IMPROVED REGRET AND VIOLATION GUARANTEES FOR CONSTRAINED CONTEXTUAL BANDITS WITH GENERAL REALIZABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper studies stochastic constrained contextual bandits under a general realizability condition. In each round, a learner observes a context, selects an action, and receives the reward and cost of the selected action. The objective is to maximize cumulative reward while ensuring that the cumulative cost remains below a prescribed threshold. In prior works, regret and violation analyses are tightly coupled, leading to regret bounds that inherit the difficulty of violation control. We propose a primal-dual learning algorithm that combines queue-normalized inverse-gap exploration with a discounted regulator. The key to this design is a regret analysis that does not require a prior bound on the dual process. Under general feasibility, for every $\theta \in [1/2, 1]$, the algorithm achieves $\widetilde{O}(T^{\theta}U^{1-\theta})$ regret and $\widetilde{O}(T^{1-\theta/2}U^{\theta/2})$ constraint violation, where U denotes the cumulative prediction error bound of the online regression oracles. When Slater's condition also holds with an unknown feasibility margin $\epsilon > 0$, we derive a novel Lyapunov analysis that improves the violation bound to $\widetilde{O}(\min\{\sqrt{TU}/\epsilon, T^{1-\theta/2}U^{\theta/2}\})$ while preserving regret guarantee. These results improve upon state-of-the-art guarantees under the same regression oracle assumption in both feasibility regimes. Experiments on learning-to-rank and constrained bidding tasks further demonstrate the effectiveness of our method.

1 INTRODUCTION

Stochastic contextual bandits provide a general framework for sequential decision-making with partial feedback and have been widely used in applications such as crowdsourcing, recommendation, and adaptive LLM routing (Yang et al., 2020; Tran-Thanh et al., 2014; Poon et al., 2026; Ong et al., 2025). In many of these applications, satisfying operational constraints is also essential to completing the task. For example, an advertising system may have a limited resource budget, an allocation platform may need to guarantee a minimum level of service, and an LLM serving system may need to route user queries while respecting constraints on inference cost or latency. These requirements naturally give rise to the constrained contextual bandit problem. Specifically, in each round, a context is drawn from a fixed but unknown distribution. After observing the context, the learner selects an action and receives only the reward and cost associated with that action. The goal is to maximize the reward while keeping the cost within a predefined threshold.

Early work on constrained bandits primarily focuses on specific types of resource or safety requirements. Bandits with knapsacks model exhaustible resources (Agrawal & Devanur, 2014; Badanidiyuru et al., 2014; 2018), fairness bandits enforce minimum allocation or service rates (Li et al., 2019; Chen et al., 2020), and conservative bandits require performance guarantees relative to a baseline policy (Kazerouni et al., 2017; Garcelon et al., 2020). While these formulations capture important applications, they typically study different types of constraints separately. This motivates a unified framework that can handle general constraints rather than a particular constraint structure (Slivkins et al., 2023; Guo & Liu, 2024). Within such a framework, both reward maximization and constraint satisfaction depend on unknown reward and cost functions, making efficient function estimation from bandit feedback a central learning challenge. Regression oracles provide a modular approach to this challenge by reducing function estimation to online prediction with guarantees on prediction error (Foster et al., 2018; Foster & Rakhlin, 2020; Han et al., 2023). By separating function estimation from decision making, it enables a modular analysis of contextual bandits with general constraints.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of algorithms for stochastic constrained contextual bandits. Here T is the horizon and U is the cumulative prediction error bound of the regression oracles. Under Slater's condition, $\epsilon > 0$ denotes the feasibility margin. The notation $\widetilde{O}(\cdot)$ suppresses logarithmic factors in T. In the table, a single centered entry indicates that regret and violation share the same bound. "Prior" specifies the additional prior information required by each method: the margin ϵ or the context distribution $\mathcal{D}$ ("–" indicates that neither is required). "Oracle" specifies the required regression guarantee: a cumulative squared prediction error bound ("Cum.") or one-sided confidence bounds in every round ("Anytime").

Algorithm	**Without Slater's condition**		**With Slater margin $\epsilon > 0$**		**Inputs**	
	Regret	Violation	Regret	Violation	Prior	Oracle
LagrangeCBwLC	$\widetilde{O}(T^{3/4}U^{1/4})$		$\widetilde{O}(\sqrt{TU}/\epsilon)$		ϵ	Cum.
LOE2D	$\widetilde{O}(T^{3/4}U^{1/4})$		$\widetilde{O}(\min\{\sqrt{TU}/\epsilon^2, T^{3/4}U^{1/4}\})$		–	Cum.
OPTIMISTIC^3	$\widetilde{O}(\sqrt{TU})$		$\widetilde{O}(\sqrt{TU})$		$\mathcal{D}$	Anytime
QUID, $\theta \in [1/2, 1]$	$\widetilde{O}(T^{\theta}U^{1-\theta})$	$\widetilde{O}(T^{1-\theta/2}U^{\theta/2})$	$\widetilde{O}(T^{\theta}U^{1-\theta})$	$\widetilde{O}(\min\{\sqrt{TU}/\epsilon, T^{1-\theta/2}U^{\theta/2}\})$	–	Cum.
QUID, $\theta = 1/2$	$\widetilde{O}(\sqrt{TU})$	$\widetilde{O}(T^{3/4}U^{1/4})$	$\widetilde{O}(\sqrt{TU})$	$\widetilde{O}(\min\{\sqrt{TU}/\epsilon, T^{3/4}U^{1/4}\})$	–	Cum.

Building on the regression oracle framework, several studies have developed algorithms with provable performance guarantees (Slivkins et al., 2024; Guo & Liu, 2024; 2025). LagrangeCBwLC in Slivkins et al. (2024) reduces the constrained problem to unconstrained contextual bandits with dual penalty, which requires prior knowledge of the Slater constant ϵ to set a proper learning rate. LOE2D in Guo & Liu (2024) removes the requirement through a Lyapunov optimization framework. It uses an unbounded virtual queue to track constraint violations and adaptively adjust the cost penalty. A key limitation of LOE2D is the tight coupling between its regret and constraint violation analyses. As a result, the regret analysis requires first bounding the queue length, and the regret bound deteriorates when constraint violation is more difficult to control. In a separate line of work, Guo et al. (2025) propose OPTIMISTIC^3, an optimistic learning algorithm that achieves near-optimal bounds on both regret and constraint violation without Slater's condition. However, it requires prior knowledge of the context distribution and assumes that regression oracles provide one-sided confidence bounds uniformly over all contexts and actions at any time. Such uniform confidence guarantees are substantially stronger than the cumulative prediction error bounds assumed by Slivkins et al. (2024) and Guo & Liu (2024), which control errors only along the realized interaction sequence. Table 1 summarizes the performance guarantees and assumptions of the main comparison methods.

The comparison in Table 1 highlights a remaining gap: stronger guarantees and weaker assumptions are not yet available simultaneously. To address the gap, we introduce **QU**eue-normalized **I**nverse-gap exploration with **D**iscounted regulator (QUID), a modular learning algorithm for stochastic constrained contextual bandits, which resolves the problem in LOE2D that regret analysis must couple with constraint violation. We further derive a refined Lyapunov drift analysis that yields a sharper bound on constraint violation under Slater's condition. The improved results are established without strengthening the regression oracle assumption or requiring additional prior information. Our main contributions are summarized as follows.

- **Algorithm Design:** QUID is a primal-dual algorithm that applies inverse-gap weighting to estimated Lagrangian scores, with novel designs for both primal action selection and dual updates. On the primal side, quadratic queue normalization adapts inverse-gap exploration to the dual variable (virtual queue), thereby reducing the sensitivity of action selection to cost prediction errors amplified by the dual variable. On the dual side, QUID regularizes the dual update through a discounting factor. On low-cost rounds, the regularized update gradually reduces the weight assigned to the cost function, which compensates for the additional exploration cost introduced by the queue normalization. By carefully calibrating the regularization parameter, QUID establishes the regret bound without first bounding the queue length, which decouples regret analysis from violation control and yields a stronger regret guarantee.

- **Theoretical Guarantees:** Under general feasibility, for every $\theta \in [1/2, 1]$, QUID achieves $\widetilde{O}(T^{\theta}U^{1-\theta})$ regret and $\widetilde{O}(T^{1-\theta/2}U^{\theta/2})$ constraint violation. In particular, setting $\theta = 1/2$ recovers the optimal unconstrained regret rate of $\widetilde{O}(\sqrt{TU})$ while maintaining $\widetilde{O}(T^{3/4}U^{1/4})$ violation. Alternatively, setting $\theta = 2/3$ yields the balanced rate $\widetilde{O}(T^{2/3}U^{1/3})$ for both criteria, improving upon the $\widetilde{O}(T^{3/4}U^{1/4})$ guarantees of LagrangeCBwLC and LOE2D. When Slater's condition holds with an unknown margin $\epsilon > 0$, the violation bound further improves to $\widetilde{O}\Big(\min\Big\{\sqrt{TU}/\epsilon,\, T^{1-\theta/2}U^{\theta/2}\Big\}\Big)$ without sacrificing regret. Compared with LOE2D, this result improves the dependence on the Slater margin from $1/\epsilon^2$ to $1/\epsilon$ without any prior information. The improvement follows from a novel Lyapunov drift lemma that analyzes rounds with low and high prediction risk separately.
- **Experiments:** We evaluate QUID on learning-to-rank and constrained bidding tasks using XGBoost as the online regression oracles. The reported results show that QUID achieves higher rewards and lower costs than baseline algorithms, demonstrating its effectiveness in balancing reward maximization and constraint satisfaction in stochastic constrained contextual bandits.

2 Model and Preliminaries

We consider stochastic constrained contextual bandits over T rounds. Let $\mathcal{X}$ be the context space and $\mathcal{A}$ a finite action set with $K := |\mathcal{A}|$ actions. A stationary policy $\pi : \mathcal{X} \to \Delta(\mathcal{A})$ maps each context x to an action distribution $\pi(\cdot \mid x)$. Let Π denote the class of all such policies.

At each round $t \in [T]$, the environment independently draws a tuple (x_t, r_t, c_t) from a fixed but unknown joint distribution. Let $\mathcal{D}$ denote the context distribution. The reward and cost functions satisfy $r_t : \mathcal{A} \to [0, F]$ and $c_t : \mathcal{A} \to [-G, G]$. After observing x_t, the learner samples an action $a_t \sim \pi_t(\cdot \mid x_t)$ and observes only $r_t(a_t)$ and $c_t(a_t)$. To specify the information available at each stage, let $\mathcal{F}_{t-1}$ contain the interaction history and algorithm state before x_t is observed, and define $\mathcal{H}_t := \mathcal{F}_{t-1} \vee \sigma(x_t)$ and $\mathcal{I}_t := \mathcal{H}_t \vee \sigma(a_t)$. The action is sampled using fresh randomness that is conditionally independent of (r_t, c_t) given $\mathcal{H}_t$. We assume the following realizability condition.

Assumption 2.1. *There exist measurable functions $f : \mathcal{X} \times \mathcal{A} \to [0, F]$ and $g : \mathcal{X} \times \mathcal{A} \to [-G, G]$ such that*

$$\mathbb{E}[r_t(a) \mid x_t = x] = f(x, a), \quad \mathbb{E}[c_t(a) \mid x_t = x] = g(x, a), \quad \forall\, x \in \mathcal{X},\ a \in \mathcal{A}.$$

Assumption 2.1 models rewards and costs as unbiased noisy observations of fixed mean functions, which is standard in stochastic contextual bandits.

For a policy $\pi \in \Pi$, define its expected reward and cost by

$$f(\pi) := \mathbb{E}_{x\sim\mathcal{D},\, a\sim\pi(\cdot|x)}[f(x, a)], \qquad g(\pi) := \mathbb{E}_{x\sim\mathcal{D},\, a\sim\pi(\cdot|x)}[g(x, a)]. \tag{1}$$

We assume that the feasible set $\{\pi \in \Pi : g(\pi) \le 0\}$ is nonempty and that an optimal solution exists for the following offline problem:

$$\pi^\star \in \arg\max_{\pi\in\Pi} f(\pi) \qquad \text{subject to} \quad g(\pi) \le 0. \tag{2}$$

The optimal policy $\pi^\star$ may be randomized, unlike in the unconstrained setting, where a deterministic policy is sufficient for optimality. And note that the constraint function g can represent a broad class of constraints, e.g., knapsack constraints (Agrawal & Devanur, 2014; Li & Stoltz, 2022) and fairness constraints (Joseph et al., 2016; Chen et al., 2020; Patil et al., 2021), which provides a general framework for constrained bandits. For ease of exposition, we focus on a single scalar constraint in this paper and the extension to any finite number of constraints is immediate.

We evaluate a learning algorithm by its expected regret relative to $\pi^\star$ and expected cumulative constraint violation over T rounds:

$$\mathrm{Reg}_T := T f(\pi^\star) - \mathbb{E}\left[\sum_{t=1}^{T} f(x_t, a_t)\right], \tag{3}$$

$$\mathrm{Viol}_T := \mathbb{E}\left[\sum_{t=1}^{T} g(x_t, a_t)\right]. \tag{4}$$

The regret Reg_T measures the cumulative reward loss relative to the best stationary feasible policy, while Viol_T measures the cumulative constraint violation. The objective is to achieve sublinear regret and violation, i.e., $\mathrm{Reg}_T = o(T)$ and $\mathrm{Viol}_T = o(T)$, so that the learner approaches the optimal average reward while asymptotically satisfying the constraint. We next introduce the regression oracle assumption used throughout the analysis and the additional strict feasibility condition used to refine the violation guarantee.

Online regression oracles. We assume that the learner has access to two online regression oracles, one for rewards and one for costs. At the beginning of round t, based on the data $\{(x_s, a_s, r_s(a_s), c_s(a_s))\}_{s<t}$ collected until round $t-1$, the oracles output predictors $\widehat{f}_t : \mathcal{X} \times \mathcal{A} \to [0, F]$ and $\widehat{g}_t : \mathcal{X} \times \mathcal{A} \to [-G, G]$ for the reward and cost functions, respectively. We quantify their predictive accuracy through the following cumulative squared error guarantee.

Assumption 2.2. *For every $p \in (0,1)$, there exist bounds $U_f(T,p)$ and $U_g(T,p)$ such that, with probability at least $1-p$, the following inequalities hold simultaneously:*

$$\sum_{t=1}^{T} \big(\widehat{f}_t(x_t, a_t) - f(x_t, a_t)\big)^2 \le U_f(T,p) \qquad \textit{and} \qquad \sum_{t=1}^{T} \big(\widehat{g}_t(x_t, a_t) - g(x_t, a_t)\big)^2 \le U_g(T,p).$$

We set $p = T^{-2}$ and write $U := \max\{U_f(T,p), U_g(T,p), 1\}$. This assumption is standard in contextual bandits based on regression oracles and can be satisfied by efficient oracles for many common function classes (Foster & Rakhlin, 2020). For a finite function class, the exponential weights forecaster yields $U(T,p) = O(\log(|\mathcal{F}|/p))$ (Vovk, 1995); for a d-dimensional linear class, online ridge regression yields $U(T,p) = O(d\log(T/p))$ (Azoury & Warmuth, 2001). Related guarantees extend to more general function classes (Rakhlin & Sridharan, 2014). Note that Assumption 2.2 is much weaker than the oracle condition in Guo et al. (2025), which requires uniform pointwise optimistic and pessimistic confidence bounds to hold uniformly over all pairs of contexts and actions throughout the T rounds. Constructing such bounds can be difficult for non-parametric function classes (Sarkar & Sinha, 2026). In contrast, our analysis requires only a cumulative bound on squared prediction errors along the learner's realized interaction sequence.

Assumption 2.3. *There exist a stationary policy $\pi^0 \in \Pi$ and a margin $\epsilon > 0$ such that $g(\pi^0) \le -\epsilon$.*

Assumption 2.3 imposes a strict feasibility requirement. Rather than merely requiring a policy with nonpositive expected cost, it assumes the existence of a policy whose expected cost lies at least ϵ below the constraint boundary. This condition is commonly known as Slater's condition, and ϵ is referred to as the *Slater constant* or *feasibility margin*. Note that we do not require prior knowledge of ϵ in the algorithm presented next, and the margin is shown only in the analysis.

3 Algorithm Design

In this section, we introduce *Queue-Normalized Inverse-Gap Exploration with Discounted Regulator* (QUID), a primal-dual algorithm for stochastic constrained contextual bandits. In each round, the algorithm constructs estimated Lagrangian scores from reward and cost predictions, selects an action through inverse-gap weighting, and updates a dual variable (virtual queue) to track cumulative constraint violation. Algorithm 1 summarizes the procedure, and we explain the design of each component below.

- **Lagrangian Score Estimation.** At the beginning of round t, online regression oracles construct reward and cost predictors $\widehat{f}_t$ and $\widehat{g}_t$ from past observations. After observing the context x_t, QUID computes an estimated Lagrangian score for each action $a \in \mathcal{A}$:

$$\widehat{L}_t(x_t, a) = \widehat{f}_t(x_t, a) - q_t\, \widehat{g}_t(x_t, a), \tag{5}$$

  where $q_t := Q_t/V \ge 0$ denotes the virtual queue Q_t scaled by a positive parameter V. The scaled queue acts as an adaptive Lagrange multiplier that balances reward maximization and constraint satisfaction. As constraint violation accumulates, q_t increases, strengthening the penalty on the predicted cost $\widehat{g}_t$ and favoring actions with lower predicted costs. Conversely, a smaller queue reduces the cost penalty and places greater emphasis on reward maximization.

Algorithm 1 Queue-Normalized Inverse-Gap Exploration with Discounted Regulator (QUID)

Require: Horizon T, action set $\mathcal{A}$ with $K = |\mathcal{A}|$, cumulative oracle error bound U, tradeoff parameter $\theta \in [1/2, 1]$, and two regression oracles.

Initialization: $\gamma \leftarrow (K+1)(T/U)^{\theta}, \quad V \leftarrow T^{1-\theta}U^{\theta}, \quad \rho \leftarrow \dfrac{K+1}{\gamma V + K + 1}, \quad Q_1 \leftarrow 0.$

for $t = 1, \ldots, T$ **do**

Prediction with Regression Oracles: Obtain $\widehat{f}_t, \widehat{g}_t$ from the regression oracles.

Lagrangian Score Estimation: Observe x_t and set $q_t \leftarrow Q_t/V$.

Compute

$$\widehat{L}_t(x_t, a) \leftarrow \widehat{f}_t(x_t, a) - q_t \widehat{g}_t(x_t, a), \qquad \forall a \in \mathcal{A}. \tag{6}$$

Queue-Normalized Inverse-Gap Exploration: Choose $\widehat{a}_t \in \arg\max_{a \in \mathcal{A}} \widehat{L}_t(x_t, a)$.

Set $\beta_t \leftarrow 1/(1+q_t^2)$ and sample $a_t \sim \pi_t(\cdot \mid x_t)$ with

$$\pi_t(a \mid x_t) := \frac{1}{\lambda_t + 2\gamma\beta_t\big(\widehat{L}_t(x_t, \widehat{a}_t) - \widehat{L}_t(x_t, a)\big)}, \qquad \forall a \in \mathcal{A}, \tag{7}$$

where $\lambda_t > 0$ is chosen to ensure $\sum_a \pi_t(a \mid x_t) = 1$.

Reward and Cost Learning: Observe $r_t(a_t), c_t(a_t)$ and update both regression oracles.

Regularized Virtual Queue Update:

$$Q_{t+1} \leftarrow [(1-\rho)Q_t + c_t(a_t)]_+. \tag{8}$$

end for

- **Quadratic Queue Normalization.** After computing the estimated Lagrangian scores, QUID uses inverse-gap weighting to construct an action distribution. Let $\widehat{a}_t$ denote an action with the highest estimated Lagrangian score. As shown in (7), the distribution assigns greater probability to actions with smaller score gaps relative to $\widehat{a}_t$, while retaining strictly positive probability for every action to ensure continued exploration. While the inverse-gap weighting technique has been widely used in previous works (Han et al., 2023; Slivkins et al., 2024; Guo & Liu, 2024), the key distinction of QUID lies in the exploration parameter β_t. Recall that a larger q_t increases the weight of cost function to favor low cost actions, however, since the cost function is estimated, it also amplifies the cost prediction error. To address the issue, we adopt a quadratic queue normalization $\beta_t = (1+q_t^2)^{-1}$ in QUID. The normalization reduces the action distribution's sensitivity to errors in cost predictions $\hat{g}_t$ and maintains exploration when the queue is large, preventing the learner from assigning excessively high probabilities to actions whose predicted costs may be inaccurate.
- **Regularized Queue Update.** After selecting a_t, the learner observes only the reward $r_t(a_t)$ and cost $c_t(a_t)$ of the selected action and feeds the feedback to online regression oracles to update predictors. The observed cost is also used to update the virtual queue according to (8), which summarizes past constraint violations and determines the cost penalty through $q_t = Q_t/V$. Unlike the standard queue update, it introduces *regularization* through a discount controlled by ρ, which discounts the accumulated queue before incorporating the current cost. The discount complements quadratic queue normalization by gradually reducing the influence of past violations. When subsequent costs no longer maintain a large queue, regularization allows the queue to decay and relaxes the normalization of score gaps. By carefully calibrating the regularization factor, the two designs jointly achieve a better exploration–exploitation tradeoff.

In summary, queue-normalized exploration and dual regularization form the two central components of QUID. In the next section, we show how these designs, together with a novel Lyapunov drift analysis, yield improved regret and constraint violation guarantees.

4 Theoretical Analysis

In this section, we establish the regret and constraint violation guarantees of QUID and present the key steps of the analysis.

4.1 MAIN GUARANTEES

Theorem 4.1. *Under Assumptions 2.1 and 2.2,* QUID *achieves the following regret and constraint violation bounds:*

$$\mathrm{Reg}_T = \widetilde{O}\big(T^{\theta}U^{1-\theta}\big)\,, \qquad \mathrm{Viol}_T = \widetilde{O}\Big(T^{1-\theta/2}U^{\theta/2}\Big)\,. \tag{9}$$

If Slater's condition in Assumption 2.3 additionally holds, the violation bound improves to

$$\mathrm{Viol}_T = \widetilde{O}\left(\min\left\{\frac{\sqrt{TU}}{\epsilon},\ T^{1-\theta/2}U^{\theta/2}\right\}\right), \tag{10}$$

where $\widetilde{O}(\cdot)$ suppresses logarithmic factors in T.

Theorem 4.1 provides a tunable regret–violation tradeoff parameterized by $\theta \in [1/2, 1]$. Under general feasibility, where a strictly feasible policy need not exist, setting $\theta = 1/2$ yields $\widetilde{O}(\sqrt{TU})$ regret and $\widetilde{O}(T^{3/4}U^{1/4})$ violation. Hence, QUID matches the optimal regret rate for unconstrained contextual bandits with regression oracles (Foster & Rakhlin, 2020), while retaining sublinear constraint violation. Alternatively, choosing $\theta = 2/3$ balances the two criteria, yielding $\widetilde{O}(T^{2/3}U^{1/3})$ regret and constraint violation. This improves upon the balanced $\widetilde{O}(T^{3/4}U^{1/4})$ guarantees of Slivkins et al. (2024) and Guo & Liu (2024).

When Slater's condition holds, QUID additionally achieves a $\widetilde{O}(\sqrt{TU}/\epsilon)$ violation bound without sacrificing the regret guarantee. Combining the bound with the general feasibility result gives (10). In particular, setting $\theta = 1/2$ yields $\widetilde{O}(\sqrt{TU})$ regret and $\widetilde{O}(\sqrt{TU}/\epsilon)$ constraint violation when $\epsilon \gtrsim (U/T)^{1/4}$. Compared with Guo & Liu (2024), QUID removes the dependence on ϵ from the regret bound and improves the dependence of the constraint violation bound on the feasibility margin from $1/\epsilon^2$ to $1/\epsilon$. Although the method of Slivkins et al. (2024) achieves the same $1/\epsilon$ dependence for violation, it requires prior knowledge of ϵ and has a regret bound that also depends on ϵ. The improvements are important when the Slater margin is small.

Remark 4.2. *Guo & Liu (2024) conjectured a lower bound of $\Omega\Big(\min\Big\{\sqrt{TU}/\epsilon,\ T^{3/4}U^{1/4}\Big\}\Big)$ on $\max\{\mathrm{Reg}_T, \mathrm{Viol}_T\}$ in stochastic contextual bandit with general constraint. In our algorithm, with $\theta = 1/2$,* QUID *attains this conjectured rate up to logarithmic factors, while achieving the stronger regret guarantee $\mathrm{Reg}_T = \widetilde{O}(\sqrt{TU})$. The improvement follows from decoupling the regret and constraint violation analyses, which makes the regret bound independent of the feasibility margin ϵ, and from a refined Lyapunov analysis that improves the dependence of violation bound on ϵ under Slater's condition. These guarantees require only cumulative regression error bounds and no prior knowledge of the context distribution, whereas the faster rates of Guo et al. (2025) rely on stronger oracle assumptions and such prior knowledge. Determining whether a matching lower bound holds under the weaker oracle assumption remains an open problem, which we leave for future work.*

We next outline three main steps of the analysis. First, we bound the instantaneous regret and the one-step change in a quadratic Lyapunov function, referred to as the Lyapunov drift. Combining these bounds yields a key regret-plus-drift inequality that controls regret directly, without first establishing a bound on constraint violation. Second, under general feasibility, we use the same inequality to control cumulative Lyapunov drift and establish a high-probability bound on the maximum queue length. We then convert queue bound into a violation guarantee through regularized queue update. Finally, under Slater's condition, we refine the Lyapunov analysis by separating rounds according to prediction risk. This allows us to exploit negative drift on low-risk rounds while controlling the cumulative effect of high-risk rounds. Combining these controls yields an improved violation guarantee.

4.2 REGRET ANALYSIS

Recall that $\mathcal{F}_{t-1}$ contains the information available before observing x_t. For a comparison policy $\pi \in \Pi$, define the regret at round t by

$$\mathrm{reg}_t(\pi) := f(\pi) - \mathbb{E}[f(x_t, a_t) \mid \mathcal{F}_{t-1}]\,. \tag{11}$$

We also define the conditional squared prediction error at round t as

$$z_t := \mathbb{E}\Big[\big(\widehat{f}_t(x_t,a_t) - f(x_t,a_t)\big)^2 + \big(\widehat{g}_t(x_t,a_t) - g(x_t,a_t)\big)^2 \;\Big|\; \mathcal{F}_{t-1}\Big]. \tag{12}$$

The first lemma bounds the one-step regret of QUID.

Lemma 4.3. *For any policy $\pi \in \Pi$ and round t, the following inequality holds:*

$$\mathrm{reg}_t(\pi) \le q_t(g(\pi) - \mathbb{E}[g(x_t,a_t) \mid \mathcal{F}_{t-1}]) + \frac{K+1}{2\gamma}(1+q_t^2) + 2\gamma z_t. \tag{13}$$

The lemma decomposes regret into three terms. The first term captures difference in expected cost between the comparison policy and the learner. The second is exploration penalty and the third term accounts for errors in the reward and cost predictions. The key property in Lemma 4.3 is that the design of quadratic queue normalization in QUID removes q_t from the coefficient of z_t, preventing the dual variable from amplifying the cost prediction error. In contrast to the algorithm design and analysis of LOE2D, this term can therefore be controlled directly by the cumulative regression oracle guarantee. However, the normalization also introduces an additional exploration cost q_t^2, which retains a quadratic dependence on the queue.

We control the additional cost through the Lyapunov drift induced by dual regularization. Let $\Phi_t := Q_t^2/2$ denote the quadratic Lyapunov function and $\Delta_t := \Phi_{t+1} - \Phi_t$ denote the one-step drift. The regularized queue update yields the following bound.

Lemma 4.4. *At every round t, the virtual queue update in* QUID *satisfies*

$$\Delta_t \le -\rho\left(1 - \frac{\rho}{2}\right) Q_t^2 + (1-\rho) Q_t c_t(a_t) + \frac{G^2}{2}. \tag{14}$$

Lemma 4.4 bounds the Lyapunov drift by a queue-weighted cost and a negative quadratic term induced by dual regularization. Taking conditional expectation and using $q_t = Q_t/V$, we see that these terms cancel the learner's queue-weighted cost and absorb the additional exploration cost in Lemma 4.3, respectively. Thus, combining the two lemmas yields the following key inequality.

Lemma 4.5. *For every policy $\pi \in \Pi$ and round t,*

$$\mathrm{reg}_t(\pi) + \frac{\mathbb{E}[\Delta_t \mid \mathcal{F}_{t-1}]}{(1-\rho)V} \le q_t g(\pi) + \frac{K+1}{2\gamma} + \frac{G^2}{2(1-\rho)V} + 2\gamma z_t. \tag{15}$$

Lemma 4.5 establishes the regret-plus-drift inequality. It formalizes the complementary roles of queue normalization and dual regularization. Normalization removes queue dependence from the prediction error, while regularization compensates for the additional exploration cost. The only remaining explicit queue-dependent term is $q_t g(\pi)$, which is nonpositive for any feasible comparison policy. Summing the inequality therefore bounds regret without first controlling the virtual queue. Conversely, moving the regret term to the right-hand side gives a drift inequality that can be used to control the queue length and establish violation guarantees. Thus, the single inequality supports both analyses while decoupling regret from the violation analysis.

Applying Lemma 4.5 with $\pi = \pi^\star$, summing over $t = 1, \dots, T$, and taking total expectations gives the regret bound. Since $g(\pi^\star) \le 0$ and $q_t \ge 0$, we have:

$$\begin{aligned}
\mathrm{Reg}_T &\le -\frac{1}{(1-\rho)V}\mathbb{E}\left[\sum_{t=1}^T \Delta_t\right] + \frac{(K+1)T}{2\gamma} + \frac{G^2 T}{2(1-\rho)V} + 2\gamma\mathbb{E}\left[\sum_{t=1}^T z_t\right] \\
&\le \frac{(K+1)T}{2\gamma} + \frac{G^2 T}{2(1-\rho)V} + 4\gamma U + \frac{2\gamma(F^2+4G^2)}{T} \\
&\le \widetilde{O}\big(T^\theta U^{1-\theta}\big).
\end{aligned} \tag{16}$$

The second inequality uses the initialization $Q_1 = 0$, which gives $\sum_{t=1}^T \Delta_t = (Q_{T+1}^2 - Q_1^2)/2 \ge 0$, and the cumulative prediction error is controlled by the regression oracle bound in Assumption 2.2. Specifically, the tower property gives $\mathbb{E}[\sum_{t=1}^T z_t] \le 2U + (F^2+4G^2)/T$, where the second term follows from the bounded prediction ranges and the oracle failure probability of at most T^{-2}. The last inequality follows by substituting the choices of γ, V, and ρ from Algorithm 1, which completes the proof of the regret.

4.3 CONSTRAINT VIOLATION UNDER GENERAL FEASIBILITY

To obtain the constraint violation bound, we first establish an inequality from the regularized virtual queue update. For every $n \leq T$,

$$\sum_{t=1}^{n} c_t(a_t) \leq Q_{n+1} + \rho \sum_{t=1}^{n} Q_t \leq (1+\rho n) \max_{s \leq n+1} Q_s \leq 2 \max_{s \leq n+1} Q_s \tag{17}$$

where we use $\rho = (K+1)/(\gamma V + K + 1) = 1/(T+1)$. Thus, the maximum virtual queue length serves as a surrogate for cumulative constraint violation, with an upper bound on the virtual queue length directly yielding a violation bound up to a constant factor.

We next establish the queue bound through Lyapunov drift analysis, a classical tool in networking for controlling workload accumulation and ensuring queue stability (Tassiulas & Ephremides, 1990; Neely, 2010). The method bounds queue lengths by controlling the drift of a potential function. We adopt this perspective for Q_t in this paper. Specifically, we control cumulative drift under general feasibility and further exploit negative drift under Slater's condition to obtain a sharper queue bound. We first present the result under general feasibility.

Lemma 4.6. *Under Assumptions 2.1 and 2.2, the maximum virtual queue length satisfies the following bound with high probability:*

$$\max_{t \leq T+1} Q_t = \widetilde{O}\Big(\sqrt{TV + \gamma VU}\Big) = \widetilde{O}\Big(T^{1-\theta/2} U^{\theta/2}\Big). \tag{18}$$

Proof sketch. Applying Lemma 4.5 with $\pi = \pi^\star$ and summing over $t = 1, \ldots, n$, we obtain $\sum_{t=1}^{n} \Delta_t = \widetilde{O}(TV + \gamma VU)$ for all $n \leq T$ with high probability. Since $\sum_{t=1}^{n} \Delta_t = (Q_{n+1}^2 - Q_1^2)/2 = Q_{n+1}^2/2$, the maximum queue bound follows.

4.4 CONSTRAINT VIOLATION UNDER SLATER'S CONDITION

We next use Slater's condition to obtain a sharper violation guarantee through a refined Lyapunov stability analysis. The key is that a strictly feasible policy introduces a negative contribution to the drift, providing additional control of the virtual queue. Applying Lemma 4.5 with the strictly feasible policy π^0 yields

$$\frac{\mathbb{E}[\Delta_t \mid \mathcal{F}_{t-1}]}{(1-\rho)V} \leq -\epsilon q_t + 2\gamma z_t + B, \tag{19}$$

where $B := F + \frac{K+1}{2\gamma} + \frac{G^2}{2(1-\rho)V}$. The Slater condition introduces a negative term, $-\epsilon q_t$, whose magnitude increases with the queue length. In standard Lyapunov analysis, it can absorb the remaining terms once the queue length exceeds a certain threshold, yielding a negative expected drift. However, in constrained contextual bandits, the regression error z_t poses an additional challenge. Assumption 2.2 controls only the cumulative prediction error and does not ensure that z_t is small in every round. Guo & Liu (2024) address the issue by amortizing the regression error z_t over $\Theta(1/\epsilon)$ rounds and bounding constraint violation through a multi-step Lyapunov drift analysis. However, the approach yields a constraint violation bound with a worse dependence on the Slater parameter ϵ.

We address the difficulty through *decoupled Lyapunov analysis*. The key is to exploit negative drift on rounds with small prediction error while controlling the cumulative effect of rounds with large errors. We define a risk threshold $r > 0$ and classify each round t as *low-risk* if $z_t \leq r$ and *high-risk* otherwise. Specifically, on low-risk rounds, standard Lyapunov analysis uses negative drift to control queue stability, which contributes $O\Big(\frac{VB}{\epsilon} + \frac{\gamma V r}{\epsilon}\Big)$ to the queue bound. For high-risk rounds, we control cumulative effect on the queue without requiring negative drift. Since the queue update ensures $Q_{t+1} - Q_t \leq G$ and each high-risk round contributes more than r to the cumulative prediction error, the number of such rounds satisfies

$$\sum_{t=1}^{T} \mathbf{1}\{z_t > r\} \leq \frac{1}{r} \sum_{t=1}^{T} z_t = \widetilde{O}\left(\frac{U}{r}\right), \tag{20}$$

which contributes $\widetilde{O}(GU/r)$ to the queue bound. Thus, choosing $r = \Theta(\sqrt{GU\epsilon/\gamma V})$ to balance the two effects yields the following bound.

Lemma 4.7. *Suppose Assumptions 2.1–2.3 hold. Taking* $r = \Theta(\sqrt{GU\epsilon/\gamma V})$*, with high probability,*

$$\max_{t \le T+1} Q_t = \widetilde{O}\left(\frac{VB}{\epsilon} + \frac{\gamma V r}{\epsilon} + \frac{U}{r}\right) = \widetilde{O}\left(\frac{\sqrt{TU}}{\epsilon}\right), \tag{21}$$

which provides a stronger result under Slater's condition. Note that the risk threshold r is used only in the analysis, and we do not need to identify high-risk rounds in the algorithm. Finally, combining the translation formula (17), Lemma 4.6 and Lemma 4.7 completes the proof.

5 Experiments

We evaluate QUID on learning-to-rank and constrained bidding tasks against LagrangeCBwLC, LOE2D, and Optimistic^3, using XGBoost as the online regression oracle for all methods. Both environments have a small Slater margin $\epsilon = 10^{-3}$, requiring a careful balance between reward maximization and constraint satisfaction. We set $\theta = 1/2$ for QUID and run each trial for $T = 5{,}000$ rounds. Figure 1 reports the cumulative average reward and cost.

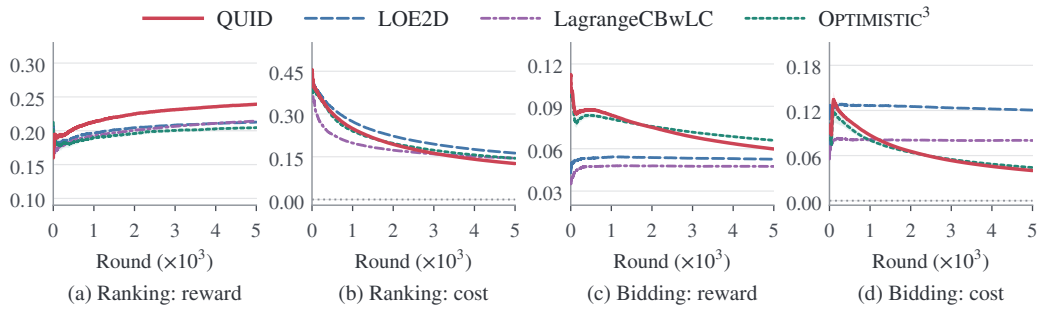


Figure 1: Performance on learning-to-rank and constrained bidding tasks over $T = 5,000$ rounds, averaged across 50 seeds. Panels (a,b) report cumulative average reward and cost for ranking; panels (c,d) show the corresponding quantities for bidding. Dotted lines indicate the constraint threshold.

Learning-to-Rank. We construct a constrained contextual bandit task from the Microsoft MSLR dataset. Each round samples a query uniformly from 5,000 queries, and the learner selects one of 20 candidate documents. Rewards are determined by document relevance, while costs are sampled from $[-1, 1]$; both observations are perturbed by Gaussian noise with standard deviation 0.05. Figure 1(a) and (b) show that QUID achieves the highest reward and eventually lowest cost among the evaluated methods. The results are consistent with the improved theoretical guarantees of QUID in the small Slater margin setting. The empirical performance of Optimistic^3 does not exhibit its theoretical advantage in the reported experiment. We think the reason is that Optimistic^3 relies on valid confidence bounds for oracle predictions in every round as well as prior knowledge of the context distribution, which XGBoost does not ensure.

Constrained Bidding. We also consider repeated first-price auctions with a return-on-investment constraint. The private value v_t and highest competing bid d_t are drawn independently from $\mathcal{N}(0.6, 0.1^2)$ and $\mathcal{N}(0.4, 0.1^2)$. Given v_t, the learner selects $b_t \in \{0, v_t/19, \ldots, v_t\}$. A winning bid gets reward $r_t = (v_t - b_t)\mathbf{1}\{b_t \ge d_t\}$ and payment $p_t = b_t\mathbf{1}\{b_t \ge d_t\}$. We define the cost as $c_t = (\eta p_t - r_t)/2$, where $\eta = 1.7$ is the ROI threshold. Figure 1(c) and (d) show that QUID achieves higher reward and lower cost than LOE2D and LagrangeCBwLC. It also achieves the lowest cost while retaining reward close to that of Optimistic^3, further demonstrating the effectiveness of QUID.

6 Conclusion

In this paper, we propose QUID, a novel algorithm that combines queue-normalized inverse-gap exploration with dual regularization. Under a cumulative regression error assumption, QUID improves existing regret and violation guarantees under both general feasibility and Slater's condition. Experiments on ranking and bidding tasks further demonstrate its empirical effectiveness.

AI USE STATEMENT

We used generative AI tools to assist with manuscript preparation, including improving clarity and readability, refining wording, and checking LaTeX formatting against the conference guidelines. We also used these tools to assist with code development, including drafting, debugging, and refining experimental scripts. The authors critically evaluated all AI-generated suggestions and verified the accuracy of the content and code incorporated into the work. We take full responsibility for the final paper, including its mathematical claims, experimental results, and conclusions.

REFERENCES

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*, pp. 1638–1646. PMLR, 2014.

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

Shipra Agrawal, Nikhil R Devanur, and Lihong Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In *Conference on Learning Theory*, pp. 4–18. PMLR, 2016.

Katy S Azoury and Manfred K Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine learning*, 43(3):211–246, 2001.

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

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

Martino Bernasconi, Matteo Castiglioni, and Andrea Celli. No-regret is not enough! bandits with general constraints through adaptive regret minimization. *arXiv preprint arXiv:2405.06575*, 2024a.

Martino Bernasconi, Matteo Castiglioni, Andrea Celli, and Federico Fusco. Beyond primal-dual methods in bandits with stochastic and adversarial constraints. *arXiv preprint arXiv:2405.16118*, 2024b.

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

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*, pp. 169–178. AUAI Press, 2011. ISBN 9780974903972.

Dylan Foster and Alexander Rakhlin. Beyond ucb: Optimal and efficient contextual bandits with regression oracles. In *International Conference on Machine Learning*, pp. 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*, pp. 1539–1548. PMLR, 2018.

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*, pp. 3962–3969, 2020.

Hengquan Guo and Xin Liu. Stochastic constrained contextual bandits via lyapunov optimization based estimation to decision framework. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 2204–2231. PMLR, 2024.

Hengquan Guo and Xin Liu. On stochastic contextual bandits with knapsacks in small budget regime. In *Proceedings of the International Conference on Learning Representations*, 2025.

Hengquan Guo, Lingkai Zu, and Xin Liu. Triple-optimistic learning for stochastic contextual bandits with general constraints. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21252–21276. PMLR, 2025. URL https://proceedings.mlr.press/v267/guo25v.html.

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*, pp. 5011–5035. PMLR, 2023.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

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

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.

Zhen Li and Gilles Stoltz. Contextual bandits with knapsacks for a conversion model. *Advances in Neural Information Processing Systems*, 35:35590–35602, 2022.

Michael Neely. *Stochastic network optimization with application to communication and queueing systems*. Morgan & Claypool Publishers, 2010.

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, Mohammed Kadous, and Ion Stoica. Routellm: Learning to route llms from preference data. In *International Conference on Learning Representations*, volume 2025, pp. 34433–34448, 2025.

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.

Manhin Poon, XiangXiang Dai, Xutong Liu, Fang Kong, John CS Lui, and Jinhang Zuo. Online multi-llm selection via contextual bandits under unstructured context evolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 24855–24863, 2026.

Alexander Rakhlin and Karthik Sridharan. Online non-parametric regression. In *Conference on Learning Theory*, pp. 1232–1264. PMLR, 2014.

Karthik Abinav Sankararaman and Aleksandrs Slivkins. Bandits with knapsacks beyond the worst case. In *Advances in Neural Information Processing Systems*, volume 34, pp. 23191–23204, 2021.

Dhruv Sarkar and Abhishek Sinha. Constrained contextual bandits with adversarial contexts. *arXiv preprint arXiv:2605.06190*, 2026.

Abhishek Sinha. Banditq: Fair bandits with guaranteed rewards. In *The 40th Conference on Uncertainty in Artificial Intelligence*, 2024.

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.

Aleksandrs Slivkins, Xingyu Zhou, Karthik Abinav Sankararaman, and Dylan J Foster. Contextual bandits with packing and covering constraints: A modular lagrangian approach via regression. *Journal of Machine Learning Research*, 25(394):1–37, 2024.

Leandros Tassiulas and Anthony Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. In *29th IEEE Conference on Decision and Control*, pp. 2130–2132. IEEE, 1990.

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.

Vladimir G Vovk. A game of prediction with expert advice. In *Proceedings of the eighth annual conference on Computational learning theory*, pp. 51–60, 1995.

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

Mengyue Yang, Qingyang Li, Zhiwei Qin, and Jieping Ye. Hierarchical adaptive contextual bandits for resource constraint based recommendation. In *Proceedings of the web conference 2020*, pp. 292–302, 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.

A Related Work

A.1 Constrained Multi-Armed Bandits

Constrained multi-armed bandits have been studied under resource budgets, fairness requirements, safety constraints, and general long-term constraints. For resource allocation, bandits with knapsacks model actions consuming limited resources and admit regret guarantees that are optimal up to logarithmic factors (Badanidiyuru et al., 2018); instance-dependent logarithmic regret has also been characterized (Sankararaman & Slivkins, 2021). Fairness constraints include merit-based fairness (Joseph et al., 2016), minimum selection rates (Patil et al., 2021; Li et al., 2019), and minimum reward guarantees (Sinha, 2024). Conservative bandits instead require cumulative performance to remain close to a baseline throughout learning (Wu et al., 2016). More recent work considers general long-term constraints. Bernasconi et al. (2024b) obtain $\widetilde{O}(\sqrt{T})$ regret under stochastic constraints without Slater's condition, while Bernasconi et al. (2024a) develop a method based on primal and dual regret minimization that controls the dual iterates without knowing the Slater margin. These works extend constrained bandits beyond specific resource or fairness structures to general constraints.

A.2 Constrained Contextual Bandits

Early computationally efficient contextual bandit methods often rely on policy or classification oracles. Dudik et al. (2011) use a cost sensitive classification oracle to obtain optimal regret guarantees, while Agarwal et al. (2014) develop an efficient oracle method with statistically optimal guarantees for general policy classes. This framework has also been extended to contextual bandits with knapsack constraints (Badanidiyuru et al., 2014; Agrawal et al., 2016). Regression oracles provide an alternative interface by learning reward functions rather than optimizing directly over policies (Foster et al., 2018). SquareCB (Foster & Rakhlin, 2020) gives a general reduction from contextual bandits to online regression, converting a cumulative squared prediction error bound U into $\widetilde{O}(\sqrt{KTU})$ regret. This regression oracle framework naturally extends to constrained problems, where reward and cost functions can be learned by separate oracles. Examples include contextual bandits with knapsack constraints (Han et al., 2023) and more general packing and covering constraints (Slivkins et al., 2023; 2024). The works most closely related to ours combine regression oracles with Lagrangian or queue based constraint control (Slivkins et al., 2024; Guo & Liu, 2024). Under general feasibility, these methods obtain $\widetilde{O}(T^{3/4}U^{1/4})$ regret and constraint violation, with improved guarantees under Slater's condition. OPTIMISTIC^3 (Guo et al., 2025) achieves $\widetilde{O}(\sqrt{TU})$ regret and constraint violation without Slater's condition, but relies on stronger oracle guarantees based on confidence bounds. More recently, Sarkar & Sinha (2026) consider adversarial contexts under related feasibility notions. However, in the same setting considered in this paper, it also achieves $\widetilde{O}(T^{3/4}U^{1/4})$ regret and constraint violation as LOE2D and LagrangeCBwLC. In contrast, QUID improves the tradeoff between regret and constraint violation while retaining only cumulative regression error guarantees.

B Regret Analysis

In this section, we prove the regret guarantee in Theorem 4.1. The constraint violation bounds under general feasibility and Slater's condition are proved in Sections C and D, respectively.

Throughout the appendix, without loss of generality, let $T \geq 2$, $1 \leq U \leq T$ and $p = T^{-2}$.

B.1 Proof of Lemma 4.3

We first establish an inverse-gap bound for the Lagrangian scores at a fixed context. Conditioning on $\mathcal{H}_t$ fixes the queue, predictions, and sampling distribution, so only the action sampling remains to be averaged. Define

$$L_t(x,a) := f(x,a) - q_t g(x,a), \qquad \widehat{L}_t(x,a) := \widehat{f}_t(x,a) - q_t \widehat{g}_t(x,a). \tag{22}$$

Lemma B.1. *In Algorithm 1, for every stationary comparison policy $\pi \in \Pi$, we have,*

$$\begin{aligned}&\mathbb{E}_{a\sim\pi(\cdot|x_t)}[L_t(x_t,a) \mid \mathcal{H}_t] - \mathbb{E}_{a_t\sim\pi_t(\cdot|x_t)}[L_t(x_t,a_t) \mid \mathcal{H}_t]\\ &\le \frac{K+1}{2\gamma\beta_t} + \gamma\beta_t\mathbb{E}_{a_t\sim\pi_t}\left[\left(\widehat{L}_t(x_t,a_t) - L_t(x_t,a_t)\right)^2 \mid \mathcal{H}_t\right].\end{aligned} \tag{23}$$

Proof. Conditionally on $\mathcal{H}_t$, decompose the score difference as

$$\begin{aligned}&\mathbb{E}_{a\sim\pi(\cdot|x_t)}[L_t(x_t,a) \mid \mathcal{H}_t] - \mathbb{E}_{a_t\sim\pi_t(\cdot|x_t)}[L_t(x_t,a_t) \mid \mathcal{H}_t]\\ &= \underbrace{\mathbb{E}_{a_t\sim\pi_t}[\widehat{L}_t(x_t,\widehat{a}_t) - \widehat{L}_t(x_t,a_t) \mid \mathcal{H}_t]}_{\text{(I)}} + \underbrace{\mathbb{E}_{a_t\sim\pi_t}[\widehat{L}_t(x_t,a_t) - L_t(x_t,a_t) \mid \mathcal{H}_t]}_{\text{(II)}}\\ &\quad + \underbrace{\mathbb{E}_{a\sim\pi}[L_t(x_t,a) - \widehat{L}_t(x_t,\widehat{a}_t) \mid \mathcal{H}_t]}_{\text{(III)}}.\end{aligned} \tag{24}$$

Term (I) is the exploration loss in estimated scores, whereas (II) is the prediction error under the learner's sampling distribution. In (III), retaining the negative estimated gap allows the inverse-probability term introduced by Young's inequality to cancel. For the first term, the inverse-gap sampling rule gives

$$\begin{aligned}\text{(I)} &= \mathbb{E}_{a_t\sim\pi_t}[\widehat{L}_t(x_t,\widehat{a}_t) - \widehat{L}_t(x_t,a_t) \mid \mathcal{H}_t]\\ &= \sum_{a\in\mathcal{A}}\pi_t(a \mid x_t)\big(\widehat{L}_t(x_t,\widehat{a}_t) - \widehat{L}_t(x_t,a)\big)\\ &= \sum_{a\in\mathcal{A}}\pi_t(a \mid x_t)\left(\frac{1}{2\gamma\beta_t\pi_t(a \mid x_t)} - \frac{\lambda_t}{2\gamma\beta_t}\right)\\ &= \frac{1}{2\gamma\beta_t}\left(K - \lambda_t\sum_{a\in\mathcal{A}}\pi_t(a \mid x_t)\right)\\ &= \frac{K-\lambda_t}{2\gamma\beta_t}.\end{aligned} \tag{25}$$

The third equality substitutes the inverse-gap rule directly into the summation, and the last uses $\sum_a \pi_t(a \mid x_t) = 1$. For the estimated error of sampled action, we have,

$$\begin{aligned}\text{(II)} &= \mathbb{E}_{a_t\sim\pi_t}\left[\widehat{L}_t(x_t,a_t) - L_t(x_t,a_t) \mid \mathcal{H}_t\right]\\ &\le \mathbb{E}_{a_t\sim\pi_t}\left[\frac{1}{2\gamma\beta_t} + \frac{\gamma\beta_t}{2}\big(\widehat{L}_t(x_t,a_t) - L_t(x_t,a_t)\big)^2 \mid \mathcal{H}_t\right]\\ &= \frac{1}{2\gamma\beta_t} + \frac{\gamma\beta_t}{2}\mathbb{E}_{a_t\sim\pi_t}\left[\big(\widehat{L}_t(x_t,a_t) - L_t(x_t,a_t)\big)^2 \mid \mathcal{H}_t\right].\end{aligned} \tag{26}$$

where the inequality is because of Young's inequality: $u \le \frac{1}{2\eta} + \frac{\eta}{2}u^2$. For term (III), separate the prediction error from the estimated gap and apply the same inequality with the action-dependent weight $\gamma\beta_t\pi_t(a \mid x_t)$, we have,

$$\begin{aligned}\text{(III)} &= \sum_{a\in\mathcal{A}}\pi(a \mid x_t)\Big[L_t(x_t,a) - \widehat{L}_t(x_t,a) - \big(\widehat{L}_t(x_t,\widehat{a}_t) - \widehat{L}_t(x_t,a)\big)\Big]\\ &\le \sum_{a\in\mathcal{A}}\pi(a \mid x_t)\left[\frac{1}{2\gamma\beta_t\pi_t(a \mid x_t)} + \frac{\gamma\beta_t\pi_t(a \mid x_t)}{2}\big(\widehat{L}_t(x_t,a) - L_t(x_t,a)\big)^2\right.\\ &\qquad\qquad\left. - \left(\frac{1}{2\gamma\beta_t\pi_t(a \mid x_t)} - \frac{\lambda_t}{2\gamma\beta_t}\right)\right]\\ &= \frac{\lambda_t}{2\gamma\beta_t} + \frac{\gamma\beta_t}{2}\sum_{a\in\mathcal{A}}\pi(a \mid x_t)\pi_t(a \mid x_t)\big(\widehat{L}_t(x_t,a) - L_t(x_t,a)\big)^2\end{aligned}$$

$$\leq \frac{\lambda_t}{2\gamma\beta_t} + \frac{\gamma\beta_t}{2}\sum_{a\in\mathcal{A}}\pi_t(a \mid x_t)\big(\widehat{L}_t(x_t,a) - L_t(x_t,a)\big)^2$$
$$= \frac{\lambda_t}{2\gamma\beta_t} + \frac{\gamma\beta_t}{2}\mathbb{E}_{a_t\sim\pi_t}\Big[\big(\widehat{L}_t(x_t,a_t) - L_t(x_t,a_t)\big)^2 \mid \mathcal{H}_t\Big]. \tag{27}$$

The first inequality combines Young's inequality with the inverse-gap rule. Substituting (25)–(27) into (24) cancels the λ_t term and completes the proof of Lemma B.1. □

Lemma B.2 (Restatement of Lemma 4.3)**.** *For every stationary comparison policy $\pi \in \Pi$,*

$$\mathrm{reg}_t(\pi) \leq q_t\Big(g(\pi) - \mathbb{E}[g(x_t,a_t) \mid \mathcal{F}_{t-1}]\Big) + \frac{K+1}{2\gamma}(1+q_t^2) + 2\gamma z_t. \tag{28}$$

Proof. Using the definition of Lagrangian score and $\beta_t = (1+q_t^2)^{-1}$, the estimation term in Lemma B.1 satisfies

$$\begin{aligned}\gamma\beta_t\mathbb{E}_{a_t\sim\pi_t}\Big[\big(\widehat{L}_t - L_t\big)^2 \mid \mathcal{H}_t\Big] &= \gamma\mathbb{E}_{a_t\sim\pi_t}\left[\frac{\big[(\widehat{f}_t - f) - q_t(\widehat{g}_t - g)\big]^2}{1+q_t^2} \;\Big|\; \mathcal{H}_t\right] \\ &\leq 2\gamma\mathbb{E}_{a_t\sim\pi_t}\left[\frac{(\widehat{f}_t - f)^2 + q_t^2(\widehat{g}_t - g)^2}{1+q_t^2} \;\Big|\; \mathcal{H}_t\right] \\ &\leq 2\gamma\mathbb{E}_{a_t\sim\pi_t}\Big[(\widehat{f}_t - f)^2 + (\widehat{g}_t - g)^2 \mid \mathcal{H}_t\Big]. \end{aligned}\tag{29}$$

where the first inequality separates the reward and cost errors using $(u-v)^2 \leq 2u^2 + 2v^2$ and the second uses that both $1/(1+q_t^2)$ and $q_t^2/(1+q_t^2)$ belong to $[0,1]$. Then taking expectation of (23) given $\mathcal{F}_{t-1}$ and applying (29) gives

$$\begin{aligned}\mathrm{reg}_t(\pi) &= f(\pi) - \mathbb{E}[f(x_t,a_t) \mid \mathcal{F}_{t-1}] \\ &\leq q_t\Big(g(\pi) - \mathbb{E}[g(x_t,a_t) \mid \mathcal{F}_{t-1}]\Big) + \frac{K+1}{2\gamma}(1+q_t^2) \\ &\qquad + 2\gamma\mathbb{E}\Big[(\widehat{f}_t(x_t,a_t) - f(x_t,a_t))^2 + (\widehat{g}_t(x_t,a_t) - g(x_t,a_t))^2 \mid \mathcal{F}_{t-1}\Big] \\ &= q_t\Big(g(\pi) - \mathbb{E}[g(x_t,a_t) \mid \mathcal{F}_{t-1}]\Big) + \frac{K+1}{2\gamma}(1+q_t^2) + 2\gamma z_t.\end{aligned}$$

Hence the proof of Lemma 4.3 is completed. □

B.2 Proof of Lemma 4.4

We now derive the one-step change in the quadratic Lyapunov function. Recall the regularized queue update and quadratic Lyapunov drift:

$$Q_{t+1} = \big[(1-\rho)Q_t + c_t(a_t)\big]_+, \qquad \Delta_t := \frac{Q_{t+1}^2 - Q_t^2}{2}. \tag{30}$$

Lemma B.3 (Restatement of Lemma 4.4)**.** *For every round t, the following bound holds on every sample path:*

$$\Delta_t \leq -\rho\left(1 - \frac{\rho}{2}\right)Q_t^2 + (1-\rho)Q_t c_t(a_t) + \frac{G^2}{2}. \tag{31}$$

Consequently,

$$\mathbb{E}[\Delta_t \mid \mathcal{F}_{t-1}] \leq -\rho\left(1 - \frac{\rho}{2}\right)Q_t^2 + (1-\rho)Q_t\mathbb{E}[g(x_t,a_t) \mid \mathcal{F}_{t-1}] + \frac{G^2}{2}. \tag{32}$$

Proof. The projection inequality $[y]_+^2 \le y^2$ gives

$$\begin{aligned}\Delta_t &\le \frac{\big((1-\rho)Q_t + c_t(a_t)\big)^2 - Q_t^2}{2} \\ &= \frac{(1-2\rho+\rho^2)Q_t^2 - Q_t^2}{2} + (1-\rho)Q_t c_t(a_t) + \frac{c_t(a_t)^2}{2} \\ &= -\rho\left(1-\frac{\rho}{2}\right)Q_t^2 + (1-\rho)Q_t c_t(a_t) + \frac{c_t(a_t)^2}{2} \\ &\le -\rho\left(1-\frac{\rho}{2}\right)Q_t^2 + (1-\rho)Q_t c_t(a_t) + \frac{G^2}{2}.\end{aligned}$$

where the first inequality comes from the regularized virtual queue update in QUID and the last inequality uses $|c_t(a_t)| \le G$. Taking conditional expectation given $\mathcal{F}_{t-1}$ Assumption 2.1 gives $\mathbb{E}[c_t(a_t) \mid \mathcal{F}_{t-1}] = \mathbb{E}[g(x_t, a_t) \mid \mathcal{F}_{t-1}]$, which proves the drift bound. □

B.3 Proof of Lemma 4.5

The preceding bounds have complementary queue-dependent terms. Thus we are ready to prove the key regret-plus-drift lemma.

Proof. Combining Lemma B.2 and Lemma B.3, we have,

$$\begin{aligned}&\mathrm{reg}_t(\pi) + \frac{\mathbb{E}[\Delta_t \mid \mathcal{F}_{t-1}]}{(1-\rho)V} \\ &\le q_t\Big(g(\pi) - \mathbb{E}[g(x_t,a_t) \mid \mathcal{F}_{t-1}]\Big) + \frac{K+1}{2\gamma}(1+q_t^2) + 2\gamma z_t \\ &\qquad + \frac{-\rho(1-\rho/2)Q_t^2 + (1-\rho)Q_t\mathbb{E}[g(x_t,a_t) \mid \mathcal{F}_{t-1}] + G^2/2}{(1-\rho)V} \\ &= q_t g(\pi) + \frac{K+1}{2\gamma} + \frac{G^2}{2(1-\rho)V} + 2\gamma z_t + \left[\frac{K+1}{2\gamma} - \frac{\rho(1-\rho/2)V}{1-\rho}\right] q_t^2 \\ &= q_t g(\pi) + \frac{K+1}{2\gamma} + \frac{G^2}{2(1-\rho)V} + 2\gamma z_t + \left[\frac{K+1}{2\gamma} - \frac{K+1}{\gamma}\left(1-\frac{\rho}{2}\right)\right] q_t^2 \\ &= q_t g(\pi) + \frac{K+1}{2\gamma} + \frac{G^2}{2(1-\rho)V} + 2\gamma z_t - \frac{(K+1)V}{2(\gamma V + K + 1)} q_t^2 \qquad (33) \\ &\le q_t g(\pi) + \frac{K+1}{2\gamma} + \frac{G^2}{2(1-\rho)V} + 2\gamma z_t. \qquad (34)\end{aligned}$$

The equality uses $q_t = Q_t/V$ and $\rho = (K+1)/(\gamma V + K + 1)$. Since $q_t g(\pi)$ is nonpositive for any feasible comparison policy, the regret can be easily derived without first bounding the queue length as in previous works. □

B.4 Proof of the Regret Guarantee in Theorem 4.1

To bound the prediction error of estimated reward and cost functions, define the realized cumulative error and its success event by

$$S_T := \sum_{t=1}^{T}\Big[\big(\widehat{f}_t(x_t,a_t) - f(x_t,a_t)\big)^2 + \big(\widehat{g}_t(x_t,a_t) - g(x_t,a_t)\big)^2\Big], \qquad \mathcal{E}_p := \{S_T \le 2U\}.$$

Assumption 2.2 implies $\mathbb{P}(\mathcal{E}_p^c) \le p$ and the prediction ranges also imply $S_T \le (F^2 + 4G^2)T$ on every trajectory.

Proof. Applying $\pi = \pi^*$ in Lemma 4.5, summing over $t = 1, \dots, T$, and taking total expectations gives

$$
\begin{aligned}
\mathrm{Reg}_T &= \sum_{t=1}^{T} \mathbb{E}[\mathrm{reg}_t(\pi^\star)] \\
&\overset{(a)}{\leq} \sum_{t=1}^{T} \left[-\frac{\mathbb{E}[\Delta_t]}{(1-\rho)V} + \frac{K+1}{2\gamma} + \frac{G^2}{2(1-\rho)V} + 2\gamma\mathbb{E}[z_t] \right] \\
&= -\frac{\mathbb{E}[\sum_{t=1}^{T}(Q_{t+1}^2 - Q_t^2)]}{2(1-\rho)V} + \frac{(K+1)T}{2\gamma} + \frac{G^2 T}{2(1-\rho)V} + 2\gamma\mathbb{E}\left[\sum_{t=1}^{T} z_t\right] \\
&= -\frac{\mathbb{E}[Q_{T+1}^2]}{2(1-\rho)V} + \frac{(K+1)T}{2\gamma} + \frac{G^2 T}{2(1-\rho)V} + 2\gamma\mathbb{E}\left[\sum_{t=1}^{T} z_t\right] \\
&\overset{(b)}{\leq} \frac{(K+1)T}{2\gamma} + \frac{G^2 T}{2(1-\rho)V} + 2\gamma\mathbb{E}[S_T \mid \mathcal{E}_p] \cdot \mathbb{P}(\mathcal{E}_p) + 2\gamma\mathbb{E}[S_T \mid \mathcal{E}_p^c] \cdot \mathbb{P}(\mathcal{E}_p^c) \qquad (35) \\
&\leq \frac{(K+1)T}{2\gamma} + \frac{G^2 T}{2(1-\rho)V} + 4\gamma U \cdot \mathbb{P}(\mathcal{E}_p) + 2\gamma(F^2+4G^2)T \cdot \mathbb{P}(\mathcal{E}_p^c) \qquad (36) \\
&\leq \frac{(K+1)T}{2\gamma} + \frac{G^2 T}{2(1-\rho)V} + 4\gamma U + \frac{2\gamma(F^2+4G^2)}{T} \\
&= \frac{V}{2} + \frac{G^2(T+1)}{2V} + \frac{4(K+1)TU}{V} + \frac{2(K+1)(F^2+4G^2)}{V} \\
&\leq \widetilde{O}\left(\frac{TU}{V}\right) = \widetilde{O}\left(T^\theta U^{1-\theta}\right). \qquad (37)
\end{aligned}
$$

Step (a) uses feasibility and the tower property. The drift sum telescopes because $Q_1 = 0$. Step (b) uses the identity $\mathbb{E}[\sum_{t=1}^{T} z_t] = \mathbb{E}[S_T]$, and total expectation over $\mathcal{E}_p$ and its complement. The final bound follows by substituting V, γ, and ρ and using $V^2/(TU) = (U/T)^{2\theta-1} \leq 1$, $T+1 \leq 2TU$, and $TU \geq 1$ give the stated rate. In particular, the argument never requires an upper bound on Q_t or a strict feasibility margin. □

C Constraint Violation under General Feasibility

The regret analysis discards the cumulative drift since the contribution is nonpositive. In this section, we instead use it to control queue length. Specifically, the regularized update relates constraint violation to queue length, while the cumulative oracle guarantee and a stopped martingale bound control the predictable and random parts of the drift, respectively.

C.1 From Queue Length to Constraint Violation

Lemma C.1. *Under the regularized queue update, for every $n \leq T$, we have,*

$$\sum_{t=1}^{n} c_t(a_t) \leq 2 \max_{s \leq n+1} Q_s.$$

Proof. The inequality $[y]_+ \geq y$ and the queue update imply $c_t(a_t) \leq Q_{t+1} - (1-\rho)Q_t$. Summing over $1, \cdots, n$ and using $Q_1 = 0$ gives

$$
\begin{aligned}
\sum_{t=1}^{n} c_t(a_t) &\leq \sum_{t=1}^{n} (Q_{t+1} - Q_t) + \rho \sum_{t=1}^{n} Q_t \\
&= Q_{n+1} + \rho \sum_{t=1}^{n} Q_t \\
&\leq (1+\rho n) \max_{s \leq n+1} Q_s
\end{aligned}
$$

$$\leq 2 \max_{s \leq n+1} Q_s,$$

where the last inequality uses the definition of ρ and $\gamma V = (K+1)T$. Thus, a maximum of queue bound immediately gives an upper bound for constraint violation. □

C.2 Auxiliary Lemmas

Since the proof of high probability queue length requires control of both cumulative conditional prediction error and fluctuations of the realized drift, we provide several auxiliary lemmas here. We first relate the conditional risk to the error covered by the oracle guarantee. Define

$$X_t := \left(\widehat{f}_t(x_t, a_t) - f(x_t, a_t)\right)^2 + \left(\widehat{g}_t(x_t, a_t) - g(x_t, a_t)\right)^2. \tag{38}$$

The prediction ranges and the definition of z_t imply

$$0 \leq X_t \leq b := F^2 + 4G^2, \qquad \mathbb{E}[X_t \mid \mathcal{F}_{t-1}] = z_t, \qquad S_T = \sum_{t=1}^{T} X_t.$$

Lemma C.2. *Fix $\delta \in (0,1)$, and define*

$$\overline{U} := 4U + 2b \log \frac{3}{\delta}, \qquad \mathcal{E}_U := \left\{ \sum_{t=1}^{T} z_t \leq \overline{U} \right\}. \tag{39}$$

Then

$$\mathbb{P}(\mathcal{E}_U^c) \leq p + \frac{\delta}{3}. \tag{40}$$

Proof. The claim is immediate if $b = 0$. For $b > 0$, consider the exponential process

$$Z_n := \exp\left(\frac{1}{2b} \sum_{t=1}^{n} z_t - \frac{1}{b} \sum_{t=1}^{n} X_t \right), \qquad Z_0 = 1. \tag{41}$$

For $u \in [0,1]$, convexity gives $e^{-u} \leq 1 - (1 - e^{-1})u \leq 1 - u/2$. Therefore

$$\begin{aligned} \mathbb{E}[Z_n \mid \mathcal{F}_{n-1}] &= Z_{n-1} e^{z_n/(2b)} \mathbb{E}[e^{-X_n/b} \mid \mathcal{F}_{n-1}] \\ &\leq Z_{n-1} e^{z_n/(2b)} \left(1 - \frac{z_n}{2b}\right) \leq Z_{n-1}, \end{aligned} \tag{42}$$

where the last step uses $1 - v \leq e^{-v}$. Thus (Z_n) is a nonnegative supermartingale and Ville's inequality implies

$$\mathbb{P}\left(\max_{n \leq T} Z_n \geq 3/\delta \right) \leq \delta/3.$$

On the complementary event, taking logarithms gives,

$$\sum_{t=1}^{n} z_t \leq 2 \sum_{t=1}^{n} X_t + 2b \log \frac{3}{\delta}, \qquad \forall n \in [T]. \tag{43}$$

Hence the cumulative conditional risk is controlled by the realized prediction error, up to a constant factor and an additive logarithmic term. On $\mathcal{E}_p$, the realized error over any prefix is at most $2U$. Intersecting the two events and applying a union bound establishes (40). □

Lemma C.3. *Let (M_t) be a martingale difference sequence with respect to $(\mathcal{F}_t)$, and suppose $|M_t| \leq b$ almost surely for some $b \geq 0$. For every $\delta_0 \in (0,1)$,*

$$\mathbb{P}\left(\max_{n \leq T} \sum_{t=1}^{n} M_t > b\sqrt{2T \log(1/\delta_0)} \right) \leq \delta_0. \tag{44}$$

Proof. The case $b = 0$ is immediate. Otherwise, let $S_n := \sum_{t=1}^{n} M_t$ and define $Z_n(\lambda) := \exp(\lambda S_n - \lambda^2 b^2 n/2)$ for $\lambda > 0$, with $Z_0(\lambda) = 1$. Conditional Hoeffding's lemma gives

$$\begin{aligned}\mathbb{E}[Z_n(\lambda) \mid \mathcal{F}_{n-1}] &= Z_{n-1}(\lambda) e^{-\lambda^2 b^2/2} \mathbb{E}[e^{\lambda M_n} \mid \mathcal{F}_{n-1}] \\ &\le Z_{n-1}(\lambda) e^{-\lambda^2 b^2/2} e^{\lambda^2 b^2/2} = Z_{n-1}(\lambda).\end{aligned}$$

Thus $(Z_n(\lambda))$ is a nonnegative supermartingale. Set $h := b\sqrt{2T \log(1/\delta_0)}$ and $\lambda := h/(b^2 T)$. If $S_n > h$ for some $n \le T$, then $Z_n(\lambda) > \exp(\lambda h - \lambda^2 b^2 T/2)$. Consequently,

$$\begin{aligned}\mathbb{P}\left(\max_{n \le T} S_n > h\right) &\le \mathbb{P}\left(\max_{n \le T} Z_n(\lambda) > \exp\left(\lambda h - \frac{\lambda^2 b^2 T}{2}\right)\right) \\ &= \mathbb{P}\left(\max_{n \le T} Z_n(\lambda) > \exp\left(\frac{h^2}{2b^2 T}\right)\right) \\ &= \mathbb{P}\left(\max_{n \le T} Z_n(\lambda) > \frac{1}{\delta_0}\right) \le \delta_0.\end{aligned}$$

where the last inequality uses Ville's inequality and completes the proof. □

Lemma C.4. *Conditionally on $\mathcal{F}_{t-1}$, the range length of Δ_t is at most $2G(Q_t + G)$. Consequently, almost surely,*

$$\left|\Delta_t - \mathbb{E}[\Delta_t \mid \mathcal{F}_{t-1}]\right| \le 2G(Q_t + G). \tag{45}$$

Proof. Fix $Q_t = s$ and write $u := (1-\rho)s \ge 0$. As a function of the observed cost $c \in [-G, G]$, the drift is $\phi(c) := ([u + c]_+^2 - s^2)/2$. Since ϕ is nondecreasing,

$$\begin{aligned}\sup_{c \in [-G,G]} \phi(c) - \inf_{c \in [-G,G]} \phi(c) &= \phi(G) - \phi(-G) \\ &= \frac{(u+G)^2 - [u-G]_+^2}{2} \\ &\le 2G(s+G).\end{aligned}$$

For $u \ge G$, the range length is $2uG \le 2sG$; for $u < G$, it is at most $2G^2$. Since the conditional mean lies in the same interval, the centered drift satisfies (45). □

C.3 Proof of Lemma 4.6

Proof. Now we are ready to prove the queue length under general feasibility. For convenience, define

$$B := F + \frac{K+1}{2\gamma} + \frac{G^2}{2(1-\rho)V}. \tag{46}$$

Applying Lemma 4.5 with $\pi = \pi^\star$ and using $q_t g(\pi^\star) \le 0$ and $-\operatorname{reg}_t(\pi^\star) \le F$ yields

$$\mathbb{E}[\Delta_t \mid \mathcal{F}_{t-1}] \le (1-\rho)V(B + 2\gamma z_t). \tag{47}$$

We choose a deterministic threshold that balances the cumulative drift bound with the martingale fluctuations before a first crossing. Define

$$L_\delta := \sqrt{2T \log \frac{3}{\delta}}, \tag{48}$$

$$H_0(\delta) := 2GL_\delta + \sqrt{4G^2 L_\delta^2 + 2\Big(T(1-\rho)VB + 2(1-\rho)\gamma V \overline{U} + 2G^2 L_\delta\Big)}. \tag{49}$$

Write $H := H_0(\delta)$ and let $\tau_H := \inf\{t \in \{1, \dots, T+1\} : Q_t > H\}$. Also define

$$\tau_U := \min\left(\left\{t \in \{1, \dots, T\} : \sum_{s=1}^{t} z_s > \overline{U}\right\} \cup \{T+1\}\right).$$

Since the risks are nonnegative and z_t is $\mathcal{F}_{t-1}$-measurable, for every $t \le T$,

$$\{t < \tau_U\} = \left\{ \sum_{s=1}^{t} z_s \le \overline{U} \right\} \in \mathcal{F}_{t-1}.$$

For $t \le T$, set

$$J_t := \mathbf{1}\{t < \tau_U \wedge \tau_H\}, \qquad M_t := J_t\big(\Delta_t - \mathbb{E}[\Delta_t \mid \mathcal{F}_{t-1}]\big). \tag{50}$$

Both indicators are $\mathcal{F}_{t-1}$-measurable, so (M_t) is a martingale-difference sequence. On a retained round, $Q_t \le H$ and Lemma C.4 gives $|M_t| \le 2G(H+G)$. Lemma C.3, with $\delta_0 = \delta/3$, therefore shows that the event

$$\mathcal{A}_H := \left\{ \max_{n \le T} \sum_{t=1}^{n} M_t \le 2G(H+G)L_\delta \right\} \tag{51}$$

has probability at least $1 - \delta/3$. On $\mathcal{A}_H \cap \mathcal{E}_U$, suppose that $\tau_H \le T+1$. Since $\tau_U = T+1$ on $\mathcal{E}_U$, every increment before the first crossing is retained. Telescoping the quadratic potential gives

$$\begin{aligned}
\frac{H^2}{2} < \frac{Q_{\tau_H}^2}{2} &= \sum_{t<\tau_H} \Delta_t \\
&= \sum_{t<\tau_H} \mathbb{E}[\Delta_t \mid \mathcal{F}_{t-1}] + \sum_{t<\tau_H} M_t \\
&\le \sum_{t<\tau_H} (1-\rho)V\big(B + 2\gamma z_t\big) + 2G(H+G)L_\delta \\
&\le T(1-\rho)VB + 2(1-\rho)\gamma V\overline{U} + 2GHL_\delta + 2G^2L_\delta = \frac{H^2}{2}.
\end{aligned} \tag{52}$$

The bounds use (47), the risk bound on $\mathcal{E}_U$, and the martingale bound on $\mathcal{A}_H$. The final equality is the defining quadratic identity for H. The contradiction excludes a crossing on their intersection, and hence

$$\begin{aligned}
\mathbb{P}\left(\max_{t \le T+1} Q_t > H_0(\delta) \right) &\le \mathbb{P}(\mathcal{E}_U^c) + \mathbb{P}(\mathcal{A}_H^c) \\
&\le p + \frac{\delta}{3} + \frac{\delta}{3} = p + \frac{2\delta}{3}.
\end{aligned} \tag{53}$$

Finally, substituting $\gamma V = (K+1)T$ and $V = T^{1-\theta}U^\theta$ into (49) with $\delta = T^{-2}$ gives

$$\begin{aligned}
H_0(\delta) &= 2GL_\delta + \Bigg[4G^2L_\delta^2 + 2T\left\{ (1-\rho)V\left(F + \frac{V}{2T}\right) + \frac{G^2}{2} \right\} \\
&\qquad\qquad + 4(1-\rho)(K+1)T\overline{U} + 4G^2L_\delta \Bigg]^{1/2} \\
&\le 2GL_\delta + \Big[4G^2L_\delta^2 + (2F+1)TV + G^2T + 4(K+1)T\overline{U} + 4G^2L_\delta \Big]^{1/2} \\
&= O\left(\sqrt{TV} + \sqrt{T\overline{U}} + \sqrt{T \log \frac{3}{\delta}} \right) \\
&= \widetilde{O}(T^{1-\theta/2}U^{\theta/2}).
\end{aligned} \tag{54}$$

The inequality uses $1 - \rho \le 1$, $V \le T$, $V \ge 1$, and $L_\delta \ge 1$. Together with (53), we complete the proof of queue length under general feasibility.

□

C.4 Proof of Violation Guarantee under General Feasibility

Proof. Take $p = \delta = T^{-2}$ and let $\mathcal{S}_0 := \mathcal{E}_U \cap \mathcal{A}_H$. On $\mathcal{S}_0$, Lemma C.1 and the maximum-queue bound imply, simultaneously for all $n \le T$,

$$\begin{aligned}\sum_{t=1}^{n} c_t(a_t) &\le (1+\rho n) \max_{s\le n+1} Q_s \\ &\le \left(1 + \frac{n}{T+1}\right) H_0(\delta) \\ &\le 2H_0(\delta). \end{aligned} \tag{55}$$

Here $\rho = 1/(T+1)$ follows from $\gamma V = (K+1)T$. On $\mathcal{S}_0^c$, the observed cumulative cost is at most GT. Splitting the expectation over the two events yields

$$\begin{aligned}\mathrm{Viol}_T &= \mathbb{E}\left[\sum_{t=1}^{T} c_t(a_t)\mathbf{1}\{\mathcal{S}_0\}\right] + \mathbb{E}\left[\sum_{t=1}^{T} c_t(a_t)\mathbf{1}\{\mathcal{S}_0^c\}\right] \\ &\le 2H_0(\delta)\mathbb{P}(\mathcal{S}_0) + GT\mathbb{P}(\mathcal{S}_0^c) \\ &\le 2H_0(\delta) + GT\left(p + \frac{2\delta}{3}\right) \qquad (56) \\ &= \widetilde{O}(T^{1-\theta/2}U^{\theta/2}) \qquad (57)\end{aligned}$$

The final bound follows from (53) and (54), which completes constraint violation bound under the general feasibility in Theorem 4.1. □

D Constraint Violation under Slater's Condition

Slater's condition contributes a negative queue-dependent term to the drift bound, but the cumulative oracle guarantee does not ensure that this term dominates the prediction error in every round. We therefore separate rounds according to the prediction risk. Low-risk rounds provide negative drift above a suitable queue threshold, while the cumulative risk bounds the number of high-risk rounds.

The key step is to compensate the queue for the possible growth on high-risk rounds. The resulting process satisfies a one-step negative-drift condition throughout the horizon, allowing a standard drift bound to control its upper tail. Adding back the compensation yields the refined queue estimate without requiring pointwise prediction guarantees. The construction is confined to the violation analysis and leaves both the algorithm and its regret guarantee unchanged.

D.1 Proof of Lemma 4.7

Before formally proving the lemma, we present a key one-step bounded-increment drift bound, in the form underlying the analyses of (Yu et al., 2017) and (Guo & Liu, 2024). It converts negative conditional drift above a threshold into an exponential upper-tail estimate, which plays a key role in the bound of queue length. We include the proof for completeness.

Lemma D.1. *Let (Y_t) be a real-valued process with Y_t measurable with respect to $\mathcal{F}_{t-1}$ and $Y_1 \le \vartheta$. Suppose that, for constants $d > 0$ and $0 < \zeta \le d$, the following conditions hold whenever the indices lie in the horizon:*

$$|Y_{t+1} - Y_t| \le d \qquad \textit{almost surely}, \tag{58}$$

$$\mathbb{E}[Y_{t+1} - Y_t \mid \mathcal{F}_{t-1}] \le \begin{cases} d, & Y_t < \vartheta, \\ -\zeta, & Y_t \ge \vartheta. \end{cases} \tag{59}$$

Define

$$\Gamma(d,\zeta) := \frac{4d^2}{\zeta}\log\left(1 + \frac{8d^2}{\zeta^2}e^{\zeta/(4d)}\right).$$

For every t in the horizon and every $\mu \in (0,1)$,

$$\mathbb{E}[Y_t] \le \vartheta + \Gamma(d,\zeta), \tag{60}$$

$$\mathbb{P}\left(Y_t \ge \vartheta + \Gamma(d,\zeta) + \frac{4d^2}{\zeta}\log\frac{1}{\mu}\right) \le \mu. \tag{61}$$

Proof. Set $\lambda := \zeta/(4d^2)$, $\kappa := 1 - \zeta^2/(8d^2)$, and $\eta_t := Y_{t+1} - Y_t$. The increment bound gives $|\lambda\eta_t| \le \zeta/(4d) \le 1/4$. On the event $\{Y_t \ge \vartheta\}$,

$$\begin{aligned}
\mathbb{E}[e^{\lambda Y_{t+1}} \mid \mathcal{F}_{t-1}] &= e^{\lambda Y_t}\mathbb{E}[e^{\lambda\eta_t} \mid \mathcal{F}_{t-1}] \\
&\le e^{\lambda Y_t}\mathbb{E}\big[1 + \lambda\eta_t + 2\lambda^2\eta_t^2 \mid \mathcal{F}_{t-1}\big] \\
&= e^{\lambda Y_t}\left(1 + \lambda\mathbb{E}[\eta_t \mid \mathcal{F}_{t-1}] + 2\lambda^2\mathbb{E}[\eta_t^2 \mid \mathcal{F}_{t-1}]\right) \\
&\le e^{\lambda Y_t}\left(1 - \lambda\zeta + 2\lambda^2 d^2\right) \\
&= e^{\lambda Y_t}\left(1 - \frac{\zeta^2}{4d^2} + \frac{\zeta^2}{8d^2}\right) = \kappa e^{\lambda Y_t}.
\end{aligned} \tag{62}$$

The first inequality uses Taylor's theorem: for $|x| \le 1$, the second-order remainder in e^x is at most $ex^2/2 < 2x^2$. The next inequality uses the negative-drift branch of (59) and $\eta_t^2 \le d^2$.

On $\{Y_t < \vartheta\}$, bounded increments instead give $Y_{t+1} \le Y_t + d < \vartheta + d$. Combining the two cases yields

$$\begin{aligned}
\mathbb{E}[e^{\lambda Y_{t+1}}] &= \mathbb{E}\big[\mathbf{1}\{Y_t \ge \vartheta\}\mathbb{E}[e^{\lambda Y_{t+1}} \mid \mathcal{F}_{t-1}]\big] \\
&\quad + \mathbb{E}\big[\mathbf{1}\{Y_t < \vartheta\}\mathbb{E}[e^{\lambda Y_{t+1}} \mid \mathcal{F}_{t-1}]\big] \\
&\le \kappa\mathbb{E}\big[e^{\lambda Y_t}\mathbf{1}\{Y_t \ge \vartheta\}\big] + e^{\lambda(\vartheta+d)}\mathbb{P}(Y_t < \vartheta) \\
&\le \kappa\mathbb{E}[e^{\lambda Y_t}] + e^{\lambda(\vartheta+d)}.
\end{aligned} \tag{63}$$

Iterating (63) from $Y_1 \le \vartheta$ gives, for every t in the horizon,

$$\begin{aligned}
\mathbb{E}[e^{\lambda Y_t}] &\le \kappa^{t-1}\mathbb{E}[e^{\lambda Y_1}] + e^{\lambda(\vartheta+d)}\sum_{i=0}^{t-2}\kappa^i \\
&\le e^{\lambda\vartheta}\left(1 + \frac{e^{\lambda d}}{1-\kappa}\right) = e^{\lambda\vartheta}C,
\end{aligned} \tag{64}$$

where $C := 1 + \frac{8d^2}{\zeta^2}e^{\zeta/(4d)}$. The sum is empty when $t = 1$. If $\mathbb{E}[Y_t] = -\infty$, the mean bound is immediate. Otherwise, Jensen's inequality gives

$$\begin{aligned}
\mathbb{E}[Y_t] &\le \frac{1}{\lambda}\log\mathbb{E}[e^{\lambda Y_t}] \\
&\le \frac{1}{\lambda}\log(e^{\lambda\vartheta}C) = \vartheta + \frac{4d^2}{\zeta}\log\left(1 + \frac{8d^2}{\zeta^2}e^{\zeta/(4d)}\right) \\
&= \vartheta + \Gamma(d,\zeta).
\end{aligned}$$

For the tail bound, apply Markov's inequality to the same exponential moment:

$$\begin{aligned}
&\mathbb{P}\left(Y_t \ge \vartheta + \Gamma(d,\zeta) + \frac{4d^2}{\zeta}\log\frac{1}{\mu}\right) \\
&= \mathbb{P}\left(e^{\lambda Y_t} \ge \frac{e^{\lambda\vartheta}C}{\mu}\right) \\
&\le \frac{\mu\mathbb{E}[e^{\lambda Y_t}]}{e^{\lambda\vartheta}C} \le \mu.
\end{aligned}$$

Since $\lambda^{-1}\log C = \Gamma(d,\zeta)$ and $\lambda^{-1} = 4d^2/\zeta$, these estimates establish (60) and (61). □

Negative drift on low-risk rounds. We first derive negative drift on rounds with small prediction error. Applying (34) to the strictly feasible policy π^0 gives

$$\begin{aligned}\mathbb{E}[\Delta_t \mid \mathcal{F}_{t-1}] &\le (1-\rho)V\left[-\operatorname{reg}_t(\pi^0) + \frac{Q_t}{V}g(\pi^0) + \frac{K+1}{2\gamma} + 2\gamma z_t\right] + \frac{G^2}{2} \\ &\le (1-\rho)V\left[F - \frac{\epsilon Q_t}{V} + \frac{K+1}{2\gamma} + 2\gamma z_t\right] + \frac{G^2}{2} \\ &= (1-\rho)V(B+2\gamma z_t) - (1-\rho)\epsilon Q_t. \end{aligned} \tag{65}$$

The second inequality uses $-\operatorname{reg}_t(\pi^0) \le F$ and $g(\pi^0) \le -\epsilon$. The equality uses (46). The negative term grows with Q_t, whereas the term of prediction error is controlled only cumulatively. To separate these effects, fix a deterministic threshold $r > 0$ and define

$$\zeta_\epsilon := \frac{(1-\rho)\epsilon}{2}, \qquad \Theta_\epsilon(r) := \frac{2V(B+2\gamma r)}{\epsilon}. \tag{66}$$

Call round t low-risk if $z_t \le r$ and high-risk otherwise. Since z_t is $\mathcal{F}_{t-1}$-measurable, the classification is predictable. The threshold $\Theta_\epsilon(r)$ ensures that the Slater term dominates the positive drift contributions on low-risk rounds. To obtain a bound on the queue increment, write $\xi_t = Q_{t+1} - Q_t$ and use

$$\Delta_t = Q_t\xi_t + \xi_t^2/2 \ge Q_t\xi_t.$$

Consequently, on a low-risk round with $Q_t \ge \Theta_\epsilon(r) > 0$,

$$\begin{aligned}\mathbb{E}[Q_{t+1} - Q_t \mid \mathcal{F}_{t-1}] &\le \frac{\mathbb{E}[\Delta_t \mid \mathcal{F}_{t-1}]}{Q_t} \\ &\le \frac{(1-\rho)V(B+2\gamma r)}{Q_t} - (1-\rho)\epsilon \\ &\le -\frac{(1-\rho)\epsilon}{2} = -\zeta_\epsilon. \end{aligned} \tag{67}$$

Thus the queue has conditional drift at most $-\zeta_\epsilon$ on low-risk rounds whenever $Q_t \ge \Theta_\epsilon(r)$. Increasing r includes more rounds in this estimate, but also raises the queue threshold. The high-risk contribution below determines the complementary cost.

Compensation for high-risk rounds. High-risk rounds need not satisfy the preceding negative-drift condition. Instead, we bound their possible contribution using the queue increments. The update and $Q_1 = 0$ imply

$$0 \le Q_t \le \min\{G(t-1), G/\rho\}. \tag{68}$$

The bounds follow by induction from $Q_{t+1} \le (1-\rho)Q_t + G$ and $Q_{t+1} \le Q_t + G$. Combining them with $[y]_+ \ge y$ gives

$$-2G \le -\rho Q_t + c_t(a_t) \le Q_{t+1} - Q_t \le G - \rho Q_t \le G. \tag{69}$$

In particular, each high-risk round increases the queue by at most G, irrespective of the prediction error. The lower increment bound controls the fluctuations of the compensated process.

Let $N_t(r)$ count the preceding high-risk rounds and define the compensated process by

$$N_t(r) := \sum_{s=1}^{t-1} \mathbf{1}\{z_s > r\}, \qquad Y_t(r) := Q_t - (G+\zeta_\epsilon)N_t(r). \tag{70}$$

On a high-risk round, subtracting $G + \zeta_\epsilon$ offsets the largest possible queue increase and leaves an increment of at most $-\zeta_\epsilon$. The correction is predictable and is used only in the analysis; it does not alter the queue update in QUID.

Since $g(\pi^0) \ge -G$, Slater's condition implies $0 < \epsilon \le G$, and hence $0 < \zeta_\epsilon \le G/2$. Combining (69) with (70) yields

$$|Y_{t+1}(r) - Y_t(r)| \le 4G. \tag{71}$$

If $Y_t(r) \geq \Theta_\epsilon(r)$, then $Q_t \geq Y_t(r) \geq \Theta_\epsilon(r)$ because the correction is nonnegative. Therefore

$$\begin{aligned}
&\mathbb{E}[Y_{t+1}(r) - Y_t(r) \mid \mathcal{F}_{t-1}] \\
&= \mathbb{E}[Q_{t+1} - Q_t \mid \mathcal{F}_{t-1}] - (G + \zeta_\epsilon)\mathbf{1}\{z_t > r\} \\
&\leq \begin{cases} -\zeta_\epsilon, & z_t \leq r, \\ G - (G + \zeta_\epsilon), & z_t > r, \end{cases} \qquad (72) \\
&= -\zeta_\epsilon. \qquad (73)
\end{aligned}$$

The low-risk case follows from (67), since $Q_t \geq Y_t(r)$, while the high-risk case uses only the pathwise upper increment bound. Hence the compensated process satisfies a single negative-drift condition on every round above $\Theta_\epsilon(r)$. Together with (71) and $Y_1(r) = 0$, this verifies Lemma D.1 with $d = 4G$, $\zeta = \zeta_\epsilon$, and $\vartheta = \Theta_\epsilon(r)$.

Recovering the queue bound. We control the maximum of the compensated process and then bound the correction by the cumulative conditional risk. Define the tail term

$$\begin{aligned}
D_\epsilon(\delta) &:= \Gamma(4G, \zeta_\epsilon) + \frac{64G^2}{\zeta_\epsilon} \log \frac{3(T+1)}{\delta} \\
&= \frac{64G^2}{\zeta_\epsilon} \left[\log \left(1 + \frac{128G^2}{\zeta_\epsilon^2} e^{\zeta_\epsilon/(16G)} \right) + \log \frac{3(T+1)}{\delta} \right]. \qquad (74)
\end{aligned}$$

For a fixed deterministic r, apply Lemma D.1 at each time with $\mu = \delta/(3(T+1))$. A union bound yields

$$\mathbb{P}\left(\max_{t \leq T+1} Y_t(r) > \Theta_\epsilon(r) + D_\epsilon(\delta) \right) \leq \frac{\delta}{3}. \qquad (75)$$

The concentration bound is obtained without conditioning on $\mathcal{E}_U$. On $\mathcal{E}_U$, each high-risk round contributes more than r to the cumulative conditional risk, so

$$N_{T+1}(r) = \sum_{t=1}^{T} \mathbf{1}\{z_t > r\} \leq \frac{1}{r} \sum_{t=1}^{T} z_t \leq \frac{\overline{U}}{r}. \qquad (76)$$

Since $N_t(r)$ is nondecreasing, the correction is bounded uniformly over time. On the intersection of $\mathcal{E}_U$ and the success event in (75),

$$\begin{aligned}
\max_{t \leq T+1} Q_t &\leq \Theta_\epsilon(r) + D_\epsilon(\delta) + (G + \zeta_\epsilon)\frac{\overline{U}}{r} \\
&= \frac{2VB}{\epsilon} + \frac{4\gamma V}{\epsilon} r + (G + \zeta_\epsilon)\frac{\overline{U}}{r} + D_\epsilon(\delta). \qquad (77)
\end{aligned}$$

The four terms have distinct roles. The first is the threshold needed to absorb the queue-independent drift cost. The second is the additional threshold needed to absorb prediction error on low-risk rounds. The third pays for the high-risk correction, and the last controls stochastic fluctuations of the corrected process. Only the middle two depend on r: increasing r raises the low-risk threshold but reduces the upper bound on the number of high-risk rounds.

We therefore balance these two terms, rather than require a per-round prediction guarantee. Their sum is minimized when they are equal, which gives

$$r_\epsilon := \sqrt{\frac{(G + \zeta_\epsilon)\overline{U}\epsilon}{4\gamma V}}. \qquad (78)$$

Let $\mathcal{A}_\epsilon$ be the success event in (75) for this choice. Substituting r_ϵ into (77) yields, on $\mathcal{E}_U \cap \mathcal{A}_\epsilon$,

$$\max_{t \leq T+1} Q_t \leq \frac{2VB}{\epsilon} + 4\sqrt{\frac{\gamma V(G + \zeta_\epsilon)\overline{U}}{\epsilon}} + D_\epsilon(\delta) =: M_\epsilon(\delta). \qquad (79)$$

By Lemma C.2 and (75),

$$\mathbb{P}\left(\max_{t \leq T+1} Q_t > M_\epsilon(\delta) \right) \leq p + \frac{2\delta}{3}. \qquad (80)$$

The balanced bound separates the baseline Slater threshold, the cumulative cost of prediction error, and the tail penalty. In particular, the effect of high-risk rounds is charged to a cumulative budget instead of being required to satisfy negative drift individually.

Uniform dependence on the Slater margin. To obtain the stated rate from (79), we also need to control the margin dependence in the tail penalty. Its logarithm contains $1/\zeta_\epsilon^2$, which can be large when the Slater margin is very small. We handle this explicitly: for margins at least G/T, the logarithm is bounded by a horizon-dependent term; for smaller margins, the deterministic queue cap suffices. Let

$$\overline{M}_\epsilon(\delta) := \min\{GT, M_\epsilon(\delta)\}.$$

By (68), the probability bound (80) remains valid with $M_\epsilon(\delta)$ replaced by $\overline{M}_\epsilon(\delta)$.

For $G/T \le \epsilon \le G$, the relation $1-\rho \ge 1/2$ implies $\epsilon/4 \le \zeta_\epsilon \le G/2$. Consequently,

$$D_\epsilon(\delta) \le \frac{256G^2}{\epsilon}\left[\log\left(1+2048T^2e^{1/32}\right)+\log\frac{3(T+1)}{\delta}\right]. \tag{81}$$

Here $128G^2/\zeta_\epsilon^2 \le 2048T^2$ follows from $\epsilon \ge G/T$, while $\zeta_\epsilon/(16G) \le 1/32$. Using $B = O(1)$, $\gamma V = (K+1)T$, and $\overline{U} = O(U\log T)$ at $p = \delta = T^{-2}$ in (79), we obtain

$$\begin{aligned}\overline{M}_\epsilon(\delta) &= O\left(\frac{V+\log T}{\epsilon}+\sqrt{\frac{TU\log T}{\epsilon}}\right)\\ &= \widetilde{O}\left(\frac{\sqrt{TU}}{\epsilon}\right).\end{aligned} \tag{82}$$

The first line keeps the different margin dependences visible: the baseline and tail contributions scale as $1/\epsilon$, while the balanced prediction error contribution scales as $1/\sqrt{\epsilon}$. To express them through the single rate in the theorem, the last line uses

$$V \le \sqrt{TU}, \qquad \frac{V}{\sqrt{TU}} = (U/T)^{\theta-1/2} \le 1, \qquad \frac{1}{\sqrt{\epsilon}} \le \frac{\sqrt{G}}{\epsilon}.$$

For $0 < \epsilon < G/T$, the deterministic bound gives $\overline{M}_\epsilon(\delta) \le GT \le G^2\sqrt{TU}/\epsilon$. We have therefore established, over all admissible margins,

$$\overline{M}_\epsilon(\delta) = \widetilde{O}\left(\frac{\sqrt{TU}}{\epsilon}\right). \tag{83}$$

Together with (80), we finally finish the proof of Lemma 4.7 and get the upper bound of queue length.

D.2 Proof of Constraint Violation Guarantee Under Slater's Condition

The general-feasibility bound remains valid under Slater's condition, so the final guarantee can use whichever queue estimate is smaller. Both estimates concern the same algorithm and share the conditional-risk event $\mathcal{E}_U$. We intersect their concentration events once, apply the queue-accounting inequality with the smaller threshold on that intersection, and use bounded costs on its complement. This retains the general-feasibility guarantee when the Slater margin is too small for the refined estimate to improve it.

Proof. Let $\mathcal{S} := \mathcal{E}_U \cap \mathcal{A}_H \cap \mathcal{A}_\epsilon$. The preceding probability bounds give

$$\mathbb{P}(\mathcal{S}^c) \le \mathbb{P}(\mathcal{E}_U^c) + \mathbb{P}(\mathcal{A}_H^c) + \mathbb{P}(\mathcal{A}_\epsilon^c) \le p + \delta. \tag{84}$$

In particular, the shared oracle and conditional-risk event is counted only once. Set $M := \min\{H_0(\delta), \overline{M}_\epsilon(\delta)\}$. On $\mathcal{S}$, Lemma C.1 gives

$$\sum_{t=1}^{T} c_t(a_t) \le (1+\rho T)\max_{t\le T+1} Q_t \le 2M. \tag{85}$$

On $\mathcal{S}^c$, the observed cumulative cost is at most GT. Using the metric relation and splitting the expectation over the two events, we obtain

$$\begin{aligned}\mathrm{Viol}_T &= \mathbb{E}\left[\sum_{t=1}^{T} g(x_t, a_t)\right] = \mathbb{E}\left[\sum_{t=1}^{T} c_t(a_t)\right] \\ &\leq 2M\mathbb{P}(\mathcal{S}) + GT\mathbb{P}(\mathcal{S}^c) \\ &\leq 2\min\{H_0(\delta), \overline{M}_\epsilon(\delta)\} + GT(p+\delta). \end{aligned} \tag{86}$$

The first two equalities use the definition of Viol_T and $\mathbb{E}[c_t(a_t) \mid \mathcal{I}_t] = g(x_t, a_t)$, followed by the tower property. Taking $p = \delta = T^{-2}$ and applying (54) and (83) now gives

$$\mathrm{Viol}_T = \widetilde{O}\left(\min\left\{T^{1-\theta/2}U^{\theta/2}, \frac{\sqrt{TU}}{\epsilon}\right\}\right). \tag{87}$$

The failure contribution is $2G/T$. Combining this result with the regret guarantee in Section B.4 and the violation guarantee under general feasibility in Section C.4 completes the proof of Theorem 4.1.

□