Provably Efficient Bayesian Optimization with Unknown Constraints

Anonymous Author(s)

Abstract

We study a black-box Bayesian optimization (BO) problem with unknown constraints. In contrast to hard safety-type constraints studied in existing works, we consider soft constraints that may be violated in any round as long as the cumulative violations are small, which is motivated by various practical applications. Our ultimate goal is to study how to utilize the nature of soft constraints to improve the performance of existing constrained BO algorithms in terms of regret and computational complexity. To this end, we first consider bandit feedback for the constraint. Leveraging primal-dual optimization, we propose a general framework for both algorithm design and performance analysis. This framework builds upon a novel sufficient condition, which not only is satisfied under very general exploration strategies, including *upper confidence bound* (UCB), *Thompson sampling* (TS), and new ones based on random exploration, but also enables a unified analysis for showing both sublinear regret and sublinear constraint violation. Then, we consider stochastic full information for the soft constraint. In this case, we show that the additional (noisy) information can be utilized to achieve even zero expected constraint violation. Besides provable performance guarantees, another distinguishing feature of our proposed algorithms is that they all enjoy the same computational complexity as the unconstrained case. Finally, we demonstrate the superior performance of our proposed algorithms via numerical experiments on both synthetic and real-world datasets.

1 Introduction

Bayesian optimization (BO) of a black-box objective function with unknown constraints has found many real-world applications (e.g., robot controls, wireless resource allocation, and materials design). There have been many BO algorithms developed for this important setting (see, e.g., (Eriksson and Poloczek 2021; Gelbart, Snoek, and Adams 2014; Hernández-Lobato et al. 2016) and the references therein). Although these algorithms have demonstrated good performance in various practical settings, their theoretical performance guarantees are still unclear.

Recently, there have been exciting advances in the theoretical analysis of constrained BO. In particular, while (Sui et al. 2015; Berkenkamp, Krause, and Schoellig 2016; Sui et al. 2018) propose algorithms with convergence guarantees, (Amani, Alizadeh, and Thrampoulidis 2020) is focused on establishing regret bounds for the developed algorithm. These algorithms mainly focus on BO with a *hard* constraint such as safety, i.e., the selected action in each round needs to satisfy the constraint with a high probability. To this end, compared to the unconstrained case, additional computation is often required to construct a *safe* action set in each round.

Motivations. In practice, there are also many BO applications that involve *soft* constraints that may be violated in any round. The goal is to maximize the total reward while minimizing the total constraint violations. To give a concrete example, let's consider resource configuration in cloud computing platforms where the objective is to minimize the cost while guaranteeing that the latency is below a threshold, e.g., 95% percentile latency. In this case, the latency of a job could be above the threshold as long as the fraction of violations is small, e.g., less than 5%. Another example is throughput maximization under energy constraints in wireless communications where energy consumption constraint is often a soft cumulative one. In both examples, one fundamental question is *whether the soft constraint feature can be utilized to design constrained BO algorithms with the same complexity as the unconstrained case while attaining a better reward performance compared to the hard constraints*.

In addition to soft constraints, in some applications, the unknown constraint can often be observed in a stochastic *full-information* way. For example, the stability constraint in queueing system can be estimated via the observed stochastic arrivals (see more details in the Appendix). Thus, compared to the standard bandit feedback of unknown constraint, one key question here is *whether one can leverage the additional information to achieve an improved performance*. Finally, existing theoretical works on constrained BO mainly focus on upper confidence bound (UCB) exploration, which sometimes has inferior empirical performance compared to Thompson sampling (TS) exploration. Hence, another key question is *whether one can design provably efficient constrained BO algorithms with general explorations*.

Contributions. In this paper, we take a systematic approach to affirmatively answer the above three fundamental questions. In particular, we formulate BO with soft constraints as a sequential decision-making problem where the objective is to maximize the cumulative reward while minimizing the cumulative constraint violation. Considering both bandit feedback and stochastic full information for the constraint function, we develop computationally efficient BO algorithms that enjoy provable performance guarantees under general exploration strategies. Our contributions are summarized as follows.

First, in the setting with bandit feedback for the constraint, we develop a unified framework for constrained BO based on primal-dual optimization, which can guarantee both sublinear reward regret and sublinear total constraint violation under a class of general exploration strategies, including UCB, TS, and new effective ones (e.g., random exploration). This framework builds upon a novel sufficient condition, which not only facilitates the design of new constrained BO algorithms but provides a unified view in the performance analysis. *Second*, in the setting with stochastic full-information for the constraint, we show that one can even achieve a zero constraint violation in expectation by utilizing the additional (noisy) information. This result is obtained via insights and tools from queueing theory such as Lyapunov-drift analysis and Hajek's Lemma. *Finally*, we demonstrate the superior performance of our proposed algorithms via simulations based on both synthetic and real-world data. In addition, we discuss the benefits of our algorithms in terms of various practical considerations.

Related Works. In the special cases of BO, such as multi-arm bandit (MAB) and linear bandits (e.g., BO with a linear kernel), there is a large body of work on bandits with different types of constraints, including knapsack bandits (Agrawal and Devanur 2016; Badanidiyuru, Kleinberg, and Slivkins 2013; Wu et al. 2015), conservative bandits (Wu et al. 2016; Kazerouni et al. 2016; Garcelon et al. 2020), bandits with fairness constraints (Chen et al. 2020; Li, Liu, and Ji 2019), bandits with hard safety constraints (Amani, Alizadeh, and Thrampoulidis 2019; Pacchiano et al. 2021; Moradipari et al. 2019), and bandits with cumulative soft constraints (Liu et al. 2020, 2021). Among them, the bandit setting with cumulative soft constraints is the closest to ours in that the goal is also to minimize the cumulative constraint violation. In fact, our results in the stochastic full-information case can be viewed as a nontrivial generalization of (Liu et al. 2020) on MAB and (Liu et al. 2021) on linear bandits, respectively. Note that as an extension, (Liu et al. 2021) also considers the harder case of bandit constraint feedback with linear functions (under UCB only); to the best of our understanding, however, there is a flaw in their current analysis that makes their claimed results ungrounded (see discussions in the Appendix).

Broadly speaking, our work is also related to reinforcement learning (RL) with soft constraints. In particular, our analysis for the bandit constraint feedback case is inspired by those on constrained RL (Efroni, Mannor, and Pirotta 2020; Ding et al. 2021) but has significant differences. First, while they focus on either tabular or linear function approximation settings, both objective and constraint functions we consider can be *nonlinear*. Second, in contrast to policy-based algorithms in (Ding et al. 2021), ours can be viewed as value-based algorithms. Third, while they only consider UCB exploration, our algorithms can be equipped with various exploration strategies (including UCB) and hence are more flexible.

Finally, we remark that our work is also related to online convex optimization with stochastic constraints (Yu, Neely, and Wei 2017; Wei, Yu, and Neely 2020), where Lyapunov-drift analysis is adopted to bound the cumulative constraint violation. However, different from full-information feedback they consider for the objective function, in our BO setting, only bandit feedback is available for the objective function, which makes the problem more challenging.

2 Problem Formulation and Preliminaries

We consider black-box optimization with *soft* constraints, i.e., $\max_{x\in\mathcal{X}} f(x)$ subject to $g(x) \leq 0$, where $\mathcal{X} \subset \mathbb{R}^d$ and both $f : \mathcal{X} \to \mathbb{R}$ and $g : \mathcal{X} \to \mathbb{R}$ are unknown functions[1]. To this end, we formulate it as an online sequential decision-making problem. In each round $t \in \{1, 2, \ldots, T\}$, a learning agent chooses an action $x_t \in \mathcal{X}$ and receives a bandit reward feedback $r_t = f(x_t) + \eta_t$, where η_t is a zero-mean noise. The learning agent can also observe certain information (specified below) about the unknown constraint function g. To capture the feature of soft constraints, the goal here is to maximize the cumulative reward (i.e., $\sum_{t=1}^{T} f(x_t)$) while minimizing the cumulative constraint violation (i.e., $\sum_{t=1}^{T} g(x_t)$) throughout the learning process.

Observable Constraint Information. We consider two different scenarios. The first one is the *bandit constraint information* case, where at each time t, after making the decision x_t, the learning agent also observes a noisy bandit constraint feedback $c_t = g(x_t) + \xi_t$, where ξ_t is a zero-mean noise. The second one is the *stochastic full constraint information* case, where in each round t, before making the decision, the learning agent has access to an *i.i.d.* random (noisy) sample of the true constraint function g.

Learning Problem. Define regret and constraint violation as $\mathcal{R}(T) := Tf(x^*) - \sum_{t=1}^{T} f(x_t)$ and $\mathcal{V}(T) := \left[\sum_{t=1}^{T} g(x_t)\right]_+$, respectively, where $x^* = \arg\max_{\{x\in\mathcal{X}:g(x)\leq 0\}} f(x)$ and $[\cdot]_+ := \max\{\cdot, 0\}$. The goal is to achieve both sublinear regret and sublinear constraint violation. In fact, we will establish bounds over the following stronger version of regret. Specifically, let π be a probability distribution over the set of actions $\mathcal{X}$, and let $\mathbb{E}_\pi[f(x)] := \int_{x\in\mathcal{X}} f(x)\pi(x)\, dx$ and $\mathbb{E}_\pi[g(x)] := \int_{x\in\mathcal{X}} g(x)\pi(x)\, dx$. We compare our achieved reward with the following optimization problem: $\max_\pi\{\mathbb{E}_\pi[f(x)] : \mathbb{E}_\pi[g(x)] \leq 0\}$ where both f and g are known, and π^* is its optimal solution. Now, a stronger regret is defined as $\mathcal{R}_+(T) := T\mathbb{E}_{\pi^*}[f(x)] - \sum_{t=1}^{T} f(x_t)$. Clearly, we have $\mathcal{R}(T) \leq \mathcal{R}_+(T)$.

Throughout the paper, we assume the following commonly used condition in constrained optimization (see also (Liu et al. 2021; Yu, Neely, and Wei 2017; Efroni, Mannor, and Pirotta 2020)).

[1] Our main results can be readily generalized to the multi-constraint case with a properly chosen norm.

Assumption 1 (Slater's condition)**.** *There is a constant $\delta > 0$ such that there exists a probability distribution π_0 that satisfies $\mathbb{E}_{\pi_0}[g(x)] \leq -\delta$. Without loss of generality, we assume $\delta \leq 1$.*

This is a quite mild assumption since it only requires that one can find a probability distribution over the set of actions under which the expected cost is less than a strictly negative value. This is in sharp constraint to existing BO algorithms for hard constraints that typically require the existence of an initial safe action (Sui et al. 2018; Amani, Alizadeh, and Thrampoulidis 2020).

In this paper, we consider the standard regularity assumption that is typically used in BO literature (e.g., (Chowdhury and Gopalan 2017; Srinivas et al. 2009)). Specifically, we assume that f is a fixed function in a reproducing kernel Hilbert space (RKHS) with a bounded norm. In particular, the RKHS for f is denoted by $\mathcal{H}_k$, which is completely determined by the corresponding kernel function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. Any function $h \in \mathcal{H}_k$ satisfies the *reproducing property*: $h(x) = \langle h, k(\cdot, x)\rangle_{\mathcal{H}_k}$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}_k}$ is the inner product defined on $\mathcal{H}_k$. We assume that the following boundedness property holds throughout the paper.

Assumption 2 (Boundedness)**.** *We assume that $\|f\|_{\mathcal{H}_k} \leq B$ and $k(x,x) \leq 1$ for any $x \in \mathcal{X}$ and that the noise η_t is* i.i.d. *R-sub-Gaussian.*

Gaussian Process Surrogate Model. We use a Gaussian process (GP), denoted by $\mathcal{GP}(0, k(\cdot,\cdot))$, as a prior for the unknown function f, and a Gaussian likelihood model for the noise variables η_t, which are drawn from $\mathcal{N}(0, \lambda)$ and are independent across t. Conditioned on a set of observations $H_t = \{(x_s, r_s), s \in [t] := \{1, 2, \dots, t\}\}$, by the properties of GP (Rasmussen 2003), the posterior distribution for f is $\mathcal{GP}(\mu_t(\cdot), k_t(\cdot, \cdot))$, where $\mu_t(x) = k_t(x)^T(K_t + \lambda I)^{-1}R_t$ and $k_t(x, x') = k(x, x') - k_t(x)^T(K_t + \lambda I)^{-1}k_t(x')$, in which $k_t(x) = [k(x_1, x), \dots, k(x_t, x)]^T$, $K_t = [k(x_u, x_v)]_{u,v \in [t]}$, and R_t is the (noisy) reward vector $[r_1, r_2, \dots, r_t]^T$. In particular, we also define $\sigma_t^2(x) = k_t(x, x)$. Let $K_A = [k(x, x')]_{x,x' \in A}$ for $A \subset \mathcal{X}$. We define maximum information gain as $\gamma_t(k, \mathcal{X}) := \max_{A \subset \mathcal{X}: |A| = t} \frac{1}{2} \ln |I_t + \lambda^{-1} K_A|$ where I_t is the $t \times t$ identity matrix. The maximum information gain plays a key role in the regret bounds of GP-based algorithms. While $\gamma_t(k, \mathcal{X})$ depends on the kernel k and domain $\mathcal{X}$, we simply use γ_t whenever the context is clear. For instance, if $\mathcal{X}$ is compact and convex with dimension d, then we have $\gamma_t = O((\ln t)^{d+1})$ for squared exponential kernel k_{SE}, $\gamma_t = O(t^{\frac{d(d+1)}{2\nu + d(d+1)}} \ln t)$ (where ν is a hyperparameter) for Matérn kernel $k_{\text{Matérn}}$, and $\gamma_t = O(d \ln t)$ for linear kernel (Srinivas et al. 2009). Note that this GP surrogate model is used for algorithm design only; it does not change the fact that each f is a fixed function in $\mathcal{H}_k$ and that the noise η_t can be sub-Gaussian (i.e., an *agnostic* setting (Srinivas et al. 2009)).

3 Bandit Constraint Information

We start with the scenario of bandit feedback for the constraint function. In this case, leveraging primal-dual optimization, we propose a unified framework for both algorithm design and performance analysis. In particular, we first propose a "master" algorithm called CBO (constrained BO), which can be equipped with very general exploration strategies. Then, we develop a novel sufficient condition, which not only provides a unified analysis of regret and constraint violation, but also facilitates the design of new exploration strategies (and hence new CBO algorithms).

Algorithm 1: CBO Algorithm

```
Input: $V$, $\rho$, $\phi_1 = 0$, $\mu_0(x) = \widetilde{\mu}_0(x) = 0$,
       $\sigma_0(x) = \widetilde{\sigma}_0(x) = 1, \forall x$, exploration
       strategies $\mathcal{A}_f$ and $\mathcal{A}_g$
1 for $t = 1, 2, \dots, T$ do
2     Based on posterior models, generate $f_t$ and $g_t$
        using $\mathcal{A}_f$ and $\mathcal{A}_g$, respectively
3     Truncate $f_t$ as $\bar{f}_t(x) = \text{Proj}_{[-B,B]} f_t(x)$
4     Truncate $g_t$ as $\bar{g}_t(x) = \text{Proj}_{[-G,G]} g_t(x)$
5     Pseudo-acquisition function:
      $\hat{z}_{\phi_t}(x) = \bar{f}_t(x) - \phi_t \bar{g}_t(x)$
6     Choose primal action $x_t = \arg\max_{x \in \mathcal{X}} \hat{z}_{\phi_t}(x)$;
        observe $r_t$ and $c_t$
7     Update dual variable:
      $\phi_{t+1} = \text{Proj}_{[0,\rho]} \left[\phi_t + \frac{1}{V} \bar{g}_t(x_t)\right]$
8     Posterior model: update $(\mu_t, \sigma_t)$ and $(\widetilde{\mu}_t, \widetilde{\sigma}_t)$ via
        GP regression using new data $(x_t, r_t, c_t)$
```

To begin with, since g is also unknown, we make a similar regularity assumption as for f. In particular, we assume that g is a fixed function in the RKHS defined by a kernel function $\widetilde{k}$, and the RKHS for g is denoted by $\mathcal{H}_{\widetilde{k}}$. The learning agent also uses a GP surrogate model for g, i.e., a GP prior $\mathcal{GP}(0, \widetilde{k}(\cdot, \cdot))$ and a Gaussian noise $\mathcal{N}(0, \widetilde{\lambda})$. Conditioned on a set of observations $\widetilde{H}_t = \{(x_s, c_s), s \in [t]\}$, by the properties of GPs (Rasmussen 2003), the posterior distribution for g is $\mathcal{GP}(\widetilde{\mu}_t(\cdot), \widetilde{k}_t(\cdot, \cdot))$, where $\widetilde{\mu}_t$ and $\widetilde{k}_t$ are computed in the same way as $\mu_t(\cdot)$ and $k_t(\cdot, \cdot)$. Similarly, we have the maximum information gain $\widetilde{\gamma}_t$ for g as the counterpart of γ_t for f. Then, we make the following boundedness assumption (similar to Assumption 2 for f and η_t).

Assumption 3. *We assume that $\|g\|_{\mathcal{H}_{\widetilde{k}}} \leq G$ and $\widetilde{k}(x, x) \leq 1$ for any $x \in \mathcal{X}$ and that the noise ξ_t is* i.i.d. *$\widetilde{R}$-sub-Gaussian.*

Now, we are ready to explain our "master" algorithm CBO in Algorithm 1, which is based on primal-dual optimization. Let the Lagrangian of the baseline problem $\max_\pi \{\mathbb{E}_\pi[f(x)] : \mathbb{E}_\pi[g(x)] \leq 0\}$ be $\mathcal{L}(\pi, \phi) := \mathbb{E}_\pi[f(x)] - \phi \mathbb{E}_\pi[g(x)]$ and the associated dual problem is defined as $\mathcal{D}(\phi) := \max_\pi \mathcal{L}(\pi, \phi)$ with the optimal dual variable being $\phi^* := \arg\min_{\phi \geq 0} \mathcal{D}(\phi)$. Note that since both f and g are unknown, the agent has to first generate estimates of them (i.e., f_t and g_t, respectively) based on exploration strategies $\mathcal{A}_f$ and $\mathcal{A}_g$, which capture the tradeoff between exploration and exploitation (line 2). Then, both estimates will be truncated according to the range of f and g, respectively (line 3-4) (where Proj is the projection operator). The truncation is necessary for our analysis but it does not impact the regret bound since it will not lead to loss of useful information. Then, lines 5-6 correspond to the primal optimization step that approximates $\mathcal{D}(\phi_t)$ (i.e., approximate $\mathcal{L}$ by $\bar{\mathcal{L}}$ with f and g replaced by $\bar{f}_t$ and $\bar{g}_t$). The reason behind line 6 is that one of the optimal solutions for $\max_\pi \bar{\mathcal{L}}(\pi, \phi_t)$ is simply $\arg\max_x(\bar{f}_t(x) - \phi_t \bar{g}_t(x))$. Then, line 7 is the dual update that minimizes $\mathcal{D}(\phi_t)$ with respect to ϕ by taking a projected gradient step with $1/V$ being the step size. The parameter ρ is chosen to be larger than the optimal dual variable ϕ^*, and hence the projected interval $[0, \rho]$ includes the optimal dual variable. This is achievable since the optimal dual variable is bounded under Slater's condition, and in particular, we have $\phi^* \leq (\mathbb{E}_{\pi^*}[f(x)] - \mathbb{E}_{\pi_0}[f(x)])/\delta$ by (Beck 2017, Theorem 8.42). Finally, line 8 is the posterior update via standard GP regression for both f and g.

Remark 1 (Computational complexity)**.** *CBO enjoys the same computational complexity as the unconstrained case (e.g., (Chowdhury and Gopalan 2017)) since the additional dual update is a simple projection and the primal optimization keeps the same flavor as the unconstrained case, i.e., without constructing a specific safe set as in existing constrained BO algorithms designed for the hard constraints.*

We call CBO a "master" algorithm as it allows us to employ different exploration strategies (or called *acquisition functions*) (i.e., $\mathcal{A}_f$ and $\mathcal{A}_g$). Therefore, one fundamental question is: *How to design efficient exploration strategies such that favorable performance can be guaranteed?* In the following, we take a two-step procedure to address this question. We first combine UCB-type exploration with CBO to gain useful insights. This, in turn, will facilitate the development of a novel sufficient condition, which not only is satisfied under very general exploration strategies, but also enables a unified analytical framework for showing both sublinear regret and sublinear constraint violation.

Before that, we first introduce standard UCB and TS explorations under GP as in (Chowdhury and Gopalan 2017).

Definition 1 (GP-UCB and GP-TS Explorations)**.** *Suppose the posterior distribution for a black-box function h in round t is given by $\mathcal{GP}(\hat{\mu}_{t-1}(\cdot), \hat{k}_{t-1}(\cdot,\cdot))$ and $\hat{\beta}_t$ is a time-varying sequence. (i) The estimate of h in round t under GP-UCB exploration strategy is given by $h_t(\cdot) = \hat{\mu}_{t-1}(\cdot) + \hat{\beta}_t \hat{\sigma}_{t-1}(\cdot)$, where $\hat{\sigma}_{t-1}(x) := \hat{k}_{t-1}(x, x)$ for all $x \in \mathcal{X}$. (ii) The estimate of h in round t under GP-TS exploration strategy is $h_t(\cdot) \sim \mathcal{GP}(\hat{\mu}_{t-1}(\cdot), \hat{\beta}_t^2 \hat{k}_{t-1}(\cdot,\cdot))$.*

3.1 Warm Up: CBO with GP-UCB Exploration

In this section, we instantiate CBO with GP-UCB exploration called `CBO-UCB`, as a warm-up. In particular, in `CBO-UCB`, $\mathcal{A}_f$ is a GP-UCB exploration (see Definition 1) with a positive $\hat{\beta}_t$ sequence (i.e., optimistic with respect to reward), and $\mathcal{A}_g$ is a GP-UCB exploration with a negative $\hat{\beta}_t$ sequence (i.e., optimistic with respect to cost). We show that `CBO-UCB` can guarantee sublinear regret and constraint violation simultaneously with high probability, as formally stated in the following theorem.

Theorem 1. *Suppose $\rho \geq 4B/\delta$, $V = G\sqrt{T}/\rho$, $\mathcal{A}_f$ is a GP-UCB exploration with $\hat{\beta}_t = \beta_t = B + R\sqrt{2(\gamma_{t-1} + 1 + \ln(2/\alpha))}$, and $\mathcal{A}_g$ is a GP-UCB exploration with $\hat{\beta}_t = -\widetilde{\beta}_t = -(G + R\sqrt{2(\widetilde{\gamma}_{t-1} + 1 + \ln(2/\alpha))})$. Under regularity assumptions in Assumptions 2 and 3 and Slater's condition in Assumption 1, `CBO-UCB` achieves the following bounds simultaneously with probability at least $1 - \alpha$ for any $\alpha \in (0, 1)$:*

$$\mathcal{R}_+(T) = O\left(B\sqrt{T\gamma_T} + \sqrt{T\gamma_T(\gamma_T + \ln(2/\alpha))} + \rho G\sqrt{T}\right),$$

$$\mathcal{V}(T) = O\left(\hat{\rho}\left(C\sqrt{T\hat{\gamma}_T} + \sqrt{T\hat{\gamma}_T(\hat{\gamma}_T + 2\ln(2/\alpha))}\right) + G\sqrt{T}\right),$$

where $C := \max\{B, G\}$, $\hat{\rho} = 1 + \frac{1}{\rho}$ and $\hat{\gamma}_T := \max\{\gamma_T, \widetilde{\gamma}_T\}$.

Remark 2. *The (reward) regret here is the stronger version, i.e., $\mathcal{R}_+(T)$. Compared to the unconstrained case, the regret bound has an additional term $\rho G\sqrt{T}$, which roughly captures the impact of the constraint. As in the unconstrained case, one can plug in different γ_T and $\widetilde{\gamma}_T$ to see that both regret and constraint violation are sublinear for commonly used kernels. Finally, the standard "doubling trick" can be used to design an anytime algorithm with regret and constraint violation bounds of the same order.*

Proof Sketch of Theorem 1. We first obtain the following key decomposition that holds for any $\phi \in [0, \rho]$: $\mathcal{R}_+(T) + \phi\sum_{t=1}^T g(x_t) \leq \mathcal{T}_1 + \mathcal{T}_2 + \frac{V}{2}\phi^2 + \frac{1}{2V}TG^2$, where

$$\begin{aligned}\mathcal{T}_1 = &\sum_{t=1}^T (\mathbb{E}_{\pi^*}[f(x)] - \phi_t \mathbb{E}_{\pi^*}[g(x)]) \\ &- \sum_{t=1}^T (\bar{f}_t(x_t) - \phi_t \bar{g}_t(x_t)), \qquad (1)\end{aligned}$$

$$\mathcal{T}_2 = \sum_{t=1}^T (\bar{f}_t(x_t) - f(x_t)) + \phi\sum_{t=1}^T (g(x_t) - \bar{g}_t(x_t)). \qquad (2)$$

This is achieved by utilizing the dual variable update and some necessary algebra. This bound will be the cornerstone for the analysis of both regret and constraint violation. Note that $\mathcal{T}_1 + \mathcal{T}_2$ is similar to the standard regret decomposition, with an incorporation of the constraint function weighted by ϕ_t (or ϕ). Assume that we already have a bound on it, i.e., $\mathcal{T}_1 + \mathcal{T}_2 \leq \chi(T, \phi)$ with high probability, and $\chi(T, \phi)$ is an increasing function of ϕ. This leads to the following inequality (with $V = G\sqrt{T}/\rho$) for all $\phi \in [0, \rho]$:

$$\mathcal{R}_+(T) + \phi\sum_{t=1}^T g(x_t) \leq \chi(T, \phi) + \frac{\phi^2 G\sqrt{T}}{2\rho} + \frac{\rho G\sqrt{T}}{2}. \qquad (3)$$

Then, the regret bound can be directly obtained by choosing $\phi = 0$ in (3), and hence $\mathcal{R}_+(T) = O(\chi(T, 0) + \rho G\sqrt{T})$. Inspired by (Efroni, Mannor, and Pirotta 2020), we will resort to tools from constrained convex optimization to obtain the bound on $\mathcal{V}(T)$. First, we have $\frac{1}{T}\sum_{t=1}^T f(x_t) =$

$\mathbb{E}_{\pi'}\left[f(x)\right]$ and $\frac{1}{T}\sum_{t=1}^{T} g(x_t) = \mathbb{E}_{\pi'}\left[g(x)\right]$ for some probability measure π' by the convexity of probability measure. Then, we have $\mathbb{E}_{\pi^*}\left[f(x)\right] - \mathbb{E}_{\pi'}\left[f(x)\right] + \rho\left[\mathbb{E}_{\pi'}\left[g(x)\right]\right]_+ = \frac{1}{T}\mathcal{R}_+(T) + \frac{1}{T}\phi\sum_{t=1}^{T} g(x_t) \leq \frac{\chi(T,\rho)+\rho G\sqrt{T}}{T}$, where the first equality holds by choosing $\phi = \rho$ if $\sum_{t=1}^{T} g(x_t) \geq 0$, and otherwise $\phi = 0$, and the second inequality holds by bounding RHS of (3) with $\phi = \rho$ since (3) holds for all $\phi \in [0,\rho]$ and $\chi(T,\phi)$ is increasing in ϕ. Then, based on the result above, we can apply the tool from constrained convex optimization (cf. (Beck 2017, Theorem 3.60)) to obtain $\mathcal{V}(T) \leq \frac{1}{\rho}\chi(T,\rho) + G\sqrt{T}$. The reason why we can apply this result is that $\mathbb{E}_{\pi}\left[h(x)\right]$ for any fixed h is a linear function with respect to π (and is thus convex). Finally, it remains to find $\chi(T,\phi)$ that bounds $\mathcal{T}_1 + \mathcal{T}_2$. This can be achieved by using results from unconstrained GP-UCB algorithm (cf. (Chowdhury and Gopalan 2017)). In particular, we have $\mathcal{T}_1 \leq 0$ and $\mathcal{T}_2 \leq 2\beta_T\sum_{t=1}^{T}\sigma_{t-1}(x_t) + 2\phi\widetilde{\beta}_T\sum_{t=1}^{T}\widetilde{\sigma}_{t-1}(x_t) = O(\beta_T\sqrt{\gamma_T T} + \phi\widetilde{\beta}_T\sqrt{\widetilde{\gamma}_T T})$. Then, we have $\chi(T,\phi) = O(\beta_T\sqrt{\gamma_T T} + \phi\widetilde{\beta}_T\sqrt{\widetilde{\gamma}_T T})$. Finally, plugging $\chi(T,0)$ and $\chi(T,\rho)$ into $\mathcal{R}_+(T)$ and $\mathcal{V}(T)$ yields the bounds on regret and on constraint violation, respectively, which completes the proof. □

3.2 A Class of Constrained BO Algorithms

The above analysis reveals that the key to obtaining sublinear performance guarantees is to find a sublinear bound on $\chi(T,\phi)$ that bounds $\mathcal{T}_1 + \mathcal{T}_2$. Motivated by this, in this section, we will establish a sufficient condition on the exploration strategies (i.e., $\mathcal{A}_f$ and $\mathcal{A}_g$) that guarantees a sublinear $\chi(T,\phi)$ and hence sublinear regret and sublinear constraint violation. In particular, we show that existing strategies such as GP-UCB and GP-TS both satisfy the conditon. More importantly, this condition also leads to the development of new exploration strategies (such as random exploration).

We first present the intuition behind the key part of the sufficient condition. Inspired by (Kveton et al. 2019), we will mainly focus on the following three nice events to bound $\mathcal{T}_1 + \mathcal{T}_2$ in (1)-(2):

$$\begin{aligned}
E^{est} &:= \{\forall(x,t); E_f^{est}(x,t) \cap E_g^{est}(x,t)\},\\
E_t^{conc} &:= \{\forall x; E_{f,t}^{conc}(x) \cap E_{g,t}^{conc}(x)\},\\
E_t^{anti} &:= \{E_{f,t}^{anti} \cap E_{g,t}^{anti}\}.
\end{aligned}$$

where $E_f^{est}(x,t) := |f(x) - \mu_{t-1}(x)| \leq c_{f,t}^{(1)}\sigma_{t-1}(x)$, $E_g^{est}(x,t) := |g(x) - \widetilde{\mu}_{t-1}(x)| \leq c_{f,t}^{(1)}\widetilde{\sigma}_{t-1}(x)$, $E_{f,t}^{conc}(x) := |f_t(x) - \mu_{t-1}(x)| \leq c_{f,t}^{(2)}\sigma_{t-1}(x)$, $E_{g,t}^{conc}(x) := |g_t(x) - \widetilde{\mu}_{t-1}(x)| \leq c_{g,t}^{(2)}\widetilde{\sigma}_{t-1}(x)$, $E_{f,t}^{anti} := \mathbb{E}_{\pi^*}\left[f_t(x) - \mu_{t-1}(x)\right] \geq c_{f,t}^{(1)}\mathbb{E}_{\pi^*}[\sigma_{t-1}(x^*)]$ and $E_{g,t}^{anti} := \mathbb{E}_{\pi^*}\left[g_t(x) - \widetilde{\mu}_{t-1}(x)\right] \leq -c_{g,t}^{(1)}\mathbb{E}_{\pi^*}[\widetilde{\sigma}_{t-1}(x^*)]$.

Suppose that events E^{est} and E_t^{conc} hold with high probability. Then, it is easy to see that the estimates are close to the true functions, and hence, one can derive a bound on $\mathcal{T}_2$ in (2). Now, suppose that events E^{est} and E_t^{anti} hold with some positive probability. Then, one can see that the estimates are optimistic compared to the true functions when evaluated at the optimal points. This probabilistic optimism is the key to bounding $\mathcal{T}_1$ in (1). Note that GP-UCB exploration is optimistic with probability one by definition (see Definition 1), and hence, there is always $\mathcal{T}_1 \leq 0$.

Define the filtration $\mathcal{F}_t$ as all the history up to the end of round t. Let $\mathbb{E}_t\left[\cdot\right] = \mathbb{E}\left[\cdot|\mathcal{F}_{t-1}\right]$ and $\mathbb{P}_t\left(\cdot\right) = \mathbb{P}\left[\cdot|\mathcal{F}_{t-1}\right]$. We are now ready to present the following sufficient condition for exploration strategies.

Assumption 4 (Sufficient Condition)**.** *The sufficient condition includes two parts:*

(1) (Probability condition) $\mathbb{P}\left(E^{est}\right) \geq 1 - p_1$, $\mathbb{P}_t\left(E_t^{conc}\right) \geq 1 - p_{2,t}$, *and* $\mathbb{P}_t\left(E_t^{anti}\right) \geq p_3 > 0$ *for some time-dependent sequences* $c_{f,t}^{(1)}, c_{g,t}^{(1)}, c_{f,t}^{(2)}$, *and* $c_{g,t}^{(2)}$.

(2) (Boundedness condition) (i) There exists a positive probability p_4 *such that* $1 + \frac{2}{(p_3 - p_{2,t})} \leq 1/p_4$ *holds for all* t*; (ii)* $c_{f,t}^{(1)} + c_{f,t}^{(2)} \leq c_f(T)$ *and* $c_{g,t}^{(1)} + c_{g,t}^{(2)} \leq c_g(T)$ *for all* t*; (iii)* $\sum_{t=1}^{T} p_{2,t} \leq C'$ *for some absolute constant* C'.

Remark 3. *The above sufficient condition generalizes existing similar results (Kveton et al. 2019; Vaswani et al. 2019; Kim and Tewari 2020) in several aspects. First, existing works mainly focus on the MAB or linear bandit settings, which are special cases of our GP bandit setting (e.g., choosing a linear kernel leads to linear bandit). Second, while existing works only establish bounds on the expected regret, we aim to establish a high-probability bound. As a result, we need the additional boundedness condition, which, however, is simply for technical reasons. Third, in contrast to existing works that consider the unconstrained case only, we consider the constrained case, which is more challenging. Specifically, it requires that* E_t^{anti} *hold under policy* π^* *rather than under a single optimal action* x^*.

The following theorem presents general performance bounds under the above sufficient condition.

Theorem 2. *Suppose* $\rho \geq 4B/\delta$ *and* $V = G\sqrt{T}/\rho$*. Assume that CBO is equipped with exploration strategies that satisfy the sufficient condition in Assumption 4. Then, under regularity assumptions in Assumptions 2 and 3 and Slater's condition in Assumption 1, CBO achieves the following bounds on regret and constraint violation with probability at least* $1 - \alpha - p_1$ *for any* $\alpha \in (0,1)$*, where* $\kappa := B + \rho G$*:*

$$\begin{aligned}
\mathcal{R}_+(T) = O&\left(\frac{1}{p_4}c_f(T)\sqrt{T\gamma_T} + \frac{1}{p_4}\rho c_g(T)\sqrt{T\widetilde{\gamma}_T}\right)\\
&+ O\left(\rho G\sqrt{T} + \kappa\frac{c_f(T) + \rho c_g(T)}{p_4}\sqrt{2T\ln(1/\alpha)}\right),\\
\mathcal{V}(T) = O&\left(\frac{1}{\rho p_4}c_f(T)\sqrt{T\gamma_T} + \frac{1}{p_4}c_g(T)\sqrt{T\widetilde{\gamma}_T}\right)\\
&+ O\left(G\sqrt{T} + \kappa\frac{c_f(T) + \rho c_g(T)}{\rho p_4}\sqrt{2T\ln(1/\alpha)}\right).
\end{aligned}$$

3.3 Examples of CBO Algorithms

The sufficient condition not only recovers previous exploration strategies, but also reveals new ones. First, we remark that the first event E^{est} in the probability condition can be easily obtained by standard GP concentration result. That is, by (Chowdhury and Gopalan 2017, Theorem 2), we have $\mathbb{P}(\forall x, t, |f(x) - \mu_{t-1}(x)| \leq \beta_t \sigma_{t-1}(x)) \geq 1 - \alpha_f$ for any $\alpha_f \in (0,1)$, where $\beta_t = B + R\sqrt{2(\gamma_{t-1} + 1 + \ln(1/\alpha_f))}$ (similar for g). Thus, we have $\mathbb{P}(E^{est}) \geq 1 - p_1$ with $p_1 = \alpha_f + \alpha_g$, $c_{f,t}^{(1)} = \beta_t$, and $c_{g,t}^{(1)} = \widetilde{\beta}_t = B + R\sqrt{2(\widetilde{\gamma}_{t-1} + 1 + \ln(1/\alpha_g))}$. Thus, we only need to check probability condition for the remaining two events and the boundedness condition under different exploration methods.

First, it is expected that GP-UCB exploration satisfies the sufficient condition and hence our CBO-UCB also has performance guarantees given by Theorem 2. In particular, we see that Theorem 2 enjoys the same order of constraint violation as in Theorem 1. The regret bound has the same order as Theorem 1 but with an additional term due to the unified analysis (i.e., $\rho c_g(T)\sqrt{T\widetilde{\gamma}_T}$) .

Corollary 1. *GP-UCB with $\hat{\beta}_t$ being β_t and $-\widetilde{\beta}_t$ for $\mathcal{A}_f$ and $\mathcal{A}_g$ respectively, satisfies the sufficient condition.*

We can also show that the standard GP-TS exploration in Definition 1 satisfies the sufficient condition. Here, we mainly consider the case when π^* concentrates on a single point, which allows us to apply the standard anti-concentration results. One can possibly utilize advanced anti-concentration results for multivariate Gaussian distributions to attain the same result for a general π^*. Thus, we can instantiate CBO with GP-TS explorations, called CBO-TS, that also enjoys the guarantees in Theorem 2.

Corollary 2. *GP-TS with $\hat{\beta}_t$ being β_t and $\widetilde{\beta}_t$ for $\mathcal{A}_f$ and $\mathcal{A}_g$ respectively, satisfies the sufficient condition when π^* concentrates on a single point.*

Our derived sufficient condition also allows us to design CBO algorithms with new exploration strategies. In the following, inspired by (Vaswani et al. 2019), we propose a new GP based exploration strategy, which aims to strike a balance between GP-UCB and GP-TS explorations.

Definition 2 (RandGP-UCB Exploration)**.** *Suppose that the posterior distribution for a black-box function h in round t is given by $\mathcal{GP}(\hat{\mu}_{t-1}(\cdot), \hat{k}_{t-1}(\cdot,\cdot))$. Then, the estimate of h in round t under RandGP-UCB exploration strategy is $h_t(\cdot) = \hat{\mu}_{t-1}(\cdot) + \hat{Z}_t\hat{\sigma}_{t-1}(\cdot)$, where $\hat{Z}_t \sim \hat{\mathcal{D}}$ for some distribution $\hat{\mathcal{D}}$ and $\hat{\sigma}_{t-1}(x) = \hat{k}(x,x)$ for all $x \in \mathcal{X}$.*

In contrast to GP-UCB, RandGP-UCB replaces the deterministic confidence bound by a randomized one. Compared to GP-TS, RandGP-UCB uses "coupled" noise in the sense that all the actions share the same noise $\hat{Z}_t$ rather than "decoupled" and correlated noise in GP-TS. This subtle difference will not only help to eliminate the additional factor $\sqrt{\ln(|\mathcal{X}|)}$ in GP-TS due to the use of union bound, but also allow us to deal with a general π^*. One possible disadvantage of RandGP-UCB (compared to GP-TS) is that GP-TS could be offline oracle-optimization efficient for the step in line 6 of Algorithm 1 while RandGP-UCB (also GP-UCB) is not, which shares the standard pattern as in linear bandits.

Corollary 3. *RandGP-UCB with $\hat{\mathcal{D}}$ being $\mathcal{N}(0,\beta_t^2)$ and $\mathcal{N}(0,\widetilde{\beta}_t^2)$ for $\mathcal{A}_f$ and $\mathcal{A}_g$ respectively, satisfies the sufficient condition.*

Algorithm 2: SCGP-UCB Algorithm

```
Input: $V_t$, $\epsilon_t$, $\beta_t$, $Q(1) = 0$,
       $\mu_0(x) = 0, \sigma_0(x) = 1, \forall x$
for $t = 1, 2, \ldots, T$ do
    Generate estimate $f_t(x) = \mu_{t-1}(x) + \beta_t\sigma_{t-1}(x)$
    Truncate $f_t(x)$ as $\bar{f}_t(x) = \text{Proj}_{[-B,B]} f_t(x)$
    Observe the random sample $g_t(\cdot)$
    Pseudo-acquisition function:
    $z_t(x) = \bar{f}_t(x) - \frac{1}{V_t}Q(t)g_t(x)$
    Choose action $x_t = \arg\max_{x\in\mathcal{X}} z_t(x)$; observe $r_t$
    Update virtual queue:
    $Q(t+1) = [Q(t) + g_t(x_t) + \epsilon_t]_+$
    Posterior model update: update $(\mu_t, \sigma_t)$ via GP
      regression using new data $(x_t, r_t)$
```

Thus, one can instantiate CBO with RandGP-UCB exploration to obtain a new algorithm called CBO-Rand with performance guarantees given by Theorem 2. Note that RandGP-UCB with other distributions $\hat{\mathcal{D}}$ can also satisfy the sufficient condition (as discussed in the Appendix).

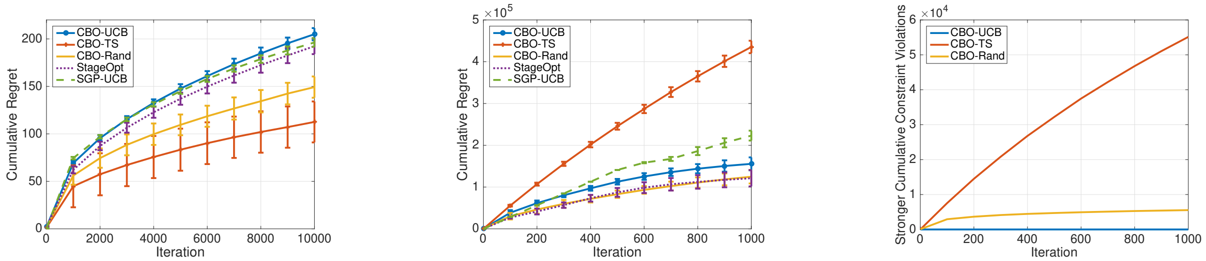
4 Stochastic Full Constraint Information

In this section, we consider a different type of feedback information for the constraint function: stochastic full constraint information. That is, before making a decision in each round, the learning agent observes a noisy sample of the true constraint function. It is not surprising that in this case, one will likely achieve a better constraint violation bound than in the previous bandit case, as additional information about the constraint is available to the learning agent. In the following, we confirm this intuition by showing that even zero (expected) constraint violation can be achieved in this case. We start with the following standard assumptions.

Assumption 5 (Randomness)**.** *Assume that $g_t(\cdot)$ at each round t is an* i.i.d. *sample of the true unknown function $g(\cdot)$ and it is independent of all other randomness.*

Assumption 6 (Boundedness)**.** *We assume that random samples of the constraint function are bounded, i.e., $|g_t(x)| \leq G$ for all $x \in \mathcal{X}$ and for all t.*

We now introduce a new algorithm named stochastic constrained GP-UCB (SCGP-UCB) as in Algorithm 2. SCGP-UCB shares the same key idea with CBO for the bandit case and is also based on primal-dual optimization, but with some key differences. We use $Q(t)$ to represent a V_t-scaled version of the dual variable since its update rule (recall that $[\cdot]_+ = \max\{\cdot, 0\}$) behaves similarly to the queue length evolution in queueing theory, and hence, we call it a virtual queue. Moreover, in order to bound constraint violation, we will resort to Lyapunov-drift analysis – a powerful tool from queueing theory and control, which fully utilizes the available constraint information. Another key difference here is that we add an additional term ϵ_t in the virtual queue update (also see works (Liu et al. 2021; Kunniyur and



(a) Regret on synthetic data

(b) Regret on real-world data

(c) Constraint violations on real-world data

Figure 1: Experimental results on BO with soft constraints.

Srikant 2001)). A properly chosen ϵ_t can lead to zero constraint violation (in expectation). Finally, we note that both V_t and ϵ_t are time dependent (in contrast to V in CBO). This modification can lead to a stronger *anytime* guarantee (in expectation), e.g., constraint violation bound for any $\tau \in [T]$. The main result is stated in the following theorem.

Theorem 3. *Let $\beta_t = B + R\sqrt{2(\gamma_{t-1} + 1 + \ln(2/\delta))}$. Under the Slater's condition in Assumption 1, and the randomness and boundedness assumptions in Assumptions 5 and 6, with $\epsilon_t = \frac{1}{\sqrt{t}}$ and $V_t = \frac{\delta}{8B}\sqrt{t}$, SCGP-UCB guarantees that for all $\tau \in [T]$,*

$$\mathbb{E}\left[\mathcal{R}_+(\tau)\right] = O\left(\frac{BG^2\sqrt{\tau}}{\delta} + \frac{BG}{\delta^3} + B\sqrt{\tau\gamma_\tau} + \gamma_\tau\sqrt{\tau}\right),$$

$$\mathbb{E}\left[\mathcal{V}(\tau)\right] = \begin{cases} O(G^2/\delta^2) & \tau \leq O(1/\delta^4) \\ 0 & \textit{otherwise} \end{cases}.$$

5 Numerical Experiments

In this section, we conduct simulations to compare the performance of our algorithms (i.e., `CBO-UCB`, `CBO-TS`, and `CBO-Rand`, that is, CBO with GP-UCB, GP-TS, and RandGP-UCB explorations, respectively) with existing safe BO algorithms based on both synthetic and real-world datasets. In particular, we consider the two most recent safe BO algorithms: `StageOpt` [28] (which has a superior performance compared to `SafeOpt` [27]) and `SGP-UCB` [3]. For a fair comparison, we mainly focus on the bandit constraint information case as in the existing works and show that our proposed CBO algorithms can trade a slight performance in constraint violation for improvement in the reward regret with a reduced computation complexity and flexible implementations.

Datasets and Settings. We consider both synthetic data and real-world data (see Appendix for details and additional results). The parameters of each algorithm are set order-wise similar to those recommended by the theorems. We run each algorithm for 50 independent trials and plot the average (and error bar) along with iterations, as shown in Figure 1.

Results. We summarize the results in terms of regret, constraint violation and practical considerations as follows.

Regret: Our three CBO algorithms achieve a better (or similar) regret performance compared to the existing safe BO algorithms (see Figures 1(a) and 1(b)). Among the three CBO algorithms, the best algorithm really depends on the applications, but `CBO-Rand` appears to have reasonably good performance at all times.

Constraint violation: Since we have $\mathcal{V}(T) = 0$ under all the algorithms, we study the total number of rounds where the constraint is violated, denoted by N. In the synthetic data setting, our proposed CBO algorithms have $N \leq 5$ over $T = 10,000$ rounds; in the real-world data setting, `CBO-UCB` enjoys $N = 0$ and `CBO-Rand` has an average $N = 38$ over a horizon $T = 1,000$. Furthermore, we plot the stronger cumulative constraint violations given by $\sum_{t=1}^{T}[g(x_t)]_+$ as shown in Figure 1(c), from which we can see that all CBO algorithms achieve sublinear performance even with respect to this stronger metric.

Practical considerations: Our proposed CBO algorithms have the same computational complexity as the unconstrained case. In particular, they scale linearly with the number of actions in the discrete-domain case. On the other hand, `StageOpt` scales quadratically due to the construction of the safe set, and `SGP-UCB` requires the additional random initialization stage, which leads to linear regret at the beginning of the learning process. Moreover, standard methods for improving the scalability of unconstrained BO can be naturally applied to our CBO algorithms. Finally, both `StageOpt` and `SGP-UCB` require the knowledge of a safe action (i.e., one that satisfies the constraint) in advance, and moreover, `StageOpt` requires f to be Lipschitz and needs to estimate the Lipschitz constant, which impacts the robustness. In contrast, CBO algorithms only require a mild Slater's condition as in Assumption 1, which does not necessarily require the existence of a safe action.

6 Conclusion

We studied BO with unknown soft constraints under both bandit and stochastic full information feedback for the constraint. In the bandit case, we presented a general framework for constrained BO with soft constraints via primal-dual optimization. Armed with our developed sufficient condition, this framework not only allows us to design provably efficient (i.e., sublinear reward regret and sublinear total constraint violation) CBO algorithms with both UCB and TS explorations, but presents a unified method to design new effective ones. We further show that with stochastic full-information for the constraint, one can judiciously achieve an improved constraint violations. We finally conduct experiments based on both synthetic and real-world data to show that in the soft-constraint setting, our proposed CBO algorithms tend to have superior regret performance compared to existing algorithms while enjoying reduced computation, improved robustness and flexible implementations.

References

Agrawal, S.; and Devanur, N. 2016. Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems*, 29: 3450–3458.

Amani, S.; Alizadeh, M.; and Thrampoulidis, C. 2019. Linear stochastic bandits under safety constraints. *arXiv preprint arXiv:1908.05814*.

Amani, S.; Alizadeh, M.; and Thrampoulidis, C. 2020. Regret Bound for Safe Gaussian Process Bandit Optimization. In *Learning for Dynamics and Control*, 158–159. PMLR.

Badanidiyuru, A.; Kleinberg, R.; and Slivkins, A. 2013. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, 207–216. IEEE.

Beck, A. 2017. *First-order methods in optimization*. SIAM.

Berkenkamp, F.; Krause, A.; and Schoellig, A. P. 2016. Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics. *arXiv preprint arXiv:1602.04450*.

Chen, Y.; Cuellar, A.; Luo, H.; Modi, J.; Nemlekar, H.; and Nikolaidis, S. 2020. Fair contextual multi-armed bandits: Theory and experiments. In *Conference on Uncertainty in Artificial Intelligence*, 181–190. PMLR.

Chowdhury, S. R.; and Gopalan, A. 2017. On kernelized multi-armed bandits. *arXiv preprint arXiv:1704.00445*.

Chowdhury, S. R.; and Gopalan, A. 2019. Bayesian optimization under heavy-tailed payoffs. In *Advances in Neural Information Processing Systems*, 13790–13801.

Ding, D.; Wei, X.; Yang, Z.; Wang, Z.; and Jovanovic, M. 2021. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, 3304–3312. PMLR.

Efroni, Y.; Mannor, S.; and Pirotta, M. 2020. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*.

Eriksson, D.; and Poloczek, M. 2021. Scalable constrained bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, 730–738. PMLR.

Garcelon, E.; Ghavamzadeh, M.; Lazaric, A.; and Pirotta, M. 2020. Improved algorithms for conservative exploration in bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 3962–3969.

Gelbart, M. A.; Snoek, J.; and Adams, R. P. 2014. Bayesian optimization with unknown constraints. *arXiv preprint arXiv:1403.5607*.

Hajek, B. 1982. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Advances in Applied probability*, 502–525.

Hernández-Lobato, J. M.; Gelbart, M. A.; Adams, R. P.; Hoffman, M. W.; and Ghahramani, Z. 2016. A general framework for constrained Bayesian optimization using information-based search.

Kazerouni, A.; Ghavamzadeh, M.; Abbasi-Yadkori, Y.; and Van Roy, B. 2016. Conservative contextual linear bandits. *arXiv preprint arXiv:1611.06426*.

Kim, B.; and Tewari, A. 2020. Randomized Exploration for Non-Stationary Stochastic Linear Bandits. In *Conference on Uncertainty in Artificial Intelligence*, 71–80. PMLR.

Kunniyur, S.; and Srikant, R. 2001. Analysis and design of an adaptive virtual queue (AVQ) algorithm for active queue management. *ACM SIGCOMM Computer Communication Review*, 31(4): 123–134.

Kveton, B.; Szepesvari, C.; Ghavamzadeh, M.; and Boutilier, C. 2019. Perturbed-history exploration in stochastic linear bandits. *arXiv preprint arXiv:1903.09132*.

Li, F.; Liu, J.; and Ji, B. 2019. Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering*, 7(3): 1799–1813.

Liu, X.; Li, B.; Shi, P.; and Ying, L. 2020. POND: Pessimistic-Optimistic oNline Dispatch. *arXiv preprint arXiv:2010.09995*.

Liu, X.; Li, B.; Shi, P.; and Ying, L. 2021. An Efficient Pessimistic-Optimistic Algorithm for Stochastic Linear Bandits with General Constraints. *arXiv preprint arXiv:2102.05295*.

Moradipari, A.; Amani, S.; Alizadeh, M.; and Thrampoulidis, C. 2019. Safe Linear Thompson Sampling with Side Information. *arXiv preprint arXiv:1911.02156*.

Pacchiano, A.; Ghavamzadeh, M.; Bartlett, P.; and Jiang, H. 2021. Stochastic bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, 2827–2835. PMLR.

Rasmussen, C. E. 2003. Gaussian processes in machine learning. In *Summer School on Machine Learning*, 63–71. Springer.

Srinivas, N.; Krause, A.; Kakade, S. M.; and Seeger, M. 2009. Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*.

Sui, Y.; Burdick, J.; Yue, Y.; et al. 2018. Stagewise safe bayesian optimization with gaussian processes. In *International Conference on Machine Learning*, 4781–4789. PMLR.

Sui, Y.; Gotovos, A.; Burdick, J.; and Krause, A. 2015. Safe exploration for optimization with Gaussian processes. In *International Conference on Machine Learning*, 997–1005. PMLR.

Vaswani, S.; Mehrabian, A.; Durand, A.; and Kveton, B. 2019. Old dog learns new tricks: Randomized ucb for bandit problems. *arXiv preprint arXiv:1910.04928*.

Wei, X.; Yu, H.; and Neely, M. J. 2020. Online primal-dual mirror descent under stochastic constraints. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 4(2): 1–36.

Wu, H.; Srikant, R.; Liu, X.; and Jiang, C. 2015. Algorithms with logarithmic or sublinear regret for constrained contextual bandits. *arXiv preprint arXiv:1504.06937*.

Wu, Y.; Shariff, R.; Lattimore, T.; and Szepesvári, C. 2016. Conservative bandits. In *International Conference on Machine Learning*, 1254–1262. PMLR.

Yu, H.; Neely, M. J.; and Wei, X. 2017. Online convex optimization with stochastic constraints. *arXiv preprint arXiv:1708.03741*.

Appendix

A Applications with stochastic constraint information

Let's walk through the following two examples.

- The first one can be seen as a bandit variant of the example considered in (Yu, Neely, and Wei 2017) (i.e., online convex optimization with stochastic constraints). We consider the optimal resource allocation (e.g.. GPU) among two computing nodes and each of them serves a population of users. The objective is to find the optimal resource allocation that minimize the total energy cost while guaranteeing stability of the system. In particular, let $x = (x_1, x_2)$ be the GPU resource allocations among two computing nodes and $f(x)$ is the bandit feedback of total energy cost. Let $\mu_i := h_i(x_i)$ be the service rate of computing node i with GPU resource x_i and we assume that $h_i(\cdot)$ is known, which is typically the case. Let $\omega(t)$ be a random number of incoming jobs (e.g., Poisson distribution) with mean λ. The stochastic full information constraint function is given by $g_t(x) = \omega(t) - \sum_{i=1}^2 \mu_i = \omega(t) - \sum_{i=1}^2 h_i(x_i)$. To guarantee stability, we needs to make sure that the cumulative arrivals are smaller than the total service capabilities, i.e., sublinear cumulative constraints. This example also demonstrates one typical use case of considering soft cumulative constraints in our paper rather than the hard constraint in previous works.
- We also briefly mention that stochastic full information constraint can be used to model a soft guidance at the beginning process of learning. In particular, in robot applications, we often aim to restrict the robot to take actions that has a similar distribution to an existing expert policy, especially in the initial stage of learning. We can model this constraint by using non-parametric maximum mean discrepancy (MMD) as the distribution distance, which can be estimated empirically based on data in an unbiased way.

B Proof of Theorem 1

Before we present the proof, we first obtain the following lemma on the dual variable.

Lemma 1. *Under the update rule of ϕ_t in Algorithm 1, we have for any $\phi \in [0, \rho]$,*

$$\sum_{t=1}^T \bar{g}_t(x_t)(\phi - \phi_t) \leq \frac{V}{2}(\phi_1 - \phi)^2 + \sum_{t=1}^T \frac{1}{2V}\bar{g}_t(x_t)^2.$$

Proof. By the dual variable update rule in Algorithm 1 and the non-expansiveness of projection to $[0, \rho]$, we have

$$\begin{aligned}(\phi_{t+1} - \phi)^2 &\leq (\phi_t + \frac{1}{V}\bar{g}_t(x_t) - \phi)^2 \\ &= (\phi_t - \phi)^2 + \frac{2}{V}\bar{g}_t(x_t)(\phi_t - \phi) + \frac{1}{V^2}\bar{g}_t(x_t)^2.\end{aligned}$$

Summing over T steps and multiplying both sides by $\frac{V}{2}$, we have

$$\frac{V}{2}(\phi_{T+1} - \phi)^2 - \frac{V}{2}(\phi_1 - \phi)^2 \leq \sum_{t=1}^T \bar{g}_t(x_t)(\phi_t - \phi) + \sum_{t=1}^T \frac{1}{2V}\bar{g}_t(x_t)^2.$$

Hence,

$$\sum_{t=1}^T \bar{g}_t(x_t)(\phi - \phi_t) \leq \frac{V}{2}(\phi_1 - \phi)^2 + \sum_{t=1}^T \frac{1}{2V}\bar{g}_t(x_t)^2, \tag{4}$$

which completes the proof. □

Now, we are ready to present the proof of Theorem 1.

Proof of Theorem 1. Under Slater condition in Assumption 1, we have the boundedness of the optimal dual solution by standard convex optimization analysis (cf. (Beck 2017, Theorem 8.42))

$$0 \leq \phi^* \leq \frac{(\mathbb{E}_{\pi^*}[f(x)] - \mathbb{E}_{\pi_0}[f(x)])}{\delta} \leq \frac{2B}{\delta},$$

where the last inequality holds by the boundedness of $f(x)$. Note that the reason why we can use convex analysis is that $\mathbb{E}_\pi[h(x)]$ for any fixed h is a linear function with respect to π (and is thus convex). Now, we turn to establish a bound over

$\mathcal{R}(T) + \phi \sum_{t=1}^{T} g(x_t)$. First, note that

$$\begin{aligned}
&\mathcal{R}(T) + \phi \sum_{t=1}^{T} g(x_t) \\
=& T\mathbb{E}_{\pi^*}\left[f(x)\right] - \sum_{t=1}^{T} f(x_t) + \phi \sum_{t=1}^{T} g(x_t) \\
=& T\mathbb{E}_{\pi^*}\left[f(x)\right] - \sum_{t=1}^{T} \bar{f}_t(x_t) + \sum_{t=1}^{T} \bar{f}_t(x_t) - f(x_t) + \phi \sum_{t=1}^{\tau} g(x_t). \qquad (5)
\end{aligned}$$

We can further bound (5) by using Lemma 1. In particular, we have

$$\begin{aligned}
& T\mathbb{E}_{\pi^*}\left[f(x)\right] - \sum_{t=1}^{T} \bar{f}_t(x_t) + \sum_{t=1}^{T} \bar{f}_t(x_t) - f(x_t) + \phi \sum_{t=1}^{\tau} g(x_t) \\
\overset{(a)}{\leq} & \sum_{t=1}^{T} \mathbb{E}_{\pi^*}\left[f(x)\right] - \phi_t \mathbb{E}_{\pi^*}\left[g(x)\right] - \sum_{t=1}^{T} \bar{f}_t(x_t) + \sum_{t=1}^{T} \bar{f}_t(x_t) - f(x_t) + \phi \sum_{t=1}^{T} g(x_t) \\
\overset{(b)}{=} & \sum_{t=1}^{T} \mathbb{E}_{\pi^*}\left[f(x)\right] - \phi_t \mathbb{E}_{\pi^*}\left[g(x)\right] - \left(\sum_{t=1}^{T} \bar{f}_t(x_t) - \phi_t \bar{g}_t(x_t) \right) + \sum_{t=1}^{T} \bar{f}_t(x_t) - f(x_t) \\
& + \phi \left(\sum_{t=1}^{T} g(x_t) - \bar{g}_t(x_t) \right) + \sum_{t=1}^{T} \bar{g}_t(x_t)(\phi - \phi_t) \\
\overset{(c)}{\leq} & \sum_{t=1}^{T} \mathbb{E}_{\pi^*}\left[f(x)\right] - \phi_t \mathbb{E}_{\pi^*}\left[g(x)\right] - \left(\sum_{t=1}^{T} \bar{f}_t(x_t) - \phi_t \bar{g}_t(x_t) \right) + \sum_{t=1}^{\tau} \bar{f}_t(x_t) - f(x_t) \\
& + \phi \left(\sum_{t=1}^{T} g(x_t) - \bar{g}_t(x_t) \right) + \frac{V}{2}(\phi_1 - \phi)^2 + \sum_{t=1}^{T} \frac{1}{2V} \bar{g}_t(x_t)^2 \\
\overset{(d)}{=} & \mathcal{T}_1 + \mathcal{T}_2 + \frac{V}{2}\phi^2 + \frac{1}{2V} T G^2, \qquad (6)
\end{aligned}$$

where (a) holds since $\phi_t \geq 0$ and $\mathbb{E}_{\pi^*}\left[g(x)\right] \leq 0$; (b) holds by adding and subtracting terms; (c) follow from Lemma 1 to bound the last term; (d) holds by the fact $\phi_1 = 0$, the boundedness of $\bar{g}_t$ and the definitions of $\mathcal{T}_1$ and $\mathcal{T}_2$, i.e.,

$$\mathcal{T}_1 = \sum_{t=1}^{T} (\mathbb{E}_{\pi^*}\left[f(x)\right] - \phi_t \mathbb{E}_{\pi^*}\left[g(x)\right]) - \sum_{t=1}^{T} (\bar{f}_t(x_t) - \phi_t \bar{g}_t(x_t)), \qquad (7)$$

$$\mathcal{T}_2 = \sum_{t=1}^{T} (\bar{f}_t(x_t) - f(x_t)) + \phi \sum_{t=1}^{T} (g(x_t) - \bar{g}_t(x_t)). \qquad (8)$$

Plugging (6) into (5), yields for any $\phi \in [0, \rho]$,

$$\mathcal{R}(T) + \phi \sum_{t=1}^{T} g(x_t) \leq \mathcal{T}_1 + \mathcal{T}_2 + \frac{V}{2}\phi^2 + \frac{1}{2V} T G^2. \qquad (9)$$

First, assume that we already have a bound on $\mathcal{T}_1 + \mathcal{T}_2$, i.e., $\mathcal{T}_1 + \mathcal{T}_2 \leq \chi(T, \phi)$ with high probability, and $\chi(T, \phi)$ is an increasing function in ϕ. This directly leads to the following inequality (with $V = G\sqrt{T}/\rho$) for any $\phi \in [0, \rho]$:

$$\mathcal{R}_{+}(T) + \phi \sum_{t=1}^{T} g(x_t) \leq \chi(T, \phi) + \frac{\phi^2 G \sqrt{T}}{2\rho} + \frac{\rho G \sqrt{T}}{2}. \qquad (10)$$

Based on this key inequality, we can analyze both regret and constraint violation.

Regret. We can simply choose $\phi = 0$ in (10), and obtain that with high probability

$$\mathcal{R}_{+}(T) = O\left(\chi(T, 0) + \rho G \sqrt{T} \right). \qquad (11)$$

Constraint violation. To obtain the bound on $\mathcal{V}(T)$, inspired by (Efroni, Mannor, and Pirotta 2020), we will resort to tools from constrained convex optimization. First, we have $\frac{1}{T}\sum_{t=1}^{T} f(x_t) = \mathbb{E}_{\pi'}\left[f(x)\right]$ and $\frac{1}{T}\sum_{t=1}^{T} g(x_t) = \mathbb{E}_{\pi'}\left[g(x)\right]$ for some probability measure π' by the convexity of probability measure. As a result, we have

$$\mathbb{E}_{\pi^*}\left[f(x)\right] - \mathbb{E}_{\pi'}\left[f(x)\right] + \rho\left[\mathbb{E}_{\pi'}\left[g(x)\right]\right]_+ = \frac{1}{T}\mathcal{R}_+(T) + \frac{1}{T}\phi\sum_{t=1}^{T} g(x_t) \leq \frac{\chi(T,\rho)+\rho G\sqrt{T}}{T}, \tag{12}$$

where $[a]_+ := \max\{0,a\}$, and the first equality holds by choosing $\phi = \rho$ if $\sum_{t=1}^{T} g(x_t) \geq 0$, and otherwise $\phi = 0$, and the second inequality holds by upper bounding RHS of (3) with $\phi = \rho$ since (10) holds for all $\phi \in [0,\rho]$ and $\chi(T,\phi)$ is increasing in ϕ.

Then, we will apply the following useful lemma, which is adapted from Theorem 3.60 in (Beck 2017).

Lemma 2. *Consider the following convex constrained problem $h(\pi^*) = \max_{\pi\in\mathcal{C}}\{h(\pi) : w(\pi) \leq 0\}$,where both h and w are convex over the convex set $\mathcal{C}$ in a vector space. Suppose $h(\pi^*)$ is finite and there exists a slater point π_0 such that $w(\pi_0) \leq -\delta$, and a constant $\rho \geq 2\kappa^*$, where κ^* is the optimal dual variable, i.e., $\kappa^* = \arg\min_{\lambda\geq 0}(\max_{\pi} h(\pi) - \kappa w(\pi))$. Assume that $\pi' \in \mathcal{C}$ satisfies*

$$h(\pi^*) - h(\pi') + \rho\left[w(\pi')\right]_+ \leq \varepsilon, \tag{13}$$

for some $\varepsilon > 0$, then we have $[w(\pi')]_+ \leq 2\varepsilon/\rho$.

Thus, since (12) satisfies (13) and $\mathbb{E}_{\pi}\left[h(x)\right]$ for any fixed h is a linear function with respect to π, by Lemma 2, we have

$$\mathcal{V}(T) = O\left(\frac{1}{\rho}\chi(T,\rho) + G\sqrt{T}\right). \tag{14}$$

We are only left to bound $\mathcal{T}_1 + \mathcal{T}_2$ by $\chi(T,\phi)$. To this end, we will resort to standard concentration results for GP bandits. First, by (Chowdhury and Gopalan 2017, Theorem 2), we have the following lemma.

Lemma 3. *Fix $\alpha \in (0,1]$, with probability at least $1-\alpha$, the followings hold simultaneously for all $t \in [T]$ and all $x \in \mathcal{X}$*

$$|f(x) - \mu_{t-1}(x)| \leq \beta_t\sigma_{t-1}(x), \quad |g(x) - \widetilde{\mu}_{t-1}(x)| \leq \widetilde{\beta}_t\widetilde{\sigma}_{t-1}(x),$$

Thus, based on this lemma and the definition of GP-UCB exploration, we have with high probability, $f_t(x) \geq f(x)$ and $g_t(x) \leq g(x)$ for all $t \in [T]$ and $x \in \mathcal{X}$. This directly implies that $\bar{f}_t(x) \geq f(x)$ and $\bar{g}_t(x) \leq g(x)$ for all $t \in [T]$ and $x \in \mathcal{X}$ (i.e., optimistic estimates), which holds by $|f(x)| \leq B$ and $|g(x)| \leq G$ and the way of truncation in Algorithm 1. Now, to bound $\mathcal{T}_1$ in (7), we have

$$\begin{aligned}
\mathcal{T}_1 &= \sum_{t=1}^{T}(\mathbb{E}_{\pi^*}\left[f(x)\right] - \mathbb{E}_{\pi^*}\left[\bar{f}_t(x)\right] + \mathbb{E}_{\pi^*}\left[\bar{f}_t(x)\right] - \bar{f}_t(x_t)) \\
&\quad + \phi_t\sum_{t=1}^{T}(\bar{g}_t(x_t) - \mathbb{E}_{\pi^*}\left[\bar{g}_t(x)\right] + \mathbb{E}_{\pi^*}\left[\bar{g}_t(x)\right] - \mathbb{E}_{\pi^*}\left[g(x)\right]) \\
&\overset{(a)}{\leq} \sum_{t=1}^{T}\left(\mathbb{E}_{\pi^*}\left[\bar{f}_t(x)\right] - \bar{f}_t(x_t)\right) + \phi_t\sum_{t=1}^{T}\left(\bar{g}_t(x_t) - \mathbb{E}_{\pi^*}\left[\bar{g}_t(x)\right]\right) \\
&= \sum_{t=1}^{T}\left(\mathbb{E}_{\pi^*}\left[\bar{f}_t(x)\right] - \phi_t\mathbb{E}_{\pi^*}\left[\bar{g}_t(x)\right] - (\bar{f}_t(x_t) - \phi_t\bar{g}_t(x_t))\right) \\
&\overset{(b)}{\leq} 0,
\end{aligned}$$

where (a) holds by the fact that estimates are optimistic, i.e., $\bar{f}_t(x) \geq f(x)$ and $\bar{g}_t(x) \leq g(x)$ for all $t \in [T]$ and $x \in \mathcal{X}$; (b) holds by the greedy selection of Algorithm 1.

Now, we turn to bound $\mathcal{T}_2$. In particular, we have

$$\begin{aligned}
\mathcal{T}_2 &\overset{(a)}{\leq} \sum_{t=1}^{T} 2\beta_t\sigma_{t-1}(x_t) + \phi\sum_{t=1}^{T} 2\widetilde{\beta}_t\widetilde{\sigma}_{t-1}(x_t) \\
&\overset{(b)}{\leq} O\left(\beta_T\sqrt{T\gamma_T} + \phi\widetilde{\beta}_T\sqrt{T\widetilde{\gamma}_T}\right),
\end{aligned} \tag{15}$$

where (a) holds by Lemma 3 and the definition of GP-UCB exploration, i.e., $f_t(x) = \mu_{t-1}(x) + \beta_t\sigma_{t-1}(x)$ and $g_t(x) = \widetilde{\mu}_{t-1}(x) - \widetilde{\beta}_t\widetilde{\sigma}_{t-1}(x)$. Note that truncation also does not affect this step; (b) holds by Cauchy-Schwartz inequality and the bound of sum of predictive variance (cf. (Chowdhury and Gopalan 2017, Lemma 4)). Note that we have also used the fact that β_t and $\widetilde{\beta}_t$ is increasing in t.

Putting the bounds on $\mathcal{T}_1$ and $\mathcal{T}_2$ together, we have obtained that with high probability

$$\mathcal{T}_1 + \mathcal{T}_2 \leq \chi(T, \phi) := O\left(\beta_T\sqrt{T\gamma_T} + \phi\widetilde{\beta}_T\sqrt{T\widetilde{\gamma}_T}\right).$$

Finally, plugging $\chi(T, 0)$ into (11), yield the regret bound as follows (note that $\beta_t = B + R\sqrt{2(\gamma_{t-1} + 1 + \ln(2/\alpha))}$)

$$\mathcal{R}_+(T) = O\left(B\sqrt{T\gamma_T} + \sqrt{T\gamma_T(\gamma_T + \ln(2/\alpha))} + \rho G\sqrt{T}\right),$$

and plugging $\chi(T, \rho)$ into (14), yields the bound on constraint violation as

$$\mathcal{V}(T) = O\left(\left(1 + \frac{1}{\rho}\right)\left(C\sqrt{T\hat{\gamma}_T} + \sqrt{T\hat{\gamma}_T(\hat{\gamma}_T + \ln(2/\alpha))}\right) + G\sqrt{T}\right),$$

where $C := \max\{B, G\}$ and $\hat{\gamma}_T := \max\{\gamma_T, \widetilde{\gamma}_T\}$. Hence, it completes the proof. □

C Proof of Theorem 2

Before we present the proof, we introduce a new notation to make the presentation easier. In particular, we let $h(\pi) := \mathbb{E}_\pi[h(x)]$ for any function h and π_t is a dirac delta function at the point x_t.

Proof of Theorem 2. As shown in the proof of Theorem 1, all we need to do is to find a high probability bound over $\mathcal{T}_1 + \mathcal{T}_2$ under the sufficient condition in Assumption 4. Under our newly introduced notation, we have

$$\mathcal{T}_1 + \mathcal{T}_2 = \sum_{t=1}^{T}\left(z_{\phi_t}(\pi^*) - \hat{z}_{\phi_t}(\pi_t) + \hat{z}_\phi(\pi_t) - z_\phi(\pi_t)\right) := \sum_{t=1}^{T} d_t, \tag{16}$$

where $z_{\phi_t}(\cdot) := f(\cdot) - \phi_t g(\cdot)$ and $\hat{z}_{\phi_t}(\cdot) := \bar{f}_t(\cdot) - \phi_t\bar{g}_t(\cdot)$, and similar definitions for z_ϕ and $\hat{z}_\phi$.

Let $\Delta_{\phi_t}(\pi) := z_{\phi_t}(\pi^*) - z_{\phi_t}(\pi) = (f(\pi^*) - \phi_t g(\pi^*)) - (f(\pi) - \phi_t g(\pi))$. Then, we define the 'undersampled' set as

$$\bar{S}_t := \{\pi \in \Pi : \alpha_{\phi_t}(\pi) := c_{f,t}\sigma_{t-1}(\pi) + \phi_t c_{g,t}\widetilde{\sigma}_{t-1}(\pi) \geq \Delta_{\phi_t}(\pi)\},$$

where $c_{f,t} = (c_{f,t}^{(1)} + c_{f,t}^{(2)})$ and $c_{g,t} = (c_{g,t}^{(1)} + c_{g,t}^{(2)})$ (similarly $\alpha_\phi(\pi) := c_{f,t}\sigma_{t-1}(\pi) + \phi c_{g,t}\widetilde{\sigma}_{t-1}(\pi)$). Let $u_t = \arg\min_{\pi\in\bar{S}_t}\alpha_{\phi_t}(\pi)$. Thus, conditioned on E^{est} and E_t^{conc}, we have

$$\begin{aligned}
d_t &= z_{\phi_t}(\pi^*) - \hat{z}_{\phi_t}(\pi_t) + \hat{z}_\phi(\pi_t) - z_\phi(\pi_t)\\
&= z_{\phi_t}(\pi^*) - z_{\phi_t}(u_t) + z_{\phi_t}(u_t) - \hat{z}_{\phi_t}(\pi_t) + \hat{z}_\phi(\pi_t) - z_\phi(\pi_t)\\
&= \Delta_{\phi_t}(u_t) + z_{\phi_t}(u_t) - \hat{z}_{\phi_t}(\pi_t) + \hat{z}_\phi(\pi_t) - z_\phi(\pi_t)\\
&\overset{(a)}{\leq} \Delta_{\phi_t}(u_t) + \hat{z}_{\phi_t}(u_t) - \hat{z}_{\phi_t}(\pi_t) + \alpha_{\phi_t}(u_t) + \alpha_\phi(\pi_t)\\
&\overset{(b)}{\leq} \Delta_{\phi_t}(u_t) + \alpha_{\phi_t}(u_t) + \alpha_\phi(\pi_t)\\
&\overset{(c)}{\leq} 2\alpha_{\phi_t}(u_t) + \alpha_\phi(\pi_t),
\end{aligned} \tag{17}$$

where (a) holds since under event $E^{est} \cap E_t^{conc}$, for all x, $|f(x) - f_t(x)| \leq (c_{f,t}^{(1)} + c_{f,t}^{(2)})\sigma_{t-1}(x)$ and $|g(x) - g_t(x)| \leq (c_{g,t}^{(1)} + c_{g,t}^{(2)})\widetilde{\sigma}_{t-1}(x)$ and the facts that $|g(x) - \bar{g}_t(x)| \leq |g(x) - g_t(x)|$ since $|g(x)| \leq G$ and $|f(x) - \bar{f}_t(x)| \leq |f(x) - f_t(x)|$ since $|f(x)| \leq B$; (b) holds by the greedy selection in Algorithm 1; (c) follows from $u_t \in \bar{S}_t$. Thus, conditioned on E^{est}, we have

$$\begin{aligned}
\mathbb{E}_t[d_t] &= \mathbb{E}_t\left[d_t I\{E_t^{conc}\}\right] + \mathbb{E}_t\left[d_t I\{\bar{E}_t^{conc}\}\right]\\
&\overset{(a)}{\leq} \mathbb{E}_t\left[r_t I\{E_t^{conc}\}\right] + (4B + 4\rho G)p_{2,t}\\
&\overset{(b)}{\leq} \mathbb{E}_t\left[\alpha_\phi(\pi_t)\right] + 2\alpha_{\phi_t}(u_t) + (4B + 4\rho G)p_{2,t}\\
&\overset{(c)}{\leq} \mathbb{E}_t\left[\alpha_\phi(\pi_t)\right] + 2\frac{\mathbb{E}_t\left[\alpha_{\phi_t}(\pi_t)\right]}{\mathbb{P}_t\left(x_t \in \bar{S}_t\right)} + (4B + 4\rho G)p_{2,t}\\
&\overset{(d)}{=} \left(1 + \frac{2}{\mathbb{P}_t\left(\pi_t \in \bar{S}_t\right)}\right)\mathbb{E}_t\left[\alpha_\rho(\pi_t)\right] + (4B + 4\rho G)p_{2,t},
\end{aligned}$$

where (a) holds by definition of $p_{2,t}$, the fact that $\phi, \phi_t \leq \rho$ and the boundedness of functions; (b) follows from Eq. (17) and the fact that given $\mathcal{F}_{t-1}$, $\alpha_{\phi_t}(u_t)$ is deterministic; (c) holds by the following argument: $\mathbb{E}_t\left[\alpha_{\phi_t}(\pi_t)\right] \geq \mathbb{E}_t\left[\alpha_{\phi_t}(\pi_t)|\pi_t \in \bar{S}_t\right] \mathbb{P}_t\left(\pi_t \in \bar{S}_t\right) \geq \alpha_{\phi_t}(u_t)\mathbb{P}_t\left(\pi_t \in \bar{S}_t\right)$, which holds by the definition of u_t and the fact that $\alpha_{\phi_t}(u_t)$ and S_t are both $\mathcal{F}_{t-1}$-measurable; (d) holds by definition $\alpha_\rho(\pi_t) := c_{f,t}\sigma_{t-1}(\pi_t) + \rho c_{g,t}\widetilde{\sigma}_{t-1}(\pi_t)$ and the fact that both ϕ, ϕ_t are bounded by ρ. Hence, the key is to find a lower bound on the probability $\mathbb{P}_t\left(\pi_t \in \bar{S}_t\right)$. In particular, conditioned on E^{est}, we have

$$
\begin{aligned}
&\mathbb{P}_t\left(\pi_t \in \bar{S}_t\right) \\
&\overset{(a)}{\geq} \mathbb{P}_t\left(\hat{z}_{\phi_t}(\pi^*) \geq \max_{\pi_j \in S_t} \hat{z}_{\phi_t}(\pi_j), E_t^{conc}\right) \\
&\overset{(b)}{\geq} \mathbb{P}_t\left(\hat{z}_{\phi_t}(\pi^*) \geq z_{\phi_t}(\pi^*), E_t^{conc}\right) \\
&\geq \mathbb{P}_t\left(\hat{z}_{\phi_t}(\pi^*) \geq z_{\phi_t}(\pi^*)\right) - \mathbb{P}_t\left(\bar{E}_t^{conc}\right) \\
&\geq \mathbb{P}_t\left(\bar{f}_t(\pi^*) \geq f(\pi^*), \bar{g}_t(\pi^*) \leq g(\pi^*)\right) - \mathbb{P}_t\left(\bar{E}_t^{conc}\right) \\
&\overset{(c)}{=} \mathbb{P}_t\left(f_t(\pi^*) \geq f(\pi^*), g_t(\pi^*) \leq g(\pi^*)\right) - \mathbb{P}_t\left(\bar{E}_t^{conc}\right) \\
&\overset{(d)}{\geq} \mathbb{P}_t\left(f_t(\pi^*) \geq \mu_{t-1}(\pi^*) + c_{f,t}^{(1)}\sigma_{t-1}(\pi^*), g_t(\pi^*) \leq \widetilde{\mu}_{t-1}(\pi^*) - c_{g,t}^{(1)}\widetilde{\sigma}_{t-1}(\pi^*)\right) - \mathbb{P}_t\left(\bar{E}_t^{conc}\right) \\
&= \mathbb{P}_t\left(E_t^{anti}\right) - \mathbb{P}_t\left(\bar{E}_t^{conc}\right) \\
&= p_3 - p_{2,t},
\end{aligned}
$$

where (a) holds by the greedy selection in Algorithm 1 and $\pi^* \in \bar{S}_t$ since $\Delta_{\phi_t}(\pi^*) = 0$. Note that S_t is the complement of the ‘undersampled’ set $\bar{S}_t$; (b) holds given $E^{est} \cap E_t^{conc}$, for all $\pi_j \in S_t$ $\hat{z}_{\phi_t}(\pi_j) \leq z_{\phi_t}(\pi_j) + \alpha_{\phi_t}(\pi_j) \leq z_{\phi_t}(\pi_j) + \Delta_{\phi_t}(\pi_j) = z_{\phi_t}(\pi^*)$; (c) holds since $|g(x)| \leq G$ for all x and $|f(x)| \leq B$ for all x; (d) holds since under E^{est}, we have $f(x) \leq \mu_{t-1}(x^*) + c_{1,f}\sigma_{t-1}(x^*)$ and $g(x) \geq \widetilde{\mu}_{t-1}(x^*) - c_{1,g}\widetilde{\sigma}_{t-1}(x^*)$ for all x.

Putting everything together, we have now arrived at that conditioned on E^{est},

$$
\begin{aligned}
\mathbb{E}_t\left[d_t\right] &\leq \mathbb{E}_t\left[\alpha_\rho(x_t)\right]\left(1 + \frac{2}{p_3 - p_{2,t}}\right) + (4B + 4\rho G)p_{2,t} \\
&\leq \frac{1}{p_4}\mathbb{E}_t\left[\alpha_\rho(x_t)\right] + (4B + 4\rho G)p_{2,t}. \qquad (18)
\end{aligned}
$$

where the last inequality follows from the boundedness condition in the sufficient condition. In order to obtain a high probability bound, inspired by (Chowdhury and Gopalan 2017), we will resort to martingale techniques. Let us define the following terms

Definition 3. *Define $Y_0 = 0$, and for all $t = 1, \ldots, T$,*

$$
\begin{aligned}
\bar{d}_t &= d_t\mathcal{I}\{E^{est}\} \\
X_t &= \bar{d}_t - \frac{1}{p_4}\alpha_\rho(x_t) - (4B + 4\rho G)p_{2,t} \\
Y_t &= \sum_{s=1}^{t} X_s,
\end{aligned}
$$

where $\mathcal{I}\{\cdot\}$ is the indicator function.

Now, we can show that $\{Y_t\}_t$ is a super-martingale with respect to filtration $\mathcal{F}_t$. To this end, we need to show that for any t and any possible $\mathcal{F}_{t-1}$, $\mathbb{E}\left[Y_t - Y_{t-1}|\mathcal{F}_{t-1}\right] \leq 0$, i.e., $\mathbb{E}_t\left[\bar{d}_t\right] \leq \frac{1}{p_4}\mathbb{E}_t\left[\alpha_\rho(x_t)\right] + (4B + 4\rho G)p_{2,t}$. For $\mathcal{F}_{t-1}$ such that E^{est} holds, we already obtained the required inequality as in Eq. (18). For $\mathcal{F}_{t-1}$ such that E^{est} does not hold, the required inequality trivially holds since the LHS is zero. Now, we turn to show that $\{Y_t\}_t$ is a bounded incremental sequence, i.e., $|Y_t - Y_{t-1}| \leq M_t$

for some constant M_t. We first note that

$$
\begin{aligned}
|Y_t - Y_{t-1}| = |X_t| &\le |\bar{d}_t| + \frac{1}{p_t}\alpha_\rho(x_t) + (4B + 4\rho G)p_{2,t} \\
&= |\bar{d}_t| + \frac{1}{p_4}\left(c_{f,t}\sigma_{t-1}(x_t) + \rho c_{g,t}\widetilde{\sigma}_{t-1}(x_t)\right) + (4B + 4\rho G)p_{2,t} \\
&\overset{(a)}{\le} (4B + 4\rho G) + \frac{1}{p_4}(c_{f,t} + \rho c_{g,t}) + (4B + 4\rho G)p_{2,t} \\
&\le \frac{1}{p_4}(c_{f,t} + \rho c_{g,t})(4B + 4\rho G) := M_t,
\end{aligned}
$$

where (a) holds since $\bar{d}_t \le d_t \le (4B + 4\rho G)$, $\sigma_{t-1}(x_t) \le \sigma_0(x_t) \le 1$ and $\widetilde{\sigma}_{t-1}(x_t) \le \widetilde{\sigma}_0(x_t) \le 1$. Thus, we can apply Azuma-Hoeffding inequality to obtain that with probability at least $1 - \alpha$,

$$
\begin{aligned}
\sum_{t=1}^{T} \bar{r}_t &\le \sum_{t=1}^{T} \frac{1}{p_4}\alpha_\rho(x_t) + \sum_{t=1}^{T}(4B + 4\rho G)p_{2,t} + \sqrt{2\ln(1/\delta)\sum_{t=1}^{T} M_t^2} \\
&\overset{(a)}{\le} \frac{1}{p_4}\sum_{t=1}^{T}\alpha_\rho(x_t) + C'(4B + 4\rho G) + \frac{(c_f(T) + \rho c_g(T))(4B + 4\rho G)}{p_4}\sqrt{2T\ln(1/\delta)},
\end{aligned}
$$

where (a) we have used the boundedness condition. Note that since E^{est} holds with probability at least $1 - p_1$ for all t and x. By a union bound, we have with probability at least $1 - \alpha - p_1$,

$$
\begin{aligned}
\sum_{t=1}^{T} d_t &\le \frac{1}{p_4}\sum_{t=1}^{T}\alpha_\rho(x_t) + C'(4B + 4\rho G) + \frac{(c_f(T) + \rho c_g(T))(4B + 4\rho G)}{p_4}\sqrt{2T\ln(1/\delta)} \\
&= O\left(\frac{1}{p_4}\sum_{t=1}^{T}(c_f(T)\sigma_{t-1}(x_t) + \rho c_g(T)\widetilde{\sigma}_{t-1}(x_t)) + \frac{(c_f(T) + \rho c_g(T))\kappa}{p_4}\sqrt{2T\ln(1/\delta)}\right) \\
&= O\left(\frac{1}{p_4}c_f(T)\sqrt{T\gamma_T} + \frac{1}{p_4}\rho c_g(T)\sqrt{T\widetilde{\gamma}_T} + \frac{(c_f(T) + \rho c_g(T))\kappa}{p_4}\sqrt{2T\ln(1/\delta)}\right), \qquad (19)
\end{aligned}
$$

where $\kappa := 4B + 4\rho G$. Plugging (19) into (16), we obtain that

$$
\begin{aligned}
\mathcal{T}_1 + \mathcal{T}_2 &\le O\left(\frac{1}{p_4}c_f(T)\sqrt{T\gamma_T} + \frac{1}{p_4}\rho c_g(T)\sqrt{T\widetilde{\gamma}_T} + \frac{(c_f(T) + \rho c_g(T))\kappa}{p_4}\sqrt{2T\ln(1/\delta)}\right) \\
&:= \chi(T, \phi).
\end{aligned}
$$

Note that here $\chi(T, \phi)$ is independent of ϕ since we have bounded it by ρ in the analysis. Finally, plugging $\chi(T, \phi)$ into (11) and (14) yields the results of Theorem 2. □

D Proofs of Corollaries

As we have mentioned before, we only need to focus on the remaining two probability conditions and the boundedness condition in the sufficient condition.

D.1 Proof of Corollary 1

Proof. By Definition 1, $f_t(x) = \mu_{t-1}(x) + \beta_t\sigma_{t-1}(x)$ and $g_t(x) = \widetilde{\mu}_{t-1}(x) - \widetilde{\beta}_t\widetilde{\sigma}_{t-1}(x)$. From this, we can directly obtain that E_t^{conc} and E_t^{anti} hold with probability one. Moreover, the boundedness condition naturally holds. □

D.2 Proof of Corollary 2

Proof. By Definition 1, we have that given the history up to the end of round $t-1$, $f_t(x) \sim \mathcal{N}(\mu_{t-1}(x), \beta_t^2\sigma_{t-1}^2(x))$ and $g_t(x) \sim \mathcal{N}(\widetilde{\mu}_{t-1}(x), \widetilde{\beta}_t^2\widetilde{\sigma}_{t-1}^2(x))$. Thus, for any fixed $x \in \mathcal{X}$, by concentration of Gaussian distribution, we have $\mathbb{P}_t(|f_t(x) - \mu_{t-1}(x)| \le 2\beta_t\sqrt{\ln t}) \ge 1 - 1/t^2$, and hence, using the union bound over all x, we obtain $\forall x$, $\mathbb{P}_t\left(E_{t,f}^{conc}(x)\right) \ge 1 - 1/t^2$ with $c_{f,t}^{(2)} = 2\beta_t\sqrt{\ln(|\mathcal{X}|t)}$. Similarly, we have $\forall x$, $\mathbb{P}_t\left(E_{t,g}^{conc}(x)\right) \ge 1 - 1/t^2$ with $c_{g,t}^{(2)} = 2\widetilde{\beta}_t\sqrt{\ln(|\mathcal{X}|t)}$. Hence, by union bound, we have $\mathbb{P}_t\left(E_t^{conc}\right) \ge 1 - p_{2,t}$ with $p_{2,t} = 2/t^2$. Moreover, when π^* concentrates on a single point, by standard anti-concentration result of Gaussian distribution (e.g., Lemma 8 in (Chowdhury and Gopalan 2017)), we have $\mathbb{P}_t(E_{t,f}^{anti}) \ge p$ with $p := \frac{1}{4e\sqrt{\pi}}$. Similarly, we also have $\mathbb{P}_t(E_{t,g}^{anti}) \ge p$. By independent sampling of f_t and g_t, we have $\mathbb{P}_t\left(E_t^{anti}\right) \ge p_3$ with $p_3 = p^2$. The boundedness condition holds due to $\sum_{t=1}^{T} p_{2,t} \le 2\sum_{t=1}^{T} 1/t^2 \le \pi^2/3 := C'$ and $p_4 = O(p^2)$. □

D.3 Proof of Corollary 3

Proof. By Definition 2, $f_t(x) = \mu_{t-1}(x) + Z_t\sigma_{t-1}(x)$, where $Z_t \sim \mathcal{N}(0, \beta_t^2)$ and $g_t(x) = \widetilde{\mu}_{t-1}(x) + \widetilde{Z}_t\widetilde{\sigma}_{t-1}(x)$, where $\widetilde{Z}_t \sim \mathcal{N}(0, \widetilde{\beta}_t^2)$. By concentration of Gaussian, we have $\mathbb{P}_t(\forall x, |f_t(x) - \mu_{t-1}(x)| \leq 2\beta_t\sqrt{\ln t}) \geq 1 - 1/t^2$ (thanks to the "coupled" noise), and hence, we have $\mathbb{P}_t\left(E_{t,f}^{conc}\right) \geq 1 - 1/t^2$ with $c_{f,t}^{(2)} = 2\beta_t\sqrt{\ln t}$. Similarly, we have $\mathbb{P}_t\left(E_{t,g}^{conc}\right) \geq 1 - 1/t^2$ with $c_{g,t}^{(2)} = 2\widetilde{\beta}_t\sqrt{\ln t}$. Thus, by the union bound, we have $\mathbb{P}_t\left(E_t^{conc}\right) \geq 1 - p_{2,t}$ with $p_{2,t} = 2/t^2$. By the anti-concentration of Gaussian, we have $\mathbb{P}_t(E_{t,f}^{anti}) \geq \mathbb{P}_t(Z_t \geq 1) \geq p$, where $p := \frac{1}{4e\sqrt{\pi}}$. Similarly, we have $\mathbb{P}_t(E_{t,g}^{anti}) \geq \mathbb{P}_t(Z_t \leq -1) \geq p$. Since the noise Z_t and $\widetilde{Z}_t$ are independent, we have $\mathbb{P}_t\left(E_t^{anti}\right) \geq p_3$ with $p_3 = p^2$. Then, the boundedness condition holds due to $C' = \pi^2/3$ and $p_4 = O(p^2)$. □

D.4 Flexible Implementations of RandGP-UCB

In this section, we will give more insights on the choices of $\hat{\mathcal{D}}$, i.e., sampling distribution for $\hat{Z}_t$. In particular, we consider the unconstrained case for useful insights with black-box function being f. By the definition of RandGP-UCB, for each t, the estimate under RandGP-UCB is given by

$$f_t(x) = \mu_{t-1(x)} + Z_t\sigma_{t-1}(x),$$

where $Z_t \sim \mathcal{D}$. First, by Lemma 3, we have with high probability

$$f(x) \leq \mu_{t-1} + \beta_t\sigma_{t-1}(x),$$

which directly implies that in order to guarantee E_t^{anti} happens with a positive probability, one needs to make sure that $\mathbb{P}(Z_t \geq \beta_t) \geq p_3 > 0$. Thus, one simple choice of $\mathcal{D}$ is a uniform discrete distribution between $[0, 2\beta_t]$ with N points. Then, it can be easily checked that $\mathbb{P}_t\left(E_t^{anti}\right) \geq p_3 > 0$ and also $\mathbb{P}_t\left(E_t^{conc}\right) = 1$ with $c_{f,t}^{(2)} = 2\beta_t$. In addition to uniform discrete distribution, one can also use discrete Gaussian distribution within a range $[L, U]$ as long as U, L are properly chosen. Of course, there are many other choices as long as the insight shown above is satisfied, and hence RandGP-UCB provides a lot of flexibility in the algorithm design.

E Proof of Theorem 3

In fact, we will prove the following more general result compared to the one stated in the main result. That is, we can vary the choices of ϵ_t and V_t to achieve different tradeoffs between regret and constraint violation guarantees (recall that $\delta \leq 1$).

Theorem 4. *Let $\beta_t = B + R\sqrt{2(\gamma_{t-1} + 1 + \ln(2/\delta))}$ in SCGP-UCB. Under the Slater's condition in Assumption 1, and the randomness and boundedness assumptions in Assumptions 5 and 6, SCGP-UCB achieves the following performance guarantees based on the choices of ϵ_t and V_t:*

(i) If $\epsilon_t = \frac{1}{\sqrt{t}}$ and $V_t = \frac{\delta}{8B}\sqrt{t}$, it obtains that for all $\tau \in [T]$,

$$\mathbb{E}\left[\mathcal{R}_+(\tau)\right] = O\left(\frac{BG^2\sqrt{\tau}}{\delta} + \frac{BG}{\delta^3} + B\sqrt{\tau\gamma_\tau} + \gamma_\tau\sqrt{\tau}\right), \mathbb{E}\left[\mathcal{V}(\tau)\right] = \begin{cases} O(G^2/\delta^2) & \textit{if } \tau \leq O(1/\delta^4), \\ 0 & \textit{otherwise.} \end{cases}$$

(ii) If $\epsilon_t = \frac{\delta}{2\sqrt{t}}$ and $V_t = \frac{\delta^2\sqrt{t}}{16B}$, it obtains that for all $\tau \in [T]$,

$$\mathbb{E}\left[\mathcal{R}_+(\tau)\right] = O\left(\frac{BG^2\sqrt{\tau}}{\delta^2} + B\sqrt{\tau\gamma_\tau} + \gamma_\tau\sqrt{\tau}\right), \mathbb{E}\left[\mathcal{V}(\tau)\right] = \begin{cases} O(\frac{G^2}{\delta}\ln(G/\delta)) & \textit{if } \tau \leq O(1/\delta^4), \\ 0 & \textit{otherwise.} \end{cases}$$

Before we present the proof for Theorem 4, we first introduce the following ϵ-tight problem.

$$\max_{\pi} \mathbb{E}_\pi\left[f(x)\right] = \int_{x\in\mathcal{X}} f(x)\pi(x)dx \tag{20}$$

$$s.t. \quad \mathbb{E}_\pi\left[g(x)\right] \leq -\epsilon \tag{21}$$

The motivation is that we will decompose the regret in terms of the solution of this problem. In particular, consider a sequence ϵ_t, let π_{ϵ_t} be a *feasible* solution to the ϵ-tight problem with $\epsilon = \min(\epsilon_t, \delta)$ and an optimal solution is denoted by $\pi_{\epsilon_t}^*$.

We then present a key result on the Laypunov drift of $\Delta(t) := L(Q(t+1)) - L(Q(t)) = \frac{1}{2}(Q(t+1))^2 - \frac{1}{2}(Q(t))^2$, which will be useful for both regret and constraint violation analysis.

Lemma 4. *Let δ_{x_t} be the dirac delta measure at point ${x_t}^2$. For any π, we have*

$$\mathbb{E}\left[\Delta(t) \mid Q(t)\right] \leq -V_t \mathbb{E}\left[\int_{x\in\mathcal{X}} \bar{f}_t(x)(\pi(x) - \delta_{x_t}(x))\, dx \mid Q(t)\right] + \frac{1}{2}(G+\epsilon_t)^2$$
$$+ Q(t)\mathbb{E}\left[\left(\int_{x\in\mathcal{X}} g_t(x)\pi(x)\, dx + \epsilon_t\right) \mid Q(t)\right].$$

Proof. Note that by the update rule of the virtual queue and non-expansiveness of projection, we have

$$\Delta(t) \leq Q(t)(g_t(x_t)+\epsilon_t) + \frac{1}{2}\left(g_t(x_t)+\epsilon_t\right)^2.$$

Now we will bound the RHS as follows.

$$\begin{aligned}
&Q(t)(g_t(x_t)+\epsilon_t) + \frac{1}{2}\left(g_t(x_t)+\epsilon_t\right)^2\\
\overset{(a)}{\leq}& Q(t)(g_t(x_t)+\epsilon_t) + \frac{1}{2}(G+\epsilon_t)^2\\
=& -V_t\bar{f}_t(x_t) + Q(t)g_t(x_t) + Q(t)\epsilon_t + V_t\bar{f}_t(x_t) + \frac{1}{2}(G+\epsilon_t)^2\\
\overset{(b)}{\leq}& -V_t\int_{x\in\mathcal{X}} \bar{f}_t(x)\pi(x)\, dx + Q(t)\int_{x\in\mathcal{X}} g_t(x)\pi(x)\, dx + Q(t)\epsilon_t + V_t\bar{f}_t(x_t) + \frac{1}{2}(G+\epsilon_t)^2,
\end{aligned}$$

where (a) holds by the boundedness of g_t; (b) holds by the greedy selection in Algorithm 2. Reorganizing the term and taking the conditional expectation, yields the required result. □

Now, we are well-prepared to present the proof of Theorem 3.

Proof of Theorem 3. We divide the proofs into two parts: regret and constraint violation.

Regret. We first have the following regret decomposition. For any $\tau \in [T]$, we have

$$\begin{aligned}
\mathcal{R}_+(\tau) &= \tau\mathbb{E}_{\pi^*}\left[f(x)\right] - \sum_{t=1}^{\tau} f(x_t)\\
&= \sum_{t=1}^{\tau}\int_{x\in\mathcal{X}} f(x)\pi^*(x)\, dx - \sum_{t=1}^{\tau}\int_{x\in\mathcal{X}} f(x)\delta_{x_t}(x)\, dx\\
&= \sum_{t=1}^{\tau}\int_{x\in\mathcal{X}} f(x)\left(\pi^*(x) - \pi^*_{\epsilon_t}(x)\right)\, dx + \sum_{t=1}^{\tau}\int_{x\in\mathcal{X}} \left(f(x) - \bar{f}_t(x)\right)\pi^*_{\epsilon_t}(x)\, dx\\
&+ \sum_{t=1}^{\tau}\int_{x\in\mathcal{X}} \bar{f}_t(x)\left(\pi^*_{\epsilon_t(x)} - \delta_{x_t}(x)\right)\, dx + \sum_{t=1}^{\tau}\int_{x\in\mathcal{X}} (\bar{f}_t(x) - f(x))\delta_{x_t}(x)\, dx.
\end{aligned} \tag{22}$$

The third term is the main difficulty. To bound it, we will utilize Lemma 4. In particular, we let $\pi = \pi^*_{\epsilon_t}$ in Lemma 4, we obtain that

$$\begin{aligned}
\mathbb{E}\left[\Delta(t) \mid Q(t)\right] &\leq -V_t\mathbb{E}\left[\int_{x\in\mathcal{X}} \bar{f}_t(x)(\pi^*_{\epsilon_t}(x) - \delta_{x_t}(x))\, dx \mid Q(t)\right] + \frac{1}{2}(G+\epsilon_t)^2\\
&+ Q(t)\mathbb{E}\left[\left(\int_{x\in\mathcal{X}} g_t(x)\pi^*_{\epsilon_t}(x)\, dx + \epsilon_t\right) \mid Q(t)\right]\\
&= -V_t\mathbb{E}\left[\int_{x\in\mathcal{X}} \bar{f}_t(x)(\pi^*_{\epsilon_t}(x) - \delta_{x_t}(x))\, dx \mid Q(t)\right] + \frac{1}{2}(G+\epsilon_t)^2\\
&+ Q(t)\left(\int_{x\in\mathcal{X}} g(x)\pi^*_{\epsilon_t}(x)\, dx + \epsilon_t\right),
\end{aligned} \tag{23}$$

where the last inequality holds by the fact that $g_t(\cdot)$ is independent of $Q(t)$ (remove the condition on $Q(t)$), Fubini's theorem (interchange integrals) and the fact that $\mathbb{E}\left[g_t(x)\right] = g(x)$ by Assumption 5. The independence holds because $Q(t)$ depends on

[2] A bit notation abuse with the slater's condition δ.

all the randomness before t, while $g_t(\cdot)$ depends on randomness at t (i.e., random sample at t), which is independent of all other randomness by Assumption 5. The last term in (23) can be divided into two parts:

$$\mathbb{1}(\epsilon_t \le \delta)Q(t)\bigg(\int_{x\in\mathcal{X}} g(x)\pi^*_{\epsilon_t}(x)\,dx + \epsilon_t\bigg) + \mathbb{1}(\epsilon_t > \delta)Q(t)\bigg(\int_{x\in\mathcal{X}} g(x)\pi^*_{\epsilon_t}(x)\,dx + \epsilon_t\bigg).$$

By the definition of $\pi^*_{\epsilon_t}$, we have the first part is non-positive, and the second part is bounded by $\mathbb{1}(\epsilon_t > \delta)Q(t)(\epsilon_t - \delta)$. Plugging the bounds on the two parts back into (23) and taking expectation over both sides, yields

$$\mathbb{E}\left[\Delta(t)\right] \le -V_t\mathbb{E}\left[\int_{x\in\mathcal{X}} \bar{f}_t(x)(\pi^*_{\epsilon_t}(x) - \delta_{x_t}(x))\,dx \mid Q(t)\right] + \mathbb{1}(\epsilon_t > \delta)\mathbb{E}\left[Q(t)\right](\epsilon_t - \delta) + \frac{1}{2}(G+\epsilon_t)^2,$$

which directly implies that

$$V_t\int_{x\in\mathcal{X}} \bar{f}_t(x)(\pi^*_{\epsilon_t}(x) - \delta_{x_t}(x))\,dx \le -\mathbb{E}\left[\Delta(t)\right] + \mathbb{1}(\epsilon_t > \delta)(G+\epsilon_t)t\epsilon_t + \frac{1}{2}(G+\epsilon_t)^2,$$

where we have used the fact that $Q(t) \le (G+\epsilon_t)t$. Hence, taking the telescope summation and dividing by V_t, yields

$$\sum_{t=1}^{\tau}\mathbb{E}\left[\int_{x\in\mathcal{X}} \bar{f}_t(x)(\pi^*_{\epsilon_t}(x) - \delta_{x_t}(x))\,dx\right] \le \frac{1}{V_t}\sum_{t=1}^{\tau} t(G+\epsilon_t)\epsilon_t\mathbb{1}(\epsilon_t > \delta) + \sum_{t=1}^{\tau}\frac{1}{2V_t}(G+\epsilon_t)^2.$$

Now, we can plug this bound back to the regret decomposition in (22) and obtain that

$$\begin{aligned}\mathbb{E}\left[\mathcal{R}(\tau)\right] \le{}& \underbrace{\mathbb{E}\left[\sum_{t=1}^{\tau}\int_{x\in\mathcal{X}} f(x)\left(\pi^*(x) - \pi^*_{\epsilon_t}(x)\right)\,dx\right]}_{\mathcal{T}_1} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{\tau}\int_{x\in\mathcal{X}} \left(f(x) - \bar{f}_t(x)\right)\pi^*_{\epsilon_t}(x)\,dx\right]}_{\mathcal{T}_2}\\ &+ \underbrace{\mathbb{E}\left[\sum_{t=1}^{\tau}\int_{x\in\mathcal{X}} (\bar{f}_t(x) - f(x))\delta_{x_t}(x)\,dx\right]}_{\mathcal{T}_3} + \frac{1}{V_t}\sum_{t=1}^{\tau} t(G+\epsilon_t)\epsilon_t\mathbb{1}(\epsilon_t > \delta) + \sum_{t=1}^{\tau}\frac{1}{2V_t}(G+\epsilon_t)^2.\end{aligned}$$

Then, we turn to bound each of the three terms. To start with, we focus on $\mathcal{T}_2$ and $\mathcal{T}_3$, which can be bounded by using GP-UCB exploration and standard concentration results in GP bandits. In particular, by the concentration result in Lemma 3, we have with high probability for all x, $f(x) \le \mu_{t-1}(x) + \beta_t\sigma_{t-1}(x) = f_t(x)$ where the second equality holds by the definition of GP-UCB exploration. Moreover, it is easy to check that truncation does not impact this term and as a result we have

$$\mathcal{T}_2 \le 0.$$

Moreover, by Lemma 3 again, we have $f_t(x) - f(x) \le 2\beta_t\sigma_{t-1}(x)$. In addition, truncation also does not impact this term. Thus, we have for some constant c

$$\mathcal{T}_3 \le 2\beta_t\sigma_{t-1}(x_t) = c\beta_\tau\sqrt{\tau\gamma_\tau},$$

where the last equality holds by the same argument as in (15).

To bound $\mathcal{T}_1$, we first note that

$$\int_{x\in\mathcal{X}} f(x)\left(\pi^*(x) - \pi^*_{\epsilon_t}(x)\right)\,dx \le \int_{x\in\mathcal{X}} f(x)\left(\pi^*(x) - \pi_{\epsilon_t}(x)\right)\,dx.$$

We can indeed construct a π_{ϵ_t} as $\pi_{\epsilon_t}(x) = (1 - \frac{\epsilon_t}{\delta})\pi^*(x) + \frac{\epsilon_t}{\delta}\pi_\delta(x)$, where π_δ is a feasible solution to the ϵ-tight problem (20)-(21) with $\epsilon = \delta$. It exists by the Slater's condition. Thus, we can show the constructed π_{ϵ_t} is indeed a feasible solution, i.e., when $\epsilon_t \le \delta$:

$$\mathbb{E}_{\pi_{\epsilon_t}}\left[g(x)\right] \le \frac{\epsilon_t}{\delta}(-\delta) = -\epsilon_t \le -\epsilon,$$

where $\epsilon = \min\{\epsilon_t, \delta\}$. Based on this, we can further bound $\mathcal{T}_1$ by using the constructed π_{ϵ_t}:

$$\begin{aligned}\mathcal{T}_1 &\le \sum_{t=1}^{\tau}\mathbb{1}(\epsilon_t \le \delta)\int_{x\in\mathcal{X}} f(x)\left(\pi^*(x) - \pi_{\epsilon_t}(x)\right)\,dx + \sum_{t=1}^{\tau}\mathbb{1}(\epsilon_t > \delta)2B\\ &= \sum_{t=1}^{\tau}\int_{x\in\mathcal{X}} f(x)\left(\pi^*(x) - (1 - \frac{\epsilon_t}{\delta})\pi^*(x) - \frac{\epsilon_t}{\delta}\pi_\delta(x)\right)\,dx + \sum_{t=1}^{\tau}\mathbb{1}(\epsilon_t > \delta)2B\\ &\le \sum_{t=1}^{\tau} 2B\frac{\epsilon_t}{\delta} + \sum_{t=1}^{\tau}\mathbb{1}(\epsilon_t > \delta)2B,\end{aligned}$$

where we have used the fact that $|f(x)| \leq B$. Finally, putting everything together, we have

$$\mathbb{E}\left[\mathcal{R}_+(\tau)\right] \leq \sum_{t=1}^{\tau} 2B\frac{\epsilon_t}{\delta} + \sum_{t=1}^{\tau} \mathbb{1}(\epsilon_t > \delta)2B + c\beta_\tau\sqrt{\tau\gamma_\tau} + \frac{1}{V_t}\sum_{t=1}^{\tau} t(G+\epsilon_t)\epsilon_t\mathbb{1}(\epsilon_t > \delta) + \sum_{t=1}^{\tau}\frac{1}{2V_t}(G+\epsilon_t)^2.$$

We consider two choices of ϵ_t and V_t. Note that $\beta_t = B + R\sqrt{2(\gamma_{t-1}+1+\ln(2/\delta))}$.

Case 1: Let $\epsilon_t = \frac{1}{\sqrt{t}}$ and $V_t = \frac{\delta}{8B}\sqrt{t}$. Then, we have

$$\mathbb{E}\left[\mathcal{R}_+(\tau)\right] = O\left(\frac{BG^2\sqrt{\tau}}{\delta} + \frac{BG}{\delta^3} + B\sqrt{\tau\gamma_\tau} + \gamma_\tau\sqrt{\tau}\right).$$

Case 2: Let $\epsilon_t = \frac{\delta}{2\sqrt{t}}$ and $V_t = \frac{\delta^2\sqrt{t}}{16B}$. Then, we have

$$\mathbb{E}\left[\mathcal{R}_+(\tau)\right] = O\left(\frac{BG^2\sqrt{\tau}}{\delta^2} + B\sqrt{\tau\gamma_\tau} + \gamma_\tau\sqrt{\tau}\right).$$

Constraint violation. We now turn to bound the constraint violation. First, by the update rule of virtual queue (along with $Q(1)=0$), we have

$$\sum_{t=1}^{\tau} g_t(x_t) \leq Q(\tau+1) - \sum_{t=1}^{\tau}\epsilon_t.$$

This directly implies that

$$\sum_{t=1}^{\tau} g(x_t) \leq Q(\tau+1) - \sum_{t=1}^{\tau}\epsilon_t + \sum_{t=1}^{\tau}(g(x_t) - g_t(x_t)).$$

Taking expectation, we have

$$\mathbb{E}\left[\mathcal{V}(\tau)\right] \leq \mathbb{E}\left[Q(\tau+1)\right] - \sum_{t=1}^{\tau}\epsilon_t.$$

since $\mathbb{E}\left[g_t(x_t)\right] = g(x_t)$ by Assumption 5. Thus, in order to bound the constraint violation, we now only need to bound $Q(\tau+1)$. To this end, we will adopt the following variant of Hajek's lemma (Hajek 1982), which is stated in (Liu et al. 2021).

Lemma 5. *Let $S(t)$ be a random process, $\Phi(t)$ be its Lyapunov function with $\Phi(0) = \Phi_0$ and $\Delta(t) = \Phi(t+1) - \Phi(t)$ be the Lyapunov drift. Given an increasing sequence $\{\phi_t\}$, ρ and $\nu_{\max}$ with $0 < \rho \leq \nu_{\max}$, if the expected drift $\mathbb{E}\left[\Delta(t) \mid S(t) = s\right]$ satisfies the following two conditions:*

1. *There exists constants $\rho > 0$ and $\phi_t > 0$ such that $\mathbb{E}\left[\Delta(t) \mid S(t) = s\right] \leq -\rho$, for all s when $\Phi(t) \geq \phi_t$ and, and*
2. *$|\Phi(t+1) - \Phi(t)| \leq \nu_{\max}$ holds with probability one.*

Then, we have

$$\mathbb{E}\left[e^{\xi\Phi(t)}\right] \leq e^{\xi\Phi_0} + \frac{2e^{\xi(\nu_{\max}+\phi_t)}}{\rho},$$

where $\xi = \frac{\rho}{\nu_{\max}^2 + \nu_{\max}\rho/3}$.

Now, we consider the Lyapunov function $\Phi(t) = Q(t)$ and $\epsilon_t \leq \frac{\delta}{2}$. Then, the second condition in Lemma 5 is directly satisfied with $\nu_{\max} = G + 1$ since $g_t(x_t) \leq G$ by our assumption and $\epsilon_t \leq \frac{\delta}{2} \leq 1$. Thus, we only need to verify the first condition. In particular, by the concavity of the function $\sqrt{x}$, we have

$$\begin{aligned}\Phi(t+1) - \Phi(t) &= Q(t+1) - Q(t)\\ &= \sqrt{(Q(t+1))^2} - \sqrt{(Q(t))^2}\\ &\leq \frac{1}{2Q(t)}\left((Q(t+1))^2 - (Q(t))^2\right)\\ &= \frac{1}{Q(t)}\Delta(t).\end{aligned}$$

Now, by letting π in Lemma 4 be π_0 (i.e., the Slater's point), we have

$$\begin{aligned}
&\mathbb{E}\left[\Delta(t) \mid Q(t)=Q\right] \\
\leq &- V_t\mathbb{E}\left[\int_{x\in\mathcal{X}} \bar{f}_t(x)\left(\pi_0(x)-\delta_{x_t}(x)\right)\,dx \mid Q\right] + Q\mathbb{E}\left[\int_{x\in\mathcal{X}} g_t(x)\pi_0(x)\,dx \mid Q\right] + Q\epsilon_t \\
&+ \frac{1}{2}(G+\epsilon_t)^2. \qquad (24)
\end{aligned}$$

Note that

$$\mathbb{E}\left[\int_{x\in\mathcal{X}} g_t(x)\pi_0(x)\,dx \mid Q\right] \overset{(a)}{=} \mathbb{E}\left[\int_{x\in\mathcal{X}} g_t(x)\pi_0(x)\,dx\right] \overset{(b)}{\leq} -\delta, \qquad (25)$$

where (a) holds by the independence between g_t and $Q(t)$; (b) holds by interchange of integral and the definition of Slater's condition under π_0. Since we only consider the case when $\epsilon_t \leq \frac{\delta}{2}$, combining (24) and (25), we have

$$\begin{aligned}
&\mathbb{E}\left[\Delta(t) \mid Q(t)=Q\right] \\
\leq &- \frac{\delta}{2}Q - V_t\mathbb{E}\left[\int_{x\in\mathcal{X}} \bar{f}_t(x)\left(\pi_0(x)-\delta_{x_t}(x)\right)\,dx \mid Q\right] + \frac{1}{2}(G+\epsilon_t)^2 \\
\leq &- \frac{\delta}{2}Q + 2BV_t + \frac{1}{2}(G+\epsilon_t)^2,
\end{aligned}$$

where we have used the fact that $|\bar{f}_t(x)| \leq B$. Let $K_t := \frac{1}{2}(G+\epsilon_t)^2$, we have that the two conditions in Lemma 5 are satisfied with $\phi_t = \frac{4(2BV_t+K_t)}{\delta}$, $\rho = \frac{\delta}{4}$, and $\nu_{\max} = G+1$ when $\epsilon_t \leq \frac{\delta}{2}$.

Then, by Lemma 5 and Jensen's inequality, we can obtain that when $\epsilon_t \leq \frac{\delta}{2}$:

$$\mathbb{E}\left[Q(t)\right] \leq \frac{12(G+1)^2}{\delta}\ln\left(\frac{8(G+1)}{\delta}\right) + (G+1) + \frac{4(2BV_t+K_t)}{\delta}.$$

Now, let us again consider two different choices of ϵ_t and V_t.

Case 1: Let $\epsilon_t = \frac{1}{\sqrt{t}}$ and $V_t = \frac{\delta}{8B}\sqrt{t}$. Then, for all $\tau \geq t' = \frac{4}{\delta^2}$, $\epsilon_\tau \leq \frac{\delta}{2}$. Note that $Q(t') \leq \sum_{t=1}^{t'}(G+\epsilon_t)$, and hence for all t,

$$\mathbb{E}\left[Q(t)\right] \leq \frac{12(G+1)^2}{\delta}\ln\left(\frac{8(G+1)}{\delta}\right) + (G+1) + \frac{4(2BV_t+K_t)}{\delta} + \sum_{t=1}^{t'}(G+\epsilon_t).$$

Thus, we have

$$\begin{aligned}
\mathbb{E}\left[\mathcal{V}(\tau)\right] &\leq \mathbb{E}\left[Q(\tau+1)\right] - \sum_{t=1}^{\tau}\epsilon_t \\
&\leq \frac{12(G+1)^2}{\delta}\ln\left(\frac{8(G+1)}{\delta}\right) + (G+1) + \frac{4(2BV_{\tau+1}+K_{\tau+1})}{\delta} + \sum_{t=1}^{t'}(G+\epsilon_t) - \sum_{t=1}^{\tau}\epsilon_t \\
&\leq \left(\frac{12(G+1)^2}{\delta}\ln\left(\frac{8(G+1)}{\delta}\right) + \frac{2G^2}{\delta^2} + \frac{2G^2+5G+7}{\delta} + 4\sqrt{2} - \sqrt{\tau}\right)^{+}.
\end{aligned}$$

Hence, we have

$$\mathbb{E}\left[\mathcal{V}(\tau)\right] = \begin{cases} O(G^2/\delta^2) & \text{if } \tau \leq O(1/\delta^4), \\ 0 & \text{otherwise.} \end{cases}$$

Case 2: Let $\epsilon_t = \frac{\delta}{2\sqrt{t}}$ and $V_t = \frac{\delta^2\sqrt{t}}{16B}$. Since $\epsilon_t \leq \frac{\delta}{2}$ for all t, we directly have

$$\mathbb{E}\left[Q(t)\right] \leq \frac{12(G+1)^2}{\delta}\ln\left(\frac{8(G+1)}{\delta}\right) + (G+1) + \frac{4(2BV_t+K_t)}{\delta}.$$

Thus, we have

$$\begin{aligned}\mathbb{E}\left[\mathcal{V}(\tau)\right] &\leq \mathbb{E}\left[Q(\tau+1)\right] - \sum_{t=1}^{\tau}\epsilon_t\\ &\leq \frac{12(G+1)^2}{\delta}\ln\left(\frac{8(G+1)}{\delta}\right) + (G+1) + \frac{4(2BV_{\tau+1}+K_{\tau+1})}{\delta} - \sum_{t=1}^{\tau}\epsilon_t\\ &\left(\frac{12(G+1)^2}{\delta}\ln\left(\frac{8(G+1)}{\delta}\right) + \frac{2G^2+G+1}{\delta} + 9\sqrt{2} - \frac{\delta}{2}\sqrt{\tau}\right)^+.\end{aligned}$$

Hence, we have

$$\mathbb{E}\left[\mathcal{V}(\tau)\right] = \begin{cases} O(\frac{G^2}{\delta}\ln(G/\delta)) & \text{if } \tau \leq O(1/\delta^4),\\ 0 \quad \text{otherwise.} \end{cases}$$

□

E.1 Discussion on the flaw in the current analysis of (Liu et al. 2021)

At the end of this section, we discuss the importance of independence assumption in the proof of Theorem 3 and pinpoint one key flaw in the current analysis of (Liu et al. 2021). In particular, (Liu et al. 2021) claims that by the similar analysis as in the stochastic full information, one can also achieve zero expected constraint violation in the bandit constraint information case (in the linear bandits with UCB exploration only in their extension part). However, this claim is not true based on its current analysis which ignores the dependence of virtue queue and the function estimates. In the following, we will provide more details on the current flaw in (Liu et al. 2021).

Recall that the key step in deriving the constraint violation bound is to bound the expected queue length, i.e., $\mathbb{E}\left[Q(t)\right]$. To this end, we utilize the important Hajek lemma, i.e., Lemma 5 by verifying the two conditions. To establish the negative-drift condition for all $Q(t)$ (i.e., condition 1 in Lemma 5), we combine (24) and (25). Note that how independence assumption plays a key role in deriving (25). With independence, we can translate the conditional expectation into unconditional expectation.

Then, one may wonder if we can use the same technique above to derive the same zero constraint violation for the bandit feedback case (i.e., g_t is also an estimate rather than i.i.d sample). This was attempted in (Liu et al. 2021) for the linear bandit case as an extension of the main contents at the end of their paper. However, the current analysis in (Liu et al. 2021) has a key flaw, which is directly related to the derivation in (25). To be more specific, the goal of (25) is to show

$$\mathbb{E}\left[\int_{x\in\mathcal{X}} g_t(x)\pi_0(x)\,dx \mid Q\right] \leq -c, \tag{26}$$

for some positive c. For the illustration purpose, we assume that π_0 is a dirac delta measure at a single point x_0. To achieve (26) in the bandit feedback case (i.e., g_t is also an estimate based on UCB exploration), (Liu et al. 2021) divide it into two parts

$$\mathbb{E}\left[g_t(x_0)|Q(t)\right] = \underbrace{\mathbb{E}\left[g_t(x_0)-g(x_0)|Q(t)\right]}_{\mathcal{T}_1} + \underbrace{\mathbb{E}\left[g(x_0)|Q(t)\right]}_{\mathcal{T}_2}.$$

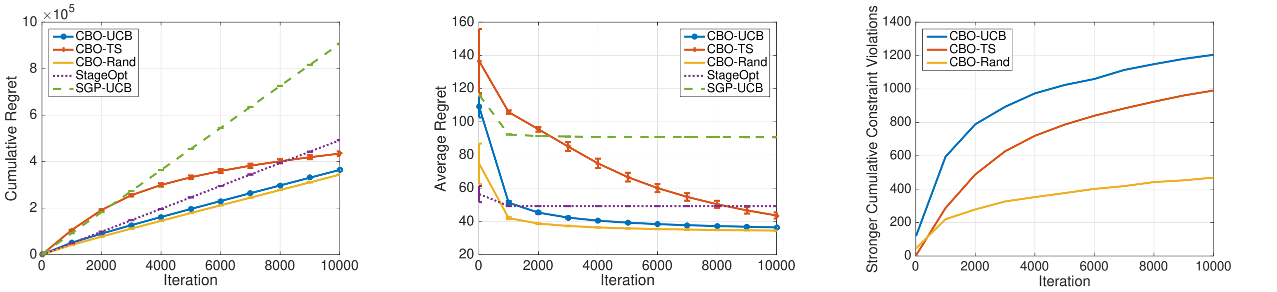
Now, for $\mathcal{T}_2$, one can easily bound it by $\mathcal{T}_2 \leq -\delta$ via Slater's condition since $g(\cdot)$ is independent of $Q(t)$. To bound $\mathcal{T}_1$, (Liu et al. 2021) resort to standard self-normalized inequality for linear bandit and the definition of UCB exploration. By these standard results, one can show that for any fixed $\delta \in (0,1]$

$$\mathbb{P}\{\forall x, \forall t, g_t(x) \leq g(x)\} \geq 1-\delta. \tag{27}$$

That is, g_t is optimistic in terms of cost. Then, by setting $\delta = 1/\tau$, (Liu et al. 2021) establishe that $\mathcal{T}_1 \leq O(1/\tau)$ (see Eq. 30 and the immediate follow-up steps in (Liu et al. 2021)). Unfortunately, the bound on $\mathcal{T}_1$ is ungrounded since it is obtained by treating the conditional expectation in $\mathcal{T}_1$ as an unconditional expectation. However, in the bandit feedback case, we cannot remove the condition on $Q(t)$ in $\mathcal{T}_1$, since g_t is not independent of $Q(t)$ as both of them depend on the randomness before time t. Given a particular $Q(t)$, it roughly means that we are taking expectation conditioned on a particular history (i.e., a sample-path). Under this particular history, (27) does not necessarily hold and moreover the concentration of g_t given $Q(t)$ is hard to compute in this case. As a result, the conditional expectation for $\mathcal{T}_1$ is hard to compute. Note that we really need to bound $\mathcal{T}_1$ for any given $Q(t)$, since this is the requirement for condition 1 in Lemma 5.

F Additional details on experiments

In this section, we will first give more details on the experiments in Section 5 and then provide additional experimental results on another real-world dataset to test the robustness of various algorithms in the presence of heavy-tailed noise.



(a) Cumulative regret on finance data (b) Average regret on finance data (c) Constraint violations on finance data

Figure 2: Additional experimental results on BO with real-world finance data.

F.1 Details on experiments in Section 5

Synthetic data. The domain $\mathcal{X}$ is generated by discretizing $[0,1]$ uniformly into 100 points. The objective function $f = \sum_{i=1}^{p} a_i k(\cdot, x_i)$ is generated by uniformly sampling $a_i \in [-1,1]$ and support points $x_i \in \mathcal{X}$ with $p = 100$. The constraint function is given by $g(\cdot) = -f(\cdot) + h$ for some threshold $h > 0$. The kernel is k_{se} with parameter $l = 0.2$. Other parameters include B, R and γ_t are set similar as in the unconstrained case (e.g., (Chowdhury and Gopalan 2017)). In addition, for the threshold value h, we consider two cases: $h = B/4$ and $h = B/2$, respectively. We perform 50 trials (each with $T = 10{,}000$) and plot the mean of the cumulative regret along with the error bars.

Results. We give more details on regret and constraint violations.

Regret: In terms of the maximization of f (i.e., Figure 1(a) for $h = B/2$), we can observe that `CBO-UCB` has similar performance (i.e., cumulative reward regret) to existing ones while both CBO-TS and CBO-Rand have a significant gain compared to existing algorithms, which both are based on UCB. Similar pattern exists in the case when $h = B/4$.

Constraint violation: The cumulative constraint violations (i.e., $\mathcal{V}(T)$) of our three algorithms are zero. In fact, we investigate a stronger requirement by looking at the total number of rounds where the constraint is violated (denoted by N). In terms of this metric, our algorithms achieve a good performance as well. For $h = B/4$, the mean total number of rounds where the constraint is violated (after $T = 10{,}000$) under `CBO-UCB`, `CBO-TS` and `CBO-Rand` are $N = 1.1$, $N = 0.7$ and $N = 1.1$, respectively, compared to $N = 0.22$ for `StageOpt` and $N = 0$ for `SGP-UCB`. For $h = B/2$, $N = 3.25$, $N = 2.9$ and $N = 5$ under `CBO-UCB`, `CBO-TS` and `CBO-Rand` respectively compared to $N = 0.3$ for `StageOpt` and $N = 0$ for `SGP-UCB`.

Summary: Thus, our proposed CBO algorithms are able to trade a slight performance in terms of constraint violation by utilizing the nature of soft constraints to achieve a significant gain in terms of reward regret.

Real-world data. We use the light sensor data collected in the CMU Intelligent Workplace in Nov 2005, which is available online as Matlab structure (http://www.cs.cmu.edu/~guestrin/Class/10708-F08/projects/) and contains locations of 41 sensors, 601 train samples and 192 test samples. We use it in the context of finding the maximum average reading of the sensors. In particular, f is set as empirical average of the test samples, with B set as its maximum, and k is set as the empirical covariance of the normalized train samples. The constraint is given by $g(\cdot) = -f(\cdot) + h$ with $h = B/2$. We perform 50 trials (each with $T = 1{,}000$) and plot the mean of the cumulative regret along with the error bars.

Results. We give results on regret and constraint violations.

Regret: In terms of reward regret (see Figure 1(b)), we can see that `CBO-Rand` and `StageOpt` has the best performance, while `CBO-TS` is worse than `CBO-UCB` in this case.

Constraint violation: The cumulative constraint violations (i.e., $\mathcal{V}(T)$) of our three algorithms are zero. In fact, CBO-UCB does not make any decisions that violate the constraint while CBO-Rand has an average of 38 rounds where the constraint is violated. We also plot stronger cumulative constraint violations given by $\sum_{t=1}^{T} [g(x_t)]_+$ as shown in Figure 1(c), from which we can see that all CBO algorithms achieve sublinear performance even with respect to this stronger metric.

F.2 Additional experimental results

In this section, we compare various constrained BO algorithms in a new real-world dataset, which demonstrates a heavy-tailed noise. Note that sub-Gaussian noise is required in all the existing theoretical works (including our work). We use this dataset to test the robustness of various constrained BO algorithms. The results tend to show that our three CBO algorithms are more robust in terms of heavy-tailed noise, which is common in practical applications.

Real-world data. This dataset is the adjusted closing price of 29 stocks from January 4th, 2016 to April 10th 2019 (which is included in our code file in the supplementary material). We use it in the context of identifying the most profitable stock in a given pool of stocks. As verified in (Chowdhury and Gopalan 2019), the rewards follows from heavy-tailed distribution. We take the empirical mean of stock prices as our objective function f and empirical covariance of the normalized stock prices as our kernel function k. The noise is estimated by taking the difference between the raw prices and its empirical mean (i.e., f), with R set as the maximum. The constraint is given by $g(\cdot) = -f(\cdot) + h$ with $h = 100$ (i.e., $h \approx B/2$). We perform 50 trials (each with $T = 10,000$) and plot the mean along with the error bars.

Results. We give results on regret and constraint violations.

Regret: We plot both cumulative regret and time-average regret in this setting (see Figures 2 (a) and (b)). We can observe that in the presence of heavy-tailed noise, our three CBO algorithms have significant performance gain over existing safe BO algorithms.
Constraint violation: We focus on the strong metric, i.e., the number of rounds where the constraint is violated, denoted by N. We have that `CBO-Rand` enjoys an average $N = 21$ and `CBO-UCB` has an average $N = 47$ within the horizon of $T = 10,000$. We also plot stronger cumulative constraint violations given by $\sum_{t=1}^{T}[g(x_t)]_+$ as shown in Figure 2(c), from which we can see that all CBO algorithms achieve sublinear performance even with respect to this stronger metric.