Rectified Pessimistic-Optimistic Learning for Stochastic Continuum-armed Bandit with Constraints

Hengquan Guo GUOHQ@SHANGHAITECH.EDU.CN
Qi Zhu ZHUQI2022@SHANGHAITECH.EDU.CN
Xin Liu LIUXIN7@SHANGHAITECH.EDU.CN
Shanghaitech University

Editors: N. Matni, M. Morari, G. J. Pappas

Abstract

This paper studies the problem of stochastic continuum-armed bandit with constraints (SCBwC), where we optimize a black-box reward function $f(x)$ subject to a black-box constraint function $g(x) \leqslant 0$ over a continuous space $\mathcal{X}$. We model reward and constraint functions via Gaussian processes (GPs) and propose a Rectified Pessimistic-Optimistic Learning (RPOL) framework, a penalty-based method incorporating optimistic and pessimistic GP bandit learning for reward and constraint functions, respectively. We consider the metric of cumulative constraint violation $\sum_{t=1}^{T}(g(x_t))^+$, which is strictly stronger than the traditional long-term constraint violation $\sum_{t=1}^{T} g(x_t)$. The rectified design for the penalty update and the pessimistic learning for the constraint function in RPOL guarantee the cumulative constraint violation is minimal. RPOL can achieve sublinear regret and cumulative constraint violation for SCBwC and its variants (e.g., under delayed feedback). These theoretical results match their unconstrained counterparts. Our experiments justify RPOL outperforms several existing baseline algorithms.[1]

Keywords: Stochastic continuum-armed bandit; Hard constraint; Bayesian optimization

1. Introduction

Stochastic continuum-armed bandit optimization is a powerful framework to model many real-world applications, (e.g., networking resource allocation Fu and Modiano (2021), online recommendation Krause and Ong (2011), clinic trials Durand et al. (2018), neural network architecture search White et al. (2021). In stochastic continuum-armed bandits, the learner aims to optimize a black-box reward/utility function over a continuous feasible set $\mathcal{X}$ by sequentially interacting with the environment. The interaction with the practical environment is often subject to a variety of operational constraints, which are also black-box and complicated. For example, in networking resource allocation, we maximize the users' quality of experience under complex resource constraints; in clinic trials, we optimize the quality of treatment while guaranteeing the side effect of patients minimal; in the neural architecture search, we search a neural network with a small generalization error while keeping the training time within the time limit. In these applications, the learner requires to optimize a black-box reward/utility function $f(x)$ while keeping the black-box constraint $(g(x) \leqslant 0)$ satisfied. The black-box problem is unsolvable in general without any regularity assumption on $f(x)$ and $g(x)$. We assume the reward and constraint functions lie in Reproducing Kernel Hilbert Space (RKHS) with a bounded norm such that $f(x)$ and $g(x)$ can be modeled via Gaussian processes.

1. The technical report can be found in Guo et al. (2022b). The corresponding author: Xin Liu.

The previous works in stochastic continuum-armed bandit with constraints (SCBwC) are classified into two categories according to the type of constraints: hard and soft constraints, respectively. For the type of hard constraints, there is a sequence of studies on safe Bayesian optimization Amani et al. (2020); Berkenkamp et al. (2021); Sui et al. (2015, 2018), where the algorithms satisfy the constraint instantaneously at each round, i.e., hard constraint. However, these results rely on the key assumption that an initial safe/feasible decision set is known apriori; otherwise, it would be impossible to guarantee the hard constraints. Moreover, the algorithms in Amani et al. (2020); Berkenkamp et al. (2021); Sui et al. (2015, 2018) suffer from high-computation complexity because they require to construct a safe decision set and search for a safe and optimal solution for each round. Without any prior information on the constraint function or safe set, the constraint violation is *unavoidable*. A recent line of work focuses on the soft constraints Ariafar et al. (2019); Shi and Eryilmaz (2022); Zhou and Ji (2022), which allow the constraints to be violated as long as they are satisfied in the long term. In other words, the soft constraint violation $\sum_{t=1}^{T} g(x_t)$ should be as small as possible. The soft constraint violation is a reasonable metric for the long-term budget or fairness constraints. However, it is improper for safety-critical applications because we may have a sequence of decisions with zero soft constraint violation and violates the constraints at every round. For example, consider a sequence of decisions $\{g_t(x_t)\}$ such that $g_t(x_t) = -1$ if t is odd and $g_t(x_t) = +1$ if t is even. For such a sequence with $T = 1000$, we have $\sum_{t=1}^{\tau} g_t(x_t) \leqslant 0$ for any $1 \leqslant \tau \leqslant T$, but the constraint violates at half of T rounds.

In this paper, we focus on stochastic continuum-armed bandit with constraints (SCBwC) via the Gaussian processes model and study the cumulative/hard constraint violation $\sum_{t=1}^{T}(g(x_t))^+$. The cumulative violation (or hard violation) is a strictly stronger metric than the soft violation because it cannot be compensated among different rounds. Our goal is to optimize a black-box reward function while keeping the cumulative violation minimal. In this paper, we propose a Rectified Pessimistic-Optimistic Learning (RPOL), an efficient penalty-based framework integrating optimistic and pessimistic estimators of reward and constraint functions into a single surrogate function. The framework acquires the information of block-box reward and constraint functions efficiently and safely, and it is flexible to achieve strong performance in SCBwC and its variants (bandits with delayed feedback). It is worth to be emphasizing that a concurrent work Xu et al. (2022) also considers the cumulative violation. However, it requires solving a complex constrained optimization problem for each round that might suffer from high computational complexity, and it is not clear if their method can be applied to bandits with delayed feedback as in our paper. Moreover, our experiments show RPOL outperforms their method w.r.t. both reward and constraint violation.

1.1. Main Contribution

Algorithm Design This paper proposes a rectified pessimistic-optimistic learning framework (RPOL) for SCBwC, where the rectified design is to avoid aggressive exploration and encourages conservative/pessimistic decisions such that it can minimize the cumulative constraint violation. The proposed framework is flexible to incorporate the classical exploration strategies in Gaussian process bandit learning (e.g., GP-UCB in Srinivas et al. (2009) or improved GP-UCB in Chowdhury and Gopalan (2017)) and provides the strong performance guarantee in regret and cumulative violation. Moreover, our framework is also readily applied to the variants of SCBwC, (e.g., bandits with delayed feedback in Section 5).

Reference	Regret	Soft Violation	Hard Violation	Design Method
Zhou and Ji (2022)	$O(\gamma_T\sqrt{T})$	$O(\gamma_T\sqrt{T}/\chi)$	N/A	Primal-dual
Xu et al. (2022)	$O(\gamma_T\sqrt{T})$	$O(\gamma_T\sqrt{T})$	$O(\gamma_T\sqrt{T})$	Constrained optimization
RPOL-UCB	$O(\gamma_T\sqrt{T})$	$O(\gamma_T\sqrt{T})$	$O(\gamma_T\sqrt{T})$	Penalty

Table 1: Our results and related work in SCBwC, where χ is a constant related to Slater's condition of the offline problem in (4) and requires to be known in Zhou and Ji (2022). Zhou and Ji (2022) and this paper can be regarded as unconstrained optimization methods, and Xu et al. (2022) is a constrained optimization-based method.

Reference	Regret	Hard Violation
Verma et al. (2022)	$O\left(\frac{\gamma_T}{\rho_m}(\sqrt{T}+m)\right)$	N/A
RPOL-CensoredUCB	$O\left(\frac{\gamma_T}{\rho_m}(\sqrt{T}+m)\right)$	$O\left(\frac{\gamma_T}{\rho_m}(\sqrt{T}+m)\right)$

Table 2: Our results and related work in SCBwC under delayed feedback.

Theoretical Results We develop a unified analysis method for RPOL framework in Theorem 1, where the regret and cumulative violation depend on the errors of optimistic or pessimistic learning. The method is quite general to be used in analyzing SCBwC and its variants, and we establish the following theoretical results (γ_T is the information gain w.r.t. the kernel used to approximate reward and constraint functions via GPs).

- For SCBwC, we instantiate RPOL with GP-UCB (RPOL-UCB) and prove it achieves $O\left(\gamma_T\sqrt{T}\right)$ regret and cumulative constraint violation. RPOL-UCB strictly improves Zhou and Ji (2022) as shown in Table 1 and achieves similar performance with an efficient penalty-based method compared to the concurrent work Xu et al. (2022), a constrained optimization-based method.
- For SCBwC with delayed feedback, we integrate RPOL with censored GP-UCB (RPOL-CensoredUCB) and show it achieves $O\left(\frac{\gamma_T}{\rho_m}(\sqrt{T}+m)\right)$ regret and cumulative violation, where m and ρ_m are the parameters related to the delay, as shown in Table 2. To the best of our knowledge, this is the first result in SCBwC with delayed feedback.

2. Problem Formulation

We study a stochastic continuum-armed bandit with constraints, where the arms/decisions are in a continuous space $\mathcal{X} \in \mathbb{R}^d$. The reward function $f : \mathcal{X} \to \mathbb{R}$ and constraint function $g : \mathcal{X} \to \mathbb{R}$ are continuous functions of the arms/decisions. Both f and g are black-box to the learner, and the learner acquires their knowledge sequentially. At each round $t \in [T]$, the learner makes decision $x_t \in \mathcal{X}$ and then observes the noisy reward and cost

$$r_t = f(x_t) + \delta_t,$$
$$c_t = g(x_t) + \varepsilon_t,$$

where noise δ_t and ε_t are random variables with zero-mean. Note r_t and c_t are bandit feedback because the leaner only observes the (noisy) version of $f(\cdot)$ and $g(\cdot)$ at x_t. Since f and g are unknown

apriori (possibly complicated and non-convex) and $\mathcal{X}$ is a continuous set with an infinity cardinality, it is infeasible in general to achieve the global optimal solution for arbitrary reward and constraint functions. We imposed the regularity assumption that $f(\cdot)$ and $g(\cdot)$ are within Reproducing Kernel Hilbert Space (RKHS). The assumption implies that a well-behaved continuous function can be represented with a properly chosen kernel function Srinivas et al. (2009) and we can model reward and constraint functions via Gaussian processes as introduced below.

Gaussian process model for f and g functions Gaussian process (GP) is a random process including a collection of random variables that follows a joint Gaussian distribution. Gaussian process $\text{GP}(\mu(x), k(x,x'))$ over $\mathcal{X}$ is specified by its mean $\mu(x)$ and covariance $k(x,x')$. For the reward function $f(x)$, we have $\text{GP}(\mu^f(x), k^f(x,x'))$ such that $\mu^f(x) = \mathbb{E}[f(x)]$ and $k^f(x,x') = \mathbb{E}[(f(x)-\mu^f(x))(f(x')-\mu^f(x'))]$. Let $\mathcal{A}_t = \{x_1, \cdots, x_{t-1}\}$ be the collection of decisions and $\{r_1, \cdots, r_{t-1}\}$ be the collection of noisy feedback until round t, respectively. The posterior distribution $\text{GP}(\mu_t^f(\cdot), k_t^f(\cdot,\cdot))$ updates at the beginning of round t

$$\mu_t^f(x) = k_t^f(x)^T (V_t^f(\lambda))^{-1} r_{1:t} \tag{1}$$

$$k_t^f(x,x') = k^f(x,x') - k_t^f(x)^T (V_t^f(\lambda))^{-1} k_t^f(x'), \tag{2}$$

$$\sigma_t^f(x) = \sqrt{k_t^f(x,x)}, \tag{3}$$

where $K_t^f := [k^f(x,x')]_{x,x' \in \{x_1,\cdots,x_{t-1}\}}$, $V_t^f(\lambda) := K_t^f + \lambda I$, $\lambda = 1 + 2/T$, $r_{1:t} = [r_1, \cdots, r_{t-1}]$, and $k_t^f(x) := [k^f(x_1,x), \cdots, k^f(x_{t-1},x)]^T$. Similarly, we define a GP model for the constraint function g to be $\mathcal{GP}(\mu_t^g(x), k_t^g(x,x'))$ with the mean $\mu_t^g(x)$, covariance $k_t^g(x,x')$ and standard deviation $\sigma^g(x)$. The model for g updates the same as in (1)-(3). The kernel function is designed by choice and one popular kernel is the square exponential (SE) kernel $k_{\text{SE}}(x,x') = e^{\frac{-\|x-x'\|^2}{2u^2}}$, where $u > 0$ is a positive hyper-parameter. We consider the SE kernel function in this paper and also use it in our experiments in Section 6.

Further, we define the information gain at round t to be $\gamma_t^f := \max_{\mathcal{A}_t \in \mathcal{X}: |\mathcal{A}_t| = t-1} \frac{1}{2} \ln |I + \lambda^{-1} K_t^f|$ and $\gamma_t^g := \max_{\mathcal{A}_t \in \mathcal{X}: |\mathcal{A}_t| = t-1} \frac{1}{2} \ln |I + \lambda^{-1} K_t^g|$, which are important parameters in GP bandits. They depend on the choice of the kernel function and the domain $\mathcal{X}$, and would play a key role in our following regret and violation analysis. For SE kernel function, we have γ_t^f and γ_t^g are $O((\ln(t))^{d+1})$ if $\mathcal{X}$ is compact and convex with dimension d.

Regret and cumulative constraint violation Given the complete knowledge of f and g, we define the following offline optimization problem

$$\max_{x \in \mathcal{X}} \ f(x) \ \text{ s.t. } \ g(x) \leqslant 0. \tag{4}$$

Let x^* be the global optimal solution to (4). We define the regret and cumulative constraint violation

$$\mathcal{R}(T) := \sum_{t=1}^{T} f(x^*) - \sum_{t=1}^{T} f(x_t), \tag{5}$$

$$\mathcal{V}(T) := \sum_{t=1}^{T} g^+(x_t). \tag{6}$$

The goal of the leaner is to develop algorithms to achieve sublinear regret and violation, i.e., $\lim_{T\to\infty} \mathcal{R}(T)/T = 0$ and $\lim_{T\to\infty} \mathcal{V}(T)/T = 0$ when f and g are modeled via Gaussian processes.

3. Rectified Pessimistic-Optimistic Learning Framework

In this section, we propose a general decision framework to tackle SCBwC with the metric of cumulative violation, called rectified pessimistic-optimistic learning framework (RPOL). The framework learns the reward function optimistically $\hat{f}_t(x)$ and the constraint function pessimistically $\check{g}_t(x)$ by a learning strategy $\mathcal{M}$ based on the model/parameters (Θ_t^f, Θ_t^g). For example, the learning strategy could be the upper confidence bound learning of Gaussian process (GP-UCB), where Θ_t^f and Θ_t^g can include $(\mu_t^f, \sigma_t^f, k_t^f)$ and $(\mu_t^g, \sigma_t^g, k_t^g)$, respectively. By imposing the rectified operator on the constraint $\check{g}_t^+(x)$, RPOL chooses the best decision to maximize a rectified surrogate function $\hat{f}_t(x) - Q_t \check{g}_t^+(x)$ in (7). After observing the noisy (possibly delayed) bandit feedback (reward and cost), we update the rectified penalty factor Q_{t+1} and the model $(\Theta_{t+1}^f, \Theta_{t+1}^g)$, according to the learning strategy $\mathcal{M}$.

RPOL Framework for SCBwC

Initialization: $Q_1 = 1$ and $\eta_t = \sqrt{t}$. Model Θ_1^f and Θ_1^g.
For $t = 1, \cdots, T$,

- **Pessimistic-optimistic learning:** estimate the reward function $\hat{f}_t(x)$ and the cost function $\check{g}_t(x)$ according to a learning strategy $\mathcal{M}$ with (Θ_t^f, Θ_t^g).
- **Rectified penalty-based decision:** choose x_t such that

$$x_t = \arg\max_{x \in \mathcal{X}} \ \hat{f}_t(x) - Q_t \check{g}_t^+(x) \tag{7}$$

- **Feedback:** noisy reward $r_t(x_t)$ and cost $c_t(x_t)$.
- **Rectified cumulative penalty update:** $Q_{t+1} = \max\left(Q_t + c_t^+(x_t), \eta_t\right)$.
- **Model update:** $\Theta_{t+1}^f = \mathcal{M}(\Theta_t^f, \{x_t, r_t, c_t\})$, $\Theta_{t+1}^g = \mathcal{M}(\Theta_t^g, \{x_t, r_t, c_t\})$.

We explain the main intuition behind the RPOL framework. The Lagrange function of the offline baseline problem in (4) is defined to be

$$L(x, \vartheta) := f(x) - \vartheta g(x),$$

where ϑ is a dual variable related to the constraint in (4). Since the reward and cost functions are approximated via Gaussian Processes, we estimate $f(x)$ with $\hat{f}_t(x)$ optimistically and $g(x)$ with $\check{g}_t(x)$ pessimistically. We impose a rectified operator $\check{g}_t^+(x)$ to associate it with the hard violation $g_t^+(x)$ at round t. Moreover, we approximate ϑ with a "rectified" penalty factor Q_{t+1}, where we first rectify the cost $c_t(x_t)$ with $c_t^+(x_t)$ and add it to Q_t such that the penalty increases when the constraint violation occurs; and then we rectify Q_{t+1} with a minimum penalty price η_t. This design adaptively controls the penalty to prevent the aggressive decision for each round. The rectified decision in (7) and rectified penalty update in (3) are the key to minimize the cumulative constraint $\sum_{t=1}^T g_t^+(x)$.

The "rectified" idea in this paper is motivated by Guo et al. (2022a) in constrained online convex optimization. However, there exists a substantial difference due to the distinct feedback model:

Guo et al. (2022a) observes the full-information feedback, imposes the rectifier on the previous constraint function, and introduces the smooth term to stabilize the learning process; this paper considers bandit feedback, learns the black-box functions (pessimistically and optimistically) directly and imposes a rectifier on the pessimistic estimator of constraint function. The "rectified" design also distinguishes our framework from the classical primal-dual approach in Zhou and Ji (2022). The work in Zhou and Ji (2022) establishes the soft constraint violation (i.e., $\sum_{t=1}^{T} g(x_t)$) by studying the bound on the virtual queue/dual variable, which relies on the assumption of Slater's condition and the knowledge of slackness constant (the information is usually not available in practical applications). However, our framework establishes the cumulative violation (i.e., $\sum_{t=1}^{T} g^+(x_t)$) directly and does not require Slater's condition.

Before presenting theoretical results for the RPOL framework, we introduce the following two assumptions on reward function, constrained function, and noise.

Assumption 1 *Let $\|\cdot\|_k$ denote the RKHS norm associated with a kernel k. For the reward function f, we assume that $\|f\|_{k^f} \leqslant B_f$ and $k^f(x,x) \leqslant 1$ for any $x \in \mathcal{X}$. For the constraint function g, we assume $\|g\|_{k^g} \leqslant B_g$ and $k^g(x,x) \leqslant 1$ for any $x \in \mathcal{X}$.*

Assumption 2 *The noise δ_t is i.i.d. R_f-sub-Gaussian and the noise ε_t is i.i.d. R_g-sub-Gaussian.*

To establish a unified analysis method for SCBwC with the cumulative violation, we introduce a critical condition on the optimistic learning of reward function $\hat{f}$ and the pessimistic learning of the constraint function $\check{g}$, respectively.

Condition 1 *Let $\{e_t^f(p,x)\}$, $\{e_t^g(p,x)\}$, and ρ be non-negative values. We have for any $x \in \mathcal{X}$ and all $t \in [T]$*

$$0 \leqslant \hat{f}_t(x) - \rho f(x) \leqslant e_t^f(p,x),$$
$$0 \leqslant \rho g(x) - \check{g}_t(x) \leqslant e_t^g(p,x),$$

hold with probability $1-p$ with $p \in (0,1)$.

Intuitively, a good learning strategy $\mathcal{M}$ should satisfy Condition 1 with small learning errors $e_t^f(x)$ and $e_t^g(x)$. These errors play important roles in regret and cumulative violation in Theorem 1.

Theorem 1 *Let Assumptions 1 and 2 hold. Under Condition 1, RPOL framework have the following regret and constraint violation*

$$\mathcal{R}(T) \leqslant \frac{1}{\rho}\sum_{t=1}^{T} e_t^f(x_t), \quad \mathcal{V}(T) \leqslant \frac{1}{\rho}\sum_{t=1}^{T} e_t^g(x_t) + \sum_{t=1}^{T} e_t^f(x_t) + 4\rho B_f\sqrt{T},$$

hold with probability $1-p$ with $p \in (0,1)$.

Remark 2 *RPOL framework is flexible to incorporate the classical learning strategies in unconstrained GP bandit learning (e.g., GP-UCB/LCB) and achieves strong performance guarantee on regret and cumulative violation for SCBwC in Theorem 1. Moreover, RPOL framework can be readily combined with dedicated learning strategies for the variants of SCBwC and establish similar performance according to Theorem 1 as in the unconstrained counterparts.*

In the following sections, we instantiate the learning strategies $\mathcal{M}$ in RPOL for SCBwC (and its variants), and establish the theoretical results according to Theorem 1.

4. Rectified Pessimistic-Optimistic Learning for SCBwC

In this section, we instantiate improved GP-UCB/LCB Chowdhury and Gopalan (2017) into RPOL framework for estimating $\hat{f}(x)$ and $\check{g}(x)$, and establish a strong performance on regret and violation according to Theorem 1.

GP-UCB/LCB The optimistic estimator of $f(x)$ and the pessimistic estimator of $g(x)$ at round t are defined by

$$\hat{f}_t(x) = \mu_t^f(x) + \beta_t^f \sigma_t^f(x),$$
$$\check{g}_t(x) = \mu_t^g(x) - \beta_t^g \sigma_t^g(x),$$

which serves the upper confidence bound for the true $f(x)$ and the lower confidence bound for the true $g(x)$ by carefully choosing $\beta_t^f(p)$ and $\beta_t^g(p)$. We consider improved GP-UCB in Chowdhury and Gopalan (2017). Let $\beta_t^f(p) = B_f + R_f\sqrt{2(\gamma_t^f + 1 + \ln(2/p))}$ and $\beta_t^g(p) = B_g + R_g\sqrt{2(\gamma_t^g + 1 + \ln(2/p))}$ with $p \in (0,1)$. To streamline the notation in the remaining sections of this paper, we will employ β_t^f and β_t^g instead. The models/parameters in GP-UCB/LCB, including $(\mu_t^f(x), \sigma_t^f(x), \mu_t^g(x), \sigma_t^g(x))$, update according to (1)-(3). The complete description of RPOL with GP-UCB/LCB (RPOL-UCB) is provided in our technical report Guo et al. (2022b) due to the limited space.

To analyze RPOL-UCB by Theorem 1, we verify Condition 1 and quantify the cumulative errors for GP-UCB/LCB in Lemmas 3 and 4, respectively. The detailed proofs are in Guo et al. (2022b).

Lemma 3 *Under Assumptions 1 and 2, the following inequalities hold for any $x \in \mathcal{X}$ and all $t \in [T]$ under RPOL-UCB*

$$0 \leqslant \hat{f}_t(x) - f(x) \leqslant 2\beta_t^f \sigma_t^f(x),$$
$$0 \leqslant g(x) - \check{g}_t(x) \leqslant 2\beta_t^g \sigma_t^g(x),$$

with probability at least $1 - p$ with $p \in (0,1)$.

Lemma 4 *Let $\{x_1, \cdots, x_T\}$ be the collection of decisions chosen by the algorithm. The cumulative standard deviation can be bounded as follows:*

$$\sum_{t=1}^{T} \beta_t^f \sigma_t^f(x_t) \leqslant \beta_T^f \sqrt{4(T+2)\gamma_T^f},$$
$$\sum_{t=1}^{T} \beta_t^g \sigma_t^g(x_t) \leqslant \beta_T^g \sqrt{4(T+2)\gamma_T^g}.$$

Based on Lemmas 3 and 4, we invoke Theorem 1 to establish the regret and violation of RPOL-UCB in Theorem 5.

Theorem 5 *RPOL-UCB achieves the following regret and constraint violation with a probability at least $1 - p$:*

$$\mathcal{R}(T) = O(\gamma_T \sqrt{T}),$$
$$\mathcal{V}(T) = O(\gamma_T \sqrt{T}),$$

where $\gamma_T = \max(\gamma_T^f, \gamma_T^g)$.

RPOL-UCB achieves a strictly stronger notation of cumulative violation compared to the soft violation in Shi and Eryilmaz (2022); Zhou and Ji (2022) and a similar performance compared to Xu et al. (2022) but with an efficient penalty approach. With the rectified design, RPOL quantifies the cumulative violation directly, which is different from the primal-dual optimization in Zhou and Ji (2022) or the penalty-based technique in Shi and Eryilmaz (2022); Lu and Paulson (2022).

5. RPOL for SCBwC with Delayed Feedback

In the previous section, we assume rewards feedback and costs/constraints feedback are available to the learner immediately. However, it might not happen in many real-world applications such as recommendation systems, clinical trials, and hyper-parameter tuning in machine learning, where the feedback is revealed to the learner after a random delay. Therefore, it motivates us to study SCBwC with stochastic delayed feedback.

At each round $t \in [T]$, the learner makes decision $x_t \in \mathcal{X}$ and observes the feedback $r_t = f(x_t) + \delta_t,\ \ c_t = g(x_t) + \varepsilon_t$ after stochastic delay d_t^f and d_t^g, respectively. We assume the delay d_t^f and d_t^g are independent and generated from an unknown distribution $\mathcal{D}$. To tackle the delayed feedback, we introduce the idea of censored feedback as in Vernade et al. (2020); Verma et al. (2022). The delayed feedback is censored by indicator functions $\mathbb{I}\{d_s^f \leqslant \min(m, t-s)\}$ and $\mathbb{I}\{d_s^g \leqslant \min(m, t-s)\}$, which indicate if reward or cost at round s are revealed by round t and the delay is within m rounds. We define the censored feedback at round s by $\tilde{r}_{s,t} := r_s \mathbb{I}\{d_s^f \leqslant \min(m, t-s)\}$ and $\tilde{c}_{s,t} := c_s \mathbb{I}\{d_s^g \leqslant \min(m, t-s)\}$ and the sequence of censored feedback by $\tilde{r}_{1:t} = [\tilde{r}_{1,t-1}, \cdots, \tilde{r}_{t-1,t-1}]^T$; and $\tilde{c}_{1:t} = [\tilde{c}_{1,t-1}, \cdots, \tilde{c}_{t-1,t-1}]^T$. We further define $\rho_m^f = \mathbb{P}\{d_s^f \leqslant m\}$ and $\rho_m^g = \mathbb{P}\{d_s^g \leqslant m\}$, which denote the probabilities of observing delayed reward feedback and cost feedback within m rounds, respectively.

Censored GP-UCB/LCB We utilize the censored feedback $\tilde{r}_{1:t}$ (instead of $r_{1:t}$ in the previous section) when estimating the reward and constraint function

$$\mu_t^f := k_t^f(x)^T (K_t^f + \lambda I)^{-1} \tilde{r}_{1:t},$$
$$\mu_t^g := k_t^f(x)^T (K_t^g + \lambda I)^{-1} \tilde{c}_{1:t}.$$

The kernel matrix and variance update exactly the same as in (2) and (3). Therefore, the optimistic and pessimistic estimators of $f(x)$ and $g(x)$ at round t are

$$\hat{f}_t(x) = \mu_t^f(x) + v_t^f \sigma_t^f(x),$$
$$\check{g}_t(x) = \mu_t^g(x) - v_t^g \sigma_t^g(x),$$

where $v_t^f = B_r \sum_{s=t-m}^{t-1} \sigma_t^f(x_s) + \beta_t^f$, $v_t^g = B_c \sum_{s=t-m}^{t-1} \sigma_t^g(x_s) + \beta_t^g$ with $B_r = B_f + R_f\sqrt{2\log T}$ and $B_c = B_g + R_g\sqrt{2\log T}$ denoting bounds for observations r_t and c_t with the probability at least $1 - 2/T$ according Assumption 2. Let $\beta_t^f = B_f + (R_f + B_r)\sqrt{2(\gamma_t^f + 1 + \ln(4/p))}$ and $\beta_t^g = B_g + (R_g + B_c)\sqrt{2(\gamma_t^g + 1 + \ln(4/p))}$, where $p \in (0, 1)$. We instantiate RPOL with Censored GP-UCB/LCB (RPOL-CensoredUCB) and defer the complete description in Guo et al. (2022b).

Similar to Section 4, we verify Condition 1 and quantify the cumulative errors for censored GP-UCB/LCB, and then invoke Theorem 1 to establish the following theorem.

Theorem 6 *RPOL with censored GP-UCB achieves the following regret and constraint violation with probability at least* $1 - p - 2/T$ *with* $p \in (0, 1 - 2/T)$*:*

$$\mathcal{R}(T) = O\left(\frac{\gamma_T}{\rho_m}(\sqrt{T} + m) + m\gamma_T\right),$$
$$\mathcal{V}(T) = O\left(\frac{\gamma_T}{\rho_m}(\sqrt{T} + m) + m\gamma_T\right),$$

where $\gamma_T = \max(\gamma_T^f, \gamma_T^g)$ *and* $\rho_m = \min(\rho_m^f, \rho_m^g)$.

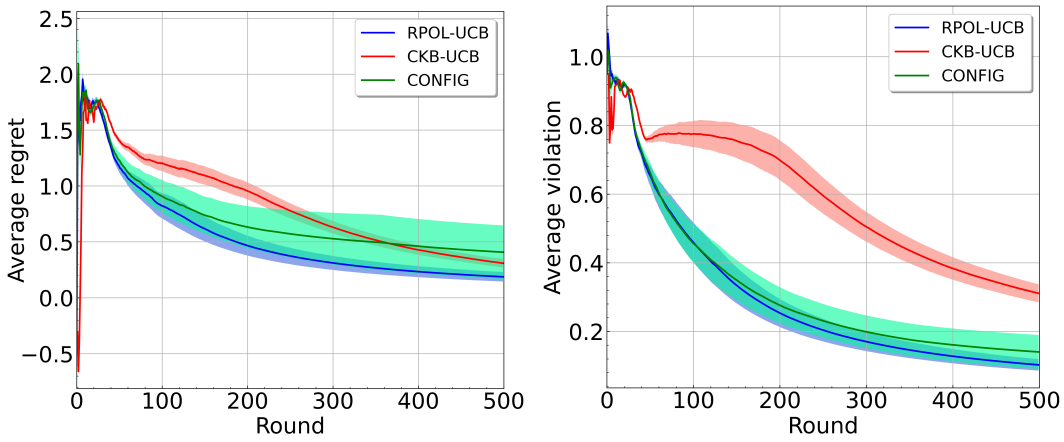
Theorem 6 shows that RPOL-CensoredUCB achieves sub-linear bounds for the regret and violation simultaneously in SCBwC with delayed feedback. The result matches the regret bound for unconstrained counterparts with delayed feedback in Verma et al. (2022).

6. Experiments

In this section, we thoroughly evaluate the efficacy of the RPOL framework by conducting a series of numerical experiments. To provide a comprehensive analysis, we compare our proposed algorithms against the currently available baseline methods in both classical environment and delayed environment. We generate graphical representations that display both the average regret and cumulative violation, denoted as $\mathcal{R}(t)/t$ and $\mathcal{V}(t)/t$ respectively. To ensure the reliability and accuracy of our findings, we calculate the results by averaging across 100 individual trials and present them with a 95% confidence interval for added precision.

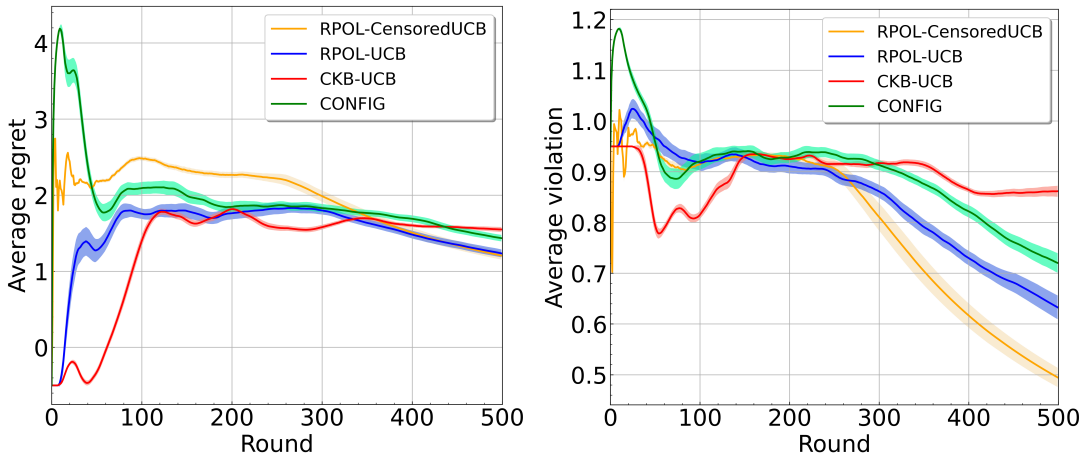
Classical SCBwC We consider the reward function $f(x) = -\sin x(1) - x(2)$ and the constraint function $g(x) = \sin x(1) \sin x(2) + 0.95$, and let $\mathcal{X} = \{x | 0 \geqslant x_1 \geqslant 6,\ 0 \geqslant x_2 \geqslant 6\}$ The constraint set $\{x \mid g(x) \leqslant 0\}$ indicates a strict region and makes the problem challenging. The observations are corrupted with Gaussian noise sampled from $\mathcal{N}(0, 0.05)$, respectively. We test RPOL-UCB and consider the baselines: CKB-UCB in Zhou and Ji (2022) and CONFIG in Xu et al. (2022). From Figure 1, we show RPOL-UCB achieves the best performance w.r.t. both regret and cumulative violation in SCBwC, where it converges to a low cumulative violation in a faster rate. The results in Figure 1 justify that our rectified design can balance the regret and cumulative violation efficiently and safely, and it is superior to handling the strict cumulative violation.

SCBwC with Delayed Feedback In this experiment, we extend the classical SCBwC problem to incorporate stochastic delayed feedback. While retaining the reward function and constraint function from the Classical SCBwC setting, we introduced stochastic delays d_t^f and d_t^g at each time slot. The delays of d_t^f and d_t^g at round t are sampled from a Poisson distribution with a mean of 15, respectively. We test the performance of RPOL-CensoredUCB and compare it to the baselines RPOL-UCB, CKB-UCB from Zhou and Ji (2022), and CONFIG from Xu et al. (2022). As shown in Figure 2, RPOL-CensoredUCB outperforms all existing baselines in terms of both regret and cumulative violation. Notably, RPOL-UCB also demonstrates strong performance compared with other baselines. These results suggest that the RPOL framework can effectively balance regret and cumulative violation, even in the presence of stochastic delayed feedback, providing a robust performance guarantee.



(*a*) Average Regret (*b*) Average Violation

Figure 1: Average Regret ($\mathcal{R}(t)/t$) and Cumulative Violation ($\mathcal{V}(t)/t$) in SCBwC.



(*a*) Average Regret (*b*) Average Violation

Figure 2: Average Regret ($\mathcal{R}(t)/t$) and Cumulative Violation ($\mathcal{V}(t)/t$) in SCBwC with delayed feedback.

7. Conclusion

In this paper, we study constrained GP bandit optimization with the cumulative constraint violation. We propose the RPOL framework and show it is flexible to be applied in variants of constrained GP bandits by incorporating the dedicated exploration techniques. Our theoretical and experimental results justify the superior of the RPOL framework.

Acknowledgment

This work was supported in part by Shanghai Sailing Program under Grant 22YF1428500.

References

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.

Setareh Ariafar, Jaume Coll-Font, Dana H Brooks, and Jennifer G Dy. Admmbo: Bayesian optimization with unknown constraints using admm. *The Journal of Machine Learning Research*, 20 (123):1–26, 2019.

Felix Berkenkamp, Andreas Krause, and Angela P Schoellig. Bayesian optimization with safety constraints: safe and automatic parameter tuning in robotics. *Machine Learning*, pages 1–35, 2021.

Sayak Ray Chowdhury and Aditya Gopalan. On kernelized multi-armed bandits. In *International Conference on Machine Learning*, pages 844–853. PMLR, 2017.

Audrey Durand, Charis Achilleos, Demetris Iacovides, Katerina Strati, Georgios D. Mitsis, and Joelle Pineau. Contextual bandits for adapting treatment in a mouse model of de novo carcinogenesis. In *Proceedings of the 3rd Machine Learning for Healthcare Conference*, volume 85 of *Proceedings of Machine Learning Research*, pages 67–82. PMLR, 17–18 Aug 2018.

Xinzhe Fu and Eytan Modiano. Learning-num: Network utility maximization with unknown utility functions and queueing delay. MobiHoc '21, New York, NY, USA, 2021. Association for Computing Machinery.

Hengquan Guo, Xin Liu, Honghao Wei, and Lei Ying. Online convex optimization with hard constraints: Towards the best of two worlds and beyond. In *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022a.

Hengquan Guo, Qi Zhu, and Xin Liu. Rectified pessimistic-optimistic learning for stochastic continuum-armed bandit with constraints. *arXiv preprint arXiv:2211.14720*, 2022b.

Andreas Krause and Cheng Ong. Contextual gaussian process bandit optimization. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.

Congwen Lu and Joel A. Paulson. No-regret bayesian optimization with unknown equality and inequality constraints using exact penalty functions. *IFAC-PapersOnLine*, 2022. 13th IFAC Symposium on Dynamics and Control of Process Systems.

Zai Shi and Atilla Eryilmaz. A bayesian approach for stochastic continuum-armed bandit with long-term constraints. In *International Conference on Artificial Intelligence and Statistics*, pages 8370–8391. PMLR, 2022.

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.

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.

Yanan Sui, Vincent Zhuang, Joel Burdick, and Yisong Yue. Stagewise safe bayesian optimization with gaussian processes. In *International conference on machine learning*, pages 4781–4789. PMLR, 2018.

Arun Verma, Zhongxiang Dai, and Bryan Kian Hsiang Low. Bayesian optimization under stochastic delayed feedback. In *International Conference on Machine Learning*, pages 22145–22167. PMLR, 2022.

Claire Vernade, Alexandra Carpentier, Tor Lattimore, Giovanni Zappella, Beyza Ermis, and Michael Brückner. Linear bandits with stochastic delayed feedback. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9712–9721. PMLR, 13–18 Jul 2020.

Colin White, Willie Neiswanger, and Yash Savani. Bananas: Bayesian optimization with neural architectures for neural architecture search. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35:10293–10301, May 2021.

Wenjie Xu, Yuning Jiang, and Colin N Jones. Constrained efficient global optimization of expensive black-box functions. *arXiv preprint arXiv:2211.00162*, 2022.

Xingyu Zhou and Bo Ji. On kernelized multi-armed bandits with constraints. In *Thirty-Sixth Conference on Neural Information Processing Systems*, 2022.