Stochastic Constrained Contextual Bandits via Lyapunov Optimization Based Estimation to Decision Framework

Hengquan Guo GUOHQ@SHANGHAITECH.EDU.CN
ShanghaiTech University

Xin Liu* LIUXIN7@SHANGHAITECH.EDU.CN
ShanghaiTech University

Editors: Shipra Agrawal and Aaron Roth

Abstract

This paper studies the problem of stochastic constrained contextual bandits (CCB) under general realizability condition where the expected rewards and costs are within general function classes. We propose LOE2D, a Lyapunov Optimization Based Estimation to Decision framework with online regression oracles for learning reward/constraint. LOE2D establishes $\tilde{O}(T^{\frac{3}{4}}U^{\frac{1}{4}})$ regret and constraint violation, which can be further refined to $\tilde{O}(\min\{\sqrt{TU}/\varepsilon^2, T^{\frac{3}{4}}U^{\frac{1}{4}}\})$ when the Slater condition holds in the underlying offline problem with the Slater "constant" $\varepsilon = \Omega(\sqrt{U/T})$, where U denotes the error bounds of online regression oracles. These results improve Slivkins et al. (2023) in two aspects: i) our results hold without any prior information while Slivkins et al. (2023) requires the knowledge of Slater constant to design a proper learning rate; ii) our results hold when $\varepsilon = \Omega(\sqrt{U/T})$ while Slivkins et al. (2023) requires a constant margin $\varepsilon = \Omega(1)$. These improvements stem from two novel techniques: violation-adaptive learning in E2D module and multi-step Lyapunov drift analysis in bounding constraint violation. The experiments further justify LOE2D outperforms the baseline algorithm.

1. Introduction

Stochastic contextual bandits (CB) is a general online learning framework for interactive decision-making in uncertain environments. It has been boasting many practical applications, including recommender systems Li et al. (2010), task scheduling in crowdsourcing Tran-Thanh et al. (2014), and clinical trials Tewari and Murphy (2017). Specifically, in a contextual bandit problem, the learner, upon observing a context x_t in period t, takes an action a_t from the decision set and then receives a stochastic reward r_t. The objective is to maximize the expected cumulative rewards $\mathbb{E}[\sum_{t=1}^{T} r_t(a_t)]$. However, many real-world applications require to take operational constraints into account. For instance, an online advertising system optimizes items display to improve click-through rates while conforming to weekly or monthly budget limits; a crowdsourcing system maximizes its utility via efficient task scheduling while satisfying fairness and laboring constraints; in clinical trials, it's crucial to design effective treatment plans according to patient conditions while adhering to scary medical resources. For such applications, stochastic constrained contextual bandits (CCB) is a more appropriate model.

Bandits with knapsacks (BwK) is a specialized class of CCB and has been extensively studied Agrawal and Devanur (2014, 2016); Badanidiyuru et al. (2014, 2018), where the interaction

* Corresponding author

terminates once the budget is exhausted. By assuming linear rewards and costs (i.e., linear realizability assumption), the contextual version of BwK has been explored in Agrawal and Devanur (2016); Badanidiyuru et al. (2018) where near-optimal regret bounds are achieved. Other categories of CCB include fairness bandits where fairness is defined as a minimum rate at which an arm is pulled Li et al. (2019); Xu et al. (2020); Patil et al. (2021) and conservative bandits Wu et al. (2016); Kazerouni et al. (2017); Garcelon et al. (2020), where conservatism is defined as a minimum requirement for anytime cumulative rewards. While most existing studies concentrate on concrete constraints and specific realizability assumptions (such as the linear class), the exploration of CCB under more general functional and constraint settings has received less attention. Han et al. (2023) studied stochastic contextual bandits with knapsacks (CBwK) under general realizability assumption, where a primal-dual framework with online regression oracles for learning rewards and costs is proposed to achieve a vanishing regret. However, the proposed algorithms are dedicated to knapsack constraints (all costs are non-negative) and it is unclear if their framework is applicable into other types of constraints. Slivkins et al. (2023) studies contextual bandits with more general constraints in the form that $\sum_{t=1}^{T} c_t(a_t) - B \leq 0$, where the costs $\{c_t(a_t)\}$ could be either positive or negative, including both packing and covering constraints. The paper also developed a primal-dual oracle-based framework similar to Han et al. (2023) and achieved a regret and constraint violation of $\tilde{O}(T^{\frac{3}{4}}U^{\frac{1}{4}})$ in a general setting, where U is the estimation error of online regression oracles. When the additional assumption of the Slater condition holds, the regret and violation bound can be improved to $\tilde{O}(\sqrt{TU})$ by tuning a key trade-off factor with the Slater constant information. However, these results in Slivkins et al. (2023) are only meaningful in the regime of "large budgets" with $B = \Omega(T)$ and a constant feasibility margin $\varepsilon = \Omega(1)$ (resemble Slater's constant). Moreover, their algorithm requires prior knowledge of the feasibility margin ε. In real-world scenarios, the system may operate in the regime of "small budget" $B = o(T)$ with a vanishing margin $\varepsilon = o(1)$, and obtaining such "margin" information can be quite challenging (if not impossible). Therefore, an open question remains:

Is there a single algorithm capable of achieving optimal performance in stochastic constrained contextual bandits without any prior knowledge, regardless of the feasibility assumption or a potentially vanishing feasibility margin?

We provide a positive answer to this question by introducing LOE2D, a Lyapunov Optimization Based Estimation to Decision. Our contributions can be summarized as follows:

- **Algorithm Design:** The design of LOE2D is motivated by the primal-dual approach based on online regression oracles in Han et al. (2023); Slivkins et al. (2023). However, we introduce novel design and perspective through Lyapunov optimization. The (primal) decision modular builds on a regression-based method for contextual bandits and inverse-gap weighting technique in Foster and Rakhlin (2020) by taking constraints into account and a new violation-adaptive exploration strategy. The decision modular can be interpreted to minimize "Regret + Lyapunov/potential drift". The dual modular includes a careful design of virtual queue (resemble scaled dual variable) that updates in a gradient-descent manner to keep track of cumulative constraint violation in each round. Note unlike the previous approaches Han et al. (2023); Slivkins et al. (2023), we do not impose any upper bound on the virtual queue. The violation-adaptive design in both modules are crucial to minimize regret and constraint violation simultaneously and establish the strong theoretical performance.

Table 1: Our results and two most related work. LOE2D achieves the universal results across three scenarios, while LagrangeCBwLC in Slivkins et al. (2023) requires customizing trade-off factors for different CCB instances (with ε information) under strict assumption, thus we present them separately. For CBwK, LOE2D is slightly looses "$1/\varepsilon$" against SquareCBwK in Han et al. (2023) and LagrangeCBWLC in Slivkins et al. (2023). However, SquareCBwK requires a dedicated learning module to learn the optimal value and set a proper learning rate, and LagrangeCBWLC requires a large budget $B = \Omega(T)$.

ALGORITHMS	CCB REGRET&VIO	CCB (SLATER CONDITION) REGRET&VIO	CBWK REGRET
LOE2D	$\tilde{O}(T^{\frac{3}{4}}U^{\frac{1}{4}})$	$\tilde{O}(\min\{\frac{\sqrt{TU}}{\varepsilon^2}, T^{\frac{3}{4}}U^{\frac{1}{4}}\})$	$\tilde{O}(\min\{\frac{\sqrt{TU}}{\varepsilon^2}, T^{\frac{3}{4}}U^{\frac{1}{4}}\})$
LAGRANGECBWLC (GENERAL SETTING)	$\tilde{O}(T^{\frac{3}{4}}U^{\frac{1}{4}})$	$\tilde{O}(T^{\frac{3}{4}}U^{\frac{1}{4}})$	$\times$
LAGRANGECBWLC (CONSTANT ε & ITS PRIOR KNOWLEDGE)	$\times$	$\tilde{O}(\sqrt{TU})$	$\times$
LAGRANGECBWLC (TAILORED LEARNING RATE)	$\times$	$\times$	$\tilde{O}(\frac{\sqrt{TU}}{\varepsilon})$
SQUARECBWK	$\times$	$\times$	$\tilde{O}(\frac{\sqrt{TU}}{\varepsilon})$

- **Theoretical Analysis:** LOE2D achieves both regret and violation within $\tilde{O}(T^{\frac{3}{4}}U^{\frac{1}{4}})$ under the general realizability assumption of reward and constraint functions. When an additional relaxed Slater's condition holds, LOE2D can guarantee $\tilde{O}(\min\{\sqrt{TU}/\varepsilon^2, T^{\frac{3}{4}}U^{\frac{1}{4}}\})$, interpolating from $\tilde{O}(\sqrt{TU})$ to $\tilde{O}(T^{\frac{3}{4}}U^{\frac{1}{4}})$ depending on the feasibility margin ε. Our result is more general than that in Slivkins et al. (2023), where $\tilde{O}(\sqrt{TU})$ is established assuming a constant ε. Moreover, unlike LagrangeCBwLC algorithm in Slivkins et al. (2023), which requires the knowledge of the Slater constant ε to determine the learning rate, LOE2D achieves these results without any prior information. Besides, LOE2D can be applied into CBwK without any modification and achieve a similar result in Han et al. (2023). The detailed comparisons are summarized in Table 1. We establish these strong results through a novel perspective and analysis of the violation/virtual queue process. Specifically, we view the virtual queue as a Markovian process and leverage a multiple-step Lyapunov drift analysis to establish its high probability upper bound. These techniques can be independent of interests and potentially applied to other constrained online learning scenarios.
- **Experiments:** We evaluate LOE2D using classification and learning-to-rank datasets with two different online regression oracles (linear regression and boosted regression trees). Our experimental results demonstrate that LOE2D significantly outperforms LagrangeCBwLC algorithm in Slivkins et al. (2023) by achieving a larger reward and smaller constraint violation.

Related works

Contextual Bandits: Multi-armed Bandits (MAB) is a classical online decision-making framework Auer et al. (2002); Bubeck et al. (2012); Lattimore and Szepesvári (2020), where the learner aims

to maximize the cumulative rewards in uncertain environments. Contextual bandits (CB) is a generalization of MAB where the contextual information is available when making decisions. There exist extensive studies on CB under linear realizability assumption Rusmevichientong and Tsitsiklis (2010); Abbasi-Yadkori et al. (2011b); Li et al. (2010); Abeille and Lazaric (2017); Agrawal and Goyal (2013). The classical exploration techniques such as Upper Confidence Bound (UCB) Abbasi-Yadkori et al. (2011a), Thompson sampling Chapelle and Li (2011), and randomized exploration Vaswani et al. (2020), have been proposed to design efficient algorithms for the contextual bandit. To relax the strict realizability assumption, CB with classification oracles was discussed in Dudik et al. (2011); Agarwal et al. (2014) where the algorithms assume access to cost-sensitive classification oracles. While classification oracles improve the computational efficiency, such classification might be still intractable for even basic hypothesis classes Klivans and Sherstov (2009). CB with regression oracles developed in Foster et al. (2018); Foster and Rakhlin (2020); Simchi-Levi and Xu (2022) are more computationally efficient compared to that with classification oracles and favorable for practical implementation.

Constrained Contextual Bandits: Constrained contextual bandits (CCB) includes various bandit scenarios such as bandits with knapsacks (BwK) Badanidiyuru et al. (2014); Agrawal and Devanur (2014); Wu et al. (2015); Agrawal and Devanur (2016); Badanidiyuru et al. (2018); Sivakumar et al. (2022), where the interaction is terminated when the budget is exhausted, fairness bandits Li et al. (2019); Xu et al. (2020); Patil et al. (2021), where each arm is required to be pulled at least a predefined times, conservative bandits Wu et al. (2016); Kazerouni et al. (2017); Garcelon et al. (2020), where learner should ensure that the cumulative rewards are not below a threshold induced by a baseline algorithm. It is also worth mentioning that another class of CCB that imposes stage-wise or anytime constraints in Amani et al. (2019); Moradipari et al. (2021); Pacchiano et al. (2021). The objective is to maximize the cumulative rewards while ensuring the constraints are satisfied in the expectation or high probability sense for each round. The proposed algorithms require solving a complicated constrained optimization for each time, suffering from high computational complexity.

2. Problem Formulation

In this section, we introduce the problem formulation and performance metric for the stochastic constrained contextual bandits.

Stochastic Constrained Contextual Bandits: We study stochastic constrained contextual bandits denoted by $\{\mathcal{X}, \mathcal{A}, \mathcal{F}, \mathcal{G}\}$, where $\mathcal{X}$ is the context set, $\mathcal{A}$ is the action set (a finite set), $\mathcal{F}$ is the reward function class, $\mathcal{G}$ is the cost function class. At period t, the learner observes a context x_t that is randomly generated from the context set $\mathcal{X}$ according to an (unknown) probability law $\mathbb{P}(\cdot)$. The learner takes an action $a_t \in \mathcal{A}$, and then receives a random reward $r_t(a_t) \in [0, F]$ and a random cost $c_t(a_t) \in [-G, G]$, where we assume $F, G \geq 1$. In this paper, we focus on the case of single constraint for a simple presentation and our results can be easily extended to the case with multiple-dimensional costs. We study a stochastic environment where the arrival of contexts and the observations for reward and cost are all drawn from unknown i.i.d. distributions. We further assume a key general realizability condition for the reward and cost functions.

Assumption 1 *There exists functions $f \in \mathcal{F}$ and $g \in \mathcal{G}$ such that $f(x, a) = \mathbb{E}[r_t(a)|x]$ and $g(x, a) = \mathbb{E}[c_t(a)|x], \forall x \in \mathcal{X}, a \in \mathcal{A}$.*

We define a policy $\pi : \mathcal{X} \to [0, 1]^A$, $(A = |\mathcal{A}|)$ which maps a context to a specific distribution over action set $\mathcal{A}$. The goal is to design a policy to optimize the cumulative rewards while satisfying

the constraint as follows

$$\max_{\pi} \mathbb{E}\left[\sum_{t=1}^{T} f(x_t, \pi(x_t))\right] \quad \text{s.t. } \mathbb{E}\left[\sum_{t=1}^{T} g(x_t, \pi(x_t))\right] \leq 0. \tag{1}$$

The constraint functions in (1) are sufficiently general to cover a wide range of constraint settings. $\{g(x_t, \cdot)\}$ could be either positive or negative for every round, allowing the formulation to capture both covering and packing constraints. For example, it can represent the knapsack constraint when $g(x_t, a) = c(x_t, a) - B/T$ with $c(x, a) \geq 0, \forall x \in \mathcal{X}, a \in \mathcal{A}$; the conservative constraint when $g(x_t, a) = \xi \cdot f(x_t, \pi_b(a)) - f(x_t, a)$ with $\xi \in (0, 1)$, where our policy should achieve at least ξ fraction of a baseline algorithm π_b or fairness constraint when $g(x_t, a) = \xi_a - \mathbb{I}(a_t = a)$ for $\xi_a \in (0, 1)$, where the action/arm a required to be pulled at least ξ_a fraction times on average.

Regret: We consider the underlying offline and relaxed problem to (1) as the baseline problem

$$\max_{\pi} \mathbb{E}_{a\sim\pi}\left[f(x, a)\right] \quad \text{s.t. } \mathbb{E}_{a\sim\pi}\left[g(x, a)\right] \leq 0. \tag{2}$$

Let π^* and ν^* be its optimal policy and value to (2), respectively. Note this offline problem serves as an upper bound to (1) and the detailed proof can be found in Agrawal and Devanur (2016). For a policy π, we define its (pseudo) regret against this baseline that

$$\mathcal{R}(T) := T\nu^* - \mathbb{E}\left[\sum_{t=1}^{T} f(x_t, a_t)\right].$$

Constraint Violation: The constraint violation is straightforward to be defined as

$$\mathcal{V}(T) := \mathbb{E}\left[\sum_{t=1}^{T} g(x_t, a_t)\right].$$

Note in CBwK, $\mathcal{V}(T) \equiv 0$ due to the "hard stopping" when the budget is exhausted.

Online Regression Oracles: We assume access to two online regression oracles $\mathcal{R}_r$ and $\mathcal{R}_c$ for reward and cost functions, respectively, where the online supervised regression problem is to minimize cumulative errors for a given loss function. Specifically, we consider the squared regression loss function that $l(\hat{y}, y) := (\hat{y} - y)^2$, where $\hat{y}$ denotes the prediction for y generated by the regression oracles. Given an instance $z_t := (x_t, a_t) \in \mathcal{X} \times \mathcal{A}$, the oracles calculate the estimated functions $\hat{f}_t$ and $\hat{g}_t$ based on the historical observations $((z_1, r_1, c_1), \cdots, (z_{t-1}, r_{t-1}, c_{t-1}))$. Online regression oracles can often be computationally efficient with provable strong theoretical guarantees Foster and Rakhlin (2020). We make the following assumption for reward and cost regression oracles.

Assumption 2 *Let $\{x_t, a_t\}$ be the trajectory generated by a policy π. Let $\hat{f}_t$ and $\hat{g}_t$ calculated by reward regression oracle $\mathcal{R}_r$ and cost regression oracle $\mathcal{R}_c$, respectively, the following error bounds*

$$\sum_{t=1}^{T}(\hat{f}_t(x_t, a_t) - f(x_t, a_t))^2 \leq U_f(p), \quad \sum_{t=1}^{T}(g(x_t, a_t) - \hat{g}_t(x_t, a_t))^2 \leq U_g(p),$$

hold with a probability $1 - p$. We further define $U(p) = \max(U_f(p), U_g(p))$.

The assumption characterizes the performance of regression oracles under the squared regression errors. An online learning algorithm that attains sublinear square loss, i.e. $U(p) = o(T)$, for the problem of predicting leads to a valid regression oracle. For example, for finite function classes, there exist learning oracles such that $U_f(p) = O(\log(|\mathcal{F}|/p))$ and $U_g(p) = O(\log(|\mathcal{G}|/p))$ Foster and Rakhlin (2023). We let $p = 1/T^2$ throughout the paper.

3. Lyapunov Optimization Based Estimation to Decision

In this section, we propose a general Lyapunov Optimization Based Estimation to Decision (LOE2D) framework to minimize the regret and constraint violation simultaneously in constrained contextual bandits. We summarize LOE2D algorithm and explain its underlying design principle and intuition.

LOE2D Framework

Initialization: $Q_1 = 0$, $V = \sqrt{TU \log T}$, $\gamma = A\sqrt{T/U}$ and $\beta_1 = 1$
For $t = 1, \cdots, T$,

- **Lyapunov Optimization Index Estimation:** Estimate the reward function $\hat{f}_t(x_t, a)$ via $\mathcal{R}_r$ and the cost function $\hat{g}_t(x_t, a)$ via $\mathcal{R}_c$. Compute the Lyapunov optimization index

$$\hat{L}_t(x_t, a) = \hat{f}_t(x_t, a) - \frac{Q_t}{V}\hat{g}_t(x_t, a). \tag{3}$$

- **Estimation to Decision:** Let $\hat{a}_t = \text{argmax}_a \hat{L}_t(x_t, a)$ and sample a_t according to the inverse gap weighting distribution of π_t that

$$\pi_t(a) = \frac{1}{\eta_t + 2\gamma\beta_t(\hat{L}_t(x_t, \hat{a}_t) - \hat{L}_t(x_t, a))}, \tag{4}$$

  where η_t is a positive term to ensure $\sum_a \pi_t(a) = 1$.

- **Feedback and Online Regression Update:** Observe noisy reward $r_t(a_t)$ and cost $c_t(a_t)$ and feed them into the oracles $\mathcal{R}_r$ and $\mathcal{R}_c$, respectively.

- **Virtual Queue Update:** Update the virtual queue and exploration parameter

$$Q_{t+1} = \max\left(Q_t + c_t(x_t, a_t), 0\right), \ \beta_{t+1} = V/(V + Q_{t+1}). \tag{5}$$

LOE2D first incorporates the Lyapunov drift optimization framework into SquareCB Foster and Rakhlin (2020), a randomization strategy that adeptly transforms the estimated values of actions into a distribution. To optimally balance exploration and exploitation within these two frameworks, we have enhanced and restructured both, introducing novel analysis techniques that ensure optimal guarantees. In detail, the LOE2D framework includes the following components:

- **Lyapunov optimization index estimation:** Upon a context x_t arrives, LOE2D estimates rewards $\hat{f}_t(x_t, \cdot)$ and costs $\hat{g}_t(x_t, \cdot)$ from the regression oracles $\mathcal{R}_r$ and $\mathcal{R}_c$, respectively. The algorithm then utilizes Lyapunov optimization to calculate "Reward $-$ Lyapunov drift" i.e.,

$V\hat{f}_t(x_t,\cdot) - \frac{1}{2}(Q_{t+1}^2 - Q_t^2)$, where the drift term is approximated by $Q_t\hat{g}_t(x_t,\cdot)$ according to the virtual queue update in (5), resulting in the Lyapunov optimization index $\hat{L}_t(x_t,a)$ as given in (3). Note the design in (3) can also be interpreted as an approximation of the Lagrange function $L(x_t,a) := f(x_t,a) - \lambda g(x_t,a)$, where the reward and cost functions are approximated by estimated ones and the dual variable is approximated by the scaled virtual queue term Q_t/V. However, our design neither imposes any upper bound on the virtual queue Q_t in (5) nor requires any information on the knowledge of optimal cumulative rewards or Slater's constant. These are different from Han et al. (2023); Slivkins et al. (2023), where the dual variables are constrained in a probability simplex and either the knowledge of the optimal cumulative rewards or Slater constant is required to scale the dual variances such that the reward and constraint violation can be balanced.

- **Inverse-gap weighting decision:** Upon estimating the Lyapunov optimization index $\hat{L}_t(x_t,\cdot)$ we employ the inverse-gap weighting technique Abe and Long (1999); Foster and Rakhlin (2020) to translate the weights into the action probabilities. Specifically, LOE2D computes the probability $\pi_t(a)$ that is inversely proportional to the gap $\hat{L}(x_t,\hat{a}) - \hat{L}(x_t,a)$ between the greedy one $\hat{a}$ and any given action a, and then sample $a_t \sim \pi_t$. Intuitively, when the estimated weight of an action is large, the algorithm tends to choose it with a high probability. The probabilistic design is to maintain a good tradeoff between information acquisition, reward maximization, and constraint violation minimization. In the inverse-gap weighting distribution in (4), we introduce a violation-adaptive (virtual queue related) exploration parameter β_t in (5) to avoid exploring too much when the constraint violation (Q_t) is large. Our design is again different with Han et al. (2023); Slivkins et al. (2023), where the exploration parameters are fixed.

- **Regression oracles and virtual queue update:** Once $r_t(a_t)$ and $c_t(a_t)$ are observed, they are used to update regression oracles and the virtual queue term in (5). Note that the (scaled) virtual queue Q_t plays a similar role with the dual price in the Lagrange function in regulating the decision. Intuitively, when the virtual queue becomes large, it would encourage a conservative decision in (3) and (4) to prevent further constraint violations; otherwise, the algorithm is optimistic in maximizing rewards. Moreover, it is worth emphasizing we provide a new perspective and analysis on the virtual queue: we treat it as a Markov process and leverage a multi-step Lyapunov drift analysis to establish its high probability upper bound.

In summary, LOE2D provides a new perspective on algorithm design and theoretical analysis for constrained contextual bandits. The Lyapunov optimization guided design in (3), violation-aware inverse-gap weighting decision in (4), and multi-step Lyapunov drift analysis on the virtual queue process are essential to establish the strong theoretical performance under mild assumption and without any prior information of the underlying offline problem.

4. Theoretical Results

In this section, we present the theoretical analysis of LOE2D framework. To state our results, we first introduce a *relaxed* Slater condition.

Assumption 3 *There exists a policy π_0 such that for a positive value $\varepsilon \geq 4\sqrt{U/T}$, $\mathbb{E}_{a\sim\pi_0}[g(x,a)] \leq -\varepsilon$ holds.*

The assumption commonly refers to the Slater condition and quantifies the degree of constraint slackness, which is standard in the optimization literature. However, rather than assuming a *constant* slackness (i.e., $\varepsilon = \Omega(1)$) as in Slivkins et al. (2023), we allow a *relaxed* Slater's condition (or a tighter problem instance) with $\varepsilon = \Omega(\sqrt{U/T})$. Now we are ready to present the following main results for LOE2D in terms of regret and constraint violation and formally answer the question raised in the introduction.

Theorem 1 *Under Assumptions 1 and 2, LOE2D achieves the following regret and constraint violation that*

$$\mathcal{R}(T) = \tilde{O}(T^{\frac{3}{4}}U^{\frac{1}{4}}),\ \mathcal{V}(T) = \tilde{O}(T^{\frac{3}{4}}U^{\frac{1}{4}}).$$

Given the additional Slater's condition in Assumption 3, LOE2D achieves that

$$\mathcal{R}(T) = \tilde{O}(\min\{\sqrt{TU}/\varepsilon^2, T^{\frac{3}{4}}U^{\frac{1}{4}}\}),\ \mathcal{V}(T) = \tilde{O}(\min\{\sqrt{TU}/\varepsilon^2, T^{\frac{3}{4}}U^{\frac{1}{4}}\}).$$

Moreover for CBwK, LOE2D achieves the following regret that

$$\mathcal{R}(T) = \tilde{O}(\min\{\sqrt{TU}/\varepsilon^2, T^{\frac{3}{4}}U^{\frac{1}{4}}\}).$$

Remark 2 *The known lower bound is only on stochastic (linear) contextual bandits with knapsacks, which is proved by reducing CBwK to unconstrained contextual bandits problem Agrawal and Devanur (2016); Han et al. (2023). Specifically, the lower bound is $\Omega(\sqrt{TU})$, as proven by Han et al. (2023). However, no existing lower bound is characterized by the feasibility margin ε. For the general CCB problem (without hard-stopping), the lower bound of CCB is even more subtle because the regret $\mathcal{R}(T)$ and constraint violation $\mathcal{V}(T)$ can trade off against each other, where we can achieve a small (even negative) regret by causing a large violation (e.g., overusing resources). However, if we consider their maximum $\max(\mathcal{R}(T), \mathcal{V}(T))$, we conjecture the lower bound is very likely to be $\Omega(\min\{\sqrt{TU}/\varepsilon, T^{\frac{3}{4}}U^{\frac{1}{4}}\})$ and we defer the formal proof to the future study.*

Remark 3 *The theorem shows that LOE2D achieves sublinear regret and constraint violation: i) when only realizability assumption holds, LOE2D guarantees both regret and violation within $\tilde{O}(T^{\frac{3}{4}}U^{\frac{1}{4}})$, consistent with the results in Slivkins et al. (2023); ii) when an additional relaxed Slater's condition holds in Assumption 3, LOE2D guarantees $\tilde{O}(\min\{\sqrt{TU}/\varepsilon^2, T^{\frac{3}{4}}U^{\frac{1}{4}}\})$, which interpolates between $\tilde{O}(\sqrt{TU})$ and $\tilde{O}(T^{\frac{3}{4}}U^{\frac{1}{4}})$ depending on the feasibility margin ε. Our result is more general than that in Slivkins et al. (2023), where $\tilde{O}(\sqrt{TU})$ is established assuming a constant ε and the regime of "large budgets" where $B = \Omega(T)$. Moreover, unlike the algorithm (named LagrangeCBwLC) in Slivkins et al. (2023), which requires the knowledge of the Slater constant ε to determine the learning rate, LOE2D achieves these results without any prior information of the environment. Besides, LOE2D can be applied into CBwK without any modification and achieve a similar result in Han et al. (2023). Finally, it's worth emphasizing that all these results are achieved using a single LOE2D algorithm, without any customization for each specific scenario. This demonstrates the flexibility and universality of the proposed framework.*

Next, we illustrate the key techniques to prove Theorem 1. We first introduce a critical lemma that bridges the regret and Lyapunov drift, which is the key to establish the constraint violation and regret in Theorem 1. We also highlight the major steps in analyzing the "hard-stopping" CBwK.

4.1. A Key Lemma of "Regret + Lyapunov Drift"

To provide a unified analysis in proving Theorem 1, we first establish the following key lemma that upper bounds the "one-step regret + Lyapunov drift" in a whole. We define the Lyapunov function as $L_t = \frac{1}{2}Q_t^2$ and its drift as $\Delta_t = L_{t+1} - L_t$. Further let the filtration $\mathcal{H}_t = [x_t, \hat{f}_t, \hat{g}_t, Q_t]$.

Lemma 4 *Under LOE2D, we have for any policy π such that*

$$\begin{aligned}
&\mathbb{E}_{a\sim\pi}\left[f(x_t,a)|\mathcal{H}_t\right] - \mathbb{E}_{a_t\sim\pi_t}\left[f(x_t,a_t)|\mathcal{H}_t\right] + \frac{1}{V}\mathbb{E}_{a_t\sim\pi_t}\left[\Delta_t|\mathcal{H}_t\right] \\
\leq&\gamma\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{f}_t(x_t,a_t) - f(x_t,a_t))^2|\mathcal{H}_t\right] + \gamma\frac{Q_t}{V}\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{g}_t(x_t,a_t) - g(x_t,a_t))^2|\mathcal{H}_t\right] \\
&+ \mathbb{E}_{a\sim\pi}\left[\frac{Q_t}{V}g(x_t,a)|\mathcal{H}_t\right] + \frac{A}{\gamma\beta_t} + \frac{G^2}{2V}.
\end{aligned} \tag{6}$$

Note the lemma above holds for any policy π, including the optimal static policy π^*. Therefore, it implies that "one-step regret + Lyapunov drift" is bounded in (6) by the (weighted) regression oracle errors, constraint satisfactory of the baseline policy π, and the remaining terms related to (β_t, γ, V). Since it is a key lemma in proving our main results, we provide highlight the key steps. It's worth noting that our analysis is a refined version compared to Slivkins et al. (2023) because we only introduce a linear form of virtual queue (dual variable) instead of a quadratic form of dual variable Slivkins et al. (2023).

Proof Sketch: Let the error-free Lyapunov optimization index be $L_t(x_t, a) := f(x_t, a) - \frac{Q_t}{V}g(x_t, a)$ and recall its approximated version $\hat{L}_t(x_t, a)$ in (3), we have the following decomposition

$$\begin{aligned}
&\mathbb{E}_{a\sim\pi}[L_t(x_t,a)|\mathcal{H}_t] - \mathbb{E}_{a_t\sim\pi_t}[L_t(x_t,a_t)|\mathcal{H}_t] \\
=&\mathbb{E}_{a_t\sim\pi_t}\left[\hat{L}_t(x_t,\hat{a}_t) - \hat{L}_t(x_t,a_t)|\mathcal{H}_t\right] + \mathbb{E}_{a_t\sim\pi_t}\left[\hat{L}_t(x_t,a_t) - L_t(x_t,a_t)|\mathcal{H}_t\right] \\
&+ \mathbb{E}_{a\sim\pi}\left[L_t(x_t,a)|\mathcal{H}_t\right] - \hat{L}_t(x_t,\hat{a}_t)
\end{aligned} \tag{7}$$

The first term represents the cost of exploration, which can be easily bounded according to the definition of inverse gap weighting distribution in (4) that

$$\mathbb{E}_{a_t\sim\pi_t}\left[\hat{L}_t(x_t,\hat{a}_t) - \hat{L}_t(x_t,a_t)|\mathcal{H}_t\right] \leq \frac{A}{2\beta_t\gamma}. \tag{8}$$

The second term follows by applying AM–GM inequality individually on the reward and constraint function that

$$\begin{aligned}
\mathbb{E}_{a_t\sim\pi_t}\left[\hat{L}_t(x_t,a_t) - L_t(x_t,a_t)|\mathcal{H}_t\right] \leq&\frac{1}{2\gamma} + \frac{\gamma}{2}\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{f}_t(x_t,a_t) - f(x_t,a_t))^2|\mathcal{H}_t\right] \\
&+ \frac{Q_t}{V}\frac{1}{2\gamma} + \frac{Q_t}{V}\frac{\gamma}{2}\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{g}_t(x_t,a_t) - g(x_t,a_t))^2|\mathcal{H}_t\right].
\end{aligned} \tag{9}$$

The last term can be further decomposed as

$$\begin{aligned}
&\mathbb{E}_{a\sim\pi}\left[L_t(x_t,a) - \hat{L}_t(x_t,\hat{a}_t)|\mathcal{H}_t\right] \\
=&\mathbb{E}_{a\sim\pi}\left[L_t(x_t,a) - \hat{L}_t(x_t,a)|\mathcal{H}_t\right] - \mathbb{E}_{a\sim\pi}\left[\hat{L}_t(x_t,\hat{a}_t) - \hat{L}_t(x_t,a)|\mathcal{H}_t\right]
\end{aligned} \tag{10}$$

$$
\begin{aligned}
=&\mathbb{E}_{a\sim\pi}\left[L_t(x_t,a)-\hat{L}_t(x_t,a)|\mathcal{H}_t\right]-\mathbb{E}_{a\sim\pi}\left[\frac{1}{2\gamma\beta_t\pi_t(a)}-\frac{\eta_t}{2\gamma\beta_t}|\mathcal{H}_t\right]\\
\leq&\frac{Q_t}{V}\frac{\gamma}{2}\mathbb{E}_{a\sim\pi}\left[\pi_t(a)(\hat{g}_t(x_t,a)-g(x_t,a))^2|\mathcal{H}_t\right]+\frac{\gamma}{2}\mathbb{E}_{a\sim\pi}\left[\pi_t(a)(\hat{f}_t(x_t,a)-f(x_t,a))^2|\mathcal{H}_t\right]\\
&+\mathbb{E}_{a\sim\pi}\left[\frac{1}{2\gamma\pi_t(a)}(1+\frac{Q_t}{V}-\frac{1}{\beta_t})|\mathcal{H}_t\right]+\frac{\eta_t}{2\beta_t\gamma}\\
\leq&\frac{Q_t}{V}\frac{\gamma}{2}\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{g}_t(x_t,a_t)-g_t(x_t,a_t))^2|\mathcal{H}_t\right]+\frac{\gamma}{2}\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{f}_t(x_t,a_t)-f_t(x_t,a_t))^2|\mathcal{H}_t\right]+\frac{A}{2\beta_t\gamma},
\end{aligned}
$$

where the second equality holds because $\hat{L}_t(x_t,\hat{a}_t)-\hat{L}_t(x_t,a)=\frac{1}{2\beta_t\gamma\pi_t(a)}-\frac{\eta_t}{2\beta_t\gamma}$ holds by (4); the third inequality again follows from AM-GM inequality, and the last inequality holds by $\beta_t=V/(V+Q_t)$ and $\eta_t\leq A$. Finally, in conjunction with the virtual queue update in (5), we have

$$
\frac{1}{V}\mathbb{E}_{a_t\sim\pi_t}\left[\Delta_t|\mathcal{H}_t\right]\leq\mathbb{E}_{a_t\sim\pi_t}\left[\frac{Q_t}{V}g(x_t,a_t)|\mathcal{H}_t\right]+\frac{G^2}{2V}. \tag{11}
$$

Now we substitute (11) into (7) and combine all inequalities (8)–(10), we complete the proof. □
Based on Lemma 4, we proceed to prove the constraint violation and regret in Theorem 1.

4.2. Constraint violation bound

According to the virtual queue update in (5), we immediately have

$$
\mathcal{V}(T):=\mathbb{E}\left[\sum_{t=1}^{T}g(x_t,a_t)\right]\leq\mathbb{E}[Q_{T+1}]. \tag{12}
$$

Now we study "Lyapunov drift" in Lemma 4 to establish the upper bound on the virtual queue. The Lyapunov drift analysis was widely used to study the stability property of the control policies in stochastic queueing networks in Hajek (1982); Tassiulas and Ephremides (1992), where a policy is called stable when its induced queue lengths are finite or bounded (a policy is usually better if its queue lengths are small). The modern analytical framework with the high-probability bounds of queue lengths can be found in Bertsimas et al. (2001); Eryilmaz and Srikant (2012). For a more comprehensive introduction on the method, the readers can refer to Neely (2022); Srikant and Ying (2014). When only the realizability assumption holds, we directly analyze the sample path of the virtual queue and establish an upper bound of $\mathbb{E}[Q_t]=\tilde{O}(T^{\frac{3}{4}}U^{\frac{1}{4}})$. When Slater's condition holds in Assumption 3, we treat the virtual queue as a Markov process and study its upper bound via multiple-step Lyapunov drift analysis, where the regression errors are amortized over multiple time slots. Specifically, from Lemma 4, we establish a "multiple-step negative drift" of Lyapunov function, implying a high probability upper bound on the virtual queue in the following lemma.

Lemma 5 *Under Assumptions 1-3, there exists a positive quantity $\delta\geq\varepsilon/4$, a absolute constants C_0 and a positive integer K such that LOE2D establishes the following Lyapunov drift*

$$
\mathbb{E}\left[Q_{t+K}^2-Q_t^2|\mathcal{H}_t\right]\leq-2K\delta Q_t+C_0(KV+\gamma VU+\gamma KU), \tag{13}
$$

and the virtual queue satisfies

$$
\mathbb{P}\left(Q_t\leq\frac{K^2\delta^2+C_0(KV+\gamma VU+\gamma KU)+12G^2K^2\log(1+16G^2T)}{K\delta}\right)\geq1-\frac{1}{T^2},\quad\forall t\in[T]. \tag{14}
$$

Intuitively, if the virtual queue process $\{Q_t\}$ already reaches the steady state, i.e., the mutiple-step drift is zero $\mathbb{E}\left[Q_{t+K}^2 - Q_t^2 | \mathcal{H}_t = h\right] = 0$ in (13), we immediately establish an upper bound of $O(K \log T/\varepsilon)$. This intuition is formally justified by (14) with slightly changing the "constant". Let $V = \sqrt{TU \log T}$, $\gamma = A\sqrt{T/U}$, and $K = 2\gamma U/\varepsilon$, we immediately prove that with a high probability,

$$Q_t = O(\sqrt{TU} \log T/\varepsilon^2), \ \forall t \in [T].$$

In the current analysis, we need to choose $K-$step Lyapunov drift (with $1/\epsilon$ dependent K multi-step) that leads to a $O(K/\varepsilon)$ upper bound of the virtual queue, resulting in the "$O(1/\varepsilon)-$gap" compared with the existing optimal results. To close this gap, we might need to develop *a new drift lemma* where the upper bound is refined to $O(K)$ rather than $O(K/\varepsilon)$, which we leave for future work. We summarize these results in the following lemma.

Lemma 6 *Under Assumptions 1 and 2, LOE2D achieves that*

$$\mathbb{E}[Q_t] = \tilde{O}(T^{\frac{3}{4}} U^{\frac{1}{4}}), \ \forall t \in [T].$$

Given the additional Slater's condition in Assumption 3, LOE2D achieves that

$$Q_t = \tilde{O}(\sqrt{TU}/\varepsilon^2), \ \forall t \in [T],$$

holds with a probability of at least $1 - 1/T^2$.

The constraint violation in Theorem 1 is then proved by directly applying Lemma 6 into (12).

4.3. Regret Bound

Let $\pi = \pi^*$ in Lemma 4, note $\mathbb{E}_{a\sim\pi^*}\left[\frac{Q_t}{V} g(x_t, a) | \mathcal{H}_t\right] \leq 0$ holds in (6) because π^* is a feasible policy such that $\mathbb{E}_{a\sim\pi^*}\left[g(x_t, a) | \mathcal{H}_t\right] \leq 0$. Taking expectation w.r.t. $\mathcal{H}_t$ and summation of the inequality from $t = 1$ to T in (6), we establish the following regret bound under the assumption that the virtual queue is bounded (i.e., $Q_t \leq Q_{\max}$):

$$\begin{aligned}
\mathcal{R}(T) \leq & -\frac{\mathbb{E}\left[\sum_{t=1}^T \Delta_t\right]}{V} + \gamma U_f + \gamma \frac{Q_{max}}{V} U_g + \sum_{t=1}^T \frac{A}{\gamma \beta_t} + \frac{G^2 T}{2V} \\
\leq & \gamma U_f + \gamma \frac{Q_{max}}{V} U_g + (1 + \frac{Q_{max}}{V}) \frac{AT}{\gamma} + \frac{G^2 T}{2V},
\end{aligned} \tag{15}$$

where the first inequality holds by Assumption 2; and the second inequality holds because of $Q_1 = 0$ and $\beta_t \geq V/(V + Q_{\max}), \forall t \in [T]$. The inequality in (15) indicates that the regret is bounded by the value of the virtual queue and the regression oracles errors. Thus the key to proving CCB regret in Theorem 1 involves determining the bound for the anytime virtual queue Q_t, which has already been shown in Lemma 5. Recall $V = \sqrt{TU \log T}$ and $\gamma = A\sqrt{T/U}$, we immediately prove the regret bound according to the virtual queue bound of $Q_{\max}$ in Lemma 6.

4.4. LOE2D for CBwK

In this section, we illustrate the key idea in analyzing LOE2D for contextual bandits with knapsack constraints. Recall the interaction stops when the budget is depleted in CBwK (i.e., hard stopping). Let $b := B/T$ and π^* and ν_b^* be the optimal policy and value to the following offline problems:

$$\max_{\pi} \ \mathbb{E}_{a\sim\pi}\left[f(x,a)\right] \quad \text{s.t.}\ \mathbb{E}_{a\sim\pi}\left[g(x,a)\right] \le b.$$

The regret is defined and decomposed as follows

$$\mathcal{R}(T) := T\nu_b^* - \mathbb{E}\left[\sum_{t=1}^{\tau} f(x_t,a_t)\right] = \mathbb{E}\left[\sum_{t=1}^{\tau} f(x_t,\pi^*(x_t)) - f(x_t,a_t)\right] + \mathbb{E}\left[(T-\tau)\nu_b^*\right],$$

where τ is the stopping time and the regret includes two parts: "regret before stopping" and "regret after stopping". For "regret before stopping", we have already established an upper bound on the difference between LOE2D and baseline algorithms in (15). For "regret after stopping", we show it can be bounded by the virtual queue. According to the virtual queue update in (5), we immediately have $Q_{\tau+1} + \tau b \ge \sum_{t=1}^{\tau} c(x_t,a_t)$, in conjunction with the definition of stopping time, we know τ satisfies $Q_{\tau+1} + \tau b \ge B := Tb$. It immediately implies "regret after stopping" is bounded by

$$\mathbb{E}[(T-\tau)\nu_b^*] \le \frac{\nu_b^*}{b}\mathbb{E}[Q_{\tau+1}].$$

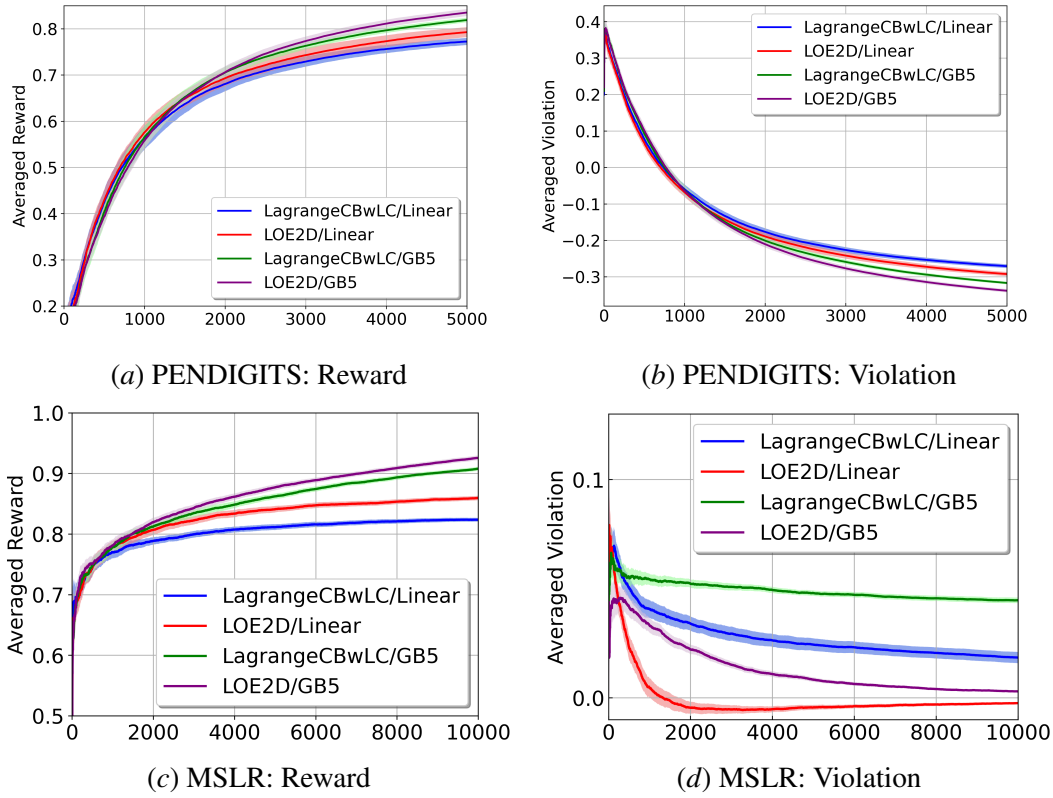
Since $\nu_b^* = \Theta(b)$, we apply Lemma 6 and prove $\tilde{O}(\min\{\sqrt{TU}/\varepsilon^2, T^{\frac{3}{4}}U^{\frac{1}{4}}\})$ regret for CBwK.

5. Experiments

In this section, we run numerical experiments to justify our algorithm with two regression oracles (including linear regression and gradient-boosted tree regression). We consider LagrangeCBwLC Slivkins et al. (2023) as the benchmark because it is most related to ours. We design two sets of experiments and plot the average reward $\sum_{s=1}^{t} f(x_s,a_s)/t$ and violation $\sum_{s=1}^{t} g(x_s,a_s)/t$. All results presented are the average of 50 trials and are reported within 95% confidence interval.

Classification dataset: We study online classification problems and customize them into stochastic CCB setting, where each context/data is randomly drawn from the dataset, one of the classes (arms) is chosen for the context, and then (noisy) reward and cost are observed. Our experiment is based on the Pendigits dataset Keller et al. (2012), where the dimension of the contextual information is 16 and there exist 10 distinct classes, i.e., $|\mathcal{A}| = 10$. When the chosen class (action) is correct, we receive a reward corrupted with Gaussian noise $r_t \sim \mathcal{N}(1, 0.05)$ and $r_t \sim \mathcal{N}(0, 0.05)$ otherwise. The constraint is imposed on the expected reward $g(x_t,a_t) := 0.5 - f(x_t,a_t)$, where at least half of the actions are required correct over the learning process. We plot the results in Figures 1(*a*) and 1(*b*). It is shown that algorithms with a sophisticated regression oracle (gradient-boosted trees) outperform those with a linear regression oracle. If we compare both algorithms w.r.t. the same learning oracle, we observe that LOE2D outperforms LagrangeCBwLC in terms of both reward and constraint violation, which justifies our LOE2D framework with violation-aware design has the advantage in achieving a good balance between rewards and constraint violations.

Learning-to-rank dataset: We also test our algorithm on a large-scale learning-to-rank dataset Microsoft MSLR-WEB30k Qin and Liu (2013). This dataset contains 31278 queries, where each



(*a*) PENDIGITS: Reward (*b*) PENDIGITS: Violation

(*c*) MSLR: Reward (*d*) MSLR: Violation

Figure 1: LOE2D v.s. LagrangeCBwLC: Averaged reward and violation

query includes a varying number of documents-query contexts with each of which has a dimension of 136 and there exists 20 documents/arms, i.e., $|\mathcal{A}| = 20$. This ranking dataset possesses inherent rewards (relevance). For each arm, we draw its expected cost uniformly randomly from $[0, 1]$ and the value remains fixed during a trial. The constraint is set $g(x, a) \leq 0.5$. The observations are also corrupted with Gaussian noise $\mathcal{N}(0, 0.05)$. We plot our results in Figures 1(*c*) and 1(*d*). It is shown that that LOE2D outperforms LagrangeCBwLC in terms of average reward and constraint violation when comparing them with the same learning oracle, respectively, and LOE2D with GB5 oracles yield the best overall performance. Interestingly, it is observed that LOE2D with linear oracle achieves the lowest constraint violation. The possible reason is that LOE2D with linear oracle has relatively inaccurate reward/cost estimation, leading to a conservative approach.

6. Conclusion

In this paper, we propose LOE2D, a general Lyapunov optimization based estimation to decision framework for stochastic constrained contextual bandits. LOE2D establishes near-optimal regret and constraint violation bounds without any prior knowledge of underlying problems regardless of the feasibility assumption or a potentially vanishing feasibility margin. These results are achieved through violation-adaptive design and a multi-step Lyapunov drift analysis. The experiments further justify our theoretical results.

Acknowledgments

The work was partly supported by the Shanghai Sailing Program 22YF1428500 and the National Nature Science Foundation of China under grant 62302305.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011a.

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011b.

Naoki Abe and Philip M Long. Associative reinforcement learning using linear probabilistic concepts. In *ICML*, pages 3–11. Citeseer, 1999.

Marc Abeille and Alessandro Lazaric. Linear thompson sampling revisited. In *Artificial Intelligence and Statistics*, pages 176–184. PMLR, 2017.

Alekh Agarwal, Daniel Hsu, Satyen Kale, John Langford, Lihong Li, and Robert Schapire. Taming the monster: A fast and simple algorithm for contextual bandits. In *International Conference on Machine Learning*, pages 1638–1646. PMLR, 2014.

Shipra Agrawal and Nikhil Devanur. Linear contextual bandits with knapsacks. *Advances in Neural Information Processing Systems*, 29, 2016.

Shipra Agrawal and Nikhil R Devanur. Bandits with concave rewards and convex knapsacks. In *Proceedings of the fifteenth ACM conference on Economics and computation*, pages 989–1006, 2014.

Shipra Agrawal and Navin Goyal. Thompson sampling for contextual bandits with linear payoffs. In *International conference on machine learning*, pages 127–135. PMLR, 2013.

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. *Advances in Neural Information Processing Systems*, 32, 2019.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47:235–256, 2002.

Ashwinkumar Badanidiyuru, John Langford, and Aleksandrs Slivkins. Resourceful contextual bandits. In *Conference on Learning Theory*, pages 1109–1134. PMLR, 2014.

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *Journal of the ACM (JACM)*, 65(3):1–55, 2018.

Dimitris Bertsimas, David Gamarnik, and John N Tsitsiklis. Performance of multiclass markovian queueing networks via piecewise linear lyapunov functions. *Annals of Applied Probability*, pages 1384–1428, 2001.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Olivier Chapelle and Lihong Li. An empirical evaluation of thompson sampling. *Advances in neural information processing systems*, 24, 2011.

Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. Efficient optimal learning for contextual bandits. In *Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence*, page 169–178. AUAI Press, 2011. ISBN 9780974903972.

Atilla Eryilmaz and R Srikant. Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Systems: Theory and Applications*, 72(3-4):311–359, 2012.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 3199–3210. PMLR, 2020.

Dylan Foster, Alekh Agarwal, Miroslav Dudík, Haipeng Luo, and Robert Schapire. Practical contextual bandits with regression oracles. In *International Conference on Machine Learning*, pages 1539–1548. PMLR, 2018.

Dylan J Foster and Alexander Rakhlin. Foundations of reinforcement learning and interactive decision making. *arXiv preprint arXiv:2312.16730*, 2023.

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.

Bruce Hajek. Hitting-time and occupation-time bounds implied by drift analysis with applications. *Advances in Applied probability*, 14(3):502–525, 1982.

Yuxuan Han, Jialin Zeng, Yang Wang, Yang Xiang, and Jiheng Zhang. Optimal contextual bandits with knapsacks under realizability via regression oracles. In *International Conference on Artificial Intelligence and Statistics*, pages 5011–5035. PMLR, 2023.

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. *Advances in Neural Information Processing Systems*, 30, 2017.

Fabian Keller, Emmanuel Muller, and Klemens Bohm. Hics: High contrast subspaces for density-based outlier ranking. In *2012 IEEE 28th international conference on data engineering*, pages 1037–1048. IEEE, 2012.

Adam R Klivans and Alexander A Sherstov. Cryptographic hardness for learning intersections of halfspaces. *Journal of Computer and System Sciences*, 75(1):2–12, 2009.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

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.

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.

Ahmadreza Moradipari, Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Safe linear thompson sampling with side information. *IEEE Transactions on Signal Processing*, 69: 3755–3767, 2021.

Michael Neely. *Stochastic network optimization with application to communication and queueing systems*. Springer Nature, 2022.

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.

Vishakha Patil, Ganesh Ghalme, Vineet Nair, and Yadati Narahari. Achieving fairness in the stochastic multi-armed bandit problem. *The Journal of Machine Learning Research*, 22(1):7885–7915, 2021.

Tao Qin and Tie-Yan Liu. Introducing LETOR 4.0 datasets. *CoRR*, abs/1306.2597, 2013. URL http://arxiv.org/abs/1306.2597.

Paat Rusmevichientong and John N Tsitsiklis. Linearly parameterized bandits. *Mathematics of Operations Research*, 35(2):395–411, 2010.

David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 47(3): 1904–1931, 2022.

Vidyashankar Sivakumar, Shiliang Zuo, and Arindam Banerjee. Smoothed adversarial linear contextual bandits with knapsacks. In *International Conference on Machine Learning*, pages 20253–20277. PMLR, 2022.

Aleksandrs Slivkins, Karthik Abinav Sankararaman, and Dylan J. Foster. Contextual bandits with packing and covering constraints: A modular lagrangian approach via regression. In *Annual Conference Computational Learning Theory*, 2023.

Rayadurgam Srikant and Lei Ying. *Communication networks: An optimization, control and stochastic networks perspective*. Cambridge University Press, 2014.

Leandros Tassiulas and Anthony Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *IEEE TRANSACTIONS ON AUTOMATIC CONTROL*, 31(12), 1992.

Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. *Mobile health: sensors, analytic methods, and applications*, pages 495–517, 2017.

Long Tran-Thanh, Sebastian Stein, Alex Rogers, and Nicholas R Jennings. Efficient crowdsourcing of unknown experts using bounded multi-armed bandits. *Artificial Intelligence*, 214:89–111, 2014.

Sharan Vaswani, Abbas Mehrabian, Audrey Durand, and Branislav Kveton. Old dog learns new tricks: Randomized ucb for bandit problems. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, 26–28 Aug 2020.

Huasen Wu, Rayadurgam Srikant, Xin Liu, and Chong Jiang. Algorithms with logarithmic or sublinear regret for constrained contextual bandits. *Advances in Neural Information Processing Systems*, 28, 2015.

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

Huanle Xu, Yang Liu, Wing Cheong Lau, and Rui Li. Combinatorial multi-armed bandits with concave rewards and fairness constraints. In *IJCAI*, pages 2554–2560, 2020.

Hao Yu, Michael J. Neely, and Xiaohan Wei. Online convex optimization with stochastic constraints. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017.

Appendix A. Proof of Lemma 4

Lemma 4 expands the inverse gap weighting technique from contextual bandits to CCB. We can take $L_t(x_t, a) := f(x_t, a) - \frac{Q_t}{V} g(x_t, a)$ as the surrogate reward function and decompose its instantaneous regret as follows

$$\begin{aligned}
&\mathbb{E}_{a\sim\pi}\left[f(x_t,a)-\frac{Q_t}{V}g(x_t,a)|\mathcal{H}_t\right]-\mathbb{E}_{a_t\sim\pi_t}\left[f(x_t,a_t)-\frac{Q_t}{V}g(x_t,a_t)|\mathcal{H}_t\right]\\
=&\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{f}_t(x_t,\hat{a}_t)-\frac{Q_t}{V}\hat{g}_t(x_t,\hat{a}_t))-(\hat{f}_t(x_t,a_t)-\frac{Q_t}{V}\hat{g}_t(x_t,a_t))|\mathcal{H}_t\right] \quad (16)\\
&+\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{f}_t(x_t,a_t)-\frac{Q_t}{V}\hat{g}_t(x_t,a_t))-(f(x_t,a_t)-\frac{Q_t}{V}g(x_t,a_t))|\mathcal{H}_t\right] \quad (17)\\
&+\mathbb{E}_{a\sim\pi}\left[f(x_t,a)-\frac{Q_t}{V}g(x_t,a)|\mathcal{H}_t\right]-(\hat{f}_t(x_t,\hat{a}_t)-\frac{Q_t}{V}\hat{g}_t(x_t,\hat{a}_t)), \quad (18)
\end{aligned}$$

The estimation-to-decision module in (4), based on inverse gap weighting distribution, facilitates converting these terms into corresponding regression oracle errors. We begin with the first part (16), which refers to the cost of exploration when the estimates for f and g are accurate. From the computation of estimation-to-decision distribution in (4), we obtain that

$$\begin{aligned}
\mathbb{E}_{a_t\sim\pi_t}\left[\hat{L}_t(x_t,\hat{a}_t)-\hat{L}_t(x_t,a)|\mathcal{H}_t\right]&=\sum_{a_t}\mathbb{E}\left[\pi_t(a)(\hat{L}_t(x_t,\hat{a}_t)-\hat{L}_t(x_t,a_t))|\mathcal{H}_t\right]\\
&=\sum_{a_t}\mathbb{E}\left[\frac{(\hat{L}_t(x_t,\hat{a}_t)-\hat{L}_t(x_t,a_t))}{\eta_t+2\gamma\beta_t(\hat{L}_t(x_t,\hat{a}_t)-\hat{L}_t(x_t,a))}|\mathcal{H}_t\right]\\
&\le\sum_{a_t}\mathbb{E}\left[\frac{(\hat{L}_t(x_t,\hat{a}_t)-\hat{L}_t(x_t,a_t))}{2\gamma\beta_t(\hat{L}_t(x_t,\hat{a}_t)-\hat{L}_t(x_t,a))}|\mathcal{H}_t\right]=\frac{A}{2\beta_t\gamma},
\end{aligned}$$

the inequality holds since η_t is positive. The second term (17) directly relates to the estimation errors of the reward and cost functions. These terms can be transformed into linear squared errors as follows,

$$\begin{aligned}
&\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{f}_t(x_t,a_t)-\frac{Q_t}{V}\hat{g}_t(x_t,a_t))-(f(x_t,a_t)-\frac{Q_t}{V}g(x_t,a_t))|\mathcal{H}_t\right]\\
=&\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{f}_t(x_t,a_t)-f(x_t,a_t))+\frac{Q_t}{V}(g(x_t,a_t)-\hat{g}_t(x_t,a_t))|\mathcal{H}_t\right],\\
\leq&\frac{1}{2\gamma}+\frac{\gamma}{2}\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{f}_t(x_t,a_t)-f(x_t,a_t))^2|\mathcal{H}_t\right]+\frac{Q_t}{V}\frac{1}{2\gamma}+\frac{Q_t}{V}\frac{\gamma}{2}\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{g}_t(x_t,a_t)-g(x_t,a_t))^2|\mathcal{H}_t\right],
\end{aligned}$$

the last inequality comes from the fact that $ab\leq(a^2+b^2)/2$. The last term (18) is equal to

$$\begin{aligned}
&\mathbb{E}_{a\sim\pi}\left[f(x_t,a)-\frac{Q_t}{V}g(x_t,a)|\mathcal{H}_t\right]-(\hat{f}_t(x_t,\hat{a}_t)-\frac{Q_t}{V}\hat{g}_t(x_t,\hat{a}_t))\\
=&\mathbb{E}_{a\sim\pi}\left[(f(x_t,a)-\hat{f}_t(x_t,a))+\frac{Q_t}{V}(\hat{g}_t(x_t,a)-g(x_t,a))-(\hat{L}_t(x_t,\hat{a}_t)-\hat{L}_t(x_t,a))|\mathcal{H}_t\right]\\
\leq&\mathbb{E}_{a\sim\pi}\left[\frac{\gamma\pi_t(a)}{2}(\hat{f}_t(x_t,a)-f(x_t,a))^2+\frac{Q_t}{V}\frac{\gamma\pi_t(a)}{2}(\hat{g}_t(x_t,a)-g(x_t,a))^2|\mathcal{H}_t\right]\\
&-\mathbb{E}_{a\sim\pi}\left[\hat{L}_t(x_t,\hat{a}_t)-\hat{L}_t(x_t,a)|\mathcal{H}_t\right]+\mathbb{E}_{a\sim\pi}\left[\frac{1}{2\gamma\pi_t(a)}(1+\frac{Q_t}{V})|\mathcal{H}_t\right]\\
\leq&\mathbb{E}_{a\sim\pi}\left[\frac{\gamma\pi_t(a)}{2}(\hat{f}_t(x_t,a)-f(x_t,a))^2+\frac{Q_t}{V}\frac{\gamma\pi_t(a)}{2}(\hat{g}_t(x_t,a)-g(x_t,a))^2|\mathcal{H}_t\right]\\
&+\mathbb{E}_{a\sim\pi}\left[\frac{1}{2\gamma\pi_t(a)}(1+\frac{Q_t}{V}-\frac{1}{\beta_t})|\mathcal{H}_t\right]+\frac{\eta_t}{2\beta_t\gamma}\\
\leq&\frac{\gamma}{2}\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{f}_t(a_t)-f_t(a_t))^2|\mathcal{H}_t\right]+\frac{Q_t}{V}\frac{\gamma}{2}\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{g}_t(x_t,a_t)-g_t(x_t,a_t))^2|\mathcal{H}_t\right]+\frac{\eta_t}{2\beta_t\gamma}\\
\leq&\frac{\gamma}{2}\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{f}_t(a_t)-f_t(a_t))^2|\mathcal{H}_t\right]+\frac{Q_t}{V}\frac{\gamma}{2}\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{g}_t(x_t,a_t)-g_t(x_t,a_t))^2|\mathcal{H}_t\right]+\frac{A}{2\beta_t\gamma}
\end{aligned}$$

where the first inequality is by employing $ab\leq(a^2+b^2)/2$ on both reward and cost estimate errors, the second one comes from inverse gap weighting distribution (4) that satisfies

$$\hat{L}_t(x_t,\hat{a}_t)-\hat{L}_t(x_t,a)=\frac{1}{2\beta_t\gamma\pi_t(a)}-\frac{\eta_t}{2\beta_t\gamma},$$

the third inequality holds since

$$1+\frac{Q_t}{V}-\frac{1}{\beta_t}=1+\frac{Q_t}{V}-(1+\frac{Q_t}{V})=0,$$

and the last one comes from the simple fact that $\eta_t\leq A$. This fact holds since the sum of the distribution would not satisfy $\sum_a\pi_t(a)=1$ if $\eta_t>A$. Combine all these terms, we have

$$\mathbb{E}_{a\sim\pi}\left[f(x_t,a)-\frac{Q_t}{V}g(x_t,a)|\mathcal{H}_t\right]-\mathbb{E}_{a_t\sim\pi_t}\left[f(x_t,a_t)-\frac{Q_t}{V}g(x_t,a_t)|\mathcal{H}_t\right]$$

$$\leq \left(1+\frac{Q_t}{V}\right)\frac{A}{\gamma}+\gamma\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{f}_t(x_t,a_t)-f(x_t,a_t))^2|\mathcal{H}_t\right]+\gamma\frac{Q_t}{V}\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{g}_t(x_t,a_t)-g(x_t,a_t))^2|\mathcal{H}_t\right].$$

Recall the virtual queue update $Q_{t+1}=\max(Q_t+c_t(x_t,a_t),0)$ and the definition that $\Delta_t=\frac{1}{2}Q_{t+1}^2-\frac{1}{2}Q_t^2$, we have

$$\begin{aligned}\mathbb{E}_{a_t\sim\pi_t}[\Delta_t|\mathcal{H}_t]=&\mathbb{E}_{a_t\sim\pi_t}\left[\frac{1}{2}Q_{t+1}^2|\mathcal{H}_t\right]-\mathbb{E}_{a_t\sim\pi_t}\left[\frac{1}{2}Q_t^2|\mathcal{H}_t\right]\\=&\mathbb{E}_{a_t\sim\pi_t}\left[\frac{1}{2}\left(\max(Q_t+c_t(x_t,a_t),0)\right)^2|\mathcal{H}_t\right]-\mathbb{E}_{a_t\sim\pi_t}\left[\frac{1}{2}Q_t^2|\mathcal{H}_t\right]\\\leq&\mathbb{E}_{a_t\sim\pi_t}\left[Q_tc_t(x_t,a_t)|\mathcal{H}_t\right]+\frac{1}{2}\mathbb{E}_{a_t\sim\pi_t}[c_t^2(x_t,a_t)|\mathcal{H}_t]\\\leq&\mathbb{E}_{a_t\sim\pi_t}\left[Q_tc_t(x_t,a_t)|\mathcal{H}_t\right]+\frac{G^2}{2},\end{aligned}$$

where the first inequality holds since $(Q_t+c_t(x_t,a_t))^2\geq 0$ and the second inequality comes from Assumption 1. Take expectation w.r.t. the feedback function and rearrange these terms, we have

$$\mathbb{E}\left[\Delta_t|\mathcal{H}_t\right]-\frac{G^2}{2}\leq\mathbb{E}\left[Q_tg(x_t,a_t)|\mathcal{H}_t\right].$$

Combining these facts and finally, we complete the proof that:

$$\begin{aligned}&\mathbb{E}_{a\sim\pi}\left[f(x_t,a)|\mathcal{H}_t\right]-\mathbb{E}_{a_t\sim\pi_t}\left[f(x_t,a_t)|\mathcal{H}_t\right]-\mathbb{E}_{a\sim\pi}\left[\frac{Q_t}{V}g(x_t,a)|\mathcal{H}_t\right]+\frac{\mathbb{E}_{a_t\sim\pi_t}\left[\Delta_t|\mathcal{H}_t\right]}{V}-\frac{G^2}{2V}\\\leq&\left(1+\frac{Q_t}{V}\right)\frac{A}{\gamma}+\gamma\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{f}_t(x_t,a_t)-f(x_t,a_t))^2|\mathcal{H}_t\right]+\gamma\frac{Q_t}{V}\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{g}_t(x_t,a_t)-g(x_t,a_t))^2|\mathcal{H}_t\right],\end{aligned}$$

Appendix B. Proof of Lemma 5

In this section, we establish a high probability bound for the virtual queue under the relaxed Slater condition. Assumption 3 suggests the existence of more robust feasible points, thus leading to an improved theoretical guarantee for cumulative violation of the CCB algorithms. However, LagrangeCBwLC Slivkins et al. (2023) requires the exact information about the key parameter ε for adjusting its learning rate. In contrast, LOE2D does not require any adjustments and can automatically detect the presence of a strong feasible point through the virtual queue update. We will prove it through the following multi-step Lyapunov drift analysis.

We first employ Lemma 4, which provides the single-step drift bound:

$$\begin{aligned}&\mathbb{E}\left[\Delta_t|\mathcal{H}_t\right]\\\leq&V\mathbb{E}_{a_t\sim\pi}\left[f(x_t,a_t)|\mathcal{H}_t\right]-V\mathbb{E}_{a\sim\pi}\left[f(x_t,a)|\mathcal{H}_t\right]+\mathbb{E}_{a\sim\pi}\left[Q_tg(x_t,a)|\mathcal{H}_t\right]+\frac{G^2}{2}+\left(1+\frac{Q_t}{V}\right)\frac{VA}{\gamma}\\&+V\gamma\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{f}_t(x_t,a_t)-f(x_t,a_t))^2|\mathcal{H}_t\right]+\gamma Q_t\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{g}_t(x_t,a_t)-g(x_t,a_t))^2|\mathcal{H}_t\right]\\\leq&2FV+\mathbb{E}_{a\sim\pi}\left[Q_tg(x_t,a)|\mathcal{H}_t\right]+\frac{G^2}{2}+\left(1+\frac{Q_t}{V}\right)\frac{VA}{\gamma}\\&+V\gamma\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{f}_t(x_t,a_t)-f(x_t,a_t))^2|\mathcal{H}_t\right]+\gamma Q_t\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{g}_t(x_t,a_t)-g(x_t,a_t))^2|\mathcal{H}_t\right],\end{aligned}$$

To calculate the multi-step drift bound, we begin by determining the conditional expectation of the drift term after K rounds from the above inequality:

$$\begin{aligned}
&\mathbb{E}\left[\Delta_{t+K}|\mathcal{H}_t\right] \\
=&\mathbb{E}\left[\mathbb{E}\left[\Delta_{t+K}|\mathcal{H}_{t+K}\right]|\mathcal{H}_t\right] \\
\leq&2FV+\frac{G^2}{2}+\frac{VA}{\gamma}+\mathbb{E}\left[\mathbb{E}_{a\sim\pi}\left[Q_{t+K}g(x_{t+K},a)|\mathcal{H}_{t+K}\right]|\mathcal{H}_t\right] \\
&+\gamma V\mathbb{E}\left[\mathbb{E}\left[(\hat{f}_{t+K}(x_{t+K},a_{t+K})-f(x_{t+K},a_{t+K}))^2|\mathcal{H}_{t+K}\right]|\mathcal{H}_t\right] \\
&+\frac{A}{\gamma}\mathbb{E}[Q_{t+K}|\mathcal{H}_t]+\gamma\mathbb{E}\left[Q_{t+K}\mathbb{E}\left[(\hat{g}_{t+K}(x_{t+K},a_{t+K})-g(x_{t+K},a_{t+K}))^2|\mathcal{H}_{t+K}\right]|\mathcal{H}_t\right]
\end{aligned}$$

From Assumption 3, we know that there exists a policy distribution π_0 that satisfies $\mathbb{E}_{a\sim\pi_0}[g(x,a)]\leq -\varepsilon$. Let $\pi=\pi_0$, we have

$$\mathbb{E}\left[\mathbb{E}_{a\sim\pi_0}\left[Q_{t+K}g(c_{t+K},a)|\mathcal{H}_{t+K}\right]|\mathcal{H}_t\right]\leq-\varepsilon\mathbb{E}[Q_{t+K}|\mathcal{H}_t].$$

The virtual queue update rule indicates that

$$Q_{t+K}=\max(Q_{t+K-1}+c_{t+K-1}(x_{t+K-1},a_{t+K-1}),0)\leq Q_{t+K-1}+G\leq Q_t+GK,$$

Integrating these two key facts back into the drift bound equation, we obtain:

$$\begin{aligned}
&\mathbb{E}\left[\Delta_{t+K}|\mathcal{H}_t\right] \\
\leq&2FV+\frac{G^2}{2}+\frac{VA}{\gamma}+(\frac{A}{\gamma}-\varepsilon)\mathbb{E}[Q_{t+K}|\mathcal{H}_t] \\
&+\gamma V\mathbb{E}\left[\mathbb{E}\left[(\hat{f}_{t+K}(x_{t+K},a_{t+K})-f(x_{t+K},a_{t+K}))^2|\mathcal{H}_{t+K}\right]|\mathcal{H}_t\right] \\
&+\gamma\mathbb{E}\left[Q_{t+K}\mathbb{E}\left[(\hat{g}_{t+K}(x_{t+K},a_{t+K})-g(x_{t+K},a_{t+K}))^2|\mathcal{H}_{t+K}\right]|\mathcal{H}_t\right] \\
\leq&2FV+\frac{G^2}{2}+\frac{VA}{\gamma}+(\frac{A}{\gamma}-\varepsilon)\mathbb{E}[Q_t+GK|\mathcal{H}_t] \\
&+\gamma V\mathbb{E}\left[\mathbb{E}\left[(\hat{f}_{t+K}(x_{t+K},a_{t+K})-f(x_{t+K},a_{t+K}))^2|\mathcal{H}_{t+K}\right]|\mathcal{H}_t\right] \\
&+\gamma\mathbb{E}\left[(Q_t+GK)\mathbb{E}\left[(\hat{g}_{t+K}(x_{t+K},a_{t+K})-g(x_{t+K},a_{t+K}))^2|\mathcal{H}_{t+K}\right]|\mathcal{H}_t\right] \\
\leq&2FV+\frac{VA}{\gamma}+\frac{G^2}{2}+\gamma V\mathbb{E}\left[\mathbb{E}\left[(\hat{f}_{t+K}(x_{t+K},a_{t+K})-f(x_{t+K},a_{t+K}))^2|\mathcal{H}_{t+K}\right]|\mathcal{H}_t\right] \\
&+\left(\frac{A}{\gamma}-\varepsilon\right)GK+\gamma GK\mathbb{E}\left[\mathbb{E}\left[(\hat{g}_{t+K}(x_{t+K},a_{t+K})-g(x_{t+K},a_{t+K}))^2|\mathcal{H}_{t+K}\right]|\mathcal{H}_t\right] \\
&+\left(\frac{A}{\gamma}-\varepsilon+\gamma\mathbb{E}\left[\mathbb{E}\left[(\hat{g}_{t+K}(x_{t+K},a_{t+K})-g(x_{t+K},a_{t+K}))^2|\mathcal{H}_{t+K}\right]|\mathcal{H}_t\right]\right)\mathbb{E}[Q_t|\mathcal{H}_t] \\
\leq&2FV+\frac{VA}{\gamma}+\frac{G^2}{2}+\gamma V\mathbb{E}\left[\mathbb{E}\left[(\hat{f}_{t+K}(x_{t+K},a_{t+K})-f(x_{t+K},a_{t+K}))^2|\mathcal{H}_{t+K}\right]|\mathcal{H}_t\right] \\
&+\gamma GK\mathbb{E}\left[\mathbb{E}\left[(\hat{g}_{t+K}(x_{t+K},a_{t+K})-g(x_{t+K},a_{t+K}))^2|\mathcal{H}_{t+K}\right]|\mathcal{H}_t\right]
\end{aligned}$$

$$+\left(\frac{A}{\gamma}-\varepsilon+\gamma\mathbb{E}\left[\mathbb{E}\left[(\hat{g}_{t+K}(x_{t+K},a_{t+K})-g(x_{t+K},a_{t+K}))^2|\mathcal{H}_{t+K}\right]|\mathcal{H}_t\right]\right)\mathbb{E}[Q_t|\mathcal{H}_t],$$

where the last inequality follows by the fact $\varepsilon \geq \sqrt{U/T} = A/\gamma$. Then we can get the multi-step drift bound by summing the above inequality over the interval $[t, t+K-1]$ and multiplying both sides by two, we obtain

$$\begin{aligned}
&\mathbb{E}\left[Q_{t+K}^2 - Q_t^2|\mathcal{H}_t\right] \\
\leq&\left(4FV+\frac{2VA}{\gamma}+G^2\right)K+2\gamma V\mathbb{E}\left[\sum_{s=t}^{t+K-1}\mathbb{E}\left[(\hat{f}_s(x_s,a_s)-f(x_s,a_s))^2|\mathcal{H}_s\right]|\mathcal{H}_t\right] \\
&+2\gamma GK\mathbb{E}\left[\sum_{s=t}^{t+K-1}\mathbb{E}\left[(\hat{g}_s(x_s,a_s)-g(x_s,a_s))^2|\mathcal{H}_s\right]|\mathcal{H}_t\right] \\
&+2\left(\frac{AK}{\gamma}-\varepsilon K+\gamma\mathbb{E}\left[\sum_{s=t}^{t+K-1}\mathbb{E}\left[(\hat{g}_s(x_s,a_s)-g(x_s,a_s))^2|\mathcal{H}_s\right]|\mathcal{H}_t\right]\right)\mathbb{E}[Q_t|\mathcal{H}_t] \\
\leq&(4FV+\frac{2VA}{\gamma}+G^2)K+2\gamma VU_f+2\gamma GKU_g-2\left[(\varepsilon-\frac{A}{\gamma})K-\gamma U_g\right]\mathbb{E}[Q_t|\mathcal{H}_t].
\end{aligned}$$

The first inequality is derived from the telescoping sum

$$\sum_{s=t}^{t+K-1}\Delta_s = \frac{Q_{t+K}^2}{2}-\frac{Q_t^2}{2},$$

the second one holds since $[t, t+K-1]$ falls within the total time horizon T, making these terms smaller than the oracles error defined by Assumption 2. With $\delta = \varepsilon - A/\gamma - \gamma U/K$ and let $C_0 = 4F + 2A + G^2 + 2 + 2G$, we have

$$\mathbb{E}\left[Q_{t+K}^2 - Q_t^2|\mathcal{H}_t\right] \leq -2K\delta Q_t + C_0(KV + \gamma VU + \gamma KU).$$

By setting $K = 2\gamma U/\varepsilon$, it follows that $\delta = \varepsilon - A/\gamma - \gamma U/K = \varepsilon/2 - \sqrt{U/T} \geq \varepsilon/4$. Next, we introduce a key lemma derived from Lemma 5 in Yu et al. (2017) to prove a high probability bound for Q_t.

Lemma 7 *Let $S(t)$ be a discrete-time stochastic process adapted to a filtration $\mathcal{F}(t)$. Suppose there exists an integer $K \geq 0$, real constant $\theta \in \mathbb{R}$, $\delta_{max} \geq 0$ and $0 \leq \zeta \leq \delta_{max}$,*

$$\begin{aligned}
|S(t+1) - S(t)| \leq& \delta_{max}, \\
\mathbb{E}\left[S(t+K) - S(t)|\mathcal{F}(t)\right] \leq& \begin{cases} K\delta_{max}, \text{ when } S(t) < \theta \\ -K\zeta, \text{ when } S(t) \geq \theta \end{cases},
\end{aligned}$$

hold $\forall t \in [T]$, Then the following holds,

- $\mathbb{E}[S(t)] \leq \theta + K\frac{4\delta_{max}^2}{\zeta}\log\left(1+\frac{8\delta_{max}^2}{\zeta^2}e^{\zeta/(4\delta_{max})}\right),\ \forall t \in [T].$
- *For any $\mu \in [0,1]$, we have $\mathbb{P}(S(t) \geq s) \leq \mu$, where $s = \theta + K\frac{4\delta_{max}^2}{\zeta}\log\left(1+\frac{8\delta_{max}^2}{\zeta^2}e^{\zeta/(4\delta_{max})}\right) + K\frac{4\delta_{max}^2}{\zeta}\log(\frac{1}{\mu})$.*

Based on the condition that $\mathbb{E}\left[Q_{t+K}^2 - Q_t^2 | \mathcal{H}_t\right] \leq -2K\delta Q_t + C_0(KV + \gamma VU + \gamma KU)$, we can get the steady state of the drift process. We then convert this into a form suitable for analyzing the virtual queue:

$$\begin{aligned}\mathbb{E}\left[Q_{t+K}^2 | \mathcal{H}_t\right] \leq & Q_t^2 - 2K\delta Q_t + C_0(KV + \gamma VU + \gamma KU) \\ \leq & Q_t^2 - K\delta Q_t + (C_0(KV + \gamma VU + \gamma KU) - K\delta Q_t)\end{aligned}$$

Then we can set $\theta = \frac{K\delta}{2} + \frac{C_0(KV+\gamma VU+\gamma KU)}{K\delta}$ and assume $Q_t \geq \theta$ to get

$$\begin{aligned}\mathbb{E}\left[Q_{t+K}^2 | \mathcal{H}_t\right] \leq & Q_t^2 - K\delta Q_t - \frac{K^2\delta^2}{2} \\ \leq & (Q_t - \frac{K\delta}{2})^2\end{aligned}$$

From the fact that $Q_t \geq \theta \geq \frac{K\delta}{2}$, we then prove the following multi-step virtual queue stability by taking square root on both sides and applying Jensen's inequality,

$$\mathbb{E}\left[Q_{t+K} | \mathcal{H}_t\right] \leq \sqrt{\mathbb{E}\left[Q_{t+K}^2 | \mathcal{H}_t\right]} \leq \mathbb{E}\left[Q_t | \mathcal{H}_t\right] - \frac{\delta K}{2}.$$

To get $|\|Q_{t+1}\| - \|Q_t\|| \leq G$, recall the virtual queue update rule that

$$\|Q_{t+1}\| \leq \|Q_t + c_t(x_t, a_t)\| \leq \|Q_t\| + \|c_t(x_t, a_t)\| \leq \|Q_t\| + G,$$

Meanwhile, from the fact that $Q_t \geq 0$, we have

$$\|Q_{t+1} - Q_t\| \leq \|c_t(x_t, a_t)\| \leq G,$$

which gives $\|Q_{t+1}\| \geq \|Q_t\| - G$ by the triangle inequality of norms. This directly leads to the result that $\mathbb{E}\left[|\|Q_{t+K}\| - \|Q_t\||\right] \leq GK$. By choosing $K = 2\gamma U/\varepsilon$ and considering $\varepsilon \geq 4\sqrt{T/U}$, we ensure that $\varepsilon \geq \delta \geq \varepsilon/4$ is non-negative. Then we can let $\mu = 1/T^2$ and apply Lemma 7 to show that with probability at least $1 - 1/T^2$,

$$\begin{aligned}Q_t \leq & \frac{K\delta}{2} + \frac{C_0(KV + \gamma VU + \gamma KU)}{K\delta} + K\frac{4G^2}{\delta}\log\left(1 + \frac{8G^2}{\delta^2}e^{\delta/(4G)}\right) + 2K\frac{4G^2}{\delta}\log(T) \\ \leq & \frac{K^2\delta^2 + C_0(KV + \gamma VU + \gamma KU) + 12G^2K^2\log(1 + 16G^2T)}{K\delta} \\ \leq & 2A\sqrt{TU} + C_0\left(\frac{4\sqrt{TU\log T}}{\varepsilon} + \frac{A\sqrt{TU\log T}}{8} + \frac{A\sqrt{TU}}{4\varepsilon}\right) + \frac{96G^2\sqrt{TU}\log(1 + 16G^2T)}{\varepsilon^2} \\ \leq & 2A\sqrt{TU} + (5C_0A + 96G^2)\log(1 + 16G^2T)\sqrt{TU}/\varepsilon^2 \\ \leq & (2A + (5C_0A + 96G^2)\log(1 + 16G^2T))\sqrt{TU}/\varepsilon^2 \\ \leq & C_1(1 + \log(1 + 16G^2T))\sqrt{TU}/\varepsilon^2,\end{aligned}$$

where the second inequality holds since $e^{\delta/(4G)} \leq 2$ and $\delta \geq \varepsilon/4 \geq \sqrt{U/T}$, the third inequality arises from the fact that $V = \sqrt{TU\log T}$ and $\gamma = A\sqrt{T/U}$, the last inequality is valid from the constant definition where $C_1 = 2A + 5C_0A + 96G^2$.

Appendix C. Proof of Lemma 6

In this section, we establish the anytime virtual queue bound in two CCB cases: the general CCB setting and CCB with the Slater condition. For the case without the Slater condition, the virtual queue bound can be derived by analyzing the drift term through Lemma 4, which provides a unified "regret plus drift" analysis. This virtual length bound would ensure the worst-case theoretical guarantee of LOE2D. Under a relaxed Slater condition, an improved bound for the virtual queue is achievable through a multi-step Lyapunov drift analysis, as demonstrated in Lemma 5.

General CCB Setting We begin with the proof without the Slater condition. Multiply both sides by V in Lemma 4 and set $\pi = \pi^*$, we have

$$
\begin{aligned}
&\mathbb{E}\left[\Delta_t|\mathcal{H}_t\right] \\
\leq& V\mathbb{E}_{a_t\sim\pi_t}\left[f(x_t,a_t)|\mathcal{H}_t\right] - V\mathbb{E}_{a\sim\pi^*}\left[f(x_t,a)|\mathcal{H}_t\right] + \mathbb{E}_{a\sim\pi^*}\left[Q_t g(x_t,a)|\mathcal{H}_t\right] + \frac{G^2}{2} + \left(1+\frac{Q_t}{V}\right)\frac{VA}{\gamma} \\
&+ V\gamma\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{f}_t(x_t,a_t) - f(x_t,a_t))^2|\mathcal{H}_t\right] + \gamma Q_t\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{g}_t(x_t,a_t) - g(x_t,a_t))^2|\mathcal{H}_t\right] \\
\leq& V\mathbb{E}_{a_t\sim\pi_t}\left[f(x_t,a_t)|\mathcal{H}_t\right] - V\mathbb{E}_{a\sim\pi^*}\left[f(x_t,a)|\mathcal{H}_t\right] + \frac{G^2}{2} + \left(1+\frac{Q_t}{V}\right)\frac{VA}{\gamma} \\
&+ V\gamma\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{f}_t(x_t,a_t) - f(x_t,a_t))^2|\mathcal{H}_t\right] + \gamma Q_t\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{g}_t(x_t,a_t) - g(x_t,a_t))^2|\mathcal{H}_t\right] \\
\leq& 2FV + \frac{G^2}{2} + \left(1+\frac{Q_t}{V}\right)\frac{VA}{\gamma} + V\gamma\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{f}_t(x_t,a_t) - f(x_t,a_t))^2|\mathcal{H}_t\right] \\
&+ \gamma Q_t\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{g}_t(x_t,a_t) - g(x_t,a_t))^2|\mathcal{H}_t\right], \qquad (19)
\end{aligned}
$$

the second inequality comes from the optimal solution definition, the last inequality holds due to Assumption 1. By summing this inequality over t and considering the fact that $\sum_{s=1}^{t}\Delta_s = (Q_{t+1}^2 - Q_1^2)/2 = Q_{t+1}^2/2$, we have

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{2}Q_{t+1}^2|\mathcal{H}_t\right] \leq& 2FVt + \frac{G^2t}{2} + \frac{VAt}{\gamma} + \gamma V\sum_{s=1}^{t}\mathbb{E}\left[\sum_{s=1}^{t}(\hat{f}_s(x_s,a_s) - f(x_s,a_s))^2|\mathcal{H}_1,\ldots,\mathcal{H}_t\right] \\
&+ \gamma\mathbb{E}\left[\sum_{s=1}^{t}Q_s(\hat{g}_s(x_s,a_s) - g(x_s,a_s))^2|\mathcal{H}_1,\ldots,\mathcal{H}_t\right] + \frac{A}{\gamma}\sum_{s=1}^{t}Q_s.
\end{aligned}
$$

Recall the update rule of Q_t and the upper bound G for costs, we can derive the worst-case bound for Q_t that

$$
Q_{t+1} = \max(Q_t + c_t(x_t,a_t), 0) \leq Q_t + |c_t(x_t,a_t)| \leq Q_t + G \leq tG.
$$

Combine the above inequalities and take expectations with historical information, we have

$$
\begin{aligned}
\mathbb{E}\left[\frac{1}{2}Q_{t+1}^2\right] \leq& 2FVt + \frac{G^2t}{2} + \frac{VAt}{\gamma} + \frac{A}{\gamma}tTG + \gamma V\mathbb{E}\left[\sum_{s=1}^{t}(\hat{f}_s(x_s,a_s) - f(x_s,a_s))^2\right] \\
&+ \gamma tG\mathbb{E}\left[\sum_{s=1}^{t}(\hat{g}_s(x_s,a_s) - g(x_s,a_s))^2\right]
\end{aligned}
$$

$$
\begin{aligned}
\leq & 2FVT + \frac{G^2T}{2} + \frac{VAT}{\gamma} + \frac{A}{\gamma}T^2G + \gamma V\mathbb{E}\left[\sum_{t=1}^{T}(\hat{f}_t(x_t,a_t) - f(x_t,a_t))^2\right] \\
& + \gamma TG\mathbb{E}\left[\sum_{t=1}^{T}(\hat{g}_t(x_t,a_t) - g(x_t,a_t))^2\right] \\
\leq & 2FVT + \frac{G^2T}{2} + \frac{VAT}{\gamma} + \frac{A}{\gamma}T^2G + \gamma VU_f + \gamma TGU_g,
\end{aligned}
$$

where the second inequality comes from replacing t with T since $t \leq T$, and the last one comes from the assumption of regression oracles. Take $V = \sqrt{TU\log T}$ and $\gamma = \sqrt{T/U}$ into the above inequality, we have for any $t \in [T]$ such that

$$
\mathbb{E}\left[Q_{t+1}^2\right] \leq \hat{B}_Q := (4FT^{\frac{3}{2}}U^{\frac{1}{2}} + 2T + 2ATU)(\log T)^{\frac{1}{2}} + G^2T + 2GT^{\frac{3}{2}}U^{\frac{1}{2}} + 2AGT^{\frac{3}{2}}U^{\frac{1}{2}}.
$$

We can obtain that $\hat{B}_Q = O(T^{\frac{3}{2}}U^{\frac{1}{2}}(\log T)^{\frac{1}{2}})$. The above inequality provides an anytime bound that $\mathbb{E}\left[Q_t\right] = \tilde{O}(T^{\frac{3}{4}}U^{\frac{1}{4}})$, $\forall t \in [T]$, which substantiates the $\tilde{O}(T^{\frac{3}{4}}U^{\frac{1}{4}})$ result of Lemma 6. Specifically, we prove that

$$
\mathbb{E}\left[Q_{t+1}\right] \leq \sqrt{\hat{B}_Q} \leq C_2T^{\frac{3}{4}}U^{\frac{1}{4}}(\log T)^{\frac{1}{4}},
$$

where $C_2 = \sqrt{4F + 2 + 2A + G^2 + 2G + 2AG}$. This result ensures the worst-case virtual queue bound of LOE2D.

Appendix D. Proof of CCB Regret and Violation in Theorem 1

In this section, we present a detailed version of the proof for both regret and violation in the general CCB. In both cases, with and without the Slater condition, we offer a comprehensive analysis framework for CCB. The following analysis will demonstrate that both regret and violation bounds can be effectively determined through the Lyapunov drift virtual queue results discussed in the previous section. We begin with the regret analysis:

Regret bound. Take expectation with historical information $\mathcal{H}_t$ and let $\pi = \pi^*$ in Lemma 4, we can rearrange the inequality and sum it over T to obtain:

$$
\begin{aligned}
& \sum_{t=1}^{T}\mathbb{E}_{a\sim\pi^*}\left[f(x_t,a)\right] - \sum_{t=1}^{T}\mathbb{E}_{a_t\sim\pi_t}\left[f(x_t,a_t)\right] \\
\leq & \gamma\mathbb{E}_{a_t\sim\pi_t}\left[\sum_{t=1}^{T}(\hat{f}_t(x_t,a_t) - f(x_t,a_t))^2\right] + \gamma\sum_{t=1}^{T}\frac{Q_t}{V}\mathbb{E}_{a_t\sim\pi_t}\left[(\hat{g}_t(x_t,a_t) - g(x_t,a_t))^2\right] \\
& + \frac{G^2T}{2V} - \frac{\mathbb{E}\left[\sum_{t=1}^{T}\Delta_t\right]}{V} + \sum_{t=1}^{T}\left(1 + \frac{Q_t}{V}\right)\frac{A}{\gamma} \\
\leq & \gamma\mathbb{E}_{a_t\sim\pi_t}\left[\sum_{t=1}^{T}(\hat{f}_t(x_t,a_t) - f(x_t,a_t))^2\right] + \gamma\frac{Q_{max}}{V}\mathbb{E}_{a_t\sim\pi_t}\left[\sum_{t=1}^{T}(\hat{g}_t(x_t,a_t) - g(x_t,a_t))^2\right]
\end{aligned}
$$

$$
\begin{aligned}
&+\frac{G^2T}{2V}-\frac{\mathbb{E}\left[\sum_{t=1}^{T}\Delta_t\right]}{V}+\left(1+\frac{Q_{max}}{V}\right)\frac{AT}{\gamma}\\
\leq&\left(1+\frac{Q_{max}}{V}\right)\frac{AT}{\gamma}+\gamma U_f+\gamma\frac{Q_{max}}{V}U_g+\frac{G^2T}{2V}-\frac{\mathbb{E}\left[\sum_{t=1}^{T}\Delta_t\right]}{V}\\
\leq&\frac{AT}{\gamma}+\gamma U+\frac{G^2T}{2V}+(\frac{AT}{\gamma V}+\frac{\gamma U}{V})Q_{max},
\end{aligned}
$$

where the first inequality holds comes from the property of optimal policy distribution, the second inequality follows by definition that $Q_{max}=\max_{t\in[T]}[Q_t]$, the third one directly use the regression oracles assumptions and the last one holds since

$$
-\sum_{t=1}^{T}\Delta_t=(Q_1^2-Q_{T+1}^2)/2\leq Q_1^2/2=0.
$$

Then substitute the learning rate setup we have

$$
\begin{aligned}
\mathcal{R}(T)\leq&\sum_{t=1}^{T}\mathbb{E}_{a\sim\pi^*}\left[f(x_t,a)\right]-\sum_{t=1}^{T}\mathbb{E}_{a_t\sim\pi_t}\left[f(x_t,a_t)\right]\\
\leq&\sqrt{TU}+A\sqrt{TU}+\frac{G^2}{2}\sqrt{\frac{T}{U\log T}}+(A+1)\frac{Q_{max}}{\sqrt{\log T}}\\
\leq&(2A+G^2)\sqrt{TU}+2A\frac{Q_{max}}{\sqrt{\log T}}.
\end{aligned}
$$

Since we have already proven the anytime virtual queue bound in Lemma 6: in general CCB, $\mathbb{E}[Q_t]=\tilde{O}(T^{\frac{3}{4}}U^{\frac{1}{4}}),\ \forall t\in[T]$; in CCB with the relaxed Slater condition, we have the high probability bound that $Q_t=\tilde{O}(\min(\sqrt{TU}/\varepsilon^2,T^{\frac{3}{4}}U^{\frac{1}{4}}))$. We prove CCB regret bound in Theorem 1. For the general setting, we have

$$
\mathcal{R}(T)\leq(2A+G^2)\sqrt{TU}+2AC_2T^{\frac{3}{4}}U^{\frac{1}{4}}.
$$

Under the relaxed Slater condition, we have

$$
\mathcal{R}(T)\leq(2A+G^2)\sqrt{TU}+2A\min\left\{C_1\left(1+\log(1+16G^2T)\right)\frac{\sqrt{TU}}{\varepsilon^2\sqrt{\log T}},C_2T^{\frac{3}{4}}U^{\frac{1}{4}}\right\}.
$$

Violation bound. Recall the virtual queue update $Q_{t+1}=\max(Q_t+c_t(x_t,a_t),0)$, we have

$$
Q_{t+1}\geq Q_t+c_t(x_t,a_t),
$$

Then we obtain that the bound of cumulative constraint violation can be directly achieved through the length of the virtual queue:

$$
\sum_{t=1}^{T}c_t(x_t,a_t)\leq Q_{T+1}
$$

Taking expectations on both sides, we have $\mathcal{V}(T) = \mathbb{E}\left[\sum_{t=1}^{T} g(x_t, a_t)\right] \leq \mathbb{E}[Q_{T+1}]$, for general setting we have

$$\mathcal{V}(T) \leq C_2 T^{\frac{3}{4}} U^{\frac{1}{4}} (\log T)^{\frac{1}{4}}.$$

Under the relaxed Slater condition, we have

$$\mathcal{V}(T) \leq \min\left\{C_1(1+\log(1+16G^2T))\sqrt{TU}/\varepsilon^2, C_2 T^{\frac{3}{4}} U^{\frac{1}{4}} (\log T)^{\frac{1}{4}}\right\}.$$

Finally, we can derive the violation bound through the anytime virtual queue length bound we proved before. Through the above analysis, we also establish the anytime violation bound $\mathcal{V}(t) \leq \mathbb{E}[Q_{t+1}]$, which highlights the strong adaptability of LOE2D in recognizing constraint violations at all times.

Appendix E. Proof of CBwK Regret in Theorem 1

As we discussed before, CBwK represents a specific variant of CCB with a hard-stopping. We present the theoretical guarantees of LOE2D for CBwK. We denote a_0 as the null arm in CBwK. The total budget is B and the budget per round is $b := B/T$. The CBwK regret is defined as

$$\mathcal{R}(T) := T\nu_b^* - \mathbb{E}\left[\sum_{t=1}^{\tau} f(x_t, a_t)\right],$$

where τ denotes the stopping time when the first $\sum_{t=1}^{\tau} c(x_t, a_t) \geq B$ holds. Let π_b^* and ν_b^* be the optimal policy and value for the relaxed problem that

$$\max_{\pi} \ \mathbb{E}_{a\sim\pi}\left[f(x,a)\right] \quad \text{s.t. } \mathbb{E}_{a\sim\pi}\left[c(x,a)\right] \leq b.$$

Intuitively, the CBwK regret can be decoupled as:

$$\mathcal{R}(T) = T\nu_b^* - \mathbb{E}\left[\sum_{t=1}^{\tau} f(x_t, a_t)\right] \tag{20}$$

$$+ (T-\tau)\nu_b^*. \tag{21}$$

The first term (20) denotes the CCB regret within τ rounds and can be easily obtained following our previous regret analysis. The second term (21) represents the loss incurred due to the hard stopping. This can be proved through the virtual queue analysis, which distinguishes our work from Han et al. (2023); Slivkins et al. (2023). We first explain the algorithm details for CBwK.

$$\hat{L}_t(x_t, a) = \hat{f}_t(x_t, a) - \frac{Q_t}{V}\hat{g}_t(x_t, a),$$
$$Q_{t+1} = \max\left(Q_t + c_t(x_t, a_t) - b, 0\right).$$

Then we can establish the proof of the CBwK regret in Theorem 1.

Proof We first prove the CCB regret within τ rounds in (20). Based on Lemma 4, we can obtain the following inequality by taking expectation w.r.t. $\mathcal{H}_t$ and let $\pi = \pi_b^*$:

$$\sum_{t=1}^{\tau} \mathbb{E}_{a\sim\pi_b^*}\left[f(x_t, a)\right] - \sum_{t=1}^{\tau} \mathbb{E}_{a_t\sim\pi_t}\left[f(x_t, a_t)\right]$$

$$
\begin{aligned}
\leq & \gamma \mathbb{E}_{a_t \sim \pi_t}\left[\sum_{t=1}^{\tau}(\hat{f}_t(x_t,a_t)-f(x_t,a_t))^2\right]+\gamma \mathbb{E}_{a_t \sim \pi_t}\left[\sum_{t=1}^{\tau}\frac{Q_t}{V}(\hat{g}_t(x_t,a_t)-g(x_t,a_t))^2\right] \\
& +\mathbb{E}_{a\sim\pi_b^*}\left[\sum_{t=1}^{\tau}\frac{Q_t}{V}g(x_t,a)\right]+\frac{G\tau}{2V}-\frac{\mathbb{E}\left[\sum_{t=1}^{\tau}\Delta_t\right]}{V}+\sum_{t=1}^{\tau}\left(1+\frac{Q_t}{V}\right)\frac{A}{\gamma} \\
\leq & \gamma \mathbb{E}_{a_t \sim \pi_t}\left[\sum_{t=1}^{\tau}(\hat{f}_t(x_t,a_t)-f(x_t,a_t))^2\right]+\gamma \mathbb{E}_{a_t \sim \pi_t}\left[\sum_{t=1}^{\tau}\frac{Q_t}{V}(\hat{g}_t(x_t,a_t)-g(x_t,a_t))^2\right]+\frac{G\tau}{2V} \\
& -\frac{\mathbb{E}\left[\sum_{t=1}^{\tau}\Delta_t\right]}{V}+\sum_{t=1}^{\tau}\left(1+\frac{Q_t}{V}\right)\frac{A}{\gamma} \\
\leq & \gamma \mathbb{E}_{a_t \sim \pi_t}\left[\sum_{t=1}^{\tau}(\hat{f}_t(x_t,a_t)-f(x_t,a_t))^2\right]+\frac{\gamma Q_{max}}{V}\mathbb{E}_{a_t \sim \pi_t}\left[\sum_{t=1}^{\tau}(\hat{g}_t(x_t,a_t)-g(x_t,a_t))^2\right]+\frac{G\tau}{2V} \\
& -\frac{\mathbb{E}\left[\sum_{t=1}^{\tau}\Delta_t\right]}{V}+\left(1+\frac{Q_{max}}{V}\right)\frac{A\tau}{2\gamma} \\
\leq & \gamma U_f+\frac{\gamma Q_{max}}{V}U_g+\frac{GT}{2V}-\frac{\mathbb{E}\left[\sum_{t=1}^{\tau}\Delta_t\right]}{V}+\left(1+\frac{Q_{max}}{V}\right)\frac{AT}{2\gamma} \\
\leq & \gamma U_f+\frac{\gamma Q_{max}}{V}U_g+\frac{GT}{2V}+\left(1+\frac{Q_{max}}{V}\right)\frac{AT}{2\gamma}.
\end{aligned}
$$

The first inequality is derived by summing up the inequality found in Lemma 4 across τ rounds; the second inequality holds because of the definition of π_b^*; the third one holds by the definition of Q_{max}; and the last inequality is justified because

$$
-\sum_{t=1}^{\tau}\Delta_t=(Q_1^2-Q_{\tau+1}^2)/2\leq Q_1^2/2=0.
$$

Then we can obtain that the CBwK regret of LOE2D satisfies:

$$
\begin{aligned}
\mathcal{R}(T) \leq & \gamma U_f+\frac{\gamma Q_{max}}{V}U_g+\frac{GT}{2V}+\left(1+\frac{Q_{max}}{V}\right)\frac{AT}{2\gamma}+(T-\tau)E_{a\sim\pi_b^*}\left[f(x_t,a)\right] \\
= & \frac{AT}{\gamma}+\gamma U+\frac{G^2T}{2V}+(\frac{AT}{\gamma V}+\frac{\gamma U}{V})Q_{max}+(T-\tau)v_b^*,
\end{aligned}
$$

where the above equality comes from Assumption 1. Then the key to proving regret lies in determining the skipping rounds $(T-\tau)$. According to the virtual queue update in (5), we have

$$
Q_{\tau+1}+\tau b\geq\sum_{t=1}^{\tau}c(x_t,a_t),
$$

in conjunction with the definition of stopping time, we know τ satisfies

$$
Q_{\tau+1}+\tau b\geq B:=Tb.
$$

It immediately implies

$$
\mathbb{E}[(T-\tau)\nu_b^*]\leq\frac{\nu_b^*}{b}\mathbb{E}[Q_{\tau+1}].
$$

The above inequality indicates that the skipping round length is bounded by the expected queue length. Since the null arm a_0 exists in CBwK, we can always construct a policy π_0^* where $\pi_0^*(a_0) = 1$ to satisfy the Slater condtion

$$\mathbb{E}_{a\sim\pi_0^*}[g(x_t, a)] - b \leq -b,$$

indicating that Assumption 3 holds with $\varepsilon = b$. Now we are good to apply Lemma 6 to obtain that with high probability

$$Q_t \leq O(\min\{\sqrt{TU}/\varepsilon^2, T^{\frac{3}{4}}U^{\frac{1}{4}}\}),\ \forall t \in [T],$$

which further guarantees the fact that $(T-\tau)\nu_b^* \leq \tilde{O}(\min\{\sqrt{TU}/\varepsilon^2, T^{\frac{3}{4}}U^{\frac{1}{4}}\})$ because $\nu_b^* = O(\varepsilon)$ in CBwK. By combining the bounds on (20) and (21), we have

$$\begin{aligned}
\mathcal{R}(T) \leq& \frac{AT}{\gamma} + \gamma U + \frac{G^2T}{2V} + (\frac{AT}{\gamma V} + \frac{\gamma U}{V})Q_{max} + (T-\tau)v_b^* \\
\leq& \frac{AT}{\gamma} + \gamma U + \frac{G^2T}{2V} + (\frac{AT}{\gamma V} + \frac{\gamma U}{V} + \frac{v_b^*}{b})Q_{max} \\
\leq& (2A + G^2)\sqrt{TU} + (\frac{2A}{\sqrt{\log T}} + \frac{v_b^*}{b})Q_{max} \\
\leq& (2A + G^2)\sqrt{TU} + (2A + \frac{v_b^*}{b})Q_{max} \\
\leq& (2A + G^2)\sqrt{TU} + (2A + \frac{v_b^*}{b})\min\left\{C_1(1+\log(1+16G^2T))\sqrt{TU}/\varepsilon^2, C_2T^{\frac{3}{4}}U^{\frac{1}{4}}(\log T)^{\frac{1}{4}}\right\},
\end{aligned}$$

which completes the proof. ■