statement that we consider two slightly different settings for the two algorithms. Specifically, for PE-GP-TS we assume the unknown prior p^* is selected arbitrarily from P whilst for HP-GP-TS we assume that the unknown prior p^* is selected from a known hyperprior distribution P_1.

4.1 Analysis of PE-GP-TS

Ziomek et al. (2025) structured the proof of the regret bound for PE-GP-UCB into 4 larger steps; First, showing that p^* is never eliminated with high probability. Second, establishing a bound on the instantaneous regret. Third, bounding the cumulative regret. Finally, the cumulative bound is re-expressed in terms of the worst-case MIG. For PE-GP-TS, we establish a new bound on the instantaneous regret and then adapt the steps of Ziomek et al. to accommodate the new bound. To bound the instantaneous regret in the lemma below, we require concentration inequalities to hold for the posteriors, the posterior samples and the noise (see Lemmas B.1 and B.2).

Lemma 4.1. *If the events of Lemmas B.1 and B.2 holds, then the following holds for the instantaneous regret of PE-GP-TS for all $t \in [T]$: $f(x^*) - f(x_t) \leq 2\sqrt{\beta_t}\sigma_{t,p^*}(x^*) + \sqrt{\beta_t}\sigma_{t,p_t}(x_t) - \eta_t + \epsilon_t$.*

Compared to the instantaneous regret bound for PE-GP-UCB, we obtain the additional term $2\sqrt{\beta_t}\sigma_{t,p^*}(x^*)$ which leads to the following regret bound:

Theorem 4.2. *Let $B_{p^*} = \beta_1 + \sup_{x\in\mathcal{X}} |\mu_{1,p^*}(x)|$ and $C = 2/\log(1+\sigma^{-2})$. If $p^* \in P$ and $f \sim \mathcal{GP}(\mu_{1,p^*}, k_{1,p^*})$, then PE-GP-TS with confidence parameters*

$\beta_t = 2\log(2|\mathcal{X}||P|\pi^2 t^2/3\delta)$ *and* $\xi_t = 2\sigma^2\log(|P|\pi^2 t^2/3\delta)$*, satisfies the following regret bound with probability at least* $1-\delta$*:*

$$R(T) \leq 2|P|B_{p^*} + 2\sqrt{\xi_T|P|T} + 2\sqrt{CT\beta_T\hat{\gamma}_T|P|} + 2\sqrt{T\beta_T\sum_{t\in[T]}\sigma^2_{t,p^*}(x^*)} \tag{1}$$

The bound of the first three terms is of order $\mathcal{O}(\sqrt{T\beta_T\hat{\gamma}_T})$ w.r.t. T which matches that of PE-GP-UCB. To our knowledge, the best lower bound for standard GP bandits in the Bayesian setting, where f is sampled from a GP, is $\Omega(\sqrt{T})$ for $d=1$ (Scarlett, 2018). This would suggest that our bound is tight up to a factor $\mathcal{O}(\sqrt{\beta_T\hat{\gamma}_T})$ when considering only the first three terms. However, note that the sublinearity of $\sum_{t\in[T]}\sigma^2_{t,p^*}(x^*)$ is not demonstrated.

4.2 Analysis of HP-GP-TS

We analyze the regret of HP-GP-TS by decomposing it into three terms and using the prior confidence technique. The initial regret decomposition is similar to Lu et al. (2023) and the prior confidence technique is first employed by Hong et al. (2022b) in standard and linear settings. However, as we discuss in Section 4.3, both of these works have fundamental issues making their theoretical analyses invalid.

First, note that HP-GP-TS inherits the probability matching property of GP-TS that $x_t|H_t \overset{d}{=} x^*|H_t$ where $\overset{d}{=}$ denotes equal in distribution. In addition, $p_t|H_t \overset{d}{=} p^*|H_t$ since p_t is sampled from the posterior distribution of p^*. Using this, one can derive the following decomposition of the regret:

$$\mathrm{BR}(T) = \sum_{t\in[T]} \mathbb{E}[\underbrace{f(x^*) - U_{t,p^*}(x^*)}_{(1)} + \underbrace{(\sqrt{\beta_t}+\sqrt{\eta_T})\sigma_{t,p_t}(x_t)}_{(2)} + \underbrace{L_{t,p_t}(x_t) - f(x_t)}_{(3)}] \tag{2}$$

where the upper confidence bound $U_{t,p}(x) = \mu_{t,p}(x) + \sqrt{\beta_t}\sigma_{t,p}(x)$ and the lower confidence bound $L_{t,p}(x) = \mu_{t,p}(x) - \sqrt{\eta_T}\sigma_{t,p}(x)$. Term (1) can be bounded using the same steps as for standard GP-TS since the confidence bound $U_{t,p^*}(x^*)$ uses the true prior p^*. The key question for term (2) is whether a tight bound for $\sum_{t\in[T]}\sigma^2_{t,p_t}(x_t)$ can be obtained. Ziomek et al. (2025) provides the bound $\sum_{p\in P}\gamma_{N_T(p),p}$ as an intermediate step in the proof of Lemma 5.3 where $N_T(p)$ is the number of times prior p is selected in total. Due to the nature of PE-GP-UCB (and similarly for PE-GP-TS) the only guarantee on $N_T(p)$ is that it is smaller than T, thus the bound $\sum_{p\in P}\gamma_{T,p}$ is used. However, we show in Lemma 4.3 that a tighter bound can be obtained for HP-GP-TS, thereby improving the dependency upon the MIG compared to the bound of Lu et al. (2023). Under a Bayesian model, we show that $\mathbb{E}[N_T(p)] = P_1(p)T$ for HP-GP-TS and by a concavity argument we provide a bound in terms of $\bar{\gamma}_T(P_1) := \sum_{p\in P}\Gamma_p(P_1(p)T)$ where $\Gamma_p(\cdot)$ is a continuous upper bound of $\gamma_{\cdot,p}$.

Lemma 4.3. *Let* $C = 2/\log(1+\sigma^{-2})$, $N_T(p) = \sum_{t\in[T]} \mathbb{1}\{p_t = p\}$ *where* $\mathbb{1}$ *is the indicator function, and* $\Gamma_p : \mathbb{R}_{\geq 0} \mapsto \mathbb{R}_{\geq 0}$ *be a concave function such that* $\Gamma_p(t) \geq \gamma_{t,p}$ *for all* $t, p \in [T] \times P$. *Then for HP-GP-TS,* $\mathbb{E}[N_T(p)] = P_1(p)T$ *and*

$$\mathbb{E}\Big[\sum_{t\in[T]} \sigma^2_{t,p_t}(x_t)\Big] \leq \sum_{p\in P} \Gamma_p(P_1(p)T) =: \bar{\gamma}_T(P_1). \tag{3}$$

To bound term (3), we define the excess reward function as $G_t(p) = \sum_{s=1}^{t-1} \mathbb{1}\{p_s = p\}\left(\mu_{s,p_s}(x_s) - \sqrt{\eta}\sigma_{s,p_s} - f(x_s) - \epsilon_s\right)$ for $\eta > 0$, similar to Lu et al. (2023). Then, we define the confidence set at time t as $\mathcal{C}_t = \{p \in P : G_s(p) \leq \xi_s(p)\ \forall s \leq t\}$ where $\xi_t(p) = \sigma\sqrt{12 N_{t-1}(p)\log(T)}$ where $N_t(p) = \sum_{s=1}^{t} \mathbb{1}\{p_s = p\}$ denotes how often the prior p was selected up to and including time t. Unlike Hong et al. (2022b); Lu et al. (2023), we impose a time-uniform requirement, i.e. a prior $p \in \mathcal{C}_t$ only if $p \in \mathcal{C}_s$ for all $s < t$, as we found their proofs uncompelling without this requirement, see Remark B.10. We show that $p^* \in \mathcal{C}_t$ with high probability in Lemma B.7 and split term (3) into two new terms:

$$\begin{aligned}(3) = \sum_{t\in[T]} \mathbb{E}\left[(L_{t,p_t}(x_t) - f(x_t))\mathbb{1}\{p_t \notin \mathcal{C}_t\}\right] \\ + \sum_{t\in[T]} \mathbb{E}\left[(L_{t,p_t}(x_t) - f(x_t))\mathbb{1}\{p_t \in \mathcal{C}_t\}\right].\end{aligned} \tag{4}$$

Since $\mathcal{C}_t$ is defined to only consider the excess reward in the past, the right term can only be bounded up to the stopping time τ_p for each prior $p \in P$. The main hurdle is therefore to bound the expectation of the stopped excess reward for each prior $\mathbb{E}[L_{\tau_p,p}(x_{\tau_p}) - f(x_{\tau_p}) - \epsilon_{\tau_p}]$. Hong et al. (2022b); Lu et al. (2023) provide incorrect bounds for this term in the linear and GP setting respectively, see Section 4.3. We first note that the stopped value of this sequence can be bounded by the maximum over the same sequence and then provide a bound for $\mathbb{E}[\max_{t\in[T]}(L_{t,p}(x_t) - f(x_t) - \epsilon_t)]$. For the left term, we know that $\mathbb{E}[\mathbb{1}\{p_t \notin \mathcal{C}_t\}] = \mathbb{P}(p^* \notin \mathcal{C}_t) = \mathcal{O}(T^{-5})$ but the factor $L_{t,p_t}(x_t) - f(x_t)$ prevents direct application of this result. Again, the bounds provided by Hong et al. (2022b); Lu et al. (2023) do not hold. We make the observation that the two factors can be separated by the Cauchy-Schwarz inequality for expected values ($\mathbb{E}[XY] \leq \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}$) and provide bounds for $\mathbb{E}[\mu^2_{t,p_t}(x_t)]$ and $\mathbb{E}[f(x_t)^2]$ in Lemmas B.11 and B.12. Finally, we are ready to state our regret bound for HP-GP-TS.

Theorem 4.4. *Let* $C = 2/\log(1+\sigma^{-2})$, $\mu_{max} = \sup_{p,x\in P\times\mathcal{X}} |\mu_{1,p}(x)|$, $M = \mathbb{E}[\sup_{x\in\mathcal{X}} |f(x)|]$, $\bar{M} = M^2 + 1 + M^2_\Delta/4$, $M_\Delta = \max_{p\in P} M_p - \min_{p\in P} M_p$, *and* $M_p = \mathbb{E}\left[\sup_{x\in\mathcal{X}} |f(x)| \middle| p^* = p\right]$. *If* $p^* \sim P_1$, $f \sim \mathcal{GP}(\mu_{1,p^*}, k_{1,p^*})$, $\beta_t = 2\log(|\mathcal{X}|t^2/\sqrt{2\pi})$, $\eta_T = 2\log|\mathcal{X}|T^6$, *then the Bayesian regret of HP-GP-TS is*

bounded by

$$
\begin{aligned}
BR(T) \leq \frac{\pi^2}{6} &+ \sqrt{CT\bar{\gamma}_T(P_1)}(\sqrt{\beta_T} + \sqrt{\eta_T}) \\
&+ \frac{\sqrt{3}}{T}\Big(\sqrt{\sigma^{-2}(\bar{M} + 2M\mu_{max} + \mu_{max}^2 + \sigma^2)} + \sqrt{\bar{M}/T}\Big) \\
&+ \sigma\sqrt{14T|P|\log T} \\
&+ |P|\left(\sigma^{-1}\sqrt{T}(M + \mu_{max} + \sigma) + M + \sigma\sqrt{2\log T}\right) \qquad (5)
\end{aligned}
$$

Note that M_p is the expected supremum of $|f(x)|$ given $p^* = p$ whereas M is the expected supremum of $|f(x)|$ for the mixture $p^* \sim P_1$. Furthermore, M_Δ denotes the spread in expected supremums and $\bar{M}$ bounds the expectaction of the squared process $\sup_x f(x)^2$, see Lemma B.12. Unlike PE-GP-TS, -UCB, and Lu et al. (2023), our regret bound of HP-GP-TS depends on the hyperprior-weighted MIG $\bar{\gamma}_T(P_1)$ rather than the worst case $|P|\hat{\gamma}_T$ which can impact the theoretical regret significantly if the complexity of the priors differ and the hyperprior is weighted towards simple priors. This is reasonable since the elimination methods assume arbitrary selection of p^* as opposed to sampling from a hyperprior. The final term in Eq. (5) is $\mathcal{O}(|P|\sqrt{T})$ whereas the term for PE-GP-TS and -UCB that is linear in $|P|$ is constant w.r.t. T. In Section 5, we empirically evaluate the dependency on $|P|$.

4.3 Comparison to MixTS and EGP-TS

Hong et al. (2022b) study MixTS, a Thompson sampling algorithm that assumes the prior is a mixture distribution, for standard and linear bandits. For the linear setting with unbounded rewards, the proof requires conditioning on the linear parameter vector θ^* to lie close to its prior mean. Under this additional event E_0, the distribution of the true parameter vector θ^*, the true prior p^* and the optimal arm x^* can shift. But, conditioned on the history H_t, MixTS is unaffected by conditioning on E_0 at time step t. Consequently, the sampled parameter vector θ_t, the selected prior p_t and the selected arm x_t maintain the same distribution. However, Hong et al. (2022b) use that $x^*, p^*|H_t, E_0 \overset{d}{=} x_t, p_t|H_t, E_0$ without proof, invalidating Theorem 1 of Hong et al. (2022b). The event E_0 bounds the maximum per-round regret by a constant, enabling $L_{t,p}(x_t) - f(x_t)$ to be conveniently bound by a constant for both terms in Eq. (4). Unfortunately, the intermediate steps contain other issues that we discuss further in Appendix C. Lu et al. (2023) study EGP-TS for sequential and parallel GP-bandit problems. Lemma 5 of Lu et al. (2023) bounds $\mathbb{E}[\mu_{t,p_t}(x_t) - f(x_t)]$ by a constant $2B$ with an incorrect proof. Even assuming a correct proof, the lemma is applied incorrectly to claim that $\mathbb{E}[(\mu_{t,p_t}(x_t) - f(x_t))\mathbb{1}\{p_t \notin \mathcal{C}_t\}] \leq \mathbb{E}[2B\mathbb{1}\{p_t \notin \mathcal{C}_t\}]$ and $\mathbb{E}[L_{\tau,p}(x_\tau) - f(x_\tau) - \epsilon_\tau] \leq 2B$. Our proof avoids the issues in previous work by separating the event $\mathbb{1}\{p_t \notin \mathcal{C}_t\}$ from the excess reward using the Cauchy-Schwarz inequality for expectations, bounding the stopped excess reward by the maximum excess reward, and bounding the expected values of $\mu_{t,p_t}(x_t)$, $\mu_{t,p}^2(x_t)$ and $f(x_t)^2$.

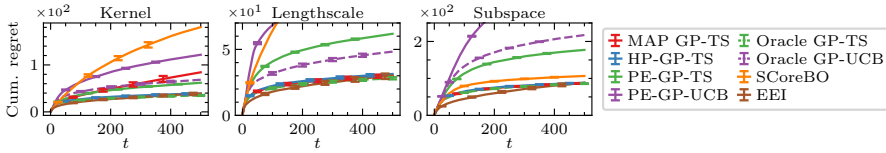


Figure 2: Cumulative regret for synthetic experiments with varying kernel, lengthscale and active subspace. The final regret for PE-GP-UCB is 114 and 389 in the lengthscale and subspace experiments, and 181 for SCoreBO in the lengthscale experiment. Errorbars correspond to ± 1 standard error.

5 Experiments

In this section, we describe our experiments based on synthetic and real-world data.

Synthetic experiments We consider three synthetic setups with different choices of priors in P. For the first setup, the priors have one of the following kernels: i) RBF kernel, ii) the rational quadratic kernel with $\alpha = 0.5$, iii) Matérn kernel with $\nu = 5/2$, iv) Matérn kernel with $\nu = 3/2$, v) periodic kernel with period $\rho = 5$, vi) linear kernel with $v = 0.05^2$. For the second setup, 8 priors use the RBF kernel with different lengthscales equidistantly spaced between $1/2$ and 4. For the third setup, the total dimensions $d = 16$ but each of the 5 priors p_i assumes $f(x)$ depends on $d_s = 4$ subdimensions. The 4 subdimensions are designed such that the priors are equally difficult to distinguish. All priors use the RBF kernel with lengthscale $\ell = 8$. For all three setups, the true prior p^* is sampled uniformly from P, the noise variance $\sigma^2 = 0.25^2$, and the horizon $T = 500$. For the first two setups, 500 arms are equidistantly spaced in $[0, 20]$ and for the third 500 arms are sampled uniformly on $[0, 20]^{16}$. All models are evaluated on 500 seeds on each setup. As baselines, we use PE-GP-UCB, SCoreBO (Hvarfner et al., 2023), fully Bayesian Expected Improvement (EEI) (Benassi et al., 2011) and Maximum A Posteriori (MAP) GP-TS. MAP GP-TS is identical to HP-GP-TS except for greedily selecting p_t from the posterior: $p_t = \arg\max_p P_t(p)$.[1] In addition, we compare against the oracle variants of PE-GP-TS and PE-GP-UCB that are only given the true prior: $P_1 = \{p^*\}$.

The cumulative regret for the three synthetic experiments is shown in Fig. 2 and the final regret is shown in Table 2 in Appendix F. Across all three experiments, we observe that HP-GP-TS and EEI has lower regret than the other methods and performs close to the oracle GP-TS. For the kernel and subspace experiments, PE-GP-TS has lower regret than the oracle GP-UCB. Hence, even if PE-GP-UCB was optimized to perform as well as the oracle, it would still not achieve the regret of the TS methods. MAP GP-TS has slightly higher regret than HP-GP-TS for the lengthscale and subspace experiments but has significantly higher regret and variance for the kernel experiment. The

[1]Note that since the hyperprior is uniform, MAP is equivalent to discrete maximum likelihood estimation.

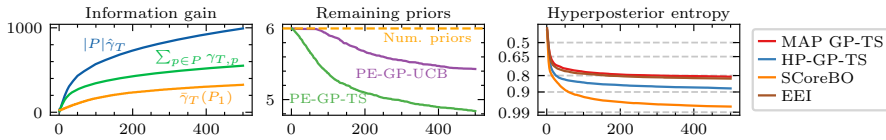


Figure 3: Analysis of the kernel experiment. Greedily maximal information gain (left). Mean number of priors remaining in P_t over time for PE-GP-UCB and -TS (middle). Entropy in the hyperposterior P_t over time for HP- and MAP GP-TS (right). The dashed reference lines correspond to entropies of discrete distributions with prob. q on one choice and prob. $\frac{1-q}{|P|-1}$ on the other $|P|-1$ choices.

greedy selection of MAP (MLE) leads to under-exploration for MAP GP-TS in certain instances. SCoreBO has the highest regret in the kernel and lengthscale experiment but has more comparable performance in the subspace experiment. In Fig. 9 in Appendix F, we report the regret for the two most competitive methods, HP-GP-TS and EEI, with an extended horizon $T = 1500$ where we observe that HP-GP-TS yields noticeably lower regret.

The maximum information gain, the number of priors remaining $|P_t|$ and the hyperposterior entropy for the kernel experiment is shown in Fig. 3. We note that $\bar{\gamma}_T(P_1)$ is significantly smaller than $|P|\hat{\gamma}_T$. The PE-methods eliminate at most one prior on average. In contrast, the final hyperposterior entropy across all algorithms is equivalent to 70-99% of the probability mass being assigned to one prior showing that the hyperposterior adapts more effectively. Across the experiments, SCoreBO has the lowest hyperposterior entropy followed by HP-GP-TS and EEI has the highest (except for the kernel experiment), see Figs. 10 and 11 in Appendix F. Thus, HP-GP-TS has similar regret to EEI but lower hyperposterior entropy.

In Fig. 4, we visualize how often the methods select the true prior p^* (or kernel) in the kernel experiment as confusion matrices. PE-GP-UCB selects the Matérn-3/2 kernel more than 96% of the rounds. The Matérn-3/2 kernel induces a distribution over functions that are less smooth compared to the other kernels and produces much wider confidence intervals outside the observed data leading to excessive optimistic exploration. PE-GP-TS also shows a bias towards the Matérn-3/2 kernel but does not select it as frequently as PE-GP-UCB – demonstrating that one layer of optimism has been removed. The overall "accuracy" of the selected priors, i.e. $\sum_{t\in[T]} \mathbb{1}\{p_t = p^*\}/T$, for the elimination-based methods is around 17% in the kernel experiment compared to 62.9% and 63.2% for MAP and HP-GP-TS respectively. For HP-GP-TS, we observe that it can easily identify the periodic and linear kernels. However, the RBF, Matérn and RQ kernels are often confused with each other. These kernels do not have as easily distinguishable characteristics and are likely to produce similar posteriors even with a small amount of data. See Fig. 12 in Appendix F for confusion matrices in the lengthscale and subspace experiments.

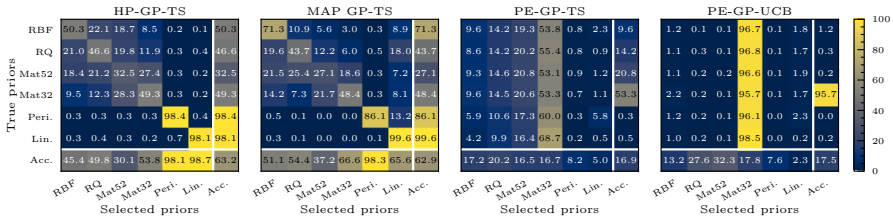


Figure 4: Confusion matrices for the true prior p^* and the selected priors p_t for the kernel experiment.

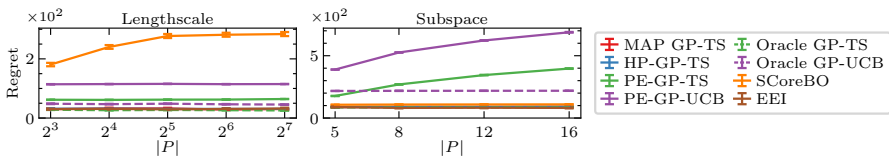


Figure 5: Total regret for the lengthscale and subspace experiments as $|P|$ increases.

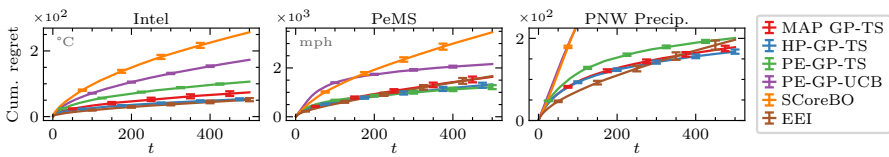


Figure 6: Cumulative regret on the real-world data experiments. Errorbars correspond to ± 1 standard error. The average final regret for SCoreBO and PE-GP-UCB is 861 and 506 on PNW.

Scaling $|P|$ We perform two experiments to understand how the regret of our algorithms scale with the number of priors. In both experiments, the average difficulty of the problem is kept constant such that the regret of the oracle models is constant. In the first experiment, we increase the discretization of the lengthscale values. The lengthscales are equidistantly spaced in $[0.5, 4]$ with $|P| \in \{8, 16, 32, 64, 128\}$. As $|P|$ increases, the difference between similar priors is reduced. In the second experiment, we increase the number of priors in the subspace experiment from 5 up to 16. Each prior can share at most 3 out of 4 dimensions with other priors which ensures the priors remain meaningfully different. The total regret as the number of priors increases is shown in Fig. 5. For the lengthscale experiment, increasing the number of priors above 8 does not affect the regret for any algorithm, likely due to the increased redundancy in the priors. The one exception is SCoreBO, whose regret increases as $|P|$ is increased to 32 but levels off beyond that. In the subspace experiment, the regret of the prior elimination algorithms scales approximately as $\sqrt{|P|}$ whilst MAP- and HP-GP-TS are consistently close to the constant regret of the oracle. The regret of EEI drops initially but is otherwise constant and SCoreBO also has constant regret.

Real-world data We perform three experiments with real-world data from the Intel Berkeley dataset (Madden et al., 2004), California Performance Measurement System (PeMS) (Chen et al., 2001; California Department of Transportation, 2024) and Pacific Northwest (PNW) daily precipitation dataset (Widmann & Bretherton, 1999, 2000). Each dataset contains measurements from a set of sensors over time. We split each dataset into a training and test set where the test set contains the last third of the data. Hence, the distribution of the test data may have shifted from the training data allowing us to test a realistic setting where the true prior is unknown but we have a set of reasonable priors. Each training set is further split into separate buckets which we use to estimate the empirical mean and covariance of the priors. See Appendix E for more details.

The cumulative regret for the experiments with real-world data is presented in Fig. 6. Across these experiments, HP-GP-TS has either the lowest regret or is within 1 standard error of the algorithm with the lowest regret. SCoreBO has significantly higher regret than all other methods. Notably for the PeMS data, PE-GP-TS has the lowest average regret whereas MAP GP-TS and EEI perform worse compared to the other experiments.

The number of priors remaining in $|P_t|$ and the hyperposterior entropy for the real-world data experiments is shown in Figs. 10 and 11 in Appendix F. Similar to the synthetic experiments, on average, the prior elimination methods eliminate less than 1 prior at best and no priors (across all 500 seeds) at worst. In contrast, the hyperposterior of HP-GP-TS concentrates to the equivalent of 60-80% of the probability mass to one prior. The relative standing in terms of reduced hyperposterior uncertainty between SCoreBO, HP-GP-TS and EEI remains consistent across all the experiments. SCoreBO reduces the hyperposterior uncertainty the most at the cost of significantly higher regret whereas HP-GP-TS provides a better balance between low regret and low

hyperposterior uncertainty.

6 Conclusion

In this paper, we have studied two algorithms for adaptive prior selection and regret minimization in GP bandits based on GP-TS. We have analyzed the algorithms theoretically, corrected and improved upon previous work, and experimentally evaluated both algorithms on synthetic and real-world data. We find that lowering the amount of optimistic exploration leads the algorithms to obtain lower or comparable regret than previous work.

Acknowledgments

The work of Jack Sandberg and Morteza Haghir Chehreghani was partially supported by the Wallenberg AI, Autonomous Systems and Software Program (WASP) funded by the Knut and Alice Wallenberg Foundation. The computations were enabled by resources provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS), partially funded by the Swedish Research Council through grant agreement no. 2022-06725.

References

Abbasi-Yadkori, Y., Pacchiano, A., and Phan, M. Regret Balancing for Bandit and RL Model Selection, June 2020. URL `https://arxiv.org/abs/2006.05491`.

Åkerblom, N., Chen, Y., and Haghir Chehreghani, M. Online Learning of Energy Consumption for Navigation of Electric Vehicles. *Artificial Intelligence*, 317: 103879, April 2023a. doi: 10.1016/j.artint.2023.103879.

Åkerblom, N., Hoseini, F. S., and Haghir Chehreghani, M. Online Learning of Network Bottlenecks via Minimax Paths. *Machine Learning*, 112(1):131–150, January 2023b. doi: 10.1007/s10994-022-06270-0.

Balandat, M., Karrer, B., Jiang, D. R., Daulton, S., Letham, B., Wilson, A. G., and Bakshy, E. BoTorch: A Framework for Efficient Monte-Carlo Bayesian Optimization. In *Advances in Neural Information Processing Systems 33*, 2020. URL `http://arxiv.org/abs/1910.06403`.

Basu, S., Kveton, B., Zaheer, M., and Szepesvari, C. No Regrets for Learning the Prior in Bandits. In *Advances in Neural Information Processing Systems*, volume 34, pp. 28029–28041. Curran Associates, Inc., 2021.

Benassi, R., Bect, J., and Vazquez, E. Robust Gaussian Process-Based Global Optimization Using a Fully Bayesian Expected Improvement Criterion. In Coello, C. A. C. (ed.), *Learning and Intelligent Optimization*, volume 6683, pp. 176–190. Springer Berlin, Heidelberg, 2011. ISBN 978-3-642-25565-6.

Berkenkamp, F., Schoellig, A. P., and Krause, A. No-Regret Bayesian Optimization with Unknown Hyperparameters. *Journal of Machine Learning Research*, 20(50):1–24, 2019. ISSN 1533-7928.

Bogunovic, I., Scarlett, J., and Cevher, V. Time-Varying Gaussian Process Bandit Optimization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 314–323. PMLR, May 2016. ISSN: 1938-7228.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, February 2013. ISBN 978-0-19-953525-5. doi: 10.1093/acprof:oso/9780199535255.001.0001. URL `https://academic.oup.com/book/26549`.

California Department of Transportation. Caltrans Performance Measurement System, 2024. URL `https://pems.dot.ca.gov/`.

California Department of Transportation. Caltrans Terms of Use, 2026. URL `https://pems.dot.ca.gov/?dnode=Help&content=help_tou#ownership`.

Chapelle, O. and Li, L. An Empirical Evaluation of Thompson Sampling. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Chen, C., Petty, K., Skabardonis, A., Varaiya, P., and Jia, Z. Freeway Performance Measurement System: Mining Loop Detector Data. *Transportation Research Record*, 1748(1):96–102, January 2001. doi: 10.3141/1748-12.

De Ath, G., Everson, R. M., and Fieldsend, J. E. How Bayesian should Bayesian optimisation be? In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, GECCO '21, pp. 1860–1869, New York, USA, July 2021. Association for Computing Machinery. ISBN 978-1-4503-8351-6.

Gardner, J. R., Pleiss, G., Weinberger, K. Q., Bindel, D., and Wilson, A. G. GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Gonzalvez, J., Lezmi, E., Roncalli, T., and Xu, J. Financial Applications of Gaussian Processes and Bayesian Optimization, 2019. URL `https://arxiv.org/abs/1903.04841`.

Hernández-Lobato, J. M., Hoffman, M. W., and Ghahramani, Z. Predictive Entropy Search for Efficient Global Optimization of Black-box Functions. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

Hong, J., Kveton, B., Zaheer, M., and Ghavamzadeh, M. Hierarchical Bayesian Bandits. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pp. 7724–7741. PMLR, May 2022a.

Hong, J., Kveton, B., Zaheer, M., Ghavamzadeh, M., and Boutilier, C. Thompson Sampling with a Mixture Prior. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, pp. 7565–7586. PMLR, May 2022b.

Hvarfner, C., Hellsten, E., Hutter, F., and Nardi, L. Self-Correcting Bayesian Optimization through Bayesian Active Learning. *Advances in Neural Information Processing Systems*, 36:79173–79199, December 2023.

Kandasamy, K., Krishnamurthy, A., Schneider, J., and Poczos, B. Parallelised Bayesian Optimisation via Thompson Sampling. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pp. 133–142. PMLR, March 2018.

Krause, A., Singh, A., and Guestrin, C. Near-Optimal Sensor Placements in Gaussian Processes: Theory, Efficient Algorithms and Empirical Studies. *Journal of Machine Learning Research*, 9(8):235–284, 2008. ISSN 1533-7928.

Kveton, B., Konobeev, M., Zaheer, M., Hsu, C.-W., Mladenov, M., Boutilier, C., and Szepesvari, C. Meta-Thompson Sampling. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 5884–5893. PMLR, July 2021.

Li, H., Liang, D., and Xie, Z. Modified Meta-Thompson Sampling for Linear Bandits and Its Bayes Regret Analysis, September 2024. URL `https://arxiv.org/abs/2409.06329`.

Lu, Q., Polyzos, K. D., Li, B., and Giannakis, G. B. Surrogate Modeling for Bayesian Optimization Beyond a Single Gaussian Process. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):11283–11296, September 2023. ISSN 1939-3539. doi: 10.1109/TPAMI.2023.3264741. URL `https://ieeexplore.ieee.org/abstract/document/10093035`.

Mackay, D. J. Introduction to Gaussian processes. In *NATO ASI Series. Series F : Computer and System Sciences*, pp. 133–165, 1998. ISBN 978-3-540-64928-1.

Madden, S. et al. Intel lab data, 2004. URL `https://db.csail.mit.edu/labdata/labdata.html`.

Matérn, B. Spatial Variation. In Brillinger, D., Fienberg, S., Gani, J., Hartigan, J., and Krickeberg, K. (eds.), *Spatial Variation*, volume 36 of *Lecture Notes in Statistics*. Springer, New York, 1986. ISBN 978-0-387-96365-5.

Mockus, J. On Bayesian Methods for Seeking the Extremum. In Marchuk, G. I. (ed.), *Optimization Techniques IFIP Technical Conference Novosibirsk, July 1–7, 1974*, pp. 400–404, Berlin, Heidelberg, 1975. Springer. ISBN 978-3-540-37497-8.

Nuara, A., Trovò, F., Gatti, N., and Restelli, M. A Combinatorial-Bandit Algorithm for the Online Joint Bid/Budget Optimization of Pay-per-Click

Advertising Campaigns. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), April 2018. doi: 10.1609/aaai.v32i1.11888.

Osborne, M. A., Garnett, R., and Roberts, S. J. Gaussian Processes for Global Optimization. In *3rd International Conference on Learning and Intelligent Optimization (LION3)*, pp. 1–15, 2009.

Pacchiano, A., Dann, C., Gentile, C., and Bartlett, P. Regret Bound Balancing and Elimination for Model Selection in Bandits and RL, December 2020. URL `https://arxiv.org/abs/2012.13045`.

Pleiss, G., Gardner, J. R., Balandat, M., and Ament, S. Linear_Operator: Structured linear algebra in pytorch. PyTorch Conference Poster, 2022. URL `https://pytorch.s3.amazonaws.com/posters/ptc2022/B05.pdf`.

Pleiss, G., Gardner, J. R., Balandat, M., et al. LinearOperator, 2025. URL `https://github.com/cornellius-gp/linear_operator`.

Russo, D. and Van Roy, B. Learning to Optimize via Posterior Sampling. *Mathematics of Operations Research*, April 2014. doi: 10.1287/moor.2014.0650.

Sandberg, J. and Haghir Chehreghani, M. Comments on "Surrogate Modeling for Bayesian Optimization Beyond a Single Gaussian Process". Under review, 2026.

Scarlett, J. Tight Regret Bounds for Bayesian Optimization in One Dimension. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 4500–4508. PMLR, July 2018.

Snoek, J., Larochelle, H., and Adams, R. P. Practical Bayesian Optimization of Machine Learning Algorithms. In *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

Srinivas, N., Krause, A., Kakade, S. M., and Seeger, M. W. Information-Theoretic Regret Bounds for Gaussian Process Optimization in the Bandit Setting. *IEEE Transactions on Information Theory*, 58(5):3250–3265, May 2012. doi: 10.1109/TIT.2011.2182033.

Turner, R., Eriksson, D., McCourt, M., Kiili, J., Laaksonen, E., Xu, Z., and Guyon, I. Bayesian Optimization is Superior to Random Search for Machine Learning Hyperparameter Tuning: Analysis of the Black-Box Optimization Challenge 2020. In *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, pp. 3–26. PMLR, August 2021.

Vakili, S., Khezeli, K., and Picheny, V. On Information Gain and Regret Bounds in Gaussian Process Bandits. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, pp. 82–90. PMLR, March 2021.

Wang, Z. and de Freitas, N. Theoretical Analysis of Bayesian Optimisation with Unknown Gaussian Process Hyper-Parameters, June 2014. URL https://arxiv.org/abs/1406.7758.

Wang, Z. and Jegelka, S. Max-value Entropy Search for Efficient Bayesian Optimization. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 3627–3635. PMLR, July 2017. ISSN: 2640-3498.

Wen, Z., Kveton, B., and Ashkan, A. Efficient Learning in Large-Scale Combinatorial Semi-Bandits. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1113–1122. PMLR, June 2015.

Widmann, M. and Bretherton, C. S. "50" km resolution daily precipitation for the Pacific Northwest, 1949-94, May 1999. URL http://research.jisao.washington.edu/data/widmann/.

Widmann, M. and Bretherton, C. S. Validation of Mesoscale Precipitation in the NCEP Reanalysis Using a New Gridcell Dataset for the Northwestern United States. *Journal of Climate*, 13(11):1936–1950, June 2000. ISSN 0894-8755, 1520-0442.

Williams, C. K. and Rasmussen, C. E. *Gaussian Processes for Machine Learning*, volume 2. MIT Press Cambridge, MA, 2006.

Ziomek, J., Adachi, M., and Osborne, M. A. Bayesian Optimisation with Unknown Hyperparameters: Regret Bounds Logarithmically Closer to Optimal. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, November 2024.

Ziomek, J., Adachi, M., and Osborne, M. A. Time-varying Gaussian Process Bandits with Unknown Prior. In *The 28th International Conference on Artificial Intelligence and Statistics*, February 2025.

A Extended discussion of related work

Plenty of previous work has proposed fully Bayesian approaches that integrate the acquisition function over the hyperposterior (Osborne et al., 2009; Benassi et al., 2011; Snoek et al., 2012; Hernández-Lobato et al., 2014; Wang & Jegelka, 2017; De Ath et al., 2021). A difficulty with such approaches is that to compute the expected acquisition function they must perform costly MCMC sampling over the hyperposterior. In contrast, HP-GP-TS optimizes a single hyperposterior sample instead of computing expected values over the hyperposterior. Hvarfner et al. (2023) proposed Self-Correcting Bayesian Optimization (SCoreBO) whose objective function balances reducing the uncertainty of (x^*, f^*) and reducing the uncertainty of the true prior p^*. Notably, SCoreBO explicitly tries to identify the prior rather than integrating out the uncertainty of the prior.

Wang & de Freitas (2014) first derived regret bounds for GP bandits with unknown lengthscale for the Expected Improvement algorithm (Mockus, 1975). However, the proposed algorithm requires a lower bound on the lengthscale and the regret bound depends on the worst-case MIG. Later work by Berkenkamp et al. (2019) introduced Adaptive GP-UCB (A-GP-UCB) that continually lowers the lengthscale parameter. Given a sufficiently small lengthscale, the function f lies within the reproducing kernel Hilbert space (RKHS) and the regular GP-UCB theory can be applied. However, A-GP-UCB lacks a stopping mechanism and will overexplore as the lengthscale continues to shrink. Recent work by Ziomek et al. (2025) introduced Prior-Elimination GP-UCB (PE-GP-UCB) for time-varying GP-bandits with unknown prior. Unlike the work before, the regret bound of PE-GP-UCB holds for arbitrary types of hyperparameters in the GP prior. PE-GP-UCB is doubly optimistic and selects the prior *and* arm with the highest upper confidence bound. PE-GP-UCB tracks the cumulative prediction error made by the selected priors and eliminates priors that exceed a threshold level.

Other works have introduced regret balancing algorithms that maintain a set of base learning algorithms and balance their selection frequency to achieve close to optimal regret (Abbasi-Yadkori et al., 2020; Pacchiano et al., 2020). Ziomek et al. (2024) built on this idea and introduced length-scale balancing GP-UCB which can adaptively explore smaller lengthscales but can return to longer ones, unlike A-GP-UCB.

In addition to Hong et al. (2022b); Lu et al. (2023), another line of work has studied Thompson sampling in standard and linear bandits with unknown prior distribution (Kveton et al., 2021; Basu et al., 2021; Hong et al., 2022a; Li et al., 2024). In their setting (meta or hierarchical bandits), the agent plays multiple bandit instances, either simultaneously or sequentially. The unknown means are sampled from the same (unknown) prior and by gathering knowledge across instances, the agent can solve later instances more efficiently once it has identified the prior. In contrast, in this paper we consider the setting where the agent can only access information from the instance it is facing.

B Proofs

In the following section, we state and prove the results shown in the main text.

B.1 PE-GP-TS

First, we state and prove concentration inequalities for $f(x)$ and $\tilde{f}_{t,p}(x)$. Lemma B.1 is based on Lemma 5.1 of Srinivas et al. (2012) but adapted to TS by specifying that it holds for any sequence of $x_1, \dots, x_T$, as discussed by Russo & Van Roy (2014). Additionally, we add Eq. (7) which can be shown through the same steps and an additional union bound over P.

Lemma B.1. *If $f(x) \sim \mathcal{GP}(\mu_{1,p^*}, k_{1,p^*})$ and $\beta_t = 2\log\left(\frac{|\mathcal{X}||P|\pi^2 t^2}{3\delta}\right)$. Then, with probability at least $1-\delta$, the following holds for all $t, x, p \in [T] \times \mathcal{X} \times P$:*

$$|f(x) - \mu_{t,p^*}(x)| \leq \sqrt{\beta_t}\sigma_{t,p^*}(x), \tag{6}$$

$$|\tilde{f}_{t,p}(x) - \mu_{t,p}(x)| \leq \sqrt{\beta_t}\sigma_{t,p}(x). \tag{7}$$

Proof. Follows by the same steps as Lemma 5.1 of Srinivas except we condition on the complete history H_t instead of only $\mathbf{y}_{1:t-1}$. Additionally, for Eq. (7) we must take an additional union bound over $p \in P$.

Fix $t, x, p \in [T] \times \mathcal{X} \times P$. Given the history H_t, $\tilde{f}_{t,p}(x) \sim \mathcal{N}(\mu_{t,p}(x), \sigma^2_{t,p}(x))$. Using that $\mathbb{P}(Z > c) \leq 1/2e^{-c^2/2}$ for $Z \sim \mathcal{N}(0,1)$, we get that

$$\mathbb{P}\left(\left|\frac{\tilde{f}_{t,p}(x) - \mu_{t,p}(x)}{\sigma_{t,p}(x)}\right| > \sqrt{\beta_t}\right) \leq \exp(-\beta_t/2) \tag{8}$$

$$= \frac{3\delta}{|\mathcal{X}||P|\pi^2 t^2} \tag{9}$$

Note that $\sum_{t\geq 1} \frac{1}{t^2} = \frac{\pi^2}{6}$. By taking the union bound over $\mathcal{X}$, P and $t \geq 1$, Eq. (7) holds w.p. at least $1-\delta/2$. By the same reasoning and skipping the union bound over P, Eq. (6) holds w.p. at least $1-\delta/2$. Thus, both events hold w.p. at least $1-\delta$. □

Next, we state three lemmas from Ziomek et al. (2025) that are used in the proof of our regret bound.

Lemma B.2. *(Lemma 5.1 of Ziomek et al. (2025)) If $\xi_t = 2\sigma^2 \log\left(\frac{|P|\pi^2 t^2}{6\delta}\right)$, then the following holds with probability at least $1-\delta$:*

$$\left|\sum_{i \in S_{t,p}} \epsilon_i\right| \leq \sqrt{\xi_t |S_{t,p}|} \quad \forall t, p \in [T] \times P. \tag{10}$$

Lemma B.3. *(Lemma 5.2 of Ziomek et al. (2025)) Let $B_{p^*} = \beta_1 + \sup_{x\in\mathcal{X}} |\mu_{1,p^*}(x)|$, then if μ_{1,p^*} and k_{1,p^*} satisfy $|\mu_{1,p^*}(\cdot)| < \infty$ and $k_{1,p^*}(\cdot,\cdot) \leq 1$ and Lemma B.1 holds, then*

$$\sup_{x\in\mathcal{X}} |f(x)| \leq B_{p^*}. \tag{11}$$

Lemma B.4. *(Lemma 5.3 of Ziomek et al. (2025)) For* $C = 2/\log(1+\sigma^{-2})$, $\sum_{t\notin\mathcal{C}} \sqrt{\beta_t}\sigma_{t,p_t}(x_t) \leq \sqrt{CT\beta_T\hat{\gamma}_T|P|}$ *where* $\beta_T = \max_{p\in P}\beta_T$ *and* $\hat{\gamma}_T = \max_{p\in P}\gamma_{T,p}$.

Lemma B.5. *If the events of Lemmas B.1 and B.2 hold, then PE-GP-TS never eliminates the true prior* p^*.

Proof. For any $t \in [T]$,

$$\begin{aligned}
\left|\sum_{i\in S_{t,p^*}} \eta_i\right| &= \left|\sum_{i\in S_{t,p^*}} (y_i - f(x_i) + f(x_i) - \mu_{i,p^*}(x_i)\right| \\
&\leq \left|\sum_{i\in S_{t,p^*}} \epsilon_i\right| + \sum_{i\in S_{t,p^*}} |f(x_i) - \mu_{i,p^*}(x_i)| \qquad \text{(Triangle ineq.)} \\
&\leq \sqrt{\xi_t |S_{t,p^*}|} + \sum_{i\in S_{t,p^*}} \sqrt{\beta_i}\sigma_{i,p^*}(x_i). \qquad \text{(Lemmas B.1 and B.2)}
\end{aligned} \tag{12}$$

(13)

(14)

Therefore, the elimination criteria on line 9 in Algorithm 4, $|\sum_{i\in S_{t,p_t}}\eta_i| > V_t$, always evaluates to `false` for $p_t = p^*$. □

Then, we state and prove the new instantaneous regret bound for PE-GP-TS.

Lemma 4.1. *If the events of Lemmas B.1 and B.2 holds, then the following holds for the instantaneous regret of PE-GP-TS for all* $t \in [T]$*:* $f(x^*) - f(x_t) \leq 2\sqrt{\beta_t}\sigma_{t,p^*}(x^*) + \sqrt{\beta_t}\sigma_{t,p_t}(x_t) - \eta_t + \epsilon_t$.

Proof. First, we upper bound $f(x^*)$ as follows

$$\begin{aligned}
f(x^*) &\leq \mu_{t,p^*}(x^*) + \sqrt{\beta_t}\sigma_{t,p^*}(x^*) && \text{(Eq. (6))} && (15)\\
&\leq \tilde{f}_{t,p^*}(x^*) + 2\sqrt{\beta_t}\sigma_{t,p^*}(x^*) && \text{(Eq. (7))} && (16)\\
&\leq \tilde{f}_{t,p_t}(x_t) + 2\sqrt{\beta_t}\sigma_{t,p^*}(x^*). && && (17)
\end{aligned}$$

For the final step, we use the TS selection rule and that $p^* \in P_t$ by Lemma B.5. Then, we lower bound $f(x_t)$

$$\begin{aligned}
f(x_t) &= \mu_{t,p_t}(x_t) + \eta_t - \epsilon_t && \text{(Def. of } \eta_t) && (18)\\
&\geq \tilde{f}_{t,p_t}(x_t) - \sqrt{\beta_t}\sigma_{t,p_t}(x_t) + \eta_t - \epsilon_t. && \text{(Eq. (7))} && (19)
\end{aligned}$$

Combining, Eqs. (17) and (19) we obtain

$$f(x^*) - f(x_t) \leq 2\sqrt{\beta_t}\sigma_{t,p^*}(x^*) + \sqrt{\beta_t}\sigma_{t,p_t}(x_t) - \eta_t + \epsilon_t. \tag{20}$$

□

Finally, we state and prove the cumulative regret bound for PE-GP-TS.

Theorem 4.2. *Let $B_{p^*} = \beta_1 + \sup_{x\in\mathcal{X}} |\mu_{1,p^*}(x)|$ and $C = 2/\log(1+\sigma^{-2})$. If $p^* \in P$ and $f \sim \mathcal{GP}(\mu_{1,p^*}, k_{1,p^*})$, then PE-GP-TS with confidence parameters $\beta_t = 2\log(2|\mathcal{X}||P|\pi^2 t^2/3\delta)$ and $\xi_t = 2\sigma^2\log(|P|\pi^2 t^2/3\delta)$, satisfies the following regret bound with probability at least $1-\delta$:*

$$R(T) \leq 2|P|B_{p^*} + 2\sqrt{\xi_T|P|T} + 2\sqrt{CT\beta_T\hat{\gamma}_T|P|} + 2\sqrt{T\beta_T\textstyle\sum_{t\in[T]}\sigma^2_{t,p^*}(x^*)} \tag{1}$$

Proof. To establish a bound on the cumulative regret, we separate out the rounds where priors are eliminated. Hence, define the set of critical iterations as

$$\mathcal{C} = \left\{ t \in [T] : \left| \sum_{i\in S_{t,p_t}} \eta_i \right| > \sqrt{\xi_t S_{t,p_t}} + \sum_{i\in S_{t,p_t}} \sqrt{\beta_i}\sigma_{i,p_t}(x_i) \right\}. \tag{21}$$

Note that $|\mathcal{C}| \leq |P|$. Using Lemma B.3 and Eq. (20), we can bound the cumulative regret as follows:

$$\begin{aligned} R(T) &= \sum_{t\in\mathcal{C}} f(x^*) - f(x_t) + \sum_{t\notin\mathcal{C}} f(x^*) - f(x_t) && (22)\\ &\leq 2|P|B_{p^*} + \sum_{t\notin\mathcal{C}} 2\sqrt{\beta_t}\sigma_{t,p^*}(x^*) + \sum_{t\notin\mathcal{C}} \sqrt{\beta_t}\sigma_{t,p_t}(x_t) \\ &\qquad + \sum_{p\in P}\sum_{t\in S_{T,p}\setminus\mathcal{C}} (\epsilon_t - \eta_t). && (23) \end{aligned}$$

where $B_{p^*} := \beta_1 + \sup_{x\in\mathcal{X}} |\mu_{1,p^*}(x)|$. If $t \notin \mathcal{C}$, line 9 in Algorithm 4 evaluates to `false` and hence

$$\sum_{p\in P}\sum_{t\in S_{T,p}\setminus\mathcal{C}} -\eta_t \leq \sum_{p\in P}\sqrt{\xi_T|S_{T,p}|} + \sum_{p\in P}\sum_{t\in S_{T,p}\setminus\mathcal{C}} \sqrt{\beta_t}\sigma_{t,p}(x_t). \tag{24}$$

Additionally, using Lemma B.2, we can bound the Gaussian noise:

$$\begin{aligned} \sum_{p\in P}\sum_{t\in S_{T,p}\setminus\mathcal{C}} \epsilon_t &\leq \sum_{p\in P}\left|\sum_{t\in S_{T,p}\setminus\mathcal{C}} \epsilon_t\right| && (25)\\ &\leq \sum_{p\in P}\sqrt{\xi_T|S_{T,p}\setminus\mathcal{C}|} && \text{(Lemma B.2)} \quad (26)\\ &\leq \sum_{p\in P}\sqrt{\xi_T|S_{T,p}|} && (27)\\ &\leq \sqrt{\xi_T|P|T}. && \text{(Cauchy-Schwarz)} \quad (28) \end{aligned}$$

Combining the above, the cumulative regret is bounded by

$$R(T) \leq 2|P|B_{p^*} + 2\sqrt{\xi_T|P|T} + 2\sum_{t\notin\mathcal{C}} \sqrt{\beta_t}\sigma_{t,p^*}(x^*) + 2\sum_{t\notin\mathcal{C}} \sqrt{\beta_t}\sigma_{t,p_t}(x_t). \tag{29}$$

Finally, applying Lemma B.4, we obtain the result

$$R(T) \leq 2|P|B_{p^*} + 2\sqrt{\xi_T|P|T} + 2\sqrt{T\beta_T \sum_{t\in[T]} \sigma^2_{t,p^*}(x^*)} + 2\sqrt{CT\beta_T\hat{\gamma}_T|P|}. \tag{30}$$

□

B.2 HP-GP-TS

In this section, we state and prove our regret bound for HP-GP-TS. We begin by proving Lemma 4.3.

Lemma 4.3. *Let $C = 2/\log(1+\sigma^{-2})$, $N_T(p) = \sum_{t\in[T]} \mathbb{1}\{p_t = p\}$ where $\mathbb{1}$ is the indicator function, and $\Gamma_p : \mathbb{R}_{\geq 0} \mapsto \mathbb{R}_{\geq 0}$ be a concave function such that $\Gamma_p(t) \geq \gamma_{t,p}$ for all $t, p \in [T] \times P$. Then for HP-GP-TS, $\mathbb{E}[N_T(p)] = P_1(p)T$ and*

$$\mathbb{E}\Big[\sum_{t\in[T]} \sigma^2_{t,p_t}(x_t)\Big] \leq \sum_{p\in P} \Gamma_p(P_1(p)T) =: \bar{\gamma}_T(P_1). \tag{3}$$

Remark B.6. *For the RBF and Matérn kernels, the known upper bounds for the maximum information gain are concave (Srinivas et al., 2012; Vakili et al., 2021), thereby satisfying the conditions of Lemma 4.3.*

Proof. We begin by showing that $\mathbb{E}[N_T(p)] = P_1(p)T$.

$$\begin{aligned}
\mathbb{E}[N_T(p)] &= \mathbb{E}\Big[\sum_{t\in[T]} \mathbb{1}\{p_t = p\}\Big] && (31)\\
&= \sum_{t\in[T]} \mathbb{E}_{H_t}\left[\mathbb{E}\left[\mathbb{1}\{p_t = p\}\middle| H_t\right]\right] \qquad \text{(Tower rule)} && (32)\\
&= \sum_{t\in[T]} \mathbb{E}_{H_t}\left[\mathbb{E}\left[\mathbb{1}\{p^* = p\}\middle| H_t\right]\right] \qquad (p_t|H_t \overset{d}{=} p^*|H_t) && (33)\\
&= \sum_{t\in[T]} \mathbb{E}\left[\mathbb{1}\{p^* = p\}\right] && (34)\\
&= \sum_{t\in[T]} P_1(p) = P_1(p)T. \qquad (P_1(p) = \mathbb{P}(p^* = p)) && (35)
\end{aligned}$$

Then, by the intermediate steps of Lemma 5.3 of Ziomek et al. (2025) $\sum_{t\in[T]} \sigma^2_{t,p_t}(x_t) \leq C\sum_{p\in P}\gamma_{N_T(p),p}$. We include the proof here for completeness

and introduce some helpful notation. Let A be a multiset over $\mathcal{X}$ s.t. $|A| < \infty$, we define

$$\sigma^2_{A,p}(x) = k_{1,p}(x,x) - \mathbf{k}_{A,p}(x)^T(\mathbf{K}_{A,p} + \sigma^2 I)^{-1}\mathbf{k}_{A,p}(x), \tag{36}$$

where $\mathbf{K}_{A,p} = [k_{1,p}(x,x')]_{x,x' \in A}$ and $\mathbf{k}_{A,p} = [k_{1,p}(x,x')]_{x' \in A}$ with elements repeated by their multiplicity in A. Then, let $A_{t,p} = \{x_i : i \in [t-1], p_i = p\}$ be the multiset of arms queried whilst selecting prior p. For any two multisets S, S' such that $S \subseteq S'$, we have that $\sigma^2_{S',p}(x) \leq \sigma^2_{S,p}(x)$ for all $x \in \mathcal{X}$. Since $A_{t,p}$ is a subset of the arms collected over the history H_t, we have that $\sigma^2_{t,p_t}(x_t) \leq \sigma^2_{A_{t,p},p}(x_t)$. By the proof of Lemma 5.4 of Srinivas et al. (2012), $\sigma^2_{A_{t,p},p}(x_t) \leq C \log(1+\sigma^{-2}\sigma^2_{A_{t,p},p}(x_t))$ for $C = 2/\log(1+\sigma^{-2})$. By reorganizing the sum over $t \in [T]$ into a sum over $p \in P$, and applying Lemma 5.3 of Srinivas et al. (2012) yields that

$$\sum_{t\in[T]} \sigma^2_{t,p_t}(x_t) \leq \sum_{p\in P} \sum_{t\in[T]:p_t=p} \sigma^2_{A_{t,p},p}(x_t) \leq C \sum_{p\in P} \gamma_{|A_{T,p}|,p} = C \sum_{p\in P} \gamma_{N_T(p),p}. \tag{37}$$

Finally, we combine the results above through a concavity argument.

$$\begin{aligned}
\mathbb{E}\left[\sum_{t\in[T]} \sigma^2_{t,p_t}(x_t)\right] &\leq C \sum_{p\in P} \mathbb{E}[\gamma_{N_T(p),p}] && \text{(Eq. (37))} && (38)\\
&\leq C \sum_{p\in P} \mathbb{E}[\Gamma_p(N_T(p))] && \begin{pmatrix}\gamma_{t,p} \leq \Gamma_p(t),\\ \forall t, p \in [T]\times P\end{pmatrix} && (39)\\
&\leq C \sum_{p\in P} \Gamma_p(\mathbb{E}[N_T(p)]) && \begin{pmatrix}\Gamma_p(t) \text{ concave, Jensen's}\\ \text{ineq.}\end{pmatrix} && (40)\\
&\leq C \sum_{p\in P} \Gamma_p(P_1(p)T). && \text{(Eq. (35))} && (41)
\end{aligned}$$

□

Next, we prove that the true prior is in the confidence set with high probability. Recall that we define the excess reward for prior p at time t as

$$G_t(p) = \sum_{s=1}^{t-1} \mathbb{1}\{p_s = p\} \left(\mu_{s,p_s}(x_s) - \sqrt{\eta_T}\sigma_{s,p_s} - f(x_s) - \epsilon_s\right) \tag{42}$$

where $\eta_T = 2\log|\mathcal{X}|T^6$. Let $\xi_t(p) = \sigma\sqrt{14 N_{t-1}(p)\log(T)}$ where $N_t(p) = \sum_{s=1}^{t} \mathbb{1}\{p_s = p\}$ denotes how often the prior p was selected up to and including time t. Then, we define the confidence set at time t as

$$\mathcal{C}_t = \{p \in P : G_\tau(p) \leq \xi_\tau(p)\ \forall \tau \leq t\}. \tag{43}$$

For notational convenience, we consider the history $H_t = (p_i, x_i, y_i)_{i=1}^{t-1}$ with the selected priors $(p_i)_{i=1}^t$ augmented such that $\mathcal{C}_t$ and $(N_{t-1}(p))_{p\in P}$ are deterministic conditioned on H_t.

Lemma B.7. *For any $t \in [T]$, $\mathbb{P}(p^* \notin \mathcal{C}_t) \leq \frac{3}{T^5}$.*

Proof. Note that $\mathcal{C}_t$ is monotonically decreasing due to the time-uniform definition of $\mathcal{C}_t$, i.e. $\mathcal{C}_s \supseteq \mathcal{C}_t$ for any $s < t$. Thus, $\mathbb{P}(p^* \notin \mathcal{C}_t) \leq \mathbb{P}(p^* \notin \mathcal{C}_T)$ and we focus on bounding $\mathbb{P}(p^* \in \mathcal{C}_T)$.

Let $E = \cap_{t=1}^{T-1} E_t$ where $E_t = \{|f(x) - \mu_{t,p^*}(x)| \leq \sqrt{\eta_T}\sigma_{t,p^*}(x), \forall x \in \mathcal{X}\}$. Then, by the law of total probability

$$\mathbb{P}(p^* \notin \mathcal{C}_T) = \underbrace{\mathbb{P}(p^* \notin \mathcal{C}_T | E^c)}_{\leq 1} \underbrace{\mathbb{P}(E^c)}_{\leq 1/T^5} + \underbrace{\mathbb{P}(p^* \notin \mathcal{C}_T | E)\mathbb{P}(E)}_{\leq 2/T^5} \tag{44}$$

where the bounds for $\mathbb{P}(E^c)$ and $\mathbb{P}(p^* \notin \mathcal{C}_T|E)\mathbb{P}(E)$ are shown below.

$$\begin{aligned}
\mathbb{P}(E^c) &= \mathbb{P}\left(\exists t \in [T-1], x \in \mathcal{X} : |f(x) - \mu_{t,p^*}(x)| > \sqrt{\eta_T}\sigma_{t,p^*}(x)\right) && (45)\\
&\leq \sum_{t\in[T-1]} \sum_{x\in\mathcal{X}} \mathbb{P}\left(|f(x) - \mu_{t,p^*}(x)| > \sqrt{\eta_T}\sigma_{t,p^*}(x)\right) && (46)\\
&\leq \sum_{t\in[T-1]} \sum_{x\in\mathcal{X}} \mathbb{E}_{H_t,p^*}\left[\mathbb{P}\left(\frac{|f(x) - \mu_{t,p^*}(x)|}{\sigma_{t,p^*}(x)} > \sqrt{\eta_T}\middle| H_t, p^* = p\right)\right] && (47)\\
&\leq \sum_{t\in[T-1]} \sum_{x\in\mathcal{X}} \mathbb{E}_{H_t,p^*}\left[\exp\left(-\frac{\eta_T}{2}\right)\right] \quad \left(\begin{matrix}\mathbb{P}(|r| > \sqrt{c}) \leq \exp(-c/2)\\ \text{for } r \sim \mathcal{N}(0,1),\, c \geq 0\end{matrix}\right) && (48)\\
&= \sum_{t\in[T-1]} \sum_{x\in\mathcal{X}} \mathbb{E}_{H_t,p^*}\left[\frac{1}{|\mathcal{X}|T^6}\right] \qquad \left(\eta_T = 2\log(|\mathcal{X}|T^6)\right) && (49)\\
&= \frac{1}{T^5}. && (50)
\end{aligned}$$

Next, we bound the right term $\mathbb{P}(p^* \notin \mathcal{C}_t|E)$. Recall that $p^* \notin \mathcal{C}_T$ is equivalent to $\exists t \in [T]$ such that $G_t(p^*) > \xi_t(p^*)$. Hence,

$$\begin{aligned}
\mathbb{P}(p^* \notin \mathcal{C}_T|E) &= \mathbb{P}(\exists t \in [T] : G_t(p^*) > \xi_t(p^*)|E) && (51)\\
&\leq \sum_{t\in[T]} \mathbb{P}(G_t(p^*) > \xi_t(p^*)|E) \qquad \text{(Union bound)} && (52)\\
&= \sum_{t\in[T]} \mathbb{P}\Bigg(\sum_{s=1}^{t-1} \mathbb{1}\{p_s = p^*\}\Big(\mu_{s,p^*}(x_s) - \sqrt{\eta_T}\sigma_{t,p^*}(x_s)\\
&\qquad\qquad - f(x_s) - \epsilon_s\Big) > \xi_t(p^*)\Big| E\Bigg). && (53)
\end{aligned}$$

Given E, $\mu_{s,p^*}(x_s) - \sqrt{\eta_T}\sigma_{s,p^*}(x_s) - f(x_s) \leq 0$, $\forall s \in [T-1]$ and therefore

$$\mathbb{P}(p^* \notin \mathcal{C}_T|E)\mathbb{P}(E) \leq \sum_{t\in[T]} \mathbb{P}\left(\sum_{s=1}^{t-1} \mathbb{1}\{p_s = p^*\}(-\epsilon_s) > \xi_t(p^*)\Big|E\right)\mathbb{P}(E) \tag{54}$$

$$\leq \sum_{t\in[T]} \mathbb{P}\left(\left|\sum_{s=1}^{t-1} \mathbb{1}\{p_s = p^*\}(-\epsilon_s)\right| > \xi_t(p^*)\Big|E\right)\mathbb{P}(E) \tag{55}$$

$$= \sum_{t\in[T]} \mathbb{P}\left(\left|\sum_{s=1}^{t-1} \mathbb{1}\{p_s = p^*\}(-\epsilon_s)\right| > \xi_t(p^*), E\right) \tag{56}$$

$$\leq \sum_{t\in[T]} \mathbb{P}\left(\left|\sum_{s=1}^{t-1} \mathbb{1}\{p_s = p^*\}(-\epsilon_s)\right| > \xi_t(p^*)\right) \tag{57}$$

$$= \sum_{t\in[T]}\sum_{p\in P} \mathbb{P}\left(\left|\sum_{s=1}^{t-1} \mathbb{1}\{p_s = p\}\epsilon_s\right| > \sigma\sqrt{14N_{t-1}(p)\log T}\,\Big|\, p^* = p\right) \cdot \mathbb{P}(p^* = p) \tag{58}$$

$$\leq \sum_{t\in[T]}\sum_{p\in P} \frac{2}{T^6}\mathbb{P}(p^* = p) \qquad \text{(Lemma B.8)} \tag{59}$$

$$\leq \frac{2}{T^5} \tag{60}$$

□

Next, we prove the self-normalizing concentration inequality for the sum of Gaussian noises as pulled by each prior that we used in Eq. (58).

Lemma B.8. *Let $S_{t,p} = \sum_{s=1}^{t-1} \mathbb{1}\{p_s = p\}\epsilon_s$ and $\alpha > 0$, then*

$$\mathbb{P}\left(|S_{t,p}| > \sigma\sqrt{2(\alpha+1)N_{t-1}(p)\log(T)}\,\Big|\, p^* = p\right) \leq \frac{2}{T^\alpha}, \quad \forall p \in P. \tag{61}$$

Remark B.9. *Note that similar results have been shown by Hong et al. (2022b, Proof of Lemma 3), Lu et al. (2023, Lemma 4), and (Ziomek et al., 2025, Lemma 5.1). We found the arguments in the proofs of Hong et al. (2022b); Lu et al. (2023) uncompelling due to their brevity. Whilst the proof of Ziomek et al. (2025) is clearer, we provide a proof using a martingale technique as a complement.*

Proof. Fix $t \in [T]$ and $p^* = p$, for the remainder of this proof all probabilities condition on $p^* = p$. We begin by defining the event

$$\mathcal{F} := \left\{S_{t,p} > \sigma\sqrt{2(\alpha+1)N_{t-1}(p)\log(T)}\right\} \tag{62}$$

$$= \bigcup_{k=1}^{t-1} \underbrace{\left\{S_{t,p} > \sigma\sqrt{2(\alpha+1)k\log(T)} \cap N_{t-1} = k\right\}}_{\mathcal{F}_k :=}. \tag{63}$$

To bound the probability of the events $\mathcal{F}_k$, we introduce a martingale $M_t(\lambda)$ for $\lambda > 0$ into $\mathcal{F}_k$ as follows:

$$\mathcal{F}_k = \left\{\lambda S_{t,p} > \lambda\sigma\sqrt{2(\alpha+1)k\log(T)} \cap N_{t-1}(p) = k\right\} \qquad (\lambda > 0) \quad (64)$$

$$= \left\{\lambda S_{t,p} - \frac{\lambda^2\sigma^2}{2}N_{t-1}(p) > \lambda\sigma\sqrt{2(\alpha+1)k\log(T)} - \frac{\lambda^2\sigma^2}{2}k \right.$$
$$\left. \cap\ N_{t-1}(p) = k\right\} \quad (65)$$

$$= \left\{ \underbrace{\exp\left(\lambda S_{t,p} - \frac{\lambda^2\sigma^2}{2}N_{t-1}(p)\right)}_{M_t(\lambda):=} \right.$$
$$\left. > \exp\left(\lambda\sigma\sqrt{2(\alpha+1)k\log(T)} - \frac{\lambda^2\sigma^2}{2}k\right) \cap N_{t-1}(p) = k\right\}. \quad (66)$$

To tighten the bound of the probability of $\mathcal{F}_k$, we select $\lambda_k = \sqrt{\frac{2(\alpha+1)\log T}{\sigma^2 k}}$, yielding:

$$\mathcal{F}_k = \{M_t(\lambda_k) > \exp\left((\alpha+1)\log T\right) \cap N_{t-1}(p) = k\} \quad (67)$$

$$\subseteq \left\{M_t\left(\sqrt{\frac{2(\alpha+1)\log T}{\sigma^2 k}}\right) \geq T^{\alpha+1}\right\}. \quad (68)$$

Since $M_t(\lambda) \geq 0$, by Markov's inequality,

$$\mathbb{P}\left(\mathcal{F}_k\right) \leq \mathbb{P}\left(M_t\left(\lambda_k\right) \geq T^{\alpha+1}\right) \leq \frac{\mathbb{E}\left[M_t(\lambda_k)\right]}{T^{\alpha+1}}. \quad (69)$$

Next, it remains to show that $M_t(\lambda)$ is a martingale such that $\mathbb{E}[M_t(\lambda)] = 1$. Let $\mathcal{H}_{t-1} = \{p_s, \epsilon_s\}_{s=1}^{t-2}$ be the history of selected priors and noise up to and including time $t-2$, then

$$\mathbb{E}[M_t(\lambda)|\mathcal{H}_{t-1}] = M_{t-1}(\lambda) \cdot \mathbb{E}\Big[\exp\Big(\lambda\mathbb{1}\{p_{t-1} = p\}\epsilon_{t-1}$$
$$- \lambda^2\sigma^2\mathbb{1}\{p_{t-1} = p\}/2\Big)\Big|\mathcal{H}_{t-1}\Big] \quad (70)$$

$$= M_{t-1}(\lambda) \cdot \Big(\underbrace{\mathbb{E}\left[\exp(0)|p_{t-1} \neq p, \mathcal{H}_{t-1}\right]}_{=1}\mathbb{P}(p_{t-1} \neq p|\mathcal{H}_{t-1}) \quad (71)$$

$$+ \underbrace{\mathbb{E}\left[\exp(\lambda\epsilon_{t-1} - \lambda^2\sigma^2/2)|p_{t-1} = p, \mathcal{H}_{t-1}\right]}_{=1 \text{ since } \epsilon_{t-1} \perp\!\!\!\perp p_{t-1}, \mathcal{H}_{t-1}}\mathbb{P}(p_{t-1} = p, \mathcal{H}_{t-1})\Big) \quad (72)$$

$$= M_{t-1}(\lambda). \quad (73)$$

Applying the above recursively to $\mathbb{E}[M_t(\lambda)]$ and defining $M_1(\lambda) = 1$, we get that $\mathbb{E}[M_t(\lambda)] = 1$. Thus, from Eq. (69) and a union bound over $k \in [T-1]$, $\mathbb{P}(S_{t,p} > \sigma\sqrt{2(\alpha+1)N_{t-1}(p)\log(T)}) \leq 1/T^\alpha$. By symmetry of the Gaussian noise, $\mathbb{P}(|S_{t,p}| > \sigma\sqrt{2(\alpha+1)N_{t-1}(p)\log(T)}) \leq 2/T^\alpha$. □

Finally, we are ready to state and prove the regret bound for HP-GP-TS.

Theorem 4.4. *Let* $C = 2/\log(1+\sigma^{-2})$, $\mu_{max} = \sup_{p,x\in P\times\mathcal{X}} |\mu_{1,p}(x)|$, $M = \mathbb{E}[\sup_{x\in\mathcal{X}} |f(x)|]$, $\bar{M} = M^2 + 1 + M_\Delta^2/4$, $M_\Delta = \max_{p\in P} M_p - \min_{p\in P} M_p$, *and* $M_p = \mathbb{E}\left[\sup_{x\in\mathcal{X}} |f(x)| \middle| p^* = p\right]$. *If* $p^* \sim P_1$, $f \sim \mathcal{GP}(\mu_{1,p^*}, k_{1,p^*})$, $\beta_t = 2\log(|\mathcal{X}|t^2/\sqrt{2\pi})$, $\eta_T = 2\log|\mathcal{X}|T^6$, *then the Bayesian regret of HP-GP-TS is bounded by*

$$\begin{aligned} BR(T) \leq \frac{\pi^2}{6} &+ \sqrt{CT\bar{\gamma}_T(P_1)}(\sqrt{\beta_T} + \sqrt{\eta_T}) \\ &+ \frac{\sqrt{3}}{T}\Big(\sqrt{\sigma^{-2}(\bar{M} + 2M\mu_{max} + \mu_{max}^2 + \sigma^2)} + \sqrt{\bar{M}/T}\Big) \\ &+ \sigma\sqrt{14T|P|\log T} \\ &+ |P|\left(\sigma^{-1}\sqrt{T}(M + \mu_{max} + \sigma) + M + \sigma\sqrt{2\log T}\right) \end{aligned} \tag{5}$$

Proof. Recall that $p^*, x^*|H_t \overset{d}{=} p_t, x_t|H_t$ and that $U_{t,p}(x)$ is a deterministic function w.r.t. p and x conditioned on the history H_t, therefore $\mathbb{E}[U_{t,p^*}(x^*)] = \mathbb{E}[U_{t,p_t}(x_t)]$ follows by the tower rule. We begin by decomposing the Bayesian regret into three terms and show the bounds that we will later obtain for each of them.

$$\mathrm{BR}(T) = \sum_{t\in[T]} \mathbb{E}[f(x^*) - f(x_t)] \tag{74}$$

$$= \sum_{t\in[T]} \mathbb{E}[f(x^*) - U_{t,p^*}(x^*) + U_{t,p_t}(x_t) - f(x_t)] \quad \begin{pmatrix} p^*, x^*|H_t \\ \overset{d}{=} p_t, x_t|H_t \end{pmatrix} \tag{75}$$

$$= \sum_{t\in[T]} \mathbb{E}\Big[f(x^*) - U_{t,p^*}(x^*) + (\sqrt{\beta_t} + \sqrt{\eta_T})\sigma_{t,p_t}(x_t)$$

$$+ \mu_{t,p_t}(x_t) - \sqrt{\eta_T}\sigma_{t,p_t}(x_t) - f(x_t)\Big] \qquad \left(\pm\sqrt{\eta_T}\sigma_{t,p_t}(x_t)\right) \tag{76}$$

$$= \underbrace{\sum_{t\in[T]} \mathbb{E}\left[f(x^*) - U_{t,p^*}(x^*)\right]}_{A_1} + \underbrace{\sum_{t\in[T]} \mathbb{E}\left[(\sqrt{\beta_t} + \sqrt{\eta_T})\sigma_{t,p_t}(x_t)\right]}_{A_2} \tag{77}$$

$$+ \underbrace{\sum_{t\in[T]} \mathbb{E}\left[\mu_{t,p_t}(x_t) - \sqrt{\eta_T}\sigma_{t,p_t}(x_t) - f(x_t)\right]}_{A_3}. \tag{78}$$

$$\leq \underbrace{\frac{\pi^2}{6}}_{A_1} + \underbrace{\sqrt{CT\bar{\gamma}_T(P_1)}(\sqrt{\beta_T}+\sqrt{\eta_T})}_{A_2}$$
$$+ \underbrace{\frac{\sqrt{3}}{T}\left(\sqrt{\sigma^{-2}(\bar{M}+2M\mu_{\max}+\mu_{\max}^2+\sigma^2)}+\sqrt{\bar{M}/T}\right)}_{A_{3,1}}$$
$$+ \underbrace{\sigma\sqrt{14T|P|\log T}}_{A_{3,2}}$$
$$+ \underbrace{|P|\left(\sigma^{-1}\sqrt{T}(M+\mu_{\max}+\sigma)+M+\sigma\sqrt{2\log T}\right)}_{A_{3,2}}. \tag{79}$$

Next, we will prove the bounds for the terms A_1, A_2, and A_3 where the bound for A_3 is given by the sum of $A_{3,1}$ and $A_{3,2}$.

Bounding A_1 Since the upper confidence term in A_1, $U_{t,p^*}(x^*)$, corresponds to the confidence bound of the true prior p^*, the bound for A_1 follows by standard techniques (Russo & Van Roy, 2014):

$$A_1 = \sum_{t\in[T]} \mathbb{E}\left[f(x^*) - U_{t,p^*}(x^*)\right] \tag{80}$$
$$\leq \sum_{t\in[T]} \mathbb{E}\left[\left[f(x^*) - U_{t,p^*}(x^*)\right]_+\right] \qquad ([\cdot]_+ := \max(\cdot,0)) \tag{81}$$
$$\leq \sum_{t\in[T]}\sum_{x\in\mathcal{X}} \mathbb{E}\left[\left[f(x) - U_{t,p^*}(x)\right]_+\right] \qquad (x^*\in\mathcal{X}, [\cdot]_+\geq 0) \tag{82}$$
$$= \sum_{t\in[T]}\sum_{x\in\mathcal{X}} \mathbb{E}_{p^*,H_t}\left[\mathbb{E}_t\left[\left[f(x) - \mu_{t,p^*}(x) - \sqrt{\beta_t}\sigma_{t,p^*}(x)\right]_+ \mid p^*, H_t\right]\right]. \tag{83}$$

Recall that for $Z\sim\mathcal{N}(\mu,\sigma)$ with $\mu\leq 0$, $\mathbb{E}[[Z]_+] \leq \frac{\sigma}{\sqrt{2\pi}}\exp\left(\frac{-\mu^2}{2\sigma^2}\right)$. In our case, note that $f(x)|p^*, H_t \sim \mathcal{N}(\mu_{t,p^*}(x), \sigma^2_{t,p^*}(x))$ and $-\mu_{t,p^*}(x) - \sqrt{\beta_t}\sigma_{t,p^*}(x)$ is deterministic given p^*, H_t. Hence,

$$A_1 \leq \sum_{t\in[T]}\sum_{x\in\mathcal{X}} \mathbb{E}_{p^*,H_t}\left[\frac{\sigma_{t,p^*}(x)}{\sqrt{2\pi}}\exp\left(\frac{-\beta_t}{2}\right)\right] \tag{84}$$
$$\leq \sum_{t\in[T]}\sum_{x\in\mathcal{X}} \mathbb{E}_{p^*,H_t}\left[\frac{1}{\sqrt{2\pi}}\exp\left(\frac{-\beta_t}{2}\right)\right] \qquad (\sigma_{t,p^*}(x)\leq\sigma_{0,p^*}(x)\leq 1) \tag{85}$$
$$= \sum_{t\in[T]}\sum_{x\in\mathcal{X}} \frac{1}{\sqrt{2\pi}}\exp(-\beta_t/2) \tag{86}$$
$$= \sum_{t\in[T]} \frac{1}{t^2} \leq \frac{\pi^2}{6}. \qquad (\beta_t = 2\log(|\mathcal{X}|t^2/\sqrt{2\pi})) \tag{87}$$

Bounding A_2 To bound A_2, we separate it into two terms and apply Cauchy-Schwarz to each term:

$$A_2 = \mathbb{E}\left[\sum_{t\in[T]} \sqrt{\beta_t}\sigma_{t,p_t}(x_t)\right] + \mathbb{E}\left[\sum_{t\in[T]} \sqrt{\eta_T}\sigma_{t,p_t}(x_t)\right] \tag{88}$$

$$\leq \mathbb{E}\left[\sqrt{\sum_{t\in[T]} \beta_t \sum_{t\in[T]} \sigma^2_{t,p_t}(x_t)}\right] + \mathbb{E}\left[\sqrt{\sum_{t\in[T]} \eta_T \sum_{t\in[T]} \sigma^2_{t,p_t}(x_t)}\right] \text{ (CS)} \tag{89}$$

$$\leq \mathbb{E}\left[\sqrt{T\beta_T \sum_{t\in[T]} \sigma^2_{t,p_t}(x_t)}\right] + \mathbb{E}\left[\sqrt{T\eta_T \sum_{t\in[T]} \sigma^2_{t,p_t}(x_t)}\right] \begin{pmatrix}\beta_t \leq \beta_T,\\ \forall t \in [T]\end{pmatrix} \tag{90}$$

$$= \sqrt{T}(\sqrt{\beta_T} + \sqrt{\eta_T})\mathbb{E}\left[\sqrt{\sum_{t\in[T]} \sigma^2_{t,p_t}(x_t)}\right] \tag{91}$$

$$\leq \sqrt{T}(\sqrt{\beta_T} + \sqrt{\eta_T})\sqrt{\mathbb{E}\Big[\sum_{t\in[T]} \sigma^2_{t,p_t}(x_t)\Big]}. \qquad \text{(Jensen's inequality)} \tag{92}$$

By Lemma 4.3, we have that $\mathbb{E}\big[\sum_{t\in[T]} \sigma^2_{t,p_t}(x_t)\big] \leq C\bar{\gamma}_T(P_1)$ and therefore

$$A_2 \leq \sqrt{CT\bar{\gamma}_T(P_1)}\left(\sqrt{\beta_T} + \sqrt{\eta_T}\right). \tag{93}$$

Bounding A_3 We further split A_3 based on whether $p_t \in \mathcal{C}_t$ holds:

$$A_3 = \underbrace{\sum_{t\in[T]} \mathbb{E}\left[(\mu_{t,p_t}(x_t) - \sqrt{\eta_T}\sigma_{t,p_t}(x_t) - f(x_t))\mathbb{1}\{p_t \notin \mathcal{C}_t\}\right]}_{A_{3,1}} \tag{94}$$

$$+ \underbrace{\sum_{t\in[T]} \mathbb{E}\left[(\mu_{t,p_t}(x_t) - \sqrt{\eta_T}\sigma_{t,p_t}(x_t) - f(x_t))\mathbb{1}\{p_t \in \mathcal{C}_t\}\right]}_{A_{3,2}}. \tag{95}$$

Bounding $A_{3,1}$: To bound $A_{3,1}$, we apply the Cauchy-Schwarz inequality for expectations to separate the factors $\mu_{t,p_t}(x_t) - f(x_t)$ and $\mathbb{1}\{p_t \notin \mathcal{C}_t\}$ into

different expectations as follows:

$$A_{3,1} \leq \sum_{t\in[T]} \mathbb{E}\left[(\mu_{t,p_t}(x_t) - f(x_t))\mathbb{1}\{p_t \notin \mathcal{C}_t\}\right] \qquad (\sqrt{\eta_T}\sigma_{t,p_t}(x_t) \geq 0) \quad (96)$$

$$= \sum_{t\in[T]} \left(\mathbb{E}\left[\mu_{t,p_t}(x_t)\mathbb{1}\{p_t \notin \mathcal{C}_t\}\right] + \mathbb{E}\left[-f(x_t)\mathbb{1}\{p_t \notin \mathcal{C}_t\}\right]\right) \quad (97)$$

$$\leq \sum_{t\in[T]} \Bigg(\sqrt{\mathbb{E}\left[(\mu_{t,p_t}(x_t))^2\right]\mathbb{E}\left[(\mathbb{1}\{p_t \notin \mathcal{C}_t\})^2\right]} + \sqrt{\mathbb{E}\left[(f(x_t))^2\right]\mathbb{E}\left[(\mathbb{1}\{p_t \notin \mathcal{C}_t\})^2\right]}\Bigg) \qquad \left(\begin{matrix}\mathbb{E}[XY] \leq \\ \sqrt{\mathbb{E}[X^2]\mathbb{E}[Y^2]}\end{matrix}\right) \quad (98)$$

$$\leq \sum_{t\in[T]} \sqrt{\mathbb{E}\left[\mathbb{1}\{p^* \notin \mathcal{C}_t\}\right]} \left(\sqrt{\mathbb{E}\left[(\mu_{t,p_t}(x_t))^2\right]} + \sqrt{\mathbb{E}\left[\left(\sup_{x\in\mathcal{X}} |f(x)|\right)^2\right]}\right) \quad (99)$$

where the final step uses that $p^*|H_t \stackrel{d}{=} p_t|H_t$. To bound the three expectations above, we have from Lemma B.7 that $\mathbb{E}\left[\mathbb{1}\{p^* \notin \mathcal{C}_t\}\right] \leq 3T^{-5}$ and from Lemma B.11 that $\mathbb{E}\left[(\mu_{t,p_t}(x_t))^2\right] \leq \sigma^{-2}T(\bar{M} + 2M\mu_{\max} + \mu_{\max}^2 + \sigma^2)$ for all $t \in [T]$. Similarly, by Lemma B.12 we have that $\mathbb{E}\left[\left(\sup_{x\in\mathcal{X}} |f(x)|\right)^2\right] \leq \bar{M}$. Put together, we arrive at the following bound for $A_{3,1}$:

$$A_{3,1} \leq \frac{\sqrt{3}}{T}\left(\sqrt{\sigma^{-2}(\bar{M} + 2M\mu_{\max} + \mu_{\max}^2 + \sigma^2)} + \sqrt{\bar{M}/T}\right). \quad (100)$$

Bounding $A_{3,2}$: Then, $A_{3,2}$ can be bound as follows:

$$A_{3,2} = \sum_{t\in[T]} \mathbb{E}\left[(\mu_{t,p_t}(x_t) - \sqrt{\eta_T}\sigma_{t,p_t}(x_t) - f(x_t) - \epsilon_t)\,\mathbb{1}\{p_t \in \mathcal{C}_t\}\right] \qquad (\epsilon_t \perp\!\!\!\perp \mathbb{1}\{p_t \in \mathcal{C}_t\}) \quad (101)$$

$$\leq \sum_{p\in P} \mathbb{E}\Bigg[\sum_{t\in[T]} \mathbb{1}\{p_t = p\}\mathbb{1}\{p \in \mathcal{C}_t\} \cdot (\mu_{t,p}(x_t) - \sqrt{\eta_T}\sigma_{t,p}(x_t) - f(x_t) - \epsilon_t)\Bigg] \quad (102)$$

We define the final time step where prior p is selected and is in the confidence set as $\tau_p := \max\{t \in [T] : p_t = p, p \in \mathcal{C}_t\}$. Then, $\sum_{t\in[\tau_p-1]}(\mu_{t,p_t}(x_t) - \sqrt{\eta_T}\sigma_{t,p_t}(x_t) - f(x_t) - \epsilon_t)\mathbb{1}\{p_t = p\}\mathbb{1}\{p \in \mathcal{C}_t\} = G_{\tau_p}(p)$ since $\mathcal{C}_t$ is a shrinking sequence of sets. By definition of $p \in \mathcal{C}_{\tau_p}$ (Eq. (43)), $G_{\tau_p}(p) \leq \sigma\sqrt{14N_{\tau_p-1}(p)\log T}$ and

$$A_{3,2} \leq \sum_{p\in P} \mathbb{E}\Big[\sigma\sqrt{14N_{\tau_p-1}(p)\log T} + \left(\mu_{\tau_p,p}(x_t) - \sqrt{\eta_T}\sigma_{\tau_p,p}(x_{\tau_p}) - f(x_{\tau_p}) - \epsilon_{\tau_p}\right)\Big] \quad (103)$$

$$\leq \sum_{p\in P} \mathbb{E}\left[\sigma\sqrt{14N_T(p)\log T}\right] + \sum_{p\in P} \mathbb{E}\left[\left(\mu_{\tau_p,p}(x_{\tau_p}) - f(x_{\tau_p}) - \epsilon_{\tau_p}\right)\right] \quad (104)$$

since $\sqrt{\eta_T}\sigma_{t,p}(x) \geq 0$, and $N_t(p) \leq N_T(p)$, $\forall t, p, x \in [T] \times P \times \mathcal{X}$. To bound the left term in Eq. (104), we apply the Cauchy-Schwarz inequality such that $\sum_{p\in P}\sqrt{N_T(p)} \leq \sqrt{T|P|}$. For the right term in Eq. (104), we note that $\tau_p \in [T]$ and consider the maximum:

$$\mathbb{E}\left[\left(\mu_{\tau_p,p}(x_{\tau_p}) - f(x_{\tau_p}) - \epsilon_{\tau_p}\right)\right] \quad (105)$$

$$\leq \sum_{p\in P} \mathbb{E}\left[\max_{t\in[T]} \left(\mu_{t,p}(x_t) - f(x_t) - \epsilon_t\right)\right] \quad (106)$$

$$\leq \sum_{p\in P} \left(\mathbb{E}\left[\max_{t\in[T]} \mu_{t,p}(x_t)\right] + \mathbb{E}\left[\sup_{x\in\mathcal{X}} |f(x)|\right] + \mathbb{E}\left[\max_{t\in[T]} -\epsilon_t\right]\right) \quad (107)$$

$$\leq |P|\left(\sigma^{-1}\sqrt{T}(M + \mu_{\max} + \sigma) + M + \sigma\sqrt{2\log T}\right). \quad \text{(Lemma B.11)} \quad (108)$$

The bound $\mathbb{E}[\max_{t\in[T]} -\epsilon_t] \leq \sigma\sqrt{2\log T}$ follows by standard results for independent zero-mean Gaussians (Boucheron et al., 2013, Section 2.5). Combined, we get that

$$A_{3,2} \leq \sigma\sqrt{14T|P|\log T} + |P|\left(\sigma^{-1}\sqrt{T}(M + \mu_{\max} + \sigma) + M + \sigma\sqrt{2\log T}\right). \quad (109)$$

□

Remark B.10. *Unlike Hong et al. (2022b); Lu et al. (2023), our definition of the confidence set $\mathcal{C}_t$ includes a condition that the excess reward $G_s(p)$ is below the threshold $\xi_s(p)$ for all $s \leq t$, not just $s = t$. This guarantees that the sets are non-increasing in size, and therefore if $p \in \mathcal{C}_t$ then $p \in C_s$ for all $s < t$. Furthermore, $\sum_{s=1}^{t-1}(\mu_{s,p_s}(x_s) - \sqrt{\eta_T}\sigma_{s,p_s}(x_s) - y_s)\mathbb{1}\{p = p_s\}\mathbb{1}\{p \in \mathcal{C}_s\} = G_t(p)$ if $p \in \mathcal{C}_{t-1}$ which is critical to go from Eq. (102) to Eq. (103). Without the time-uniform requirement, $p \notin C_s$ could hold for some $s < t$ s.t. $\mu_{s,p_s}(x_s) - \sqrt{\eta_T}\sigma_{s,p_s}(x_s) - y_s < 0$. Then, $\sum_{s=1}^{t-1}(\mu_{s,p_s}(x_s) - \sqrt{\eta_T}\sigma_{s,p_s}(x_s) - y_s)\mathbb{1}\{p = p_s\}\mathbb{1}\{p \in \mathcal{C}_s\} > G_t(p)$ which prevents bounding $G_t(p)$ by $\xi_t(p)$.*

B.3 Auxiliary lemmas

In this section, we state and prove auxiliary lemmas that bound the expectations of $\mu_{t,p}(x_t)$ and $\sup_{x\in\mathcal{X}} |f(x)|^2$.

Lemma B.11. *Let* $\mu_{max} = \sup_{p,x\in P\times\mathcal{X}} \mu_{1,p}(x)$, $M = \mathbb{E}[\sup_{x\in\mathcal{X}} |f(x)|]$, $M_p = \mathbb{E}[\sup_{x\in\mathcal{X}} |f(x)||p^* = p]$, $M_\Delta = \max_{p\in P} M_p - \min_{p\in P} M_p$, *and* $\bar{M} = M^2 + 1 + \frac{M_\Delta^2}{4}$. *If* $k_p(x,x) : \mathcal{X}\times\mathcal{X} \mapsto [-1,1]$, $\forall p \in P$, *then*

$$\mathbb{E}[\mu_{t,p_t}(x_t)^2] \leq \frac{T}{\sigma^2}(\bar{M} + 2M\mu_{max} + \mu_{max}^2 + \sigma^2), \tag{110}$$

$$\mathbb{E}[\max_{t\in[T]} \mu_{t,p}(x_t)] \leq \frac{\sqrt{T}}{\sigma}(M + \mu_{max} + \sigma). \tag{111}$$

Proof. To begin, recall that $\mu_{t,p}(x) = \mathbf{k}_{t,p}(x)^T \left(\mathbf{K}_{t,p} + \sigma^2 I\right)^{-1} (\mathbf{f}_{1:t-1} + \boldsymbol{\epsilon}_{1:t-1} - \boldsymbol{\mu}_{1:t-1,p})$ where $\mathbf{f}_{1:t} = [f(x_1), \ldots, f(x_{t-1})]^T$, $\boldsymbol{\epsilon}_{1:t} = [\epsilon_1, \ldots, \epsilon_{t-1}]$, and $\boldsymbol{\mu}_{1:t,p} = [\mu_{1,p}(x_1), \ldots, \mu_{1,p}(x_{t-1})]$. Additionally, we note that

$$\left\|\mathbf{k}_{t,p}(x)^T \left(\mathbf{K}_{t,p} + \sigma^2 I\right)^{-\frac{1}{2}}\right\|_2^2 = \mathbf{k}_{t,p}(x)^T \left(\mathbf{K}_{t,p} + \sigma^2 I\right)^{-1} \mathbf{k}_{t,p}(x) \leq k_p(x,x) \leq 1, \tag{112}$$

$\forall t, p, x \in [T]\times P\times\mathcal{X}$ by the definition of the posterior variance $\sigma_{t,p}^2(x)$ and since the posterior variance is non-negative $\sigma_{t,p}^2(x) \geq 0$. Similarly, note that $\left\|\left(\mathbf{K}_{t,p} + \sigma^2 I\right)^{-\frac{1}{2}}\right\|_2 \leq \sigma^{-1}$ since $\mathbf{K}_{t,p}$ is positive semi-definite for any t and p. Therefore,

$$\begin{aligned}\mu_{t,p}(x) &\leq \left\|\mathbf{k}_{t,p}(x)^T \left(\mathbf{K}_{t,p} + \sigma^2 I\right)^{-\frac{1}{2}}\right\|_2 \cdot \left\|\left(\mathbf{K}_{t,p} + \sigma^2 I\right)^{-\frac{1}{2}}\right\|_2 \\ &\qquad \cdot \|\mathbf{f}_{1:t} + \boldsymbol{\epsilon}_{1:t} - \boldsymbol{\mu}_{1:t,p}\|_2 &(113)\\ &\leq \frac{1}{\sigma}\|\mathbf{f}_{1:t} + \boldsymbol{\epsilon}_{1:t} - \boldsymbol{\mu}_{1:t,p}\|_2. &(114)\end{aligned}$$

Then, we bound $\mathbb{E}[\max_{t\in[T]} \mu_{t,p}(x_t)]$.

$$\mathbb{E}[\max_{t\in[T]} \mu_{t,p}(x_t)] \leq \frac{1}{\sigma}\left(\mathbb{E}\left[\max_{t\in[T]}\|\mathbf{f}_{1:t}\|_2 + \max_{t\in[T]}\|\boldsymbol{\epsilon}_{1:t}\|_2 + \max_{t\in[T]}\|\boldsymbol{\mu}_{1:t,p}\|_2\right]\right) \tag{115}$$

$$\leq \frac{1}{\sigma}\mathbb{E}\left[\max_{t\in[T]}\sqrt{\sum_{s=1}^t f(x_s)^2} + \max_{t\in[T]}\sqrt{\sum_{s=1}^t \epsilon_s^2} + \max_{t\in[T]}\sqrt{\sum_{s=1}^t \mu_{1,p}^2(x_s)}\right] \tag{116}$$

$$\leq \frac{1}{\sigma}\left(\mathbb{E}\left[\sqrt{T\sup_{x\in\mathcal{X}} f(x)^2}\right] + \mathbb{E}\left[\sqrt{\sum_{t=1}^T \epsilon_t^2}\right] + \sqrt{T\sup_{x\in\mathcal{X}} \mu_{1,p}^2(x)}\right) \tag{117}$$

$$\leq \frac{1}{\sigma}\left(\sqrt{T}\mathbb{E}\left[\sup_{x\in\mathcal{X}}|f(x)|\right] + \sqrt{\mathbb{E}\left[\sum_{t=1}^T \epsilon_t^2\right]} + \sqrt{T}\mu_{\max}\right) \text{ (Jensen's ineq.)} \tag{118}$$

$$\leq \frac{\sqrt{T}}{\sigma}\left(M + \sigma + \mu_{\max}\right). \qquad \left(\textstyle\sum_{t\in[T]} \epsilon_t^2 \sim \sigma^2\chi_T^2\right) \tag{119}$$

Similarly, we bound $\mathbb{E}[\mu_{t,p}(x_t)^2]$:

$$\mathbb{E}[\mu_{t,p_t}(x_t)^2] \leq \frac{1}{\sigma^2}\mathbb{E}\left[\|\mathbf{f}_{1:t} + \boldsymbol{\epsilon}_{1:t} - \boldsymbol{\mu}_{1:t,p_t}\|_2^2\right] \tag{120}$$

$$\leq \frac{1}{\sigma^2}\mathbb{E}\left[\sum_{t\in[T]}(f(x_t) + \epsilon_t - \mu_{1,p_t}(x_t))^2\right] \tag{121}$$

$$= \frac{1}{\sigma^2}\mathbb{E}\Bigg[\sum_{t\in[T]} f(x_t)^2 + \epsilon_t^2 + \mu_{1,p_t}^2(x_t) + 2(f(x_t)\epsilon_t - f(x_t)\mu_{1,p_t}(x_t) - \epsilon_t\mu_{1,p_t}(x_t))\Bigg] \tag{122}$$

$$= \frac{1}{\sigma^2}\mathbb{E}\left[\sum_{t\in[T]} f(x_t)^2 + \epsilon_t^2 + \mu_{1,p_t}^2(x_t) - 2f(x_t)\mu_{1,p_t}(x_t)\right] \quad (f(x_t), \mu_{1,p_t}(x_t) \perp\!\!\!\perp \epsilon_t, \text{ and } \mathbb{E}[\epsilon_t] = 0) \tag{123}$$

$$\leq \frac{1}{\sigma^2}\mathbb{E}\Bigg[\sum_{t\in[T]} \sup_{x\in\mathcal{X}} f(x)^2 + \epsilon_t^2 + \mu_{1,p_t}^2(x_t) + 2\sup_{x\in\mathcal{X}}|f(x)||\mu_{1,p_t}(x)|\Bigg] \tag{124}$$

$$\leq \frac{1}{\sigma^2}T\left(\mathbb{E}\left[\left(\sup_{x\in\mathcal{X}}|f(x)|\right)^2\right] + \sigma^2 + \mu_{\max}^2 + 2M\mu_{\max}\right) \quad \left(\textstyle\sum_{t\in[T]}\epsilon_t^2 \sim \sigma^2\chi_T^2\right) \tag{125}$$

$$\leq \frac{1}{\sigma^2}T\left(\bar{M} + \sigma^2 + \mu_{\max}^2 + 2M\mu_{\max}\right). \quad \text{(Lemma B.12)} \tag{126}$$

□

Lemma B.12. *Let* $M = \mathbb{E}[\sup_{x\in\mathcal{X}}|f(x)|]$, $M_p = \mathbb{E}[\sup_{x\in\mathcal{X}}|f(x)|\big|p^* = p]$, *and* $M_\Delta = \max_{p\in P} M_p - \min_{p\in P} M_p$. *If* $k_p(x,x) : \mathcal{X}\times\mathcal{X} \mapsto [-1,1]$, $\forall p \in P$, *then*

$$\mathbb{E}\left[\left(\sup_{x\in\mathcal{X}}|f(x)|\right)^2\right] \leq M^2 + 1 + \frac{M_\Delta^2}{4} =: \bar{M}. \tag{127}$$

Proof. First, by the variance formula $\mathbb{V}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$,

$$\mathbb{E}[(\sup_{x\in\mathcal{X}}|f(x)|)^2] = M^2 + \mathbb{V}\left(\sup_{x\in\mathcal{X}}|f(x)|\right). \tag{128}$$

The variance $\mathbb{V}[\sup_{x\in\mathcal{X}}|f(x)|]$ can be bounded by the law of total variance as

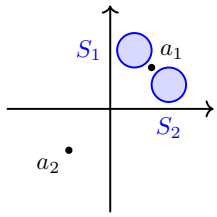


Figure 7: Potential counterexample to Eq (9) in Hong et al. (2022b). The blue regions represent when the event E_0 holds and the black dots represent the two arms a_1 and a_2.

follows:

$$\mathbb{V}\left[\sup_{x\in\mathcal{X}}|f(x)|\right] = \mathbb{E}_{p^*}\left[\mathbb{V}\left(\sup_{x\in\mathcal{X}}|f(x)|\big|p^*\right)\right] + \mathbb{V}_{p^*}\left(\underbrace{\mathbb{E}\left[\sup_{x\in\mathcal{X}}|f(x)|\big|p^*\right]}_{M_{p^*}:=}\right) \tag{129}$$

$$\overset{(a)}{\leq} \mathbb{E}_{p^*}\left[\sup_{x\in\mathcal{X}}\sigma^2_{1,p^*}(x)\right] + \mathbb{V}_{p^*}(M_{p^*}) \tag{130}$$

$$\leq 1 + \frac{(\max_p M_p - \min_p M_p)^2}{4} \tag{131}$$

where the final step follows by $\sigma^2_{1,p}(x) \leq 1$ and Popoviciu's inequality. Note that $\mathbb{V}\left(\sup_{x\in\mathcal{X}}|f(x)|\,\middle|\,p^*\right) \leq \sup_{x\in\mathcal{X}}\sigma^2_{1,p^*}(x)$, used in (a), follows from the Gaussian Poincaré inequality applied to $\sup_{x\in\mathcal{X}}|f(x)|$, see Boucheron et al. (2013, Theorem 3.20 and Exercise 3.24). □

C Technical issues with MixTS regret bound in the linear setting

Theorem 1 in Hong et al. (2022b) provides a regret bound for MixTS in the linear setting. The linear setting assumes that the true parameter $\theta^*|S_* \sim \mathcal{N}(\theta_{0,S_*}, \Sigma_{0,S_*})$ where the latent state S_* is sampled from a discrete prior P_1. The proof of Theorem 1 in Hong et al. (2022b) contains non-obvious steps that seem difficult to motivate. We use the notation of Hong et al. (2022b).

First, Eq. (9) in Hong et al. (2022b) uses the TS property that the true prior and optimal arm is equal in distribution to the selected prior and selected arm given the history: $A^\top_{t,*}\bar{\theta}_{t,S_*}|H_t \overset{d}{=} A_t^\top\bar{\theta}_{t,S_t}|H_t$ (equivalent to $\mu_{t,p^*}(x^*)|H_t \overset{d}{=} \mu_{t,p_t}(x_t)|H_t$ in our notation). However, Eq. (9) additionally conditions on the event $E_0 = \{\|\theta_* - \theta_{0,S_*}\|_{\Sigma^{-1}_{0,S_*}} \leq \sqrt{2d\log(dn)}\}$ where θ_* lies close to its prior mean. The TS property does not hold under this event since it modifies the distribution of the linear parameter θ^* but not the sampled parameters θ_t, thus changing the distribution of the optimal arm $A_{t,*}$ but not the selected arm

A_t. Consider the example in Fig. 7, if E_0 holds then θ_* lies in the blue regions and thus a_2 is optimal w.p. 0. If E_0^c holds, then a_2 is optimal with a non-zero probability. However, MixTS is oblivious to E_0 given the history and thus $A_{t,*}|H_t, E_0 \stackrel{d}{\neq} A_t|H_t$. This counterexample illustrates the overall idea but we have not validated that the scale of the arms and the blue regions are feasible.

Second, five lines above Eq. (9) in Hong et al. (2022b), it is stated that the *regret* is upper-bounded by a constant M whenever E_0 occurs. However, from Eq (9) to the first term in step 3 of their analysis (page 15), the bound of M is applied implicitly to $A_t^\top \bar{\theta}_{t,S_t} - A_t^\top \theta_*$ without motivation. For the setting with bounded rewards, then $\bar{\theta}_{t,S_t}$ is also bounded but for Gaussian rewards $\bar{\theta}_{t,S_t}$ can be unbounded.

Third, the second term in Eq. (9) contains the indicator function $\mathbf{1}\{E_0\}$: $\mathbb{E}[(A_t^\top \bar{\theta}_{t,S_t} - A_t^\top \theta_*)\mathbf{1}\{E_0\}]$. In step 3 (page 15), this indicator function is dropped without motivation: $\mathbb{E}[\langle A_t^\top \bar{\theta}_{t,S_t} - A_t^\top \theta_* \rangle_M]$ where $\langle \cdot \rangle_M = \min(\cdot, M)$ for the bound M. If the expression inside is non-negative w.p. 1, then this step would be valid but this is not the case.

Fourth, from our understanding, the final equation on page 15 adds and subtracts the confidence bound and adds a zero-mean Gaussian inside a minimum. However, adding a zero-mean Gaussian inside a minimum reduces the expectation but the analysis seems to assume that it would increase the expectation. I.e. it is seemingly assumed that $\mathbb{E}[\min(M, X)] \leq \mathbb{E}[\min(M, X + \epsilon_t)$ for a constant M and random variable X. However, the reverse inequality is true.

D Description of kernels

The RBF kernel, $k(x, \tilde{x}) = \exp(-||x - \tilde{x}||^2/\ell^2)$ guarantees that f is smooth. The lengthscale parameter $\ell > 0$ determines how quickly f changes, smaller values lead to more fluctuations. The rational quadratic (RQ) kernel $k(x, \tilde{x}) = \left(1 + \frac{||x-\tilde{x}||^2}{2\alpha\ell^2}\right)^{-\alpha}$ where $\alpha > 0$ is a mixture of RBF kernels with varying lengthscales. The Matérn kernel (Matérn, 1986) $k(x, \tilde{x}) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}||x-\tilde{x}||}{\ell}\right)^\nu \cdot K_\nu \left(\frac{\sqrt{2\nu}||x-\tilde{x}||}{\ell}\right)$ where $\nu > 0$ is the smoothness parameter that imposes that f is k-times differentiable if $\nu > k$ for integer k. The functions $\Gamma(\nu)$ and K_v correspond to the gamma function and a modified Bessel function (Williams & Rasmussen, 2006). The periodic kernel $k(x, \tilde{x}) = \exp\Big(-\frac{1}{2}\sum_{i=1}^d \sin^2\big(\frac{\pi}{\rho}(x_i - \tilde{x}_i)\big)/\ell\Big)$ generates smooth and periodic functions with period $\rho > 0$ (Mackay, 1998). The linear kernel $k(x, \tilde{x}) = vx^\top \tilde{x}$ generates linear functions where v is the variance parameter.

E Additional experimental details

In this section, we provide some additional details about the experiments. All experiments were run on a compute cluster with a mix of GPUs (Nvidia A100, A40, T4 and V100). The GPU used was decided based on availability at the

time and no implementation depends on a specific GPU. The algorithms were run in parallel in a single job for each seed. Each job in the synthetic and real-world data experiments ran for approximately 5 minutes. With the 500 seeds, this leads to a combined 250 GPU-hours. Running all algorithms for one seed in the lengthscale scaling experiment with $|P| = 128$ took approximately 40 minutes, and is in total equivalent to around 330 GPU-hours.

E.1 Synthetic experiments

For the kernel experiment, all kernels use a lengthscale of 1.0 and are scaled s.t. $k(x, \tilde{x}) \leq 1$. In addition, the mean function for all priors is zero everywhere. For the subspace experiment, the total dimensions $d = 16$ but each prior p_i assumes $f(x)$ depends on $d_s = 4$ subdimensions: $[i, i+1, i+2, i+3]$ for $i \in [5]$. Dimensions larger than 5 are wrapped around 1, i.e. $((j-1) \mod 5)+1$, such that the priors are equally difficult to distinguish and optimize. The prior elimination methods use $\delta = 0.05$ across all experiments, including the oracle methods. During every iteration t, SCoreBO samples M priors from the hyperposterior P_t and samples N optimizers x^*, f^* for each prior sampled through posterior sampling. In all experiments, we use $M = 16$ and $N = 12$ for SCoreBO. While our M value matches that of Hvarfner et al. (2023, Table 3), we increase the N value from 8 to 12. We use the implementation of the SCoreBO acquisition function in BoTorch (Community) (Balandat et al., 2020). To make the implementation fast with GPUs, we set `linear_operator.settings.stable_qr_cpu_threshold` to 8 in order to avoid QR-factorization being performed on CPU (Gardner et al., 2018; Pleiss et al., 2022, 2025). To avoid out of memory issues, we replace the default `torch.matmul` in `DefaultPredictionStrategy._exact_predictive_covar_inv_quad_form_root` (from `gpytorch.models.exact_prediction_strategies`) with an equivalent `torch.einsum` (Gardner et al., 2018). Since the priors in our experiments are discrete, we compute the hyperposterior exactly and sample from it directly. Similarly, the expectation with respect to the hyperposterior is computed exactly for EEI.

E.2 Real-world data experiments

As discussed in Section 5, each dataset is split into a training and test set. The training sets are split into separate buckets to define our priors. For each bucket p, we compute the empirical mean $\hat{\mu}_p$ and covariance $\hat{\Sigma}_p$ which defines the prior $\mathcal{GP}(\hat{\mu}_p, \hat{\Sigma}_p)$. The buckets in the Intel data corresponds to the 12 days in the training dataset. For the PeMS data, each hour between 06:00 and 13:00 defines one bucket, giving 7 priors. For the daily precipitation data, each month in the year constitutes a bucket, yielding 12 priors. When running the experiments, we select a measurement of all sensors from the test data uniformly at random. The selected measurements correspond to the unknown function $f(x)$ where x is the sensor index and the goal is then to identify sensors measuring large temperatures, small speeds or high precipitation respectively for the three datasets. When the algorithms select an arm to evaluate, we add Gaussian noise with variance σ^2 around 5% of the signal variance, similar to

Srinivas et al. (2012); Bogunovic et al. (2016).

For all the real-world datasets, sensors containing any null measurements are filtered out.

The Intel Berkeley dataset consists of measurements from 46 temperature sensors across 19 days. The training set consists of the first 12 days of measurements and the remaining 7 days constitute the test set. The noise variance is set to $\sigma^2 = 0.7^2$.

The PeMS data is considered in the public domain (California Department of Transportation, 2026) and consists of measurements from 211 sensors along the I-880 highway from all of 2023. The goal is to find the sensors with low speeds to identify congestions. We negate the speed values to obtain a maximization problem. We use the 5-min averages provided by PeMS. Data between 2023-01-01 and 2023-09-01 is put into the training set whilst the data until 2023-12-31 is put into the test set. The noise variance is set to $\sigma^2 = 2.25^2$.

The PNW precipitation data consists of daily precipitation data from 1949 to 1994 across 167 50×50 km regions in the Pacific Northwest. The goal is to find the region with the highest precipitation for any given day. The training data consists of the measurements made prior to 1980 and the test data consists of the measurements between 1980 and 1994. The original data is stated to be given in mm/day however the data seems to be off by a factor of 10. We rescale the data to a log-scale using $\log(\cdot/10 + 0.1)$, similar to Krause et al. (2008). The noise variance is set to $\sigma^2 = 0.41^2$.

In the Intel experiment, we removed one outlier seed. All methods had a final cumulative regret around 6000°C on this instance, note that the average for the worst performing model across the other seeds was $\approx$ 250°C. The outlier is shown in Fig. 8. We can see that one of the sensors display very high temperatures compared to all other sensors, which is why all methods performed poorly on this seed. It should be noted that many of the sensors in the Intel data logged degrees above 100°C after a certain time - likely due to sensor failure rather than boiling temperatures in an office environment. Also note that these days were excluded from both our training and test data. The outlier could be an indication that this particular sensor was starting to fail earlier than others.

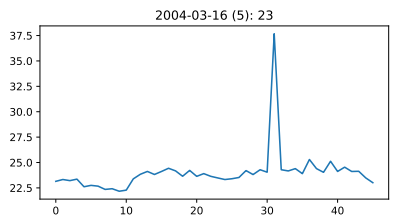


Figure 8: Removed sample from the test data in the Intel experiment. One of the sensors displays very high temperatures.

Table 2: Average total regret and ± 1 standard error for the synthetic and real-world data experiments. The algorithms with the lowest regret (excluding oracle algorithms) are highlighted in bold.

Algorithm	Synthetic			Real-world data		
	Kernel	Lengthscale	Subspace	Intel	PeMS	PNW Precip.
MAP GP-TS	84.3 ± 8.4	**30.2 ± 1.2**	**87.2 ± 1.0**	73.8 ± 7.7	1635.0 ± 129.3	**178.4 ± 6.9**
HP-GP-TS	**39.2 ± 1.4**	**31.4 ± 1.0**	**88.3 ± 0.9**	**54.1 ± 3.0**	**1327.8 ± 107.9**	**167.7 ± 5.2**
PE-GP-TS	62.0 ± 0.6	61.8 ± 0.5	177.1 ± 1.4	106.5 ± 2.1	**1214.2 ± 81.5**	200.9 ± 4.0
PE-GP-UCB	121.6 ± 1.2	114.2 ± 0.6	389.0 ± 1.5	173.0 ± 2.7	2159.2 ± 48.4	506.2 ± 2.6
Oracle GP-TS	35.0 ± 1.1	28.1 ± 0.8	86.0 ± 1.0			
Oracle GP-UCB	68.5 ± 1.9	48.3 ± 1.2	217.3 ± 1.0			
SCoreBO	180.4 ± 7.7	180.8 ± 5.8	106.6 ± 0.9	256.8 ± 9.3	3460.2 ± 163.7	861.3 ± 21.0
EEI	**39.0 ± 2.6**	**30.1 ± 2.1**	**88.3 ± 4.2**	**51.6 ± 4.8**	1664.1 ± 137.7	196.5 ± 12.3

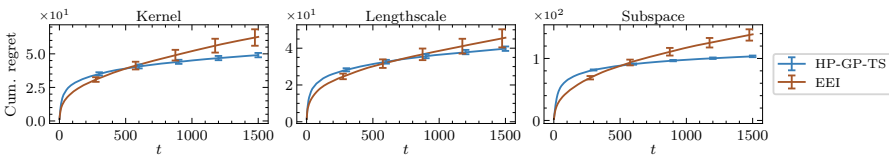


Figure 9: Cumulative regret for synthetic experiments extended time horizon $T = 1500$ with varying kernel (left), lengthscale (center) and mean function (right). Errorbars correspond to ±1 standard error.

F Additional experimental results

In this section, we provide some additional experimental results.

First, we provide the average total regret for the synthetic and real-world data experiments in Table 2. We observe that HP-GP-TS either has the lowest regret or is within 1 standard error of the algorithm with the lowest regret across all the experiments. In Fig. 2, it can be noted that EEI had low regret early in the synthetic experiment but HP-GP-TS either catches up or almost catches up later in the experiments. We compare both algorithms with an extended time horizon of $T = 1500$, the results are shown in Fig. 9 and Table 3. With the extended time horizon, HP-GP-TS achieves the lowest regret across all synthetic experiments. Although, EEI is still within 1 standard error on the lengthscale experiment.

Next, we include the mean number of priors in P_t for all experiments in Fig. 10. Similarly, we include the average entropy of the hyperposterior for

Table 3: Average total regret and ± 1 standard error for the synthetic experiments with longer horizon $T = 1500$. The algorithms with the lowest regret (excluding oracle algorithms) are highlighted in bold.

Algorithm	Kernel	Lengthscale	Subspace
HP-GP-TS	**49.1 ± 1.6**	**39.7 ± 1.2**	**103.4 ± 1.3**
EEI	62.6 ± 6.2	**45.7 ± 5.0**	138.9 ± 9.2

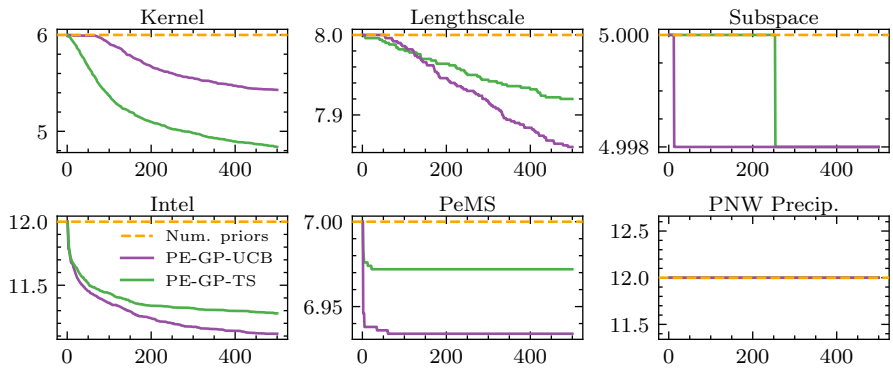


Figure 10: Mean number of priors remaining in P_t over time for PE-GP-UCB and -TS.

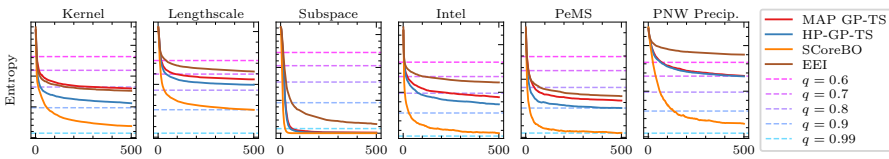


Figure 11: Average entropy in the hyperposterior P_t over time for HP- and MAP GP-TS. The dashed reference values correspond to entropies of discrete distributions with prob. q on one choice and prob. $\frac{1-q}{|P|-1}$ on the other $|P|-1$ choices.

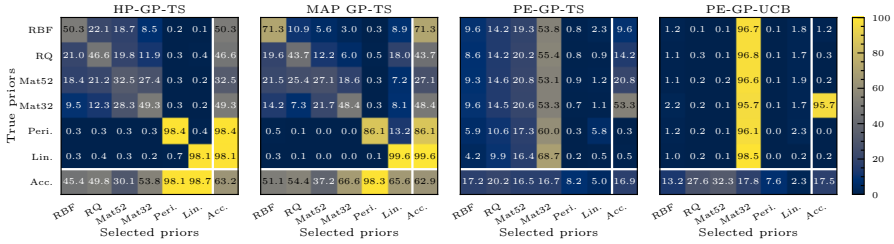
Table 4: Average total regret and ±1 standard error for the lengthscale experiment as $|P|$ increases. The algorithms with the lowest regret (excluding oracle algorithms) are highlighted in bold.

Algorithm	Lengthscales, $\|P\|$				
	8	16	32	64	128
MAP GP-TS	**30.2 ± 1.2**	**32.4 ± 2.5**	**32.5 ± 2.1**	**28.7 ± 1.1**	**30.8 ± 1.9**
HP-GP-TS	**31.4 ± 1.0**	**31.7 ± 0.9**	**30.8 ± 0.8**	**30.7 ± 1.0**	**31.0 ± 1.4**
PE-GP-TS	61.8 ± 0.5	61.3 ± 0.5	62.2 ± 0.5	62.4 ± 0.4	64.3 ± 0.4
PE-GP-UCB	114.2 ± 0.6	114.8 ± 0.6	115.5 ± 0.6	114.5 ± 0.6	114.8 ± 0.6
Oracle GP-TS	28.1 ± 0.8	26.4 ± 0.8	27.3 ± 0.8	26.5 ± 0.7	25.7 ± 0.7
Oracle GP-UCB	48.3 ± 1.2	46.9 ± 1.1	48.4 ± 1.1	46.5 ± 1.0	45.6 ± 1.0
SCoreBO	180.8 ± 5.8	240.3 ± 6.3	277.2 ± 6.8	281.6 ± 6.7	283.9 ± 6.8
EEI	**30.1 ± 2.1**	**30.9 ± 2.2**	**29.4 ± 2.2**	**30.8 ± 2.6**	**32.1 ± 2.8**

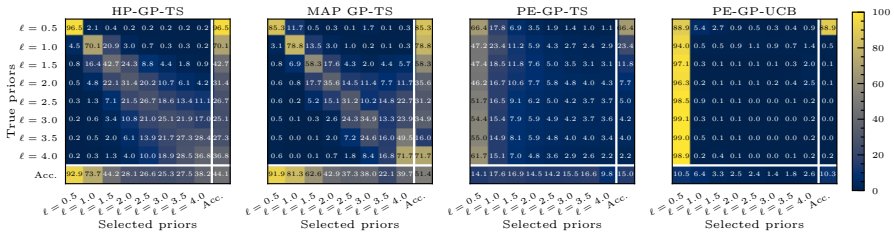
all experiments in Fig. 11. For the lengthscale, subspace, PeMS and PNW precipitation experiments, hardly any priors are eliminated. In contrast, the hyperposterior entropy concentrates rapidly across all experiments with the subspace and PNW precipitation having the most and least concentrated hyperposteriors.

We include the full set of confusion matrices for the lengthscale and subspace experiments in Fig. 12. In the lengthscale experiments, we observe that PE-GP-UCB and -TS oversample the shortest lengthscale. This is similar to the kernel experiment where the Matérn 3/2 kernel was also oversampled. However, we see that HP-GP-TS and MAP GP-TS do not suffer from this optimistic bias. In the subspace experiment, HP- and MAP GP-TS have an accuracy of around 96% whereas PE-GP-TS and -UCB have accuracies 30% and 36% respectively. Even though PE-GP-UCB has a higher accuracy than PE-GP-TS, it still has significantly higher regret. Additionally, the priors are equivalent up to coordinate permutations and therefore generate functions that are equally difficult to optimize. Unlike the kernel and lengthscale experiments, the PE-methods do not oversample any specific prior but commit too much time to exploring along the irrelevant dimensions.

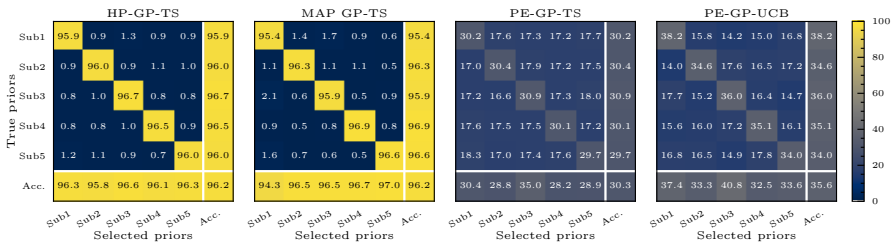
In Tables 4 and 5, the total regret for the lengthscale and subspace scaling experiments are shown.



(a) Kernel experiment



(b) Lengthscale experiment



(c) Subspace experiment

Figure 12: Confusion matrices for the true prior p^* and p_t across all time steps of the synthetic experiments.

Table 5: Average total regret and ± 1 standard error for the subspace experiment as $|P|$ increases. The algorithms with the lowest regret (excluding oracle algorithms) are highlighted in bold.

Algorithm	Subspaces, $\lvert P\rvert$			
	5	8	12	16
MAP GP-TS	**87.2 ± 1.0**	89.9 ± 1.1	89.1 ± 0.9	90.9 ± 1.2
HP-GP-TS	**88.3 ± 0.9**	88.8 ± 0.9	89.5 ± 0.9	90.8 ± 0.9
PE-GP-TS	177.1 ± 1.4	269.5 ± 1.9	344.7 ± 2.3	396.9 ± 2.5
PE-GP-UCB	389.0 ± 1.5	526.0 ± 1.8	622.4 ± 2.3	688.0 ± 2.7
Oracle GP-TS	86.0 ± 1.0	84.1 ± 0.9	84.6 ± 1.0	84.8 ± 1.0
Oracle GP-UCB	217.3 ± 1.0	218.2 ± 1.0	218.6 ± 1.0	218.9 ± 0.9
SCoreBO	106.6 ± 0.9	108.2 ± 0.8	108.9 ± 0.7	109.5 ± 0.7
EEI	**88.3 ± 4.2**	**81.3 ± 3.8**	**82.5 ± 3.6**	**81.3 ± 3.8**