Regret Bounds for Safe Gaussian Process Bandit Optimization

Sanae Amani
UCLA, samani@ucla.edu

Mahnoosh Alizadeh
UCSB, alizadeh@ucsb.edu

Christos Thrampoulidis
UBC, cthrampo@ece.ubc.ca

***Abstract*—Many applications require a learner to make sequential decisions given uncertainty regarding both the system's payoff function and safety constraints. In safety-critical systems, it is paramount that the learner's actions do not violate the safety constraints at any stage of the learning process. In this paper, we study a stochastic bandit optimization problem where the unknown payoff and constraint functions are sampled from Gaussian Processes (GPs) first considered in [1]. We develop a safe variant of GP-UCB called SGP-UCB, with necessary modifications to respect safety constraints at every round. The algorithm has two distinct phases. The first phase seeks to estimate the set of safe actions in the decision set, while the second phase follows the GP-UCB decision rule. Our main contribution is to derive the first sub-linear regret bounds for this problem. We numerically compare SGP-UCB against existing safe Bayesian GP optimization algorithms.**

A full version referred to as supplementary material (SM) of this paper is accessible at: https://arxiv.org/pdf/2005.01936.pdf

I. Introduction

The application of stochastic bandit optimization in safety-critical systems requires the learner to select actions that satisfy a number of *unknown* safety constraints at each round. This setting has recently found many applications in medical trials and robotics, e.g., [2], [3], [4], [5]. In this paper, we consider a stochastic bandit optimization problem where both the reward function f and the constraint function g are samples from Gaussian Processes. We require that the learner's chosen actions respect safety constraints at every round in spite of uncertainty about safe actions. This setting was first studied in [6] in the specific case of a single safety constraint of the form $f(\mathbf{x}) \geq h$. The proposed safe algorithm in [6] guarantees that the system's performance never falls below a critical value; that is, safety is defined based on the reward function. Later, in [7], the authors studied the more general case of $g(\mathbf{x}) \geq h$ as adopted in our paper. The reason for this generalization is that coupling the system's performance and safety requirements is often not desirable in applications such as robotics. For example, high-gain controllers might have great performance by achieving low average tracking error, however, they can overshoot and violate input constraints which is not desirable [5].

In the presence of safety constraints, the learner hopes to overcome the two-fold challenge of keeping the cumulative regret as small as possible while ensuring that selected actions respect the safety constraints at each round of the algorithm. We present SGP-UCB, which is a safety-constrained variant of GP-UCB proposed by [1]. To ensure constraint satisfaction, SGP-UCB restricts the learner to choose actions from a conservative inner-approximation of the safe decision set that is known to satisfy safety constraints with high probability given the algorithm's history. The cumulative regret bound of our proposed algorithm (given in Section III as our main theoretical result) implies that SGP-UCB is a *no-regret* algorithm. This is the main difference of our results compared to the algorithms studied in [6], [7] that only come with convergence-but, no regret- guarantees.

Notation. $\|\mathbf{x}\|_2$ denotes the Euclidean norm of a vector $\mathbf{x}$. The weighted 2-norm of a vector $\mathbf{v} \in \mathbb{R}^d$ with respect to $A \in \mathbb{R}^{d\times d}$ is defined by $\|\mathbf{v}\|_{\mathbf{A}} = \sqrt{\mathbf{v}^T\mathbf{A}\mathbf{v}}$. We denote the minimum and maximum eigenvalue of A by $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$. The maximum of two numbers α, β is denoted $\alpha \vee \beta$. For a positive integer n, $[n]$ denotes the set $\{1, 2, \ldots, n\}$.

A. Problem Statement

The learner is given a finite decision set $\mathcal{D}_0 \subset \mathbb{R}^d$. At each round t, she chooses an action $\mathbf{x}_t \in \mathcal{D}_0$ and observes a noise-perturbed value of an unknown reward function $f : \mathcal{D}_0 \to \mathbb{R}$, i.e. $y_t := f(\mathbf{x}_t) + \eta_t$. At every round, the learner must ensure that the chosen action $\mathbf{x}_t$ satisfies the following safety constraint:

$$g(\mathbf{x}_t) \geq h, \tag{1}$$

where $g : \mathcal{D}_0 \to \mathbb{R}$ is an unknown function and h is a known constant.[1] We define the safe set from which the learner is allowed to take action as:

$$\mathcal{D}_0^{\mathrm{S}} := \{\mathbf{x} \in \mathcal{D}_0 \ : \ g(\mathbf{x}) \geq h\}. \tag{2}$$

Since g is unknown, the learner cannot identify $\mathcal{D}_0^{\mathrm{S}}$. As such, the best she can do is to choose actions $\mathbf{x}_t$ that are in $\mathcal{D}_0^{\mathrm{S}}$ *with high probability*. We assume that at every round, the learner also receives noise-perturbed feedback on the safety constraint, i.e. $z_t := g(\mathbf{x}_t) + \zeta_t$.

Goal. Since our knowledge of g comes from noisy observations, we are not able to fully identify the true safe set $\mathcal{D}_0^{\mathrm{S}}$ and infer $g(\mathbf{x})$ exactly, but only up to some statistical confidence $g(\mathbf{x}) \pm \epsilon$ for some $\epsilon > 0$. Hence, we consider the optimal action through an ϵ-reachable safe set for some $\epsilon > 0$:

$$\mathcal{D}_\epsilon^{\mathrm{S}} := \{\mathbf{x} \in \mathcal{D}_0 : g(\mathbf{x}) \geq h + \epsilon\}, \tag{3}$$

[1] Our results can be simply extended to the settings with several safety constraints, i.e., set of g_i's and h_i's, however, for the sake of brevity we focus on one constraint function.

as our benchmark. A natural performance metric in this context is *cumulative pseudo-regret* [8] over the course of T rounds, which is defined by $R_T = \sum_{t=1}^{T} f(\mathbf{x}_\epsilon^*) - f(\mathbf{x}_t)$, where $\mathbf{x}_\epsilon^*$ is the optimal *safe* action that maximizes the reward in expectation over the $\mathcal{D}_\epsilon^{\mathrm{S}}$, i.e., $\mathbf{x}_\epsilon^* \in \arg\max_{\mathbf{x}\in\mathcal{D}_\epsilon^{\mathrm{S}}} f(\mathbf{x})$. For the rest of this paper, we simply use regret to refer to the pseudo-regret R_T and drop the subscript ϵ from $\mathbf{x}_\epsilon^*$. The goal of the learner is to follow a no-regret algorithm, i.e. such that $R_T/T \to 0$ as T grows, while ensuring all actions she chooses are safe with high probability over the the entire time horizon $[T]$.

Regularity Assumptions. The above specified goal cannot be achieved unless certain assumptions are made on f and g. We model the reward function f and the constraint function g as a sample from a Gaussian Process (GP) [9]. We now present necessary standard terminology and notations on GPs. A $GP(\mu(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$ is a probability distribution across a class of smooth functions, which is parameterized by a kernel function $k(\mathbf{x}, \mathbf{x}')$ that characterizes the smoothness of the function. The Bayesian algorithm we analyze uses $GP(0, k_f(\mathbf{x}, \mathbf{x}'))$ and $GP(0, k_g(\mathbf{x}, \mathbf{x}'))$ as prior distributions over f and g, respectively, where k_f and k_g are positive semi-definite kernel functions. Moreover, we assume bounded variance $k_f(\mathbf{x}, \mathbf{x}) \leq 1$ and $k_g(\mathbf{x}, \mathbf{x}) \leq 1$. For a noisy sample $\mathbf{y}_t = [y_1, \ldots, y_t]^T$, with i.i.d Gaussian noise $\eta_t \sim \mathcal{N}(0, \sigma^2)$ the posterior over f is also a GP with the mean $\mu_{f,t}(\mathbf{x})$ and variance $\sigma_{f,t}^2(\mathbf{x})$:

$$\mu_{f,t}(\mathbf{x}) = \mathbf{k}_{f,t}(\mathbf{x})^T(\mathbf{K}_{f,t} + \sigma^2\mathbf{I})^{-1}\mathbf{y}_t$$
$$\sigma_{f,t}^2(\mathbf{x}) = k_{f,t}(\mathbf{x}, \mathbf{x})$$

where $k_{f,t}(\mathbf{x}, \mathbf{x}') = k_f(\mathbf{x}, \mathbf{x}') - \mathbf{k}_{f,t}(\mathbf{x})^T(\mathbf{K}_{f,t} + \sigma^2\mathbf{I})^{-1}\mathbf{k}_{f,t}(\mathbf{x}')$, $\mathbf{k}_{f,t}(\mathbf{x}) = [k_f(\mathbf{x}_1, \mathbf{x}), \ldots, k_f(\mathbf{x}_t, \mathbf{x})]^T$ and $\mathbf{K}_{f,t}$ is the positive definite kernel matrix $[k_f(\mathbf{x}, \mathbf{x}')]_{\mathbf{x},\mathbf{x}'\in\{\mathbf{x}_1,\ldots,\mathbf{x}_t\}}$. Associated with g with i.i.d Gaussian noise $\zeta_t \sim \mathcal{N}(0, \sigma^2)$, the mean $\mu_{g,t}(\mathbf{x})$ and variance $\sigma_{g,t}^2(\mathbf{x})$ are defined similarly.

B. Comparison to Related work

In this section we discuss the most closely related works: [5], [6], [7] and [10]. See SM for a broader discussion.

As mentioned our GP model for the reward and constraints is motivated by [6], [7], [5]. First, contrary to us, the aforementioned papers seek to identify a safe decision with the highest possible reward given a limited number of trials; i.e., their goal is to provide *best-arm identification* with convergence guarantees by [11]. Instead, our paper focuses on a long-term performance characterized through cumulative regret bounds. Second, implementing the algorithms of [6], [7] requires the knowledge (or at least some estimate) of the Lipschitz constant L of f and g over the decision set. In contrast, our algorithm does *not* use Lipschitzness of either f or g, hence, it avoids the need for a processing step that estimates L.

Our algorithm can be seen as an extension of Safe-LUCB proposed by [10] to safe GPs. Specifically, in Section III-A, we show that our algorithm and guarantees are similar to those in [10] for linear kernels. While [10] studies a frequentist setting, our results hold for a rich class of kernels beyond linear kernel.

II. A Safe GP-UCB Algorithm

We start with a description of SGP-UCB, which is summarized in Algorithm 1. Similar to a number of previous works (e.g., [7], [10]), SGP-UCB proceeds in two phases to balance the goal of expanding the safe set and controlling the regret. Prior to designing the decision rule, the algorithm requires a proper expansion of $\mathcal{D}_0^{\mathrm{S}}$. Hence, in the first phase, it takes actions at random from a given safe seed set $\mathcal{D}^w$ until the safe set has sufficiently expanded. In the second phase, the algorithm exploits GP properties to make predictions of f from past noisy observations y_t. It then follows the *Upper Confidence Bound* (UCB) machinery to select the action. In the absence of constraint (1), UCB-based algorithms select action $\mathbf{x}_t$ that maximizes a high probability upper bound of $f(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{D}_0$. However, the safety constraint (1) requires the algorithm to have a more delicate sampling rule as follows. Thus, the algorithm exploits the noisy constraint observations z_t to further establish confidence intervals for the unknown constraint function, which allow us to design an inner approximation $\mathcal{D}_t^{\mathrm{S}}$ of the safe set (see (6)). The chosen actions belong to $\mathcal{D}_t^{\mathrm{S}}$ which guarantees that the safety constraint (1) is met with high probability.

A key difference in the analysis of SGP-UCB compared to the classical GP-UCB is that $\mathbf{x}^*$ may not lie within the estimated safe set $\mathcal{D}_t^{\mathrm{S}}$ (see (6)) at each round. To see what changes, consider the standard decomposition of the instantaneous regret $r_t := f(\mathbf{x}^*) - f(\mathbf{x}_t)$ in two terms as follows: $r_t = (f(\mathbf{x}^*) - u_{f,t}(\mathbf{x}_t)) + (u_{f,t}(\mathbf{x}_t) - f(\mathbf{x}_t))$, where, $\mathbf{x}_t$ is the optimistic action at round t, i.e. the solution to the maximization in Eqn. (7). On the one hand, controlling the second term, is more or less standard and closely follows previous such bounds on UCB-type algorithms [1]. On the other hand, controlling the first term is more delicate. This complication lies at the heart of the new formulation with additional safety constraints. When safety constraints are absent, classical GP-UCB guarantees that the first term is non-positive. Unfortunately, this is *not* the case here since $\mathbf{x}^*$ does *not* necessarily belong to $\mathcal{D}_t^{\mathrm{S}}$ (see (6)). Our main contribution towards establishing regret guarantees is to add the extra pure-exploration phase with duration T', where T' is chosen such that for all $t \geq T' + 1$, the first term above is non-positive with high probability.

Exploration phase: The exploration phase aims to reach a sufficiently expanded safe subset of $\mathcal{D}_0$. The algorithm starts exploring by choosing actions from $\mathcal{D}^w$ at random and it stops after T' rounds once it has reached an approximate safe set within which $\mathbf{x}^*$ lies with high probability.

Safe exploration-exploitation phase: In the second phase, the algorithm follows an approach similar to GP-UCB [1] in order to balance exploration and exploitation and guarantee the no-regret property. At rounds $t = T' + 1, \ldots, T$, SGP-UCB uses previous observations to estimate $\mathcal{D}_0^{\mathrm{S}}$ and predict f. It creates the following confidence interval for $f(\mathbf{x})$:

Algorithm 1 SGP-UCB

Input: δ, ϵ, $\mathcal{D}_0$, $\mathcal{D}^w$, $\lambda_-(\tilde{\lambda}_-)$, T', T
Pure exploration phase:
for $t = 1 \ldots, T'$ **do**
  Randomly choose $\mathbf{x}_t \in \mathcal{D}^w$ and observe y_t and z_t.
end for
Safe exploration-exploitation phase:
for $t = T' + 1 \ldots, T$ **do**
  Compute $\ell_{f,t}(\mathbf{x})$, $u_{f,t}(\mathbf{x})$, and $\ell_{g,t}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{D}_0$ using (4) and (5) and β_t specified in Theorem 1.
  Create $\mathcal{D}_t^{\mathrm{S}}$ as in (6).
  Choose $\mathbf{x}_t = \arg\max_{\mathbf{x} \in \mathcal{D}_t^{\mathrm{S}}} u_{f,t}(\mathbf{x})$ and observe y_t, z_t.
end for

$Q_{f,t}(\mathbf{x}) := [\ell_{f,t}(\mathbf{x}), u_{f,t}(\mathbf{x})]$, where,

$$\ell_{f,t}(\mathbf{x}) = \mu_{f,t-1}(\mathbf{x}) - \beta_t^{1/2}\sigma_{f,t-1}(\mathbf{x}), \tag{4}$$

$$u_{f,t}(\mathbf{x}) = \mu_{f,t-1}(\mathbf{x}) + \beta_t^{1/2}\sigma_{f,t-1}(\mathbf{x}). \tag{5}$$

Confidence intervals $Q_{g,t}(\mathbf{x})$ corresponding to $g(\mathbf{x})$ are defined in a similar way. We choose β_t according to Theorem 1 to guarantee $f(\mathbf{x}) \in Q_{f,t}(\mathbf{x})$ and $g(\mathbf{x}) \in Q_{g,t}(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{D}_0$ and $t > 0$ with high probability.

Theorem 1 (Confidence Intervals, [1]). *Pick $\delta \in (0,1)$ and set $\beta_t = 2\log(2|\mathcal{D}_0|t^2\pi^2/6\delta)$, then, $f(\mathbf{x}) \in Q_{f,t}(\mathbf{x})$, $g(\mathbf{x}) \in Q_{g,t}(\mathbf{x})$, $\forall \mathbf{x} \in \mathcal{D}_0, t > 0$, with probability at least $1 - \delta$.*

Using the above defined confidence intervals $Q_{f,t}(\mathbf{x})$ and $Q_{g,t}(\mathbf{x})$, the algorithm is able to act conservatively to ensure that safety constraint (1) is satisfied. Specifically, at the beginning of each round $t = T' + 1, \ldots, T$, SGP-UCB forms the following so-called *safe decision sets* based on the mentioned confidence bounds:

$$\mathcal{D}_t^{\mathrm{S}} := \{\mathbf{x} \in \mathcal{D}_0 : \ell_{g,t}(\mathbf{x}) \geq h\}. \tag{6}$$

Recall that $g(\mathbf{x}) \geq \ell_{g,t}(\mathbf{x})$ for all $t > 0$ with high probability. Therefore, $\mathcal{D}_t^{\mathrm{S}}$ is guaranteed to be a set of safe actions with the same probability. After creating safe decision sets in the second phase, the algorithm follows a similar decision rule as in GP-UCB algorithm in [1]:

$$\mathbf{x}_t = \arg\max_{\mathbf{x} \in \mathcal{D}_t^{\mathrm{S}}} u_{f,t}(\mathbf{x}). \tag{7}$$

III. Regret Analysis of SGP-UCB

Consider decomposing the cumulative regret as

$$R_T = \sum_{t=1}^{T'} r_t + \sum_{t=T'+1}^{T} r_t = \text{Term I} + \text{Term II}. \tag{8}$$

Bounding Term II. In the following sections, we show how T' is appropriately chosen such that $\mathbf{x}^* \in \mathcal{D}_t^{\mathrm{S}}$ with high probability for all $t \geq T' + 1$. Once this is accomplished, we bound the second term of (8) using the standard regret analysis in [1]. Specifically, the bound depends on the so-called *information gain* γ_t which quantifies how fast f can be learned in an information theoretic sense. See [1] and SM for more details.
Bounding Term I. Since for the first T' rounds actions are selected at random, the bound on Term I is linear in T'. In other words, the upper bound on the first term is of the form BT', where $B := C\sqrt{2\ell d}\ \mathrm{diam}(\mathcal{D}_0)/\delta$ for some $C > 0$ if k_f is an RBF kernel with parameter ℓ, otherwise $B := 2\sqrt{2\log(2|\mathcal{D}_0|)}/\delta$ such that (see SM for details): $\max_{\mathbf{x},\mathbf{y} \in \mathcal{D}_0} |f(\mathbf{x}) - f(\mathbf{y})| < B$, with probability at least $1 - \delta$.
Determining T'. We need to find the value of T' such that with high probability $\mathbf{x}^* \in \mathcal{D}_t^{\mathrm{S}}$ for all $t \geq T' + 1$. The following lemma, proved in the SM establishes a sufficient condition for $\mathbf{x}^* \in \mathcal{D}_t^{\mathrm{S}}$ which is more convenient to work with.

Lemma 1 ($\mathbf{x}^* \in \mathcal{D}_t^{\mathrm{S}}$). *With probability at least $1 - \delta$, it holds that $\mathbf{x}^* \in \mathcal{D}_t^{\mathrm{S}}$ for any $t > 0$ that satisfies:*

$$\frac{\epsilon^2}{4\beta_t} \geq \sigma_{g,t-1}^2(\mathbf{x}^*), \tag{9}$$

From Lemma 1, it suffices to establish an appropriate upper bound on the RHS of (9) to determine the duration of the first phase, i.e., T'. We do this by expressing $\sigma_{g,t-1}^2(\mathbf{x}^*)$ in terms of the associated feature maps as follows. Recall that a positive semi-definite kernel function $k_g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is associated with a feature map $\varphi_g : \mathbb{R}^d \to \mathcal{H}_{k_g}$ that maps the vectors in the primary space to a reproducing kernel Hilbert space (RKHS). In terms of the mapping φ_g, the kernel function k_g is defined by: $k_g(\mathbf{x}, \mathbf{x}') = \varphi_g(\mathbf{x})^T\varphi_g(\mathbf{x}')$, $\forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$. Let d_g denote the dimension of $\mathcal{H}_{k_g}$ (potentially infinite) and define the $t \times d_g$ matrices $\mathbf{\Phi}_{g,t} := [\varphi_g(\mathbf{x}_1), \ldots, \varphi_g(\mathbf{x}_t)]^T$ at each round t. Using this notation, we can rewrite (see SM):

$$\sigma_{g,t}^2(\mathbf{x}) = \sigma^2\varphi_g(\mathbf{x})^T(\mathbf{\Phi}_{g,t}^T\mathbf{\Phi}_{g,t} + \sigma^2\mathbf{I})^{-1}\varphi_g(\mathbf{x}). \tag{10}$$

A. Constraint with finite-dimensional RKHS

In this section we consider g with finite dimensional RKHS. For example, for linear kernel $k_g(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T\mathbf{y}$ and polynomial kernel $k_g(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T\mathbf{y} + 1)^p$, the corresponding d_g is d and $\binom{d+p}{d}$, respectively [12].

Let $\bar{\mathbf{x}} \sim \mathrm{Unif}(\mathcal{D}^w)$ be a d-dimensional random vector uniformly distributed in $\mathcal{D}^w$. We denote the covariance matrix of $\varphi_g(\bar{\mathbf{x}})$ by $\mathbf{\Sigma}_g = \mathbb{E}[\varphi_g(\bar{\mathbf{x}})\varphi_g(\bar{\mathbf{x}})^T] \in \mathbb{R}^{d_g \times d_g}$. A key quantity in our analysis is the minimum eigenvalue of $\mathbf{\Sigma}_g$ denoted by:

$$\lambda_- := \lambda_{\min}(\mathbf{\Sigma}_g). \tag{11}$$

At rounds $t \in [T']$, SGP-UCB chooses safe iid actions $\mathbf{x}_t \overset{\text{iid}}{\sim} \mathrm{Unif}(\mathcal{D}^w)$. Regarding the definition of $\sigma_{g,t}^2(\mathbf{x})$ in (10), we show that if $\lambda_- > 0$, $\sigma_{g,t-1}^2(\mathbf{x}^*)$ can be controlled for all $t \geq T' + 1$ by appropriately lower bounding the minimum eigenvalue of the Gram matrix $\mathbf{\Phi}_{g,T'}^T\mathbf{\Phi}_{g,T'}$, which is possible due to the randomness of chosen actions in the first phase.

Lemma 2. *Assume $d_g < \infty$, $\lambda_- > 0$, and $\mathbf{x} \in \mathcal{D}_0$. Then, for any $\delta \in (0,1)$, provided $T' \geq t_\delta := \frac{8}{\lambda_-}\log(\frac{d_g}{\delta})$, the following holds with probability at least $1 - \delta$,*

$$\lambda_{\min}\left(\mathbf{\Phi}_{g,T'}^T\mathbf{\Phi}_{g,T'} + \sigma^2\mathbf{I}\right) \geq \sigma^2 + \frac{\lambda_- T'}{2}. \tag{12}$$

Consequently, $\sigma^2_{g,t-1}(\mathbf{x}^*) \leq \frac{2\sigma^2}{2\sigma^2+\lambda_- T'}$, *for all* $t \geq T'+1$.

Combining Lemmas 1 and 2 gives the desired value of T' that guarantees $\mathbf{x}^* \in \mathcal{D}_t^{\mathrm{S}}$ for all $t \geq T'+1$ with high probability. Putting these together, we conclude the following regret bound for constraint with corresponding finite-dimensional RKHS.

Theorem 2 (Regret bound for g with finite dimensional RKHS)**.** *Let the same assumptions as in Lemma 2 hold. Let* $t_\epsilon := \frac{8\sigma^2\beta_T}{\lambda_- \epsilon^2}$ *and define* $T' := t_\epsilon \vee t_\delta$. *Then for sufficiently large* T *such that* $T \geq T'$ *and any* $\delta \in (0, 1/3)$, *with probability at least* $1-3\delta$*:*

$$R_T \leq BT' + \sqrt{C_1 T \beta_T \gamma_T},$$

where $C_1 = 8/\log(1+\sigma^{-2})$.

As a special case, let f and g be associated with linear kernels. Let k be a linear kernel with mapping $\varphi_g : \mathbb{R}^d \to \mathcal{H}_k = \mathbb{R}^d$, $\mathbf{X}_t = \mathbf{\Phi}_t = [\mathbf{x}_1, \ldots, \mathbf{x}_t]^T$, and $\mathbf{y}$ be the corresponding observation vector. Therefore, we have $\mu_t(\mathbf{x}) = \mathbf{x}^T\hat{\theta}_t$ where $\hat{\theta}_t = (\mathbf{X}_t^T\mathbf{X}_t + \sigma^2\mathbf{I})^{-1}\mathbf{X}_t^T\mathbf{y}$. We derive the following from (10): $\sigma_t^2(\mathbf{x}) = \sigma^2\|\mathbf{x}\|_{\mathbf{A}_t^{-1}}$, where $\mathbf{A}_t = \mathbf{X}_t^T\mathbf{X}_t + \sigma^2\mathbf{I}$. Thus, we observe the close relation in these notations with that in Linear stochastic bandits settings, [13]. As such, our setting is an extension to [10], where linear loss and constraint functions have been studied (albeit in a frequentist setting).

B. Constraint with infinite-dimensional RKHS

In the infinite-dimensional RKHS setting, controlling $\sigma_{g,t-1}(\mathbf{x}^*)$ for $t \geq T'+1$ can be challenging. To address this issue, we focus on stationary kernels, i.e., $k_g(\mathbf{x},\mathbf{y}) = k_g(\mathbf{x}-\mathbf{y})$[2], and apply a *finite basis approximation* in our analysis. Particularly, we consider $\tilde{\varphi}_g : \mathbb{R}^d \to \mathbb{R}^{D_g}$ which maps the input to a lower-dimensional Euclidean inner product space with dimension D_g such that: $k_g(\mathbf{x},\mathbf{y}) \approx \tilde{\varphi}_g(\mathbf{x})^T\tilde{\varphi}_g(\mathbf{y})$.

Definition 1 ((ϵ_0, D_g)**-uniform approximation**)**.** *For a stationary kernel* $k_g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, *the inner product* $\tilde{\varphi}_g(\mathbf{x})^T\tilde{\varphi}_g(\mathbf{y})$ *in* $\mathbb{R}^{D_g}$, (ϵ_0, D_g)*-uniformly approximates* $k_g(\mathbf{x},\mathbf{y})$ *if and only if:*

$$\sup_{\mathbf{x},\mathbf{y}\in\mathcal{D}_0} |\tilde{\varphi}_g(\mathbf{x})^T\tilde{\varphi}_g(\mathbf{y}) - k(\mathbf{x},\mathbf{y})| \leq \epsilon_0.$$

Due to the infinite dimensionality of $\mathcal{H}_{k_g}$, there is no notion for minimum eigenvalue of $\mathbf{\Phi}_{g,T'}^T\mathbf{\Phi}_{g,T'}$. Hence, we adopt an (ϵ_0, D_g)-uniform approximation to bound $\sigma^2_{g,t-1}(\mathbf{x}^*)$ for all $t \geq T'+1$ by lower bounding the minimum eigenvalue of the approximated $D_g \times D_g$ matrix $\tilde{\mathbf{\Phi}}_{g,T'}^T\tilde{\mathbf{\Phi}}_{g,T'}$ instead. The argument follows the same procedure as in Lemma 2, other than an error bound on $\sigma^2_{g,t-1}(\mathbf{x}^*)$ caused by the (ϵ_0, D_g)-unifromly approximation is required.

We consider $\tilde{\varphi}_g(.)$ to be an (ϵ_0, D_g)-uniform approximation and denote the covariance matrix of $\tilde{\varphi}_g(\bar{\mathbf{x}})$ by $\tilde{\mathbf{\Sigma}}_g = \mathbb{E}[\tilde{\varphi}_g(\bar{\mathbf{x}})\tilde{\varphi}_g(\bar{\mathbf{x}})^T] \in \mathbb{R}^{D_g \times D_g}$ with minimum eigenvalue:

$$\tilde{\lambda}_- := \lambda_{\min}(\tilde{\mathbf{\Sigma}}_g). \tag{13}$$

[2]This property holds for a wide variety of kernels including Exponential, Gaussian, Rational quadratic, etc.

Lemma 3. *Assume that* $d_g = \infty$, k_g *is a stationary kernel, and* $\tilde{\lambda}_-$ *defined in* (13) *is positive. Fix* $\delta, \epsilon_0 \in (0,1)$. *Then, it holds with probability at least* $1-\delta$ *for all* $t \geq T'+1$,

$$\sigma^2_{g,t-1}(\mathbf{x}^*) \leq \frac{2\sigma^2}{2\sigma^2 + \tilde{\lambda}_- T'} + \frac{4t^3\epsilon_0}{\sigma^2}, \tag{14}$$

provided that $T' \geq \tilde{t}_\delta := \frac{8}{\tilde{\lambda}_-}\log(\frac{D_g}{\delta})$.

See the SM for rechnical details on how $\tilde{\varphi}_g$ analytically helps us obtain this upper bound on $\sigma^2_{g,t-1}(\mathbf{x}^*)$ for all $t \geq T'+1$ by lower bounding the minimum eigenvalue of $\tilde{\mathbf{\Phi}}_{g,T'}^T\tilde{\mathbf{\Phi}}_{g,T'}$. Putting these together leads to the following general regret bound.

Theorem 3 (Regret bound for g with infinite dimensional RKHS)**.** *Assume there exists an* (ϵ_0, D_g)*-uniform approximation of stationary kernel* k_g *with* $0 < \epsilon_0 \leq \frac{\epsilon^2\sigma^2}{32T^3\beta_T}$ *for which* $\tilde{\lambda}_-$ *defined in* (13) *is positive. Let* $\tilde{t}_\epsilon := \frac{16\sigma^2\beta_T}{\tilde{\lambda}_-\epsilon^2}$ *and* $\tilde{t}_\delta := \frac{8}{\tilde{\lambda}_-}\log(\frac{D_g}{\delta})$ *and define* $T' := \tilde{t}_\epsilon \vee \tilde{t}_\delta$. *Then, for sufficiently large* T *such that* $T \geq T'$ *and any* $\delta \in (0, 1/3)$, *with probability at least* $1-3\delta$*:*

$$R_T \leq BT' + \sqrt{C_1 T \beta_T \gamma_T}, \tag{15}$$

where $C_1 = 8/\log(1+\sigma^{-2})$.

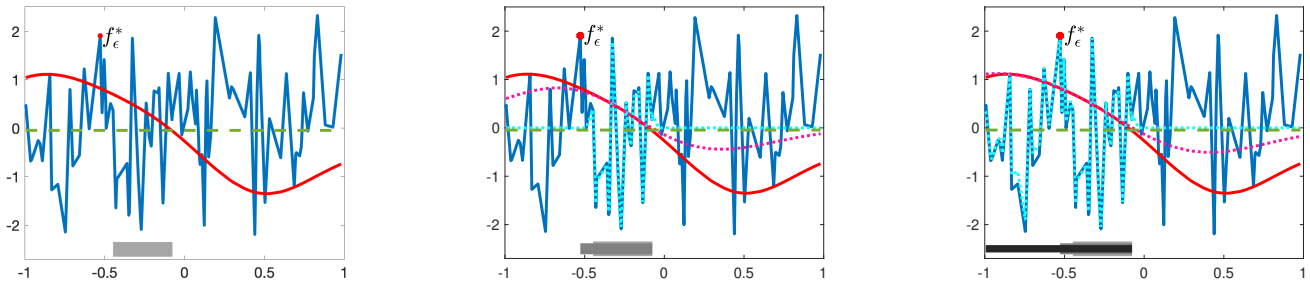
Depending on the feature map approximation $\tilde{\varphi}_g$, the dimension D_g can be appropriately chosen as a function of the algorithm's inputs ϵ, δ and d to control the accuracy of the approximation. We emphasize that our analysis is not restricted to specific approximations. We focus on the *Quadrature Fourier features* (QFF) for which [14] showed that the QFF uniform approximation error ϵ_0 decreases exponentially with D_g. Concretely, in this case, $D_g = \mathcal{O}\left((d+\log(d/\epsilon_0))^d\right)$ features are required to obtain an ϵ_0-accurate approximation of the SE kernel k_g.

IV. Additive Models with Low Effective Dimension

Our analysis above suggests that Safe GP learning is easy when the safety constraint is *simple* (e.g. linear/polynomial kernels). However, for instances with dimension $d_g = \infty$, the assumption $\tilde{\lambda}_- > 0$ requires $\mathcal{D}^w$ to contain at least $D_g = \mathcal{O}\left(\left(d+\log(T/\epsilon)\right)^d\right)$ actions, scaling unfavorably with d. While this can be problematic in general, the use of QFF remains relevant in applications involving either small dimensional problems ($d \leq 5$), or high dimensional GPs with low effective dimension, such as *additive models* [14], [15]. Specifically, suppose the constraint function g admits an additive decomposition: [14], [15]:

$$g(\mathbf{x}) = \sum_{i=1}^{G} g_i(\mathbf{x}^{(i)}), \tag{16}$$

where $\mathbf{x}^{(i)} \in \mathcal{D}_0^{(i)}$, $\mathcal{D}_0^{(i)} \subseteq \mathcal{D}_0$ is a d_i-dimensional subspace of $\mathcal{D}_0$ with $d_i \ll d$, and $\mathcal{D}_0^{(i)} \cap \mathcal{D}_0^{(j)} = \emptyset$ for any i and $j \neq i$. Let the *effective dimension* of the model by $d_e := \max_{i\in[G]} d_i$. Under this assumption, the kernel and



(a) True f,g and given safe seed set. (b) $t = 50$ (c) $t = 200$

Figure 1: Illustration of SGP-UCB: (a) The dark blue and red solid lines denote the unknown reward function f and constraint function g, respectively. The dashed green line represents the threshold $h + \epsilon$, the gray bar shows the safe seed set $\mathcal{D}^w$, and the red star is the optimum value of f through $\mathcal{D}_\epsilon^{\mathrm{S}}$. (b,c) The dotted blue and pink lines are the estimated GP mean functions corresponding to f and g, respectively at rounds 50 and 200.

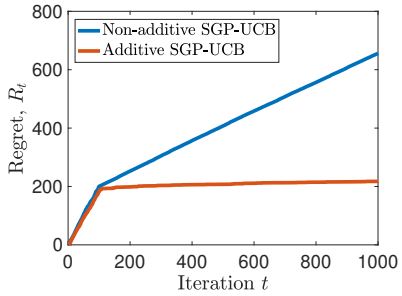


Figure 2: Regret comparison of AdditiveSGP-UCB with $G = 3, d_1 = 1, d_2 = 2, d_3 = 3$ vs Non-additive SGP-UCB with $d = 6$.

the mean function of g decompose in the same fashion: $k_g(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{G} k_{g_i}(\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$, $\mu_g(\mathbf{x}) = \sum_{i=1}^{G} \mu_{g_i}(\mathbf{x}^{(i)})$. Thus, similar to what we did in previous sections, we can define the following quantities for each $i \in [G]$: $\mu_{g_i,t}(\mathbf{x}^{(i)})$, $\sigma^2_{g_i,t}(\mathbf{x}^{(i)})$, and $\ell_{g_i,t}(\mathbf{x}^{(i)})$ for all $\mathbf{x}^{(i)} \in \mathcal{D}_0^{(i)}$ and $i \in [G]$. With these, after appropriately choosing T', the agent constructs the following safe estimated set at round $t \geq T' + 1$: $\mathcal{D}_t^{\mathrm{S}} := \{\mathbf{x} \in \mathcal{D}_0 : \sum_{i=1}^{G} \ell_{g_i,t}(\mathbf{x}^{(i)}) \geq h\}$. Thus, using the machinery of in Section III-B, we can show that the "new" $D_g := \max_{i \in [G]} D_{g_i}$, that now determines T' is $\mathcal{O}((d_e + \log(T/\epsilon))^{d_e})$, i.e., it only depends on the effective dimension d_e rather than the ambient dimension d that can be much larger $d \gg d_e$.

V. Experiments

In Figure 1, we give an illustration of SGP-UCB's performance. For the sake of visualization, we implement the algorithm in a 1-dimensional space and connect the data points since we find it instructive to also depict estimates of f and g as well as the growth of the safe sets. The algorithm starts the first phase by sampling actions at random from a given safe seed set. After 50 rounds, in Figure 1b, the safe set has sufficiently expanded such that the optimal action $\mathbf{x}^*$ lies within the $\mathcal{D}_{50}^{\mathrm{S}}$. Figure 1c shows the expansion of the safe set after 200 rounds, which still includes $\mathbf{x}^*$.

Figure 2 illustrates the superiority SGP-UCB's performance when g admits an additive model, as above, compared to its performance when faced with g of the same total dimension d but no additive structure.

In (Figure 2 of) the SM, we present additional experiments comparing SGP-UCB's performance against other existing algorithms: 1) StageOpt [7]; 2) SafeOpt-MC [5]; 3) A heuristic variant of GP-UCB, which proceeds the same as SGP-UCB except that there is no exploration phase, i.e., $T' = 0$; 4) The standard GP-UCB with oracle access to the safe set. In summary we find that when the safe seed set contains enough actions (in line with the requirements of our theorems), SGP-UCB outperforms SafeOpt-MP and StageOpt. We also implemented SGP-UCB for settings where $\mathcal{D}^w$ has relatively small number of safe actions. The results show the poor performance of SGP-UCB which is expected since $\mathcal{D}^w$ is not large enough to explore the whole space for the purpose of safe set expansion.

Please refer to Section 4 in the SM for details on implementation and additional discussions.

VI. Future work

Several issues remain to be studied. While our algorithm is the first providing regret guarantees for safe GP optimization, it is not clear whether it is the best to apply. The answer could depend on the application. Hence, numerical comparisons on real application-specific data are worth investigating. A related important issue that needs to be addressed is that the existing guarantees (either in terms of cumulative regret, simple regret or optimization gap) for all safe-GP optimization algorithms, suffer from loose constants that make such comparisons hard. Indeed evaluating the performances of all these four algorithms in numerical experiments requires us to resort to empirical tuning of parameters like T' [3], which is an important challenge to overcome.

[3]For all our implementations, we stopped the first pure-exploration phase when the safe region plateaued for at least 20 iterations, and also hard capped T' at 100 iterations (a similar approach was adopted by [7]).

REFERENCES

[1] Niranjan Srinivas, Andreas Krause, Sham Kakade, and Matthias Seeger. Gaussian process optimization in the bandit setting: no regret and experimental design. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 1015–1022. Omnipress, 2010.

[2] Felix Berkenkamp, Angela P Schoellig, and Andreas Krause. Safe controller optimization for quadrotors with gaussian processes. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 491–496. IEEE, 2016.

[3] Anayo K Akametalu, Jaime F Fisac, Jeremy H Gillula, Shahab Kaynama, Melanie N Zeilinger, and Claire J Tomlin. Reachability-based safe learning with gaussian processes. In *53rd IEEE Conference on Decision and Control*, pages 1424–1431. IEEE, 2014.

[4] Chris J Ostafew, Angela P Schoellig, and Timothy D Barfoot. Robust constrained learning-based nmpc enabling reliable mobile robot path tracking. *The International Journal of Robotics Research*, 35(13):1547–1563, 2016.

[5] Felix Berkenkamp, Andreas Krause, and Angela P Schoellig. Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics. *arXiv preprint arXiv:1602.04450*, 2016.

[6] Yanan Sui, Alkis Gotovos, Joel W. Burdick, and Andreas Krause. Safe exploration for optimization with gaussian processes. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 997–1005. JMLR.org, 2015.

[7] Yanan Sui, Joel Burdick, Yisong Yue, et al. Stagewise safe bayesian optimization with gaussian processes. In *International Conference on Machine Learning*, pages 4788–4796, 2018.

[8] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

[9] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[10] Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. In *Advances in Neural Information Processing Systems*, pages 9252–9262, 2019.

[11] Shahin Shahrampour, Mohammad Noshad, and Vahid Tarokh. On sequential elimination algorithms for best-arm identification in multi-armed bandits. *IEEE Transactions on Signal Processing*, 65(16):4281–4292, 2017.

[12] Ninh Pham and Rasmus Pagh. Fast and scalable polynomial kernels via explicit feature maps. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 239–247. ACM, 2013.

[13] Varsha Dani, Thomas P Hayes, and Sham M Kakade. Stochastic linear optimization under bandit feedback. 2008.

[14] Mojmir Mutny and Andreas Krause. Efficient high dimensional bayesian optimization with additivity and quadrature fourier features. In *Advances in Neural Information Processing Systems*, pages 9005–9016, 2018.

[15] Kirthevasan Kandasamy, Jeff Schneider, and Barnabás Póczos. High dimensional bayesian optimisation and bandits via additive models. In *International conference on machine learning*, pages 295–304, 2015.