# On Kernelized Multi-Armed Bandits with Constraints

**Xingyu Zhou**
Electrical and Computer Engineering
Wayne State University
Detroit, MI, USA
xingyu.zhou@wayne.edu

**Bo Ji**
Computer Science
Virginia Tech
Blacksburg, VA, USA
boji@vt.edu

## Abstract

We study a stochastic bandit problem with a general unknown reward function and a general unknown constraint function. Both functions can be non-linear (even non-convex) and are assumed to lie in a reproducing kernel Hilbert space (RKHS) with a bounded norm. In contrast to safety-type hard constraints studied in prior works, we consider soft constraints that may be violated in any round as long as the cumulative violations are small. Our ultimate goal is to study how to utilize the nature of soft constraints to attain a finer complexity-regret-constraint trade-off in the kernelized bandit setting. To this end, 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 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 or even zero constraint violation. We demonstrate the superior performance of our proposed algorithms via numerical experiments based on both synthetic and real-world datasets. Along the way, we also make the first detailed comparison between two popular methods for analyzing constrained bandits and Markov decision processes (MDPs) by discussing the key difference and some subtleties in the analysis, which could be of independent interest to the communities.

## 1 Introduction

Stochastic bandit optimization of an unknown function $f$ has recently gained increasing popularity due to its widespread real-life applications such as recommendations [1], cloud resource configurations [2], and wireless power control [3]. At each time $t$, an action $x_t$ is selected and then a (noisy) bandit reward feedback $y_t$ is observed. The goal is to maximize the cumulative reward, or equivalently minimize the total regret due to not choosing the optimal action in hindsight. To capture a generic function, researchers have turned to nonparametric models of $f$ via Gaussian process or reproducing kernel Hilbert space (RKHS), which are able to uniformly approximate an arbitrary continuous function over a compact set [4]. In this paper, as in [5, 6], we consider the agnostic setting (i.e., frequentist-type), where $f$ is assumed to be a fixed function in an RKHS with a bounded norm (i.e., a measure of smoothness). We call this setting *frequentist-type kernelized bandits* (KB).

In addition to a generic non-linear or even non-convex function $f$, another common feature in practical applications is that there often exist additional constraints in the decision-making process such as hard-constraint like safety or soft-constraint like cost. To this end, there have been exciting recent advances in the theoretical analysis of constrained kernelized bandits. In particular, [7–9] propose algorithms with convergence guarantees, while [10], to the best our knowledge, is the first

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

work that establishes regret bounds for their developed algorithm, although under the Bayesian-type[1] setting. These algorithms mainly focus on KB with a *hard* constraint such as safety, i.e., the selected action in *each round* needs to satisfy the constraint with a high probability. Hence, compared to the unconstrained case, additional computation is required to construct a *safe* action set in each round, which not only incurs additional complexity burdens, but often leads to conservative performance.

**Motivations.** In practice, there are also many 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. For example, in a wireless networking system, the reward could be the throughput while the constraint is that the average energy consumption is below a threshold. In this case, the constraint can be violated at any time, and moreover, using less energy at one time instant allows one to use more in the next time instant, i.e., "compensation" across time (see a formal definition in Eq.(1)) Furthermore, existing provably efficient algorithms [7–10] largely build on upper confidence bound (UCB) exploration, which often has inferior empirical performance compared to Thompson sampling (TS) exploration. Hence, another key question is whether one can design provably efficient CKB algorithms with general explorations. In summary, the following fundamental theoretical question remains open:

*Can a finer complexity-regret-constraint trade-off be attained in CKB under general explorations?*

**Contributions.** In this paper, we take a systematic approach to affirmatively answer the above fundamental question. In particular, we tackle the complexity-regret-constraint trade-off by formulating KB under soft constraints as a stochastic bandit problem where the objective is to maximize the cumulative reward while minimizing the cumulative constraint violations and maintaining the same computation complexity as in the unconstrained case. Our detailed contributions are as follows.

- We develop a unified framework for CKB 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) under the same complexity as the unconstrained case. We also show that by introducing slackness in the dual update, one can trade regret to achieve bounded or even zero constraint violation. This framework builds upon a novel sufficient condition, which not only facilitates the design of new CKB algorithms but provides a unified view in the performance analysis.
- We demonstrate the superior performance of our proposed algorithms via numerical experiments based on both synthetic and real-world data. In addition, we discuss the benefits of our algorithms in terms of various practical considerations such as low complexity, scalability, robustness, and flexibility.
- We also provide the first detailed comparison between two popular methods for analyzing constrained bandits in general. Specifically, the first one is based on convex optimization tool as in [11, 12], which is also the inspiration for our paper. The other one is based on Lyapunov-drift argument as in [13–15]. We discuss the key difference in terms of regret and constraint violation analysis in these two methods and highlight the subtlety in applying a standard queueing technique (i.e., Hajek lemma [16]) to bound the constraint violation in the second method. We believe this provides a clear picture on the methodology, which is of independent interest to the communities.

### 1.1 Related Work

In the special cases of KB, including multi-armed bandits (MAB) and linear bandits (e.g., KB with a linear kernel), there is a large body of work on bandits with different types of constraints, including knapsack bandits [17–19], conservative bandits [20–22], bandits with fairness constraints [23, 24], bandits with hard safety constraints [25–27], and bandits with cumulative soft constraints [28, 13]. Among them, the bandit setting with cumulative soft constraints is the closest to ours. In particular, [13] considers linear bandits under UCB exploration and a zero constraint violation is attained via the Lyapunov-drift method. However, it is unclear how to generalize it to handle general exploration strategies such as TS; see a further discussion in Section 4.

[1] In the Bayesian-type KB, $f$ is assumed to be a sample from a Gaussian process and the observation noise is Gaussian. In contrast, in our considered frequentist-type KB, $f$ is a fixed function in an RKHS and the noise can be any sub-Gaussian. A better regret bound can often be achieved in the easier Bayesian-type KB setting [6].

Broadly speaking, our work is also related to reinforcement learning (RL) with soft constraints, i.e., constrained MDPs. In particular, our analysis is inspired by those on constrained MDPs [11, 12] (which is another popular method to handle constrained bandits and MDPs via convex optimization tools), but has significant differences. First, in those works, the constraint violation is $\widetilde{O}(\sqrt{T})$. In contrast, ours can attain bounded and even zero constraint violations by introducing the slackness in the dual update. Second, they only consider UCB-type exploration, but our algorithms can be equipped with various exploration strategies (including UCB), thanks to our general sufficient condition. Third, they focus on either tabular or linear function approximation settings. In contrast, both objective and constraint functions we consider can be *nonlinear*. There are also recent works on constrained MDPs that claim to achieve bounded or zero constraint violation [14, 15] based on the Lyapunov-drift method. However, as in the bandit case, it is unclear how to generalize it to handle general explorations beyond UCB. Finally, the theoretical understanding of unconstrained KB (a.k.a. Gaussian process bandit) is well-established, including regret upper bounds under UCB and TS [5, 29, 6] and lower bounds [30, 31]. We remark that our work is mainly a theory-guided study. In a more practical area of KB, i.e., Bayesian optimization (BO), there have been many BO algorithms developed for the constrained setting; see [32–34] and the references therein. Although these algorithms enjoy good performance in various practical settings, their theoretical performance guarantees are still unclear.

## 2 Problem Formulation and Preliminaries

We consider a stochastic bandit optimization problem with *soft* constraints. In particular, 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 $\eta_t$ is a zero-mean noise. The learning agent also observes a bandit constraint feedback $c_t = g(x_t) + \xi_t$, where $\xi_t$ is a zero-mean noise. By soft constraints, the goal here is to maximize the cumulative reward while minimizing the cumulative total violation.

**Learning Problem.** Define cumulative regret and constraint violation as follows:

$$\mathcal{R}(T) := Tf(x^*) - \sum_{t=1}^{T} f(x_t), \quad \mathcal{V}(T) := \left[\sum_{t=1}^{T} g(x_t)\right]_+, \tag{1}$$

where $x^* := \operatorname{argmax}_{\{x \in \mathcal{X}: g(x) \leq 0\}} f(x)$ and $[\cdot]_+ := \max\{\cdot, 0\}$. The goal of the learning agent is to achieve both sublinear regret and sublinear constraint violation. In fact, we will establish bounds on the following stronger version of regret. Specifically, let $\pi$ be a probability distribution over the set of actions $\mathcal{X}$ (i.e., stochastic policy), 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 baseline optimization problem: $\max_\pi \{\mathbb{E}_\pi[f(x)] : \mathbb{E}_\pi[g(x)] \leq 0\}$ where both $f$ and $g$ are known, and $\pi^*$ is its optimal solution. Now, our stronger regret is defined as

$$\mathcal{R}_+(T) := T\mathbb{E}_{\pi^*}[f(x)] - \sum_{t=1}^{T} f(x_t). \tag{2}$$

Clearly, we have $\mathcal{R}(T) \leq \mathcal{R}_+(T)$[2]. Throughout the paper, we assume the following commonly used condition in constrained optimization literature; see also [13, 35, 11].

**Assumption 1** (Slater's condition)**.** *There is a constant $\delta > 0$ such that there exists a probability distribution $\pi_0$ 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 stochastic policy under which the expected cost is less than a strictly negative value. This is in sharp constraint to existing KB algorithms for hard constraints that typically require the existence of an initial safe action [9, 10].

In this paper, we consider the frequentist-type regularity assumption that is typically used in unconstrained KB works (e.g., [5, 6]). Specifically, we assume that $f$ is a fixed function in an 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 $\mathcal{H}_k$. Similarly, for the unknown constraint function $g$, we assume that $g$ is a fixed function in the RKHS defined by a

[2]This stronger regret also allows us to use convex optimization tool since $\mathbb{E}_\pi[f(x)] := \int_{x \in \mathcal{X}} f(x)\pi(x)\, dx$ is a convex function with respect to $\pi$, which will be important for our analysis.

kernel function $\widetilde{k}$, and the RKHS for $g$ is denoted by $\mathcal{H}_{\widetilde{k}}$. 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 $\eta_t$ is* i.i.d. *$R$-sub-Gaussian. Similarly, 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 $\xi_t$ is* i.i.d. *$\widetilde{R}$-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 $\eta_t$, which are drawn from $\mathcal{N}(0,\lambda)$ and are independent across $t$. Note that this GP surrogate model is used for algorithm design only; it does not change the fact that $f$ is a fixed function in $\mathcal{H}_k$ and that the noise $\eta_t$ can be sub-Gaussian (i.e., an *agnostic* setting [6]). Let $[t] := \{1, 2, \ldots, t\}$. Conditioned on a set of observations $H_t = \{(x_s, r_s), s \in [t]\}$, by the properties of GP [36], the posterior distribution for $f$ is $\mathcal{GP}(\mu_t(\cdot), k_t(\cdot,\cdot))$, where

$$\mu_t(x) := k_t(x)^T (K_t + \lambda I)^{-1} R_t, \tag{3}$$

$$k_t(x,x') := k(x,x') - k_t(x)^T (K_t + \lambda I)^{-1} k_t(x'), \tag{4}$$

in which $k_t(x) := [k(x_1,x), \ldots, k(x_t,x)]^T$, $K_t := [k(x_u,x_v)]_{u,v\in[t]}$, and $R_t$ is the (noisy) reward vector $[r_1, r_2, \ldots, r_t]^T$. In particular, we also define $\sigma_t^2(x) := k_t(x,x)$. Let $K_A := [k(x,x')]_{x,x'\in A}$ for $A \subseteq \mathcal{X}$. We define the maximum information gain as $\gamma_t(k,\mathcal{X}) := \max_{A\subseteq\mathcal{X}:|A|=t} \frac{1}{2}\ln|I_t + \lambda^{-1}K_A|$ 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 $\gamma_t$ 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 $k_{\text{SE}}$, $\gamma_t = O(d \ln t)$ for linear kernel [6] and $\gamma_t = \tilde{O}(T^{\frac{d}{2\nu+d}})$ (where $\nu$ is a hyperparameter) for Matérn kernel $k_{\text{Matérn}}$ [37]. Similarly, 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 $\widetilde{H}_t = \{(x_s, c_s), s \in [t]\}$, 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.

## 3 A Unified Framework for Constrained Kernelized Bandits

In this section, leveraging primal-dual optimization, we develop a unified framework for both algorithm design and performance analysis in constrained kernelized bandits. In particular, we first propose a "master" algorithm called CKB (constrained KB), 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 CKB algorithms) with rigorous performance guarantees.

**Algorithm.** We first explain our "master" algorithm CKB 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 $\phi^* := \operatorname{argmin}_{\phi\geq 0}\mathcal{D}(\phi)$. 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 3). Then, both estimates will be truncated according to the range of $f$ and $g$, respectively (lines 4-5) (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 6-7 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 7 is that one of the optimal solutions for $\max_\pi \bar{\mathcal{L}}(\pi,\phi_t)$ is simply $\operatorname{argmax}_x(\bar{f}_t(x) - \phi_t\bar{g}_t(x))$. Then, line 8 is the dual update that minimizes $\mathcal{D}(\phi_t)$ with respect to $\phi$ by taking a projected gradient step with $1/V$ being the step size. The parameter $\rho$ is chosen to be larger than the optimal dual variable $\phi^*$, and hence the projected interval $[0,\rho]$ includes the optimal dual variable. This can be achieved 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 [38, Theorem 8.42]. Finally, line 8 is the posterior update via standard GP regression for both $f$ and $g$ as computed in (3) and (4) with $\sigma_t^2(x) = k_t(x,x)$ and $\widetilde{\sigma}_t^2(x) = \widetilde{k}_t(x,x)$.

**Remark 1** (Computational complexity)**.** *CKB enjoys the same computational complexity as the standard unconstrained case (e.g., [5]) since the additional dual update is a simple projection and*

**Algorithm 1** CKB Algorithm

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

*the primal optimization keeps the same flavor as the unconstrained case, i.e., without constructing a specific safe set as in existing constrained KB algorithms designed for hard constraints.*

We call CKB 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 first use UCB-type exploration as an example to gain useful insights, which in turn will facilitate the development of a novel sufficient condition. This condition 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 in constrained kernelized bandits.

We first introduce standard UCB and TS explorations under GP as in [5].

**Definition 1** (GP-UCB and GP-TS Explorations)**.** *Suppose the posterior distribution for a black-box function $h$ in round $t$ is given by $\mathcal{GP}(\widehat{\mu}_{t-1}(\cdot), \widehat{k}_{t-1}(\cdot,\cdot))$ and $\widehat{\beta}_t$ is an increasing sequence.*

*(i) The estimate of $h$ in round $t$ under GP-UCB exploration is $h_t(\cdot) = \widehat{\mu}_{t-1}(\cdot) + \widehat{\beta}_t \widehat{\sigma}_{t-1}(\cdot)$, where $\widehat{\sigma}^2_{t-1}(x) := \widehat{k}_{t-1}(x,x)$ for all $x \in \mathcal{X}$.*

*(ii) The estimate of $h$ in round $t$ under GP-TS exploration is $h_t(\cdot) \sim \mathcal{GP}(\widehat{\mu}_{t-1}(\cdot), \widehat{\beta}^2_t \widehat{k}_{t-1}(\cdot,\cdot))$.*

### 3.1 Warm Up: CKB with GP-UCB Exploration

In this section, we instantiate CKB with GP-UCB exploration called `CKB-UCB`, as a warm-up. In particular, in `CKB-UCB`, $\mathcal{A}_f$ is a GP-UCB exploration (see Definition 1) with a positive $\widehat{\beta}_t$ sequence (i.e., optimistic with respect to reward), and $\mathcal{A}_g$ is a GP-UCB exploration with a negative $\widehat{\beta}_t$ sequence (i.e., optimistic with respect to cost). This instantiation enjoys the following performance guarantee.

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

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

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

*where $C := \max\{B, G\}$ and $\widehat{\gamma}_T := \max\{\gamma_T, \widetilde{\gamma}_T\}$.*

**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 $\gamma_T$ and $\widetilde{\gamma}_T$ to see that both regret and constraint violation are sublinear for commonly used kernels. For example, for an SE kernel, both $\gamma_T$ and $\widetilde{\gamma}_T$ are on the order of $(\ln T)^{d+1}$. Finally, the standard "doubling trick" can be used to design an anytime algorithm (i.e., without the knowledge of $T$) with regret and constraint violation bounds of the same order.*

**Proof Sketch of Theorem 1.** The key step is to establish a sublinear bound on the term $\mathcal{R}_+(T) + \phi \sum_{t=1}^T g(x_t)$ for any $\phi \in [0, \rho]$. Then, with convex optimization tools, one can establish regret and constraint violation bound. To this end, By the dual update of CKB, we can first show that $\mathcal{R}_+(T) + \phi \sum_{t=1}^T g(x_t) \leq \mathcal{T}_1 + \mathcal{T}_2 + \frac{V}{2}\phi^2 + \frac{1}{2V}TG^2$, where

$$\mathcal{T}_1 := \sum_{t=1}^T (\mathbb{E}_{\pi^*}[f(x)] - \phi_t \mathbb{E}_{\pi^*}[g(x)]) - \sum_{t=1}^T (\bar{f}_t(x_t) - \phi_t \bar{g}_t(x_t)), \tag{5}$$

$$\mathcal{T}_2 := \sum_{t=1}^T (\bar{f}_t(x_t) - f(x_t)) + \phi \sum_{t=1}^T (g(x_t) - \bar{g}_t(x_t)). \tag{6}$$

Thus, the analysis for each choice of exploration strategy only differ in how to bound $\mathcal{T}_1 + \mathcal{T}_2$. Under GP-UCB exploration, the term $\mathcal{T}_1 \leq 0$ by the optimism while $\mathcal{T}_2$ can be upper bounded by $\mathcal{T}_2 \leq 2\beta_T \sum_{t=1}^T \sigma_{t-1}(x_t) + 2\phi\widetilde{\beta}_T \sum_{t=1}^T \widetilde{\sigma}_{t-1}(x_t) = O(\beta_T\sqrt{\gamma_T T} + \phi\widetilde{\beta}_T\sqrt{\widetilde{\gamma}_T T})$, i.e., standard predication error bound. Building on the bound on $\mathcal{T}_1 + \mathcal{T}_2$, one can then show the required results. □

### 3.2 A Sufficient Condition for Provably Efficient Explorations

The above analysis reveals that the key step in obtaining sublinear performance guarantees of Algorithm 1 is to find a sublinear bound on $\mathcal{T}_1 + \mathcal{T}_2$, which depends on the choice of exploration strategies. To go beyond GP-UCB exploration, we will establish a sufficient condition on general exploration strategies (i.e., $\mathcal{A}_f$ and $\mathcal{A}_g$), which guarantees a sublinear bound on $\mathcal{T}_1 + \mathcal{T}_2$ and hence sublinear regret and sublinear constraint violation.

We first present the intuition behind the key components of our sufficient condition. Inspired by [39], it suffices to focus on the following three nice events so as to bound $\mathcal{T}_1 + \mathcal{T}_2$ in (5)-(6):

$E^{est} := \{E_f^{est}(x,t) \cap E_g^{est}(x,t); \forall (x,t)\}$, $E_t^{conc} := \{E_{f,t}^{conc}(x) \cap E_{g,t}^{conc}(x); \forall x\}$, $E_t^{anti} := E_{f,t}^{anti} \cap E_{g,t}^{anti}$, in which

$E_f^{est}(x,t) := |f(x) - \mu_{t-1}(x)| \leq c_{f,t}^{(1)}\sigma_{t-1}(x)$, $E_g^{est}(x,t) := |g(x) - \widetilde{\mu}_{t-1}(x)| \leq c_{f,t}^{(1)}\widetilde{\sigma}_{t-1}(x)$,

$E_{f,t}^{conc}(x) := |f_t(x) - \mu_{t-1}(x)| \leq c_{f,t}^{(2)}\sigma_{t-1}(x)$, $E_{g,t}^{conc}(x) := |g_t(x) - \widetilde{\mu}_{t-1}(x)| \leq c_{g,t}^{(2)}\widetilde{\sigma}_{t-1}(x)$,

$E_{f,t}^{anti} := \mathbb{E}_{\pi^*}[f_t(x) - \mu_{t-1}(x)] \geq c_{f,t}^{(1)}\mathbb{E}_{\pi^*}[\sigma_{t-1}(x)]$, $E_{g,t}^{anti} := \mathbb{E}_{\pi^*}[g_t(x) - \widetilde{\mu}_{t-1}(x)] \leq -c_{g,t}^{(1)}\mathbb{E}_{\pi^*}[\widetilde{\sigma}_{t-1}(x)]$.

To see the intuition, first 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 (6). Now, suppose that events $E^{est}$ and $E_t^{anti}$ hold with some positive probability. Then, one can see that the estimates are optimistic with respect to the true functions when evaluated at the optimal points. This probabilistic optimism is the key to bounding $\mathcal{T}_1$ in (5). Note that GP-UCB exploration is optimistic with probability one by definition (see Definition 1), and hence, $\mathcal{T}_1 \leq 0$ always holds.

To formally state our result, let us define the filtration $\mathcal{F}_t$ as all the history up to the end of round $t$. Let $\mathbb{E}_t[\cdot] := \mathbb{E}[\cdot|\mathcal{F}_{t-1}]$ and $\mathbb{P}_t(\cdot) := \mathbb{P}[\cdot|\mathcal{F}_{t-1}]$. Then, we have the following sufficient condition.

**Assumption 3** (Sufficient Condition)**.** *The sufficient condition includes two parts:*

*(1) (Probability condition)* $\mathbb{P}(E^{est}) \geq 1 - p_1$, $\mathbb{P}_t(E_t^{conc}) \geq 1 - p_{2,t}$, *and* $\mathbb{P}_t(E_t^{anti}) \geq p_3 > 0$ *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* $1 + \frac{2}{(p_3 - p_{2,t})} \leq 1/p_4$ *for all* $t$*; (ii) There exist some functions of* $T$*:* $c_f(T), c_g(T)$ *such that* $c_{f,t}^{(1)} + c_{f,t}^{(2)} \leq c_f(T)$ *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 constant* $C'$.

**Remark 3.** *The above sufficient condition generalizes existing similar results [39–41] in several aspects. First, existing works mainly focus on the MAB or linear bandit settings, which are special cases of our KB 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* $E_t^{anti}$ *holds under policy* $\pi^*$ *(i.e., the expectation over* $\pi^*$*) rather than under a single optimal action* $x^*$.

With the above sufficient condition, we have the following general performance bounds for CKB.

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

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

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

In the following, we will show that Theorem 2 provides a unified view of the regret and constraint violation performance for various CKB algorithms, thanks to our sufficient condition. In particular, we first show that existing exploration strategies, such as GP-UCB and GP-TS, satisfy our sufficient condition, given by the following corollary.

**Corollary 1.** *Let $\beta_t = B + R\sqrt{2(\gamma_{t-1} + 1 + \ln(1/\alpha_f))}$ and $\widetilde{\beta}_t = B + R\sqrt{2(\widetilde{\gamma}_{t-1} + 1 + \ln(1/\alpha_g))}$.*
*(i). GP-UCB with $\widehat{\beta}_t = \beta_t$ and $\widehat{\beta}_t = -\widetilde{\beta}_t$ for $\mathcal{A}_f$ and $\mathcal{A}_g$, respectively, satisfies the sufficient condition.*

*(ii). GP-TS with $\widehat{\beta}_t = \beta_t$ and $\widehat{\beta}_t = -\widetilde{\beta}_t$ for $\mathcal{A}_f$ and $\mathcal{A}_g$, respectively, satisfies the sufficient condition when $\pi^*$ is a deterministic policy.*

The sufficient condition also enables us to design CKB algorithms with new exploration strategies. In the following, inspired by [40], 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 $\mathcal{GP}(\widehat{\mu}_{t-1}(\cdot), \widehat{k}_{t-1}(\cdot,\cdot))$. Then, the estimate of $h$ in round $t$ under RandGP-UCB exploration strategy is $h_t(\cdot) = \widehat{\mu}_{t-1}(\cdot) + \widehat{Z}_t\widehat{\sigma}_{t-1}(\cdot)$, where $\widehat{Z}_t \sim \widehat{\mathcal{D}}$ for some distribution $\widehat{\mathcal{D}}$ and $\widehat{\sigma}^2_{t-1}(x) = \widehat{k}_{t-1}(x, x)$ for all $x \in \mathcal{X}$.*

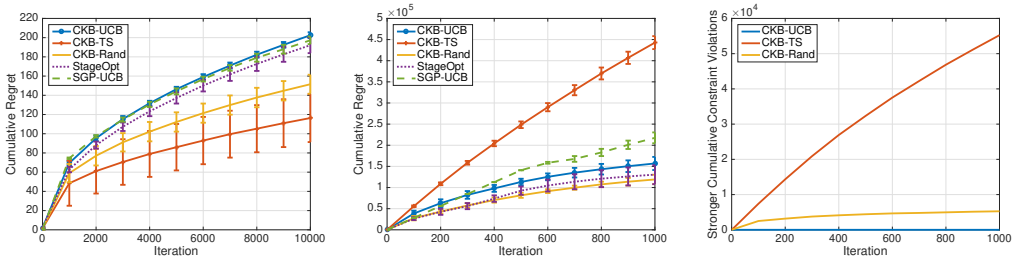
In contrast to GP-UCB, RandGP-UCB replaces the deterministic confidence bound by a randomized one. Compared to GP-TS, RandGP-UCB uses "coupled" noise in the sense that all the actions share the same noise $\widehat{Z}_t$ rather than "decoupled" and correlated noise in GP-TS. This subtle difference will 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 stochastic $\pi^*$.

**Corollary 2.** *Let $\beta_t, \widetilde{\beta}_t$ be the same as in Corollary 1. RandGP-UCB with $\widehat{\mathcal{D}} = \mathcal{N}(0, \beta_t^2)$ and $\widehat{\mathcal{D}} = \mathcal{N}(0, \widetilde{\beta}_t^2)$ for $\mathcal{A}_f$ and $\mathcal{A}_g$, respectively, satisfies the sufficient condition.*

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

## 4 Discussions

In this section, we discuss several possible questions one may have at this moment about our CKB algorithm. The first natural question to ask is whether one can further improve the constraint violation bound. It turns out that with a minor modification of CKB algorithm, one can achieve a bounded and even zero constraint violation by trading the regret slightly (but still the same order as before). The modification is to introduce a slackness given by $\varepsilon$ in the dual update in Algorithm 1, i.e., $\phi_{t+1} = \text{Proj}_{[0,\rho]}\left[\phi_t + \frac{1}{V}(\bar{g}_t(x_t) + \varepsilon)\right]$ with $\varepsilon \leq \delta/2$. Intuitively speaking, this can be viewed as if one is working on a new pessimistic constraint function. After obtaining the constraint violation under this new hypothetic constraint, one can subtract $\varepsilon T$ to find the true constraint violation under the true function $g$. The catch here is that one needs also change the baseline problem to the following one: $\max_\pi\{\mathbb{E}_\pi[f(x)] : \mathbb{E}_\pi[g(x)] + \varepsilon \leq 0\}$ so that it matches the new pessimistic constraint function. Let $\pi^*_\varepsilon$ be the optimal solution to this new problem and the obtained regret is only with respect to $\pi^*_\varepsilon$ rather than the original $\pi^*$. Thus, we need to further bound the following difference $T\mathbb{E}_{\pi^*}[f(x)] - T\mathbb{E}_{\pi^*_\varepsilon}[f(x)]$ to obtain the true regret bound, which is given in Appendix E.




(a) Regret on synthetic data (b) Regret on real-world data (c) Constraint on real-world data

Figure 1: Experimental results on constrained kernelized bandits with light-tailed data .

Another question one may ask is why we are able to go beyond UCB-type exploration compared to existing works (e.g., [13–15]) that only establish performance guarantees under UCB. The answer is that they resort to another popular method for constrained optimization based on the so-called Lyapunov-drift argument. However, this method relies heavily on the optimism under UCB to derive both regret and constraint violation bounds. Thus, it is difficult to generalize it for other exploration strategies even for simpler linear bandit problems. In contrast, our analysis relies only on the bound for $\mathcal{T}_1 + \mathcal{T}_2$, which allows us to consider more general exploration strategies. In Appendix F, we also discuss how to apply the Lyapunov-drift based method to analyze UCB exploration in our setting, through which we highlight more differences and some subtleties in Lyapunov-drift method.

## 5 Simulation Results

Although our work is mainly a theoretical study, we conduct simulations to compare the performance of our algorithms (i.e., `CKB-UCB`, `CKB-TS`, and `CKB-Rand`, that is, CKB with GP-UCB, GP-TS, and RandGP-UCB explorations, respectively) with existing safe KB algorithms based on both synthetic and real-world datasets. In particular, we consider the two most recent safe KB algorithms: `StageOpt` [28] (which has a superior performance compared to `SafeOpt` [27]) and `SGP-UCB` [3]. Our goal is to show that our proposed CKB algorithms can trade a slight performance in constraint violation (i.e., soft constraint) for improvement in the reward regret, computation complexity and flexible implementations, i.e., a better complexity-regret-constraint trade-off.

### 5.1 Synthetic Data and Light-Tailed Real-World Data

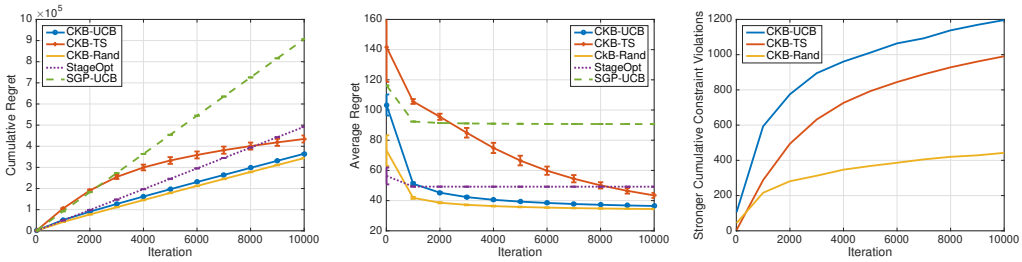
**Synthetic Data.** The domain $\mathcal{X}$ is generated by discretizing $[0, 1]$ uniformly into 100 points. The objective function $f(\cdot) = \sum_{i=1}^{p} a_i k(\cdot, x_i)$ is generated by uniformly sampling $a_i \in [-1, 1]$ and support points $x_i \in \mathcal{X}$ with $p = 100$. With the same manner, we generate the constraint function $g$. The kernel is $k_{se}$ with parameter $l = 0.2$. Other parameters include $B$, $R$ and $\gamma_t$ are set similar as in the unconstrained case (e.g., [5]).

**Light-Tailed 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[3] 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 $T = 10,000$) and plot the mean of the cumulative regret along with the error bars, as shown in Fig. 1.

**Regret.** Our three CKB 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 CKB algorithms, `CKB-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 CKB algorithms have $N \leq 5$ over $T = 10,000$ rounds; in the real-world data setting, `CKB-UCB` enjoys $N = 0$ and `CKB-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),

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




(a) Regret on finance data (b) Avg. regret on finance data (c) Constraint on finance data

Figure 2: Experimental results on constrained kernelized bandits under heavy-tailed finance data.

from which we can see that all CKB algorithms achieve sublinear performance even with respect to this stronger metric, i.e., violation cancellation across rounds is not allowed.

**Practical considerations.** Our proposed CKB 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[4]. 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 KB can be naturally applied to our CKB algorithms [42]. 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, CKB algorithms for soft constraints only require a mild Slater's condition as in Assumption 1, which does not require the existence of a safe action.

### 5.2 Heavy-Tailed Real-World Data

We further compare different constrained KB 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 KB algorithms. The experimental results tend to show that our three CKB algorithms are more robust in terms of heavy-tailed noise, which is common in practical applications. The detail of this real-world dataset is deferred to Appendix D.

**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 CKB algorithms have significant performance gain over existing safe KB 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 `CKB-Rand` enjoys an average $N = 21$ and `CKB-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 CKB algorithms achieve sublinear performance even with respect to this stronger metric.

## 6 Conclusion

We presented a general framework for constrained KB 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) CKB algorithms with both UCB and TS explorations, but presents a unified method to design new effective ones. By introducing slackness, our algorithm can also attain a bounded or even zero constraint violation while still achieving a sublinear regret. We further perform simulations on both synthetic data and real-world data that corroborate our theoretical results. Along the way, we also present the first detailed discussion on two existing methods for analyzing constrained bandits and MDPs by highlighting interesting insights. For future work, one interesting direction is to build upon recent advances in unconstrained KB that attain an improved regret (e.g., [43–46]) to study the corresponding constrained case in the hope to maintain the same improvement.

[4]For a continuous domain, as in the unconstrained case, one can resort to heuristic solvers (e.g., a combination of random sampling (cheap) and the "L-BFGS-B'" optimization method.). In fact, to attain the same order of regret bound, the solution to the acquisition maximization problem need not be exact. Instead, it only needs to maximize the acquisition function within $C/\sqrt{t}$ accuracy for some constant $C$ at each step.

## Acknowledgement

We would like to thank the anonymous reviewers for their helpful comments. XZ is partially supported by NSF grant CNS-2153220. BJ is partially supported by NSF grant CNS-2112694.

## References

[1] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670, 2010.

[2] Brijen Thananjeyan, Kirthevasan Kandasamy, Ion Stoica, Michael I Jordan, Ken Goldberg, and Joseph E Gonzalez. Pac best arm identification under a deadline. *arXiv preprint arXiv:2106.03221*, 2021.

[3] Mung Chiang, Prashanth Hande, and Tian Lan. *Power control in wireless cellular networks*. Now Publishers Inc, 2008.

[4] Charles A Micchelli, Yuesheng Xu, and Haizhang Zhang. Universal kernels. *Journal of Machine Learning Research*, 7(12), 2006.

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

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

[7] Yanan Sui, Alkis Gotovos, Joel Burdick, and Andreas Krause. Safe exploration for optimization with gaussian processes. In *International Conference on Machine Learning*, pages 997–1005. PMLR, 2015.

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

[9] Yanan Sui, Joel Burdick, Yisong Yue, et al. Stagewise safe bayesian optimization with gaussian processes. In *International Conference on Machine Learning*, pages 4781–4789. PMLR, 2018.

[10] Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Regret bound for safe gaussian process bandit optimization. In *Learning for Dynamics and Control*, pages 158–159. PMLR, 2020.

[11] Yonathan Efroni, Shie Mannor, and Matteo Pirotta. Exploration-exploitation in constrained mdps. *arXiv preprint arXiv:2003.02189*, 2020.

[12] Dongsheng Ding, Xiaohan Wei, Zhuoran Yang, Zhaoran Wang, and Mihailo Jovanovic. Provably efficient safe exploration via primal-dual policy optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 3304–3312. PMLR, 2021.

[13] Xin Liu, Bin Li, Pengyi Shi, and Lei Ying. An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. *arXiv preprint arXiv:2102.05295*, 2021.

[14] Honghao Wei, Xin Liu, and Lei Ying. A provably-efficient model-free algorithm for constrained markov decision processes. *arXiv preprint arXiv:2106.01577*, 2021.

[15] Tao Liu, Ruida Zhou, Dileep Kalathil, PR Kumar, and Chao Tian. Learning policies with zero or bounded constraint violation for constrained mdps. *arXiv preprint arXiv:2106.02684*, 2021.

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

[17] Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems*, 29:3450–3458, 2016.

[18] Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pages 207–216. IEEE, 2013.

[19] Huasen Wu, Rayadurgam Srikant, Xin Liu, and Chong Jiang. Algorithms with logarithmic or sublinear regret for constrained contextual bandits. *arXiv preprint arXiv:1504.06937*, 2015.

[20] Yifan Wu, Roshan Shariff, Tor Lattimore, and Csaba Szepesvári. Conservative bandits. In *International Conference on Machine Learning*, pages 1254–1262. PMLR, 2016.

[21] Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi-Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. *arXiv preprint arXiv:1611.06426*, 2016.

[22] Evrard Garcelon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Matteo Pirotta. Improved algorithms for conservative exploration in bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3962–3969, 2020.

[23] Yifang Chen, Alex Cuellar, Haipeng Luo, Jignesh Modi, Heramb Nemlekar, and Stefanos Nikolaidis. Fair contextual multi-armed bandits: Theory and experiments. In *Conference on Uncertainty in Artificial Intelligence*, pages 181–190. PMLR, 2020.

[24] Fengjiao Li, Jia Liu, and Bo Ji. Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering*, 7(3):1799–1813, 2019.

[25] Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. *arXiv preprint arXiv:1908.05814*, 2019.

[26] Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 2827–2835. PMLR, 2021.

[27] Ahmadreza Moradipari, Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Safe linear thompson sampling with side information. *arXiv preprint arXiv:1911.02156*, 2019.

[28] Xin Liu, Bin Li, Pengyi Shi, and Lei Ying. Pond: Pessimistic-optimistic online dispatch. *arXiv preprint arXiv:2010.09995*, 2020.

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

[30] Jonathan Scarlett, Ilija Bogunovic, and Volkan Cevher. Lower bounds on regret for noisy gaussian process bandit optimization. In *Conference on Learning Theory*, pages 1723–1742. PMLR, 2017.

[31] Xu Cai and Jonathan Scarlett. On lower bounds for standard and robust gaussian process bandit optimization. In *International Conference on Machine Learning*, pages 1216–1226. PMLR, 2021.

[32] David Eriksson and Matthias Poloczek. Scalable constrained bayesian optimization. In *International Conference on Artificial Intelligence and Statistics*, pages 730–738. PMLR, 2021.

[33] Michael A Gelbart, Jasper Snoek, and Ryan P Adams. Bayesian optimization with unknown constraints. *arXiv preprint arXiv:1403.5607*, 2014.

[34] José Miguel Hernández-Lobato, Michael A Gelbart, Ryan P Adams, Matthew W Hoffman, and Zoubin Ghahramani. A general framework for constrained bayesian optimization using information-based search. *MIT Press*, 2016.

[35] Hao Yu, Michael J Neely, and Xiaohan Wei. Online convex optimization with stochastic constraints. *arXiv preprint arXiv:1708.03741*, 2017.

[36] Carl Edward Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

[37] Sattar Vakili, Kia Khezeli, and Victor Picheny. On information gain and regret bounds in gaussian process bandits. In *International Conference on Artificial Intelligence and Statistics*, pages 82–90. PMLR, 2021.

[38] Amir Beck. *First-order methods in optimization*. SIAM, 2017.

[39] Branislav Kveton, Csaba Szepesvari, Mohammad Ghavamzadeh, and Craig Boutilier. Perturbed-history exploration in stochastic linear bandits. *arXiv preprint arXiv:1903.09132*, 2019.

[40] Sharan Vaswani, Abbas Mehrabian, Audrey Durand, and Branislav Kveton. Old dog learns new tricks: Randomized ucb for bandit problems. *arXiv preprint arXiv:1910.04928*, 2019.

[41] Baekjin Kim and Ambuj Tewari. Randomized exploration for non-stationary stochastic linear bandits. In *Conference on Uncertainty in Artificial Intelligence*, pages 71–80. PMLR, 2020.

[42] Daniele Calandriello, Luigi Carratino, Alessandro Lazaric, Michal Valko, and Lorenzo Rosasco. Gaussian process optimization with adaptive sketching: Scalable and no regret. In *Conference on Learning Theory*, pages 533–557. PMLR, 2019.

[43] Zihan Li and Jonathan Scarlett. Gaussian process bandit optimization with few batches. In *International Conference on Artificial Intelligence and Statistics*, pages 92–107. PMLR, 2022.

[44] Sudeep Salgia, Sattar Vakili, and Qing Zhao. A domain-shrinking based bayesian optimization algorithm with order-optimal regret performance. *Advances in Neural Information Processing Systems*, 34:28836–28847, 2021.

[45] Romain Camilleri, Kevin Jamieson, and Julian Katz-Samuels. High-dimensional experimental design and kernel bandits. In *International Conference on Machine Learning*, pages 1227–1237. PMLR, 2021.

[46] Sattar Vakili, Jonathan Scarlett, Da-shan Shiu, and Alberto Bernacchia. Improved convergence rates for sparse approximation methods in kernel-based learning. *arXiv preprint arXiv:2202.04005*, 2022.

[47] Sayak Ray Chowdhury and Aditya Gopalan. Bayesian optimization under heavy-tailed payoffs. In *Advances in Neural Information Processing Systems*, pages 13790–13801, 2019.

[48] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24:2312–2320, 2011.

## Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This is rather a theoretical work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes] In the appendix
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] Experiments were rather run on a personal computer.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

# Appendix

## A 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 $\phi_t$ 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,

$$\sum_{t=1}^{T} \bar{g}_t(x_t)(\phi - \phi_t) \leq \frac{V}{2}(\phi_1 - \phi)^2 + \sum_{t=1}^{T} \frac{1}{2V} \bar{g}_t(x_t)^2, \tag{7}$$

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. [38, 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 $\pi$ (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

$$\begin{aligned}&\mathcal{R}_+(T) + \phi \sum_{t=1}^{T} g(x_t) \\ =&T\mathbb{E}_{\pi^*}[f(x)] - \sum_{t=1}^{T} f(x_t) + \phi \sum_{t=1}^{T} g(x_t) \\ =&T\mathbb{E}_{\pi^*}[f(x)] - \sum_{t=1}^{T} \bar{f}_t(x_t) + \sum_{t=1}^{T} \bar{f}_t(x_t) - f(x_t) + \phi \sum_{t=1}^{\tau} g(x_t).\end{aligned} \tag{8}$$

We can further bound (8) by using Lemma 1. In particular, we have

$$
\begin{aligned}
&T\mathbb{E}_{\pi^*}\left[f(x)\right] - \sum_{t=1}^{T}\bar{f}_t(x_t) + \sum_{t=1}^{T}\bar{f}_t(x_t) - f(x_t) + \phi\sum_{t=1}^{\tau} g(x_t) \\
\overset{(a)}{\leq}& \sum_{t=1}^{T}\mathbb{E}_{\pi^*}\left[f(x)\right] - \phi_t\mathbb{E}_{\pi^*}\left[g(x)\right] - \sum_{t=1}^{T}\bar{f}_t(x_t) + \sum_{t=1}^{T}\bar{f}_t(x_t) - f(x_t) + \phi\sum_{t=1}^{T} g(x_t) \\
\overset{(b)}{=}& \sum_{t=1}^{T}\mathbb{E}_{\pi^*}\left[f(x)\right] - \phi_t\mathbb{E}_{\pi^*}\left[g(x)\right] - \left(\sum_{t=1}^{T}\bar{f}_t(x_t) - \phi_t\bar{g}_t(x_t)\right) + \sum_{t=1}^{T}\bar{f}_t(x_t) - f(x_t) \\
&+ \phi\left(\sum_{t=1}^{T} g(x_t) - \bar{g}_t(x_t)\right) + \sum_{t=1}^{T}\bar{g}_t(x_t)(\phi - \phi_t) \\
\overset{(c)}{\leq}& \sum_{t=1}^{T}\mathbb{E}_{\pi^*}\left[f(x)\right] - \phi_t\mathbb{E}_{\pi^*}\left[g(x)\right] - \left(\sum_{t=1}^{T}\bar{f}_t(x_t) - \phi_t\bar{g}_t(x_t)\right) + \sum_{t=1}^{\tau}\bar{f}_t(x_t) - f(x_t) \\
&+ \phi\left(\sum_{t=1}^{T} g(x_t) - \bar{g}_t(x_t)\right) + \frac{V}{2}(\phi_1 - \phi)^2 + \sum_{t=1}^{T}\frac{1}{2V}\bar{g}_t(x_t)^2 \\
\overset{(d)}{=}& \mathcal{T}_1 + \mathcal{T}_2 + \frac{V}{2}\phi^2 + \frac{1}{2V}TG^2,
\end{aligned} \tag{9}
$$

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.,

$$
\mathcal{T}_1 = \sum_{t=1}^{T}(\mathbb{E}_{\pi^*}\left[f(x)\right] - \phi_t\mathbb{E}_{\pi^*}\left[g(x)\right]) - \sum_{t=1}^{T}(\bar{f}_t(x_t) - \phi_t\bar{g}_t(x_t)), \tag{10}
$$

$$
\mathcal{T}_2 = \sum_{t=1}^{T}(\bar{f}_t(x_t) - f(x_t)) + \phi\sum_{t=1}^{T}(g(x_t) - \bar{g}_t(x_t)). \tag{11}
$$

Plugging (9) into (8), yields for any $\phi \in [0, \rho]$,

$$
\mathcal{R}_+(T) + \phi\sum_{t=1}^{T} g(x_t) \leq \mathcal{T}_1 + \mathcal{T}_2 + \frac{V}{2}\phi^2 + \frac{1}{2V}TG^2. \tag{12}
$$

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 $\phi$. This directly leads to the following inequality (with $V = G\sqrt{T}/\rho$) for any $\phi \in [0, \rho]$:

$$
\mathcal{R}_+(T) + \phi\sum_{t=1}^{T} g(x_t) \leq \chi(T, \phi) + \frac{\phi^2 G\sqrt{T}}{2\rho} + \frac{\rho G\sqrt{T}}{2}. \tag{13}
$$

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

**Regret**. We can simply choose $\phi = 0$ in (13), and obtain that with high probability

$$
\mathcal{R}_+(T) = O\left(\chi(T, 0) + \rho G\sqrt{T}\right). \tag{14}
$$

**Constraint violation**. To obtain the bound on $\mathcal{V}(T)$, inspired by [11], 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 $\pi'$ by the convexity of probability measure. As a result, we have

$$
\mathbb{E}_{\pi^*}\left[f(x)\right] - \mathbb{E}_{\pi'}\left[f(x)\right] + \rho\left[\mathbb{E}_{\pi'}\left[g(x)\right]\right]_+ = \frac{1}{T}\mathcal{R}_+(T) + \frac{1}{T}\phi\sum_{t=1}^{T} g(x_t) \leq \frac{\chi(T, \rho) + \rho G\sqrt{T}}{T}, \tag{15}
$$

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 (13) with $\phi = \rho$ since (13) holds for all $\phi \in [0, \rho]$ and $\chi(T, \phi)$ is increasing in $\phi$.

Then, we will apply the following useful lemma, which is adapted from Theorem 3.60 in [38].

**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 $\pi_0$ such that $w(\pi_0) \leq -\delta$, and a constant $\rho \geq 2\kappa^*$, where $\kappa^*$ is the optimal dual variable, i.e., $\kappa^* = \operatorname{argmin}_{\lambda \geq 0}(\max_\pi h(\pi) - \kappa w(\pi))$. Assume that $\pi' \in \mathcal{C}$ satisfies*

$$h(\pi^*) - h(\pi') + \rho\,[w(\pi')]_+ \leq \varepsilon, \tag{16}$$

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

Thus, since (15) satisfies (16) and $\mathbb{E}_\pi\,[h(x)]$ for any fixed $h$ is a linear function with respect to $\pi$, by Lemma 2, we have

$$\mathcal{V}(T) = O\left(\frac{1}{\rho}\chi(T, \rho) + G\sqrt{T}\right). \tag{17}$$

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 [5, 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 (10), we have

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

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

$$\begin{aligned}
\mathcal{T}_2 &\overset{(a)}{\leq} \sum_{t=1}^T 2\beta_t \sigma_{t-1}(x_t) + \phi \sum_{t=1}^T 2\widetilde{\beta}_t \widetilde{\sigma}_{t-1}(x_t) \\
&\overset{(b)}{\leq} O\left(\beta_T \sqrt{T\gamma_T} + \phi \widetilde{\beta}_T \sqrt{T\widetilde{\gamma}_T}\right),
\end{aligned} \tag{18}$$

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. [5, Lemma 4]). Note that we have also used the fact that $\beta_t$ 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 (14), 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 (17), yields the bound on constraint violation as

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

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

## B 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 $\pi_t$ 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 3. Under our newly introduced notation, we have

$$\mathcal{T}_1 + \mathcal{T}_2 = \sum_{t=1}^{T} \left(z_{\phi_t}(\pi^*) - \widehat{z}_{\phi_t}(\pi_t) + \widehat{z}_\phi(\pi_t) - z_\phi(\pi_t)\right) := \sum_{t=1}^{T} d_t, \tag{19}$$

where $z_{\phi_t}(\cdot) := f(\cdot) - \phi_t g(\cdot)$ and $\widehat{z}_{\phi_t}(\cdot) := \bar{f}_t(\cdot) - \phi_t \bar{g}_t(\cdot)$, and similar definitions for $z_\phi$ and $\widehat{z}_\phi$. 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 $u_t = \operatorname{argmin}_{\pi \in \bar{S}_t} \alpha_{\phi_t}(\pi)$. Thus, conditioned on $E^{est}$ and $E_t^{conc}$, we have

$$\begin{aligned}
d_t &= z_{\phi_t}(\pi^*) - \widehat{z}_{\phi_t}(\pi_t) + \widehat{z}_\phi(\pi_t) - z_\phi(\pi_t) \\
&= z_{\phi_t}(\pi^*) - z_{\phi_t}(u_t) + z_{\phi_t}(u_t) - \widehat{z}_{\phi_t}(\pi_t) + \widehat{z}_\phi(\pi_t) - z_\phi(\pi_t) \\
&= \Delta_{\phi_t}(u_t) + z_{\phi_t}(u_t) - \widehat{z}_{\phi_t}(\pi_t) + \widehat{z}_\phi(\pi_t) - z_\phi(\pi_t) \\
&\overset{(a)}{\leq} \Delta_{\phi_t}(u_t) + \widehat{z}_{\phi_t}(u_t) - \widehat{z}_{\phi_t}(\pi_t) + \alpha_{\phi_t}(u_t) + \alpha_\phi(\pi_t) \\
&\overset{(b)}{\leq} \Delta_{\phi_t}(u_t) + \alpha_{\phi_t}(u_t) + \alpha_\phi(\pi_t) \\
&\overset{(c)}{\leq} 2\alpha_{\phi_t}(u_t) + \alpha_\phi(\pi_t),
\end{aligned} \tag{20}$$

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

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

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. (20) 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 $\phi, \phi_t$ are bounded by $\rho$. 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

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

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 $\pi_j \in S_t$ $\widehat{z}_{\phi_t}(\pi_j) \leq z_{\phi_t}(\pi_j) + \alpha_{\phi_t}(\pi_j) \leq z_{\phi_t}(\pi_j) + \Delta_{\phi_t}(\pi_j) = z_{\phi_t}(\pi^*)$; (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 $E^{est}$,

$$\begin{aligned}
\mathbb{E}_t\left[d_t\right] &\leq \mathbb{E}_t\left[\alpha_\rho(x_t)\right]\left(1 + \frac{2}{p_3 - p_{2,t}}\right) + (4B + 4\rho G)p_{2,t} \\
&\leq \frac{1}{p_4}\mathbb{E}_t\left[\alpha_\rho(x_t)\right] + (4B + 4\rho G)p_{2,t}. \qquad (21)
\end{aligned}$$

where the last inequality follows from the boundedness condition in the sufficient condition. In order to obtain a high probability bound, inspired by [5], we will resort to martingale techniques. Let us define the following terms

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

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

*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. (21). 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

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

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

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

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 $1 - \alpha - p_1$,

$$\begin{aligned}
\sum_{t=1}^{T} d_t &\leq \frac{1}{p_4}\sum_{t=1}^{T}\alpha_\rho(x_t) + C'(4B + 4\rho G) + \frac{(c_f(T) + \rho c_g(T))(4B + 4\rho G)}{p_4}\sqrt{2T\ln(1/\delta)} \\
&= O\left(\frac{1}{p_4}\sum_{t=1}^{T}(c_f(T)\sigma_{t-1}(x_t) + \rho c_g(T)\widetilde{\sigma}_{t-1}(x_t)) + \frac{(c_f(T) + \rho c_g(T))\kappa}{p_4}\sqrt{2T\ln(1/\delta)}\right) \\
&= O\left(\frac{1}{p_4}c_f(T)\sqrt{T\gamma_T} + \frac{1}{p_4}\rho c_g(T)\sqrt{T\widetilde{\gamma}_T} + \frac{(c_f(T) + \rho c_g(T))\kappa}{p_4}\sqrt{2T\ln(1/\delta)}\right), \quad (22)
\end{aligned}$$

where $\kappa := 4B + 4\rho G$. Plugging (22) into (19), we obtain that

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

Note that here $\chi(T, \phi)$ is independent of $\phi$ since we have bounded it by $\rho$ in the analysis. Finally, plugging $\chi(T, \phi)$ into (14) and (17) yields the results of Theorem 2. □

# C Flexible Implementations of RandGP-UCB

In this section, we will give more insights on the choices of $\widehat{\mathcal{D}}$, i.e., sampling distribution for $\widehat{Z}_t$. 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.

# D Details on Heavy-Tailed Real-World Data

This dataset is the adjusted closing price of 29 stocks from January 4th, 2016 to April 10th 2019. We use it in the context of identifying the most profitable stock in a given pool of stocks. As verified in [47], 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.

# E More Details on Zero Constraint Violation

**Claim 1.** $T\mathbb{E}_{\pi^*}\left[f(x)\right] - T\mathbb{E}_{\pi_\varepsilon^*}\left[f(x)\right] \leq \frac{2BT\varepsilon}{\delta}$.

To show this, we let $\pi_\varepsilon(x) := (1 - \frac{\varepsilon}{\delta})\pi^*(x) + \frac{\varepsilon}{\delta}\pi_0(x)$, where $\pi^*$ is the optimal solution to the original baseline problem and $\pi_0$ is the Slater's policy satisfying Slater's condition. First, we note that $\pi_\varepsilon$ is a feasible solution to the new baseline problem introduced above. To see this, we note that $\pi_\varepsilon(x) \geq 0$ and

$$\mathbb{E}_{\pi_\varepsilon}\left[g(x)\right] = (1 - \frac{\varepsilon}{\delta})\mathbb{E}_{\pi^*}\left[g(x)\right] + \frac{\varepsilon}{\delta}\mathbb{E}_{\pi_0}\left[g(x)\right] \leq 0 + (-\varepsilon) = -\varepsilon.$$

Since $\pi_\varepsilon^*$ is the optimal solution while $\pi_\varepsilon$ is a feasible one, we have

$$\begin{aligned} T\mathbb{E}_{\pi^*}\left[f(x)\right] - T\mathbb{E}_{\pi_\varepsilon^*}\left[f(x)\right] &\leq T\mathbb{E}_{\pi^*}\left[f(x)\right] - T\mathbb{E}_{\pi_\varepsilon}\left[f(x)\right] \\ &= T\left(\mathbb{E}_{\pi^*}\left[f(x)\right] - (1 - \frac{\varepsilon}{\delta})\mathbb{E}_{\pi^*}\left[f(x)\right] - \frac{\varepsilon}{\delta}\mathbb{E}_{\pi_0}\left[f(x)\right]\right) \\ &\leq \frac{2BT\varepsilon}{\delta}, \end{aligned}$$

where in the last step, we use the boundedness of $f$. Therefore, one can properly choose $\varepsilon$ such that the subtraction of $\varepsilon T$ in the constraint violation can cancel the leading term $O(\sqrt{T})$ (hence bounded or even zero constraint violation) while only incurring an additional additive term of the same order in the regret.

**Remark 4.** *We believe that the same slackness trick can improve the existing $\widetilde{O}(\sqrt{T})$ constraint violation in MDPs [11, 12].*

# F Discussion on Alternative Method

To the best of our knowledge, there exist two popular methods for analyzing constrained bandits or MDPs. They are both based on primal-dual optimization and differ mainly in the analysis techniques. The first one is based on convex optimization tools as in [11, 12] and our paper. The other one is based on Lyapunov-drift arguments as in [13–15]. For simplicity, we call the first method *convex-opt* method and the second one as *Lyapunov-drift* method. Before we provide further discussion, one thing to note is that all existing works only deal with UCB-type exploration for tabular or linear functions, while our paper is the first one that studies general functions with general exploration strategies beyond UCB.

**Algorithm 2** Algorithm in the Lyapunov-drift method

**Parameters:** $V, \varepsilon, Q(1) = 0$
**for** batch $t = 1, 2, \ldots$ **do**
  Generate estimate $f_t(x), g_t(x)$ and truncate them to $\bar{f}_t, \bar{g}_t$
  Pseudo-acquisition function: $z_t(x) = \bar{f}_t(x) - \frac{1}{V} Q(t) \bar{g}_t(x)$
  Choose action $x_t = \operatorname{argmax}_{x \in \mathcal{X}} z_t(x)$; observe reward $r_t$, and cost $c_t$
  Pseudo-acquisition function: $\widehat{z}_{\phi_t}(x) = \bar{f}_t(x) - \phi_t \bar{g}_t(x)$
  Update virtual queue: $Q(t+1) = [Q(t) + \bar{g}_t(x_t) + \varepsilon]_+$
  Posterior model update: using observations to update model
**end for**

Now we first briefly explain the main idea behind the Lyapunov-drift method when applied to our setting (for the UCB exploration only). It basically has the same algorithm as the convex-opt method. One minor change is that in Lyapunov-drift method, the dual variable is not truncated by $\rho$ and is denoted by $Q(t)$, since this dual update is similar to a typical queue length update in queueing theory, i.e., truncated at zero; see Algorithm 2. To bound the regret, Lyapunov-drift method decomposes it as the following one, where $\varepsilon \leq \delta/2$ is the slackness as in the last section.

$$
\begin{aligned}
\mathcal{R}_+(T) &= T\mathbb{E}_{\pi^*}\left[f(x)\right] - \sum_{t=1}^{T} f(x_t) \\
&= \underbrace{T\mathbb{E}_{\pi^*}\left[f(x)\right] - T\mathbb{E}_{\pi^*_\varepsilon}\left[f(x)\right]}_{\text{Term 1}} + \underbrace{\sum_{t=1}^{T} \int_{x \in \mathcal{X}} \left(f(x) - \bar{f}_t(x)\right) \pi^*_\varepsilon(x)\, dx}_{\text{Term 2}} \\
&\quad + \underbrace{\sum_{t=1}^{T} \int_{x \in \mathcal{X}} \bar{f}_t(x) \pi^*_\varepsilon(x)\, dx - \bar{f}_t(x_t)}_{\text{Term 3}} + \underbrace{\sum_{t=1}^{T} \bar{f}_t(x_t) - f(x_t)}_{\text{Term 4}}. \qquad (23)
\end{aligned}
$$

From this, one can see that Term 1, Term 2, and Term 4 can be easily bounded under UCB-type exploration. In particular, by optimism and well-concentration of $\bar{f}_t$, one has Term 2 $\leq 0$ and Term 4 $= \widetilde{O}(\sqrt{T})$ (we ignore $\gamma_T$ term in this section for simplicity). Moreover, Term 1 enjoys the bound as in Claim 1. Thus, the only challenge is to bound Term 3, which cannot be naturally bounded by greedy selection as in the standard way, since in the constrained case, the greedy selection is with respect to the combined function. To handle this, one needs the following result, which not only helps to bound Term 3, but also is the key in bounding the constraint violation.

**Lemma 4.** *Let $\Delta(t) := L(Q(t+1)) - L(Q(t)) = \frac{1}{2}(Q(t+1))^2 - \frac{1}{2}(Q(t))^2$. For any $\pi$, we have*

$$
\Delta(t) \leq -V\left(\int_{x \in \mathcal{X}} \bar{f}_t(x)\pi(x)\, dx - \bar{f}_t(x_t)\right) + \frac{1}{2}(G+\varepsilon)^2 + Q(t)\left(\int_{x \in \mathcal{X}} \bar{g}_t(x)\pi(x)\, dx + \varepsilon\right). \qquad (24)
$$

*Proof.* See Appendix F.1. □

Thus, one can see that the first term on the RHS of (24) exists in Term 3 if one chooses $\pi = \pi^*_\varepsilon$. By the optimism $\bar{g}_t(x) \leq g(x)$ and the definition of $\pi^*_\varepsilon$, with a telescope summation, one can easily bound Term 3, hence the regret bound.

**Comparison in regret analysis.** Compared to the regret decomposition in our paper (i.e., (5) and (6)), (23) in the Lyapunov-drift method is more tailored to UCB-type exploration in the sense that the Term 3 is upper bounded separately using the optimism. As a result, it is unclear to us how to generalize it to handle general exploration strategies where one often need to bound Term 2 + Term 3 + Term 4 together and optimism does not hold in general. In contrast, our decomposition (5) and (6) basically keep the same pattern as in the unconstrained case, which enables us to utilize this structure to handle general exploration strategies.

We now turn to the constraint violation. By the virtual queue length update in Algorithm 2, the key step behind the constraint violation bound is to bound $Q(T+1)$. To see this, by the virtual queue length update in Algorithm 2, we have

$$Q(T+1) \geq \sum_{t=1}^{T} \bar{g}_t(x_t) + T\varepsilon = \sum_{t=1}^{T} \bar{g}_t(x_t) - g(x_t) + g(x_t) + T\varepsilon,$$

which implies that

$$\sum_{t=1}^{T} g(x_t) \leq Q(T+1) + \sum_{t=1}^{T} (g(x_t) - \bar{g}_t(x_t)) - T\varepsilon = Q(T+1) + \widetilde{O}(\sqrt{T}) - T\varepsilon,$$

where in the last step we uses the well-concentration of $\bar{g}_t$. To bound the remaining term $Q(T+1)$, the Lyapunov-drift method resorts to a classic tool in queueing theory, i.e., Hajek lemma [16], to bound the virtual queue length at time $T+1$. The idea behind it is simple: if the queue length drift $\Delta(t)$ (defined in Lemma 4) is negative whenever the queue length is large, then $Q(T+1)$ is bounded. To establish the negative drift, one resorts to (24) again by choosing $\pi = \pi_0$. By the definition of $\pi_0$ (Slater's policy), the optimism $\bar{g}_t(x) \leq g(x)$ and boundedness of $\bar{f}_t$, one can easily establish a negative drift, and hence the constraint violation.

**Comparison in constraint violation analysis.** Instead of using Hajek lemma, we directly utilize the convex optimization tool to obtain the constraint violation as in [11, 12], which is conceptually simpler. Moreover, the current constraint violation analysis in the Lyapunov-drift method also relies on the optimism of $\bar{g}_t$, which does not hold in general explorations beyond UCB. On the other hand, one possible limitation of convex-opt method is that the constraint violation depends on the maximum information gain of both $f$ and $g$ (for small $T$) while under Lyapunov-drift method and UCB exploration, the constraint violation only depends on $g$. Finally, when applying Hajek lemma to bound the virtual queue length, there exists a subtlety that makes the standard expected version of Hajek lemma fail due to the correlation of virtual queue length $Q(t)$ and $\bar{g}_t$. We give more details on this subtlety in Appendix G.

**Summary.** Both methods are able to establish sublinear regret and sublinear constraint violation under UCB. Moreover, with the aid of slackness (i.e., $\varepsilon$) in the dual update, both methods can establish bounded or even zero constraint violation. For general exploration strategies beyond UCB, we tend to believe that convex-opt method has advantages over the current analysis in the Lyapunov-drift method, since the latter explicitly relies on optimism in both regret and constraint violation analysis. On the other hand, Lyapunov-drift method can easily handle an anytime slackness, say, $\varepsilon_t = O(1/\sqrt{t})$ rather than $\varepsilon = O(1/\sqrt{T})$.

### F.1 Proof of Lemma 4

*Proof.* Note that by the update rule of the virtual queue in Algorithm 2 and non-expansiveness of projection, we have

$$\Delta(t) \leq Q(t)(\bar{g}_t(x_t) + \varepsilon) + \frac{1}{2}\left(\bar{g}_t(x_t) + \varepsilon\right)^2.$$

Now we will bound the RHS as follows.

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

where (a) holds by the boundedness of $\bar{g}_t$; (b) holds by the greedy selection in Algorithm 2. Reorganizing the term, yields the required result. □

## G Subtlety in Applying Hajek Lemma to Constraint Violation

As stated before, the key step behind the constraint violation is to establish a negative drift of the virtual queue and then by Hajek lemma, one can show that the virtual queue is bounded in expectation, which in turn can be used to establish a zero constraint violation with a proper choice of slackness variable (i.e., $\varepsilon$) in the virtual queue update. However, the negative drift condition in the standard Hajek lemma (cf. Lemma 11 in [13]) requires a conditional expectation, i.e., condition on all large enough $Q$, the expected drift is negative. Then, if one directly applies the standard Hajek lemma, she would proceed as follows. The goal is to show that $\mathbb{E}\left[\Delta(t) \mid Q(t) = Q\right] \leq -cQ$ for all large $Q$ and $c$ is some positive constant. Recall the bound on $\Delta(t)$ in (24), by the boundedness and let $\pi = \pi_0$, the key is to show that

$$\mathbb{E}\left[\int_{x\in\mathcal{X}} \bar{g}_t(x)\pi_0(x)\,dx + \varepsilon \mid Q(t) = Q\right] \leq -c. \tag{25}$$

To illustrate the idea, we simply suppose that the Slater's condition is satisfied at a single point $x_0$ and $\varepsilon = 0$. To show the above inequality, she may choose the following direction.

$$\mathbb{E}\left[\bar{g}_t(x_0)|Q(t) = Q\right] = \underbrace{\mathbb{E}\left[\bar{g}_t(x_0) - g(x_0)|Q(t) = Q\right]}_{\text{Term (i)}} + \underbrace{\mathbb{E}\left[g(x_0)|Q(t) = Q\right]}_{\text{Term (ii)}} \leq -c.$$

For Term (ii), it is easily bounded by Term (ii) $\leq -\delta$ via Slater's condition since $g(\cdot)$ is a fixed function. To bound Term (i), she may resort to the standard self-normalized inequality for linear bandits and the definition of UCB exploration (cf. [48]). By these standard results, she can show that for any fixed $\alpha \in (0, 1]$, the following holds:

$$\mathbb{P}\{\forall x, \forall t, \bar{g}_t(x) \leq g(x)\} \geq 1 - \alpha. \tag{26}$$

That is, $\bar{g}_t$ is optimistic with respect to $g$. Then, by setting $\alpha = 1/T$ and using the boundedness assumption of both $\bar{g}_t$ and $g$, she may conclude that Term (i) $= O(1/T)$. Unfortunately, the bound on Term (i) is ungrounded since it is obtained by treating the conditional expectation in Term (i) as an unconditional expectation. The subtlety here is that one cannot remove the condition on $Q(t)$ in Term (i), since $\bar{g}_t$ 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, (26) does not necessarily hold, and moreover, the concentration of $\bar{g}_t$ given $Q(t)$ is hard to compute in this case. As a result, the conditional expectation for Term (i) is hard to compute in general.

**One correct way.** Instead of applying the standard expected version of Hajek lemma, one can consider removing the expectation in Hajek lemma by directly showing that $\bar{g}_t(x_0) \leq -c$ almost surely under the "good event". This is exactly the approach used in [15] (cf. Lemma 5.6). In this way, one can show that with a high probability (i.e., under good event), a negative drift exists and hence the constraint violation bound with high probability.

## H Proof of Corollaries

First, we remark that the first event $E^{est}$ in the probability condition of Assumption 3 can be easily obtained by standard GP concentration result. That is, by [5, 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 $\mathbb{P}\left(E^{est}\right) \geq 1 - p_1$ with $p_1 = \alpha_f + \alpha_g$, $c^{(1)}_{f,t} = \beta_t$, and $c^{(1)}_{g,t} = \widetilde{\beta}_t = B + R\sqrt{2(\widetilde{\gamma}_{t-1} + 1 + \ln(1/\alpha_g))}$. Thus, we only need to check probability condition for the remaining two events and the boundedness condition under different exploration methods.

*Proof of Corollary 1.* <u>UCB:</u> 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 $E^{conc}_t$ and $E^{anti}_t$ hold with probability one. Moreover, the boundedness condition naturally holds.

<u>TS:</u> By Definition 1, we have that given the history up to the end of round $t-1$, $f_t(x) \sim \mathcal{N}(\mu_{t-1}(x), \beta_t^2\sigma^2_{t-1}(x))$ and $g_t(x) \sim \mathcal{N}(\widetilde{\mu}_{t-1}(x), \widetilde{\beta}_t^2\widetilde{\sigma}^2_{t-1}(x))$. Thus, for any fixed $x \in \mathcal{X}$, by concentration of Gaussian distribution, we have $\mathbb{P}_t(|f_t(x) - \mu_{t-1}(x)| \leq 2\sigma_{t-1}(x)\beta_t\sqrt{\ln t}) \geq 1 - 1/t^2$, and hence, using the union bound over all $x$, we obtain $\forall x$, $\mathbb{P}_t\left(E^{conc}_{t,f}(x)\right) \geq 1 - 1/t^2$ with

$c_{f,t}^{(2)} = 2\beta_t\sqrt{\ln(|\mathcal{X}|t)}$. Similarly, we have $\forall x, \mathbb{P}_t\left(E_{t,g}^{conc}(x)\right) \geq 1 - 1/t^2$ with $c_{g,t}^{(2)} = 2\widetilde{\beta}_t\sqrt{\ln(|\mathcal{X}|t)}$. Hence, by union bound, we have $\mathbb{P}_t\left(E_t^{conc}\right) \geq 1 - p_{2,t}$ with $p_{2,t} = 2/t^2$. Moreover, when $\pi^*$ concentrates on a single point, by standard anti-concentration result of Gaussian distribution (e.g., Lemma 8 in [5]), we have $\mathbb{P}_t(E_{t,f}^{anti}) \geq p$ with $p := \frac{1}{4e\sqrt{\pi}}$. Similarly, we also have $\mathbb{P}_t(E_{t,g}^{anti}) \geq p$. By independent sampling of $f_t$ and $g_t$, we have $\mathbb{P}_t\left(E_t^{anti}\right) \geq p_3$ with $p_3 = p^2$. The boundedness condition holds due to $\sum_{t=1}^T p_{2,t} \leq 2\sum_{t=1}^T 1/t^2 \leq \pi^2/3 := C'$ and $p_4 = O(p^2)$. □

*Proof of Corollary 2.* 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

$$\mathbb{P}_t(\forall x, |f_t(x) - \mu_{t-1}(x)| \leq 2\sigma_{t-1}(x)\beta_t\sqrt{\ln t}) \geq 1 - 1/t^2,$$

thanks to the "coupled" noise. 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 $\mathbb{P}_t(E_{t,f}^{anti}) \geq \mathbb{P}_t(Z_t \geq \beta_t) \geq p$, where $p := \frac{1}{4e\sqrt{\pi}}$. Similarly, we have $\mathbb{P}_t(E_{t,g}^{anti}) \geq \mathbb{P}_t(Z_t \leq -\widetilde{\beta}_t) \geq p$. 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 $p_4 = O(p^2)$. □