IMPROVED REGRET BOUNDS FOR STOCHASTIC CONTEXTUAL BANDITS WITH GENERAL CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider the contextual bandit problem with general constraints, in which the learner sequentially observes contexts and seeks to maximize cumulative reward subject to a broad class of feasibility requirements. We operate under a general realizability assumption, positing that the expected reward and cost functions belong to rich function classes. Prevailing approaches predominantly employ primal–dual methods, which generally yield $\tilde{\mathcal{O}}(\sqrt{T})$ bounds on both regret and cumulative constraint violation. By contrast, we develop a purely primal framework that exploits online regression oracles to estimate the reward and cost functions, and proceeds by sequentially solving approximations of the offline benchmark problem. We establish that the proposed algorithm attains logarithmic instance-dependent regret while simultaneously preserving $\tilde{\mathcal{O}}(\sqrt{T})$ guarantees on constraint violation. These results hinge on a refined analysis of the offline benchmark structure.

1 INTRODUCTION

Stochastic contextual bandits provide a general online learning framework for interactive decision-making in uncertain environments (Abbasi-Yadkori et al., 2011; Chu et al., 2011; Li et al., 2017; Lattimore & Szepesvári, 2020). At each step of a T-round horizon, the learner observes a context and selects an action accordingly, receiving a payoff that depends on both the context and the chosen action through an unknown function. A central challenge is to balance *exploration*, i.e., acquiring information about the unknown reward structure, with *exploitation*, i.e., leveraging current knowledge to optimize performance. Numerous extensions of the bandit framework have been proposed to address practical constraints arising in real-world applications. In clinical trials, for instance, safety requirements impose strict constraints on treatment–dosage combinations in order to balance efficacy against adverse effects (Chen et al., 2022b; Pacchiano et al., 2021; Amani et al., 2019). Budget-constrained settings give rise to *knapsack bandits*, in which the learner seeks to maximize cumulative reward subject to limited resources (Badanidiyuru et al., 2018; Chzhen et al., 2023). Fairness considerations further motivate constraints such as equitable exposure across arms (Wang et al., 2021; Li et al., 2020) or guaranteed minimum revenue levels (Ben Yahmed et al., 2026). These applications necessitate extending the classical bandit framework to incorporate reward maximization under diverse constraints.

In this work, we study stochastic contextual bandits with general constraints, accommodating a wide range of constraint structures. Rewards and costs both belong to rich function classes, and the learner seeks to maximize cumulative reward while minimizing cumulative constraint violation, relying solely on a realizability assumption and access to online regression oracles.

2 RELATED WORK AND CONTRIBUTIONS

Constrained bandit problems have been studied from two main perspectives.

Hard constraints. A first line of work considers *hard constraints*, under which constraint violations are strictly forbidden. This includes the extensive literature on *knapsack bandits* (Tran-Thanh et al., 2012; Badanidiyuru et al., 2014; Agrawal et al., 2016; Sivakumar et al., 2022; Kumar & Kleinberg, 2022; Han et al., 2023; Chzhen et al., 2023; Guo & Liu, 2025; Liu, 2025; Deb & Saha, 2023), in

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.

which the constraint corresponds to the consumption of a known initial budget B, and the interaction terminates once the budget is exhausted, thereby ensuring zero violation. Chen et al. (2024) study a related contextual decision-making framework with knapsack constraints under a different feedback structure, in which the reward and cost functions are known and depend on a random request—an external stochastic variable—and on the chosen action. Leveraging techniques from network revenue management (Chen et al., 2022a; Balseiro et al., 2024), they employ a re-solving strategy conceptually similar to ours and provide guarantees in both full- and partial-information regimes, including beyond worst-case analyses. *Conservative bandits* (Kazerouni et al., 2017; Deb et al., 2025a;b) provide another instance of hard constraints, in which at each round the agent's policy must yield a reward at least as high as that of a known baseline policy. A further stream of hard-constrained bandit work assumes prior knowledge of a feasible (safe) policy and relies on optimistic–pessimistic confidence sets to guarantee zero constraint violation while achieving $\tilde{\mathcal{O}}(\sqrt{T})$ regret. Representative examples include Amani et al. (2019) and Pacchiano et al. (2021), which focus on linear reward and cost models, and their extension to nonlinear settings in Pacchiano et al. (2025).

Soft constraints. In the absence of external information—such as a known initial budget or a pre-specified safe policy/baseline—soft-constrained bandit settings allow for round-wise constraint violations and focus on the fundamental trade-off between performance and constraint satisfaction, the two central metrics in constrained bandit problems. Notably, Ben Yahmed et al. (2026) extend the minimum-revenue constraint framework of Baudry et al. (2024) to the contextual setting and study the resulting trade-off between performance and constraint satisfaction. Gangrade et al. (2024) propose an algorithm for safe linear bandits that learns the optimal action $\mathrm{argmax}_{x\in\mathbb{R}^d} x^\top\theta^\star \quad \text{s.t.} \quad Ax \le b$, achieving poly-logarithmic regret with $\tilde{\mathcal{O}}(\sqrt{T})$ cumulative constraint violation. The constraint formulation most closely related to ours (i.e., with general functions) is studied in Slivkins et al. (2024): assuming strict feasibility characterized by a *known* Slater constant ν, the authors introduce a primal–dual algorithm and establish $\tilde{\mathcal{O}}(\sqrt{T}/\nu)$ bounds on both regret and constraint violation. Guo & Liu (2024) remove the Slater condition, obtaining $\tilde{\mathcal{O}}(T^{3/4})$ bounds for both metrics; when Slater's condition holds, their method improves to $\tilde{\mathcal{O}}(\sqrt{T}/\nu^2)$ without requiring prior knowledge of ν. More recently, Guo et al. (2025) consider the same setting and assumptions as Guo & Liu (2024) but combine a primal–dual approach with optimistic estimation to achieve $\tilde{\mathcal{O}}(\sqrt{T})$ bounds on both regret and constraint violation, without requiring Slater's condition. Table 1 summarizes and compares these results with ours under the same general constraint formulation.

Problem-dependent guarantees. Instance-dependent rates beyond $\sqrt{T}$ have so far been obtained only for specific constraint structures. In the non-contextual knapsack setting, Flajolet & Jaillet (2017) and Sankararaman & Slivkins (2021) derive logarithmic regret under non-degeneracy-type conditions on the underlying LP, the latter also characterizing the instances for which such rates are achievable, and Kumar & Kleinberg (2022) obtain logarithmic regret against the LP relaxation in a variant with non-monotonic resource consumption. Closest in spirit to our approach, Wu et al. (2015) study contextual bandits with a budget constraint, finitely many contexts, and a known context distribution, and show that an LP-based policy re-solved with optimistic reward estimates achieves logarithmic regret outside degenerate "boundary" cases, where the rate falls back to $\sqrt{T}$. In the safe-bandit literature, Chen et al. (2022b) obtain logarithmic regret and risk for non-contextual arms under round-wise constraints. Our work extends this problem-dependent perspective to general soft constraints with rewards and costs in general function classes accessed through regression oracles: the gap $\Delta^\star$ we introduce is read directly off the geometry of the policy LP, without assuming a budget structure, unit or known costs, or linearity, and yields logarithmic regret alongside $\tilde{\mathcal{O}}(\sqrt{T})$ violation.

Contributions. Existing approaches to stochastic contextual bandits with general constraints predominantly rely on primal–dual methods, yielding $\widetilde{\mathcal{O}}(\sqrt{T})$ bounds for both regret and cumulative constraint violation. Differences across methods stem mainly from how Slater's condition is assumed or relaxed. This naturally raises the following question:

Is it possible to obtain guarantees beyond the $\sqrt{T}$ regime without assuming Slater's condition?

Our work provides the first affirmative answer to this question, through a primal approach that yields problem-dependent guarantees. Our main contributions are summarized as follows. We reduce the

Reference	Regret/Violation	Known context distribution	Slater's condition free/ Slater constant agnostic
Slivkins et al. (2024)	$\tilde{\mathcal{O}}\left(\frac{\sqrt{T}}{\nu}\right), \tilde{\mathcal{O}}\left(\frac{\sqrt{T}}{\nu}\right)$	✗	✗, ✗
Guo & Liu (2024)	$\tilde{\mathcal{O}}\left(T^{\frac{3}{4}}\right), \tilde{\mathcal{O}}\left(T^{\frac{3}{4}}\right)$	✗	✓, ✓
	$\tilde{\mathcal{O}}\left(\frac{\sqrt{T}}{\nu^2}\right), \tilde{\mathcal{O}}\left(\frac{\sqrt{T}}{\nu^2}\right)$	✗	✗, ✓
Guo et al. (2025)	$\tilde{\mathcal{O}}\left(\sqrt{T}\right), \tilde{\mathcal{O}}\left(\sqrt{T}\right)$	✓	✓, ✓
OPALP (Ours)	$\mathcal{O}\left(\frac{\log T}{\Delta^\star}\right), \tilde{\mathcal{O}}(\sqrt{T})$	✓ (readily relaxed)	✓, ✓

Table 1: Comparison with related work under general constraints.

contextual bandit problem with general constraints to a linear program (LP) defined over the policy space. By exploiting the structure of this LP, we introduce a tailored notion of sub-optimality gap that identifies the optimal policy via a geometric duality between active sets and policy localization. We then propose the **OPALP** algorithm, which leverages reward and cost estimates supplied by external online regression oracles to solve a noisy approximation of the oracle LP. Through a refined analysis of how estimation errors propagate into the localization of the optimal policy, we establish problem-dependent logarithmic regret guarantees together with $\sqrt{T}$ bounds on cumulative constraint violation. We further instantiate our approach in the celebrated contextual knapsack bandit setting. We additionally discuss the computational complexity of our method and propose strategies to mitigate it, as it proves more demanding than primal–dual approaches. Finally, we corroborate our theoretical results through several numerical experiments.

In doing so, we highlight a trade-off between improved regret guarantees and computational cost, and we fill a gap in the existing literature – which has primarily addressed either very large contextual systems or non-contextual ones – by establishing sharper results for the moderately large contextual regime delineated by the available computational resources.

3 Setting

Formulation. We study a stochastic constrained contextual bandit problem defined by the tuple $(\mathcal{C}, \mathcal{A}, \mathcal{R}, \Psi)$, where $\mathcal{C}$ denotes the context set, $\mathcal{A}$ the action set, $\mathcal{R}$ a class of reward functions, and Ψ a class of cost functions. At each round $t \in [T]$, a context $c_t \in \mathcal{C}$ is observed, drawn independently from a distribution $\mathcal{P}_\mathcal{C}$. An action $a_t \in \mathcal{A}$ is then selected, after which a stochastic reward $r_t(c_t, a_t) \in [0, 1]$ and a stochastic cost $\psi_t(c_t, a_t) \in [-1, 1]$ are observed. Note that we consider a single constraint for ease of exposition, and our results are readily generalized to multiple constraints. We assume a stochastic environment in which the arrival of contexts is i.i.d., and rewards and costs are generated from i.i.d. unknown distributions conditioned on the selected context–action pair. The resulting sequential interaction protocol is summarized below.

Stochastic Constrained Contextual Bandits with General Constraints:

Initialize history $\mathcal{H}_0 \leftarrow \emptyset$.
for $t = 1$ **to** T **do**
  Observe the context c_t.
  Select an action $a_t \in \mathcal{A}$.
  Observe the reward $r_t(c_t, a_t)$ and the cost $\psi_t(c_t, a_t)$.
  Update the history $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1} \cup \{c_t, a_t, r_t(c_t, a_t), \psi_t(c_t, a_t)\}$.

We further impose a standard realizability condition on the reward and cost functions (Slivkins et al., 2024; Han et al., 2023; Guo & Liu, 2024; Chzhen et al., 2023; Guo et al., 2025).

Assumption 3.1. *There exist a mean reward function $r \in \mathcal{R}$ and a mean cost function $\psi \in \Psi$ such that, for all $c \in \mathcal{C}$, all $a \in \mathcal{A}$, and all $t \in [T]$,*

$$r(c, a) = \mathbb{E}[r_t(c, a) \mid c, a], \quad \psi(c, a) = \mathbb{E}[\psi_t(c, a) \mid c, a].$$

We denote by $\boldsymbol{r}(c)$ and $\boldsymbol{\psi}(c)$ the reward and cost vectors, respectively, over the action set $\mathcal{A}$ given context c. In the main text, following Guo et al. (2025), we assume for ease of presentation that the context distribution is known; we discuss in Section E.1 that this assumption is mild and can be readily relaxed.

Planning problem. We consider a policy space Π, where each policy is defined as a mapping $\pi : \mathcal{C} \to \Delta_{\mathcal{A}}$ that assigns to every context $c \in \mathcal{C}$ a probability distribution over the action set $\mathcal{A}$. Throughout the paper, $\boldsymbol{\pi}(c)$ denotes the probability distribution vector under context c, and $\pi(c, a)$ represents the probability of selecting action a under context c. The learner's objective is to design a (dynamic) policy π that maximizes the cumulative reward while ensuring a non-positive cumulative cost:

$$\max_{\pi} \sum_{t=1}^{T} r_t(c_t, a_t) \quad \text{s.t.} \quad \sum_{t=1}^{T} \psi_t(c_t, a_t) \leq 0. \tag{1}$$

This formulation is general enough to encompass a variety of constraints, including, for example:

(i) *Fairness constraints*, obtained by defining $\psi_t(c, a) = \lambda_k - r_t(c, a)\mathbb{1}_{[a=k]}$, in which case the cumulative constraint reads $\sum_{t=1}^{T} r_t(c_t, a_t)\, \mathbb{1}_{[a_t=k]} \geq \lambda_k T$, thereby guaranteeing a minimum cumulative revenue of $\lambda_k T$ for each arm k.

(ii) *Budget or resource constraints*, obtained by setting $\psi_t(c, a) = w_t(c, a) - \frac{B}{T}$, where $w_t(c, a)$ denotes the budget consumed by action a in context c at time t, and B is the total available budget. The cumulative constraint reads $\sum_{t=1}^{T} w_t(c_t, a_t) \leq B$. If the interaction terminates once the initial budget is exhausted, this formulation recovers the classical bandits-with-knapsacks setting, studied in detail in Section 6. Note that the standard knapsack setting (Badanidiyuru et al., 2018) involves multiple resource consumption costs subject to multiple budget constraints; as mentioned earlier, this poses no difficulty, since our approach readily generalizes to multiple constraints as discussed in Section E.2.

Offline oracle problem. To benchmark any learning strategy, we consider the offline setting in which the learner has full knowledge of the mean reward and cost functions, as well as of the context distribution. The corresponding oracle policy $\pi^\star$ is defined as the solution to the following optimization problem:

$$\max_{\pi} \mathop{\mathbb{E}}_{c \sim \mathcal{P}_{\mathcal{C}}} [\langle \boldsymbol{\pi}(c), \boldsymbol{r}(c) \rangle] \quad \text{s.t.} \quad \mathop{\mathbb{E}}_{c \sim \mathcal{P}_{\mathcal{C}}} [\langle \boldsymbol{\pi}(c), \boldsymbol{\psi}(c) \rangle] \leq 0. \tag{2}$$

Throughout the paper, we adopt the following technical assumption, which ensures the well-posedness of our problem.

Assumption 3.2. *The oracle problem* (2) *is feasible and non-degenerate.*

The offline benchmark (2) is a linear program in the decision variable π. Explicitly, it can be written as

$$\begin{aligned} \mathrm{LP}(r, \psi) := \max_{\pi \in \mathbb{R}^{|\mathcal{A}| \times |\mathcal{C}|}} \sum_{c \in \mathcal{C}} p_c \, \langle \boldsymbol{\pi}(c), \boldsymbol{r}(c) \rangle \quad \text{s.t.} \quad & \sum_{c \in \mathcal{C}} p_c \, \langle \boldsymbol{\pi}(c), \boldsymbol{\psi}(c) \rangle \leq 0, \\ & \sum_{a \in \mathcal{A}} \pi(c, a) = 1, \quad \forall\, c \in \mathcal{C}, \\ & \pi(c, a) \geq 0, \quad \forall\, c \in \mathcal{C},\ \forall\, a \in \mathcal{A}, \end{aligned} \tag{3}$$

where p_c denotes the probability of observing context c under $\mathcal{P}_{\mathcal{C}}$. With a slight abuse of notation, we let $\pi^\star \in \operatorname{argmax} \mathrm{LP}(r, \psi)$ denote the optimal oracle policy and $v^\star = \mathrm{LP}(r, \psi)$ the corresponding optimal value. As established in Lemma 3.1 (the proof is deferred to Section B.1), the quantity $v^\star$ provides an upper bound on the optimal value of the online problem (1).

Lemma 3.1. *Under Assumptions 3.1 and 3.2, let π denote the optimal policy of problem* (1)*. Then,*

$$\mathbb{E}\left[\sum_{t=1}^{T} \langle \boldsymbol{\pi}(c_t), \boldsymbol{r}(c_t) \rangle\right] \leq T\, v^\star.$$

Learning Problem. At each round t, the learner exploits the available information $\mathcal{H}_{t-1}$ to select an action according to a learned policy. To assess the performance of the learner's algorithm, we consider two key metrics: the cumulative regret and the cumulative constraint violation.

Definition 3.1. *The cumulative regret $\mathcal{R}_T$ and the cumulative constraint violation $\mathcal{V}_T$ are defined, respectively, as*

$$\mathcal{R}_T := T\,v^\star - \mathbb{E}\left[\sum_{t=1}^{T} r_t(c_t, a_t)\right], \qquad \mathcal{V}_T := \mathbb{E}\left[\sum_{t=1}^{T} \psi_t(c_t, a_t)\right].$$

The regret measures the expected cumulative loss incurred by the learner relative to an oracle that plays the optimal offline policy $\pi^\star$ at every round. The cumulative constraint violation quantifies the expected total cost accumulated along the learning trajectory, thereby capturing the extent to which the algorithm violates the prescribed constraint.

Online regression oracles. Following Foster et al. (2018); Foster & Rakhlin (2020); Chzhen et al. (2023); Slivkins et al. (2024); Guo & Liu (2025), under the realizability assumption we further assume access to efficient reward and cost estimators, which are essential for learning.

Assumption 3.3. *We have access to online regression oracles $\{\Omega_r, \Omega_\psi\}$ for the reward and cost functions, producing estimators $\hat{r}_t$ and $\check{\psi}_t$ such that, with probability at least $1-\delta$, the event*

$$\xi := \left\{ \begin{array}{l} \forall\, t \in [T],\ \forall\, c \in \mathcal{C},\ \forall\, a \in \mathcal{A}: \\ \quad 0 \le \hat{r}_t(c,a) - r(c,a) \le \varepsilon_t(c,a,\delta), \\ \quad 0 \le \psi(c,a) - \check{\psi}_t(c,a) \le \varepsilon_t(c,a,\delta) \end{array} \right\}$$

holds. Moreover, the estimation errors satisfy $\sum_{t=1}^{T} \varepsilon_t(c_t, a_t, \delta)^2 = \mathcal{O}\Big(\log\Big(\frac{\max\{|\mathcal{R}|,|\Psi|\}}{\delta}\Big)\Big)$.[1]

This is a standard assumption in prior work, appearing in essentially the same form in Chzhen et al. (2023, Assumption 3.2), Guo & Liu (2025, Assumption 2) and Guo et al. (2025, Assumption 2). Foster et al. (2018) show that, for general function classes under a realizability assumption, weighted online regression estimators satisfy this condition; moreover, the associated error bounds can be computed efficiently via a binary search procedure. For completeness, in Section A we provide a theoretical justification of Assumption 3.3. Throughout the paper, we fix $\delta = 1/T^2$.

Remark 3.1. *A detailed discussion of all assumptions—3.1, 3.2 and 3.3—including their limitations and possible relaxations, is provided in Appendix A. We reiterate that, in the main text, we restrict our attention to the case of a known context distribution and a single global constraint in order to present our contribution and technical novelty more clearly; in Section E, however, we show that our method extends naturally to the setting with an unknown context distribution and multiple constraints.*

4 Suboptimality Characterization

Previous primal–dual optimization approaches (Slivkins et al., 2024; Guo & Liu, 2024; Guo et al., 2025) primarily rely on efficiently controlling a Lyapunov drift, a technique inherited from the mirror descent literature. Such methods typically yield bounds of order $\widetilde{\mathcal{O}}(\sqrt{T})$. In contrast, we adopt a purely primal approach that consists of solving a carefully designed linear program at each round. A key theoretical contribution of our work is to exploit the structural properties of this LP and introduce a tailored notion of sub-optimality gap that distinguishes optimal from suboptimal policies, thereby enabling problem-dependent regret bounds.

The offline oracle problem $\mathrm{LP}(r, \psi)$ defined in (3) is a linear program over the policy space, constrained by the intersection of the simplex with the feasible region induced by the general constraint. For convenience, we refer to the generic constraint as **GC** and define $\mathcal{J} := \{(c,a) \mid c \in \mathcal{C},\ a \in \mathcal{A}\}$ to be the set of all context–action pairs, so that the inequality constraints of $\mathrm{LP}(r, \psi)$ are indexed

[1] Following the convention in the contextual bandit literature (Foster & Krishnamurthy, 2021), we focus on finite function classes, i.e., $\max\{|\mathcal{R}|, |\Psi|\} < \infty$. Handling infinite classes requires more involved complexity notions, such as the eluder dimension (Russo & Van Roy, 2013), which lies beyond the scope of this paper.

by $\{\mathbf{GC}\} \cup \mathcal{J}$. Our analysis rests on the classical correspondence between vertices of the feasible polytope and the sets of constraints they saturate; the relevant combinatorial objects are formalized as follows.

Definition 4.1 (Candidate sets)**.** *A* candidate set *is a subset $\mathcal{I} \subseteq \{\boldsymbol{GC}\} \cup \mathcal{J}$ of $|\mathcal{I}| = |\mathcal{C}|\,(|\mathcal{A}| - 1)$ constraint indices defining, jointly with the simplex equalities of* (3)*, a basic index set. We denote by $\mathfrak{B}$ the (finite) family of candidate sets, and by $\pi_{\mathcal{I}} \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{C}|}$ a policy that activates $\mathcal{I}$.*

Under Assumption 3.2, the optimal policy $\pi^\star$ is a non-degenerate vertex of (3) and is therefore characterized by a *unique* optimal basis $\mathcal{I}^\star \in \mathfrak{B}$ of active (or binding) constraints, interpreted as follows:

(i) If $\mathbf{GC} \in \mathcal{I}^\star$, then $\sum_{c\in\mathcal{C}} p_c \,\langle \boldsymbol{\pi}^\star(c), \boldsymbol{\psi}(c)\rangle = 0$, indicating that the policy $\pi^\star$ saturates the generic constraint.

(ii) If $(c, a) \in \mathcal{I}^\star \cap \mathcal{J}$, then $\pi^\star(c, a) = 0$, indicating that the policy $\pi^\star$ never selects action a in context c.

There is thus a geometric duality between a basic policy and the candidate set $\mathcal{I} \in \mathfrak{B}$ of constraints of $\mathrm{LP}(r, \psi)$ that it saturates. Introducing the intermediate object $\mathcal{I}$ allows us to recover a notion of gap analogous to that of standard multi-armed bandit problems. Specifically, we define $\Delta(\mathcal{I})$ to quantify the sub-optimality of a candidate set, i.e., the extent to which $\mathcal{I} \neq \mathcal{I}^\star$ within $\mathfrak{B}$. This quantity will play a central role in establishing that $\mathcal{I}^\star$ can be efficiently identified by our algorithm.

Optimality Characterization. The sub-optimality of a candidate set $\mathcal{I} \in \mathfrak{B}$ may arise from its infeasibility; that is, no policy may saturate the set under the true parameters. This notion is formalized in the following definition.

Definition 4.2 (Infeasibility gap)**.** *For any candidate set $\mathcal{I} \in \mathfrak{B}$, let $s(\mathcal{I}) := \min\{s \geq 0 : \Gamma(s, \mathcal{I}) \neq \emptyset\}$, where*

$$\Gamma(s,\mathcal{I}) := \left\{ \pi \in \Pi : \begin{array}{l} \displaystyle\sum_{c\in\mathcal{C}} p_c \,\langle \boldsymbol{\pi}(c), \boldsymbol{\psi}(c)\rangle \leq s, \\ \displaystyle\sum_{c\in\mathcal{C}} p_c \,\langle \boldsymbol{\pi}(c), \boldsymbol{\psi}(c)\rangle \geq -s \text{ if } \boldsymbol{GC} \in \mathcal{I}, \\ \pi(c,a) = 0 \quad \forall\,(c,a) \in \mathcal{J} \cap \mathcal{I} \end{array} \right\}.$$

The quantity $s(\mathcal{I})$ captures the minimum slack required to render the general constraint feasible, as well as the minimum margin needed to relax its saturation while satisfying the policy sparsity prescribed by $\mathcal{J} \cap \mathcal{I}$ (i.e., the set of pairs $(c, a) \in \mathcal{J}$ at which $\pi(c, a)$ is constrained to vanish). It is immediate that, if $s(\mathcal{I}) = 0$, then $\Gamma(0, \mathcal{I}) \neq \emptyset$, and every $\pi \in \Gamma(0, \mathcal{I})$ is feasible for the oracle problem (2). Furthermore, if $\mathbf{GC} \in \mathcal{I}$, then for every $\pi \in \Gamma(0, \mathcal{I})$, $\sum_{c\in\mathcal{C}} p_c \,\langle \boldsymbol{\pi}(c), \boldsymbol{\psi}(c)\rangle = 0$.

Beyond infeasibility, we also quantify sub-optimality from a performance standpoint.

Definition 4.3 (Performance gap)**.** *For any candidate set $\mathcal{I} \in \mathfrak{B}$, define $f(\boldsymbol{\pi}, \boldsymbol{r}) := \mathbb{E}_{c\sim\mathcal{P}_\mathcal{C}}[\langle \boldsymbol{\pi}(c), \boldsymbol{r}(c)\rangle]$. The* performance sensitivity *$\mathcal{L}(\mathcal{I})$ is given by*

$$\mathcal{L}(\mathcal{I}) := \min\Big\{\mathcal{L} \geq 0 \;:\; \forall\, s(\mathcal{I}) \leq s_1 < s_2, \quad \max_{\pi\in\Gamma(s_2,\mathcal{I})} f(\boldsymbol{\pi}, \boldsymbol{r}) - \max_{\pi\in\Gamma(s_1,\mathcal{I})} f(\boldsymbol{\pi}, \boldsymbol{r}) \leq \mathcal{L}\,(s_2 - s_1)\Big\},$$

and the performance gap *is defined as $\rho(\mathcal{I}) := \frac{v^\star - \max_{\pi\in\Gamma(s(\mathcal{I}),\mathcal{I})} f(\boldsymbol{\pi},\boldsymbol{r})}{1+\mathcal{L}(\mathcal{I})}$.*

The performance sensitivity $\mathcal{L}(\mathcal{I})$ quantifies how much the performance is affected by a relaxation of the constraint for a given set $\mathcal{I}$. Proposition C.1 in Section C.2 establishes that $\mathcal{L}(\mathcal{I})$ is finite for every candidate $\mathcal{I} \in \mathfrak{B}$. The quantity $\rho(\mathcal{I})$ measures the gap between the optimal value $v^\star$ and the maximum performance attainable over policies in $\Gamma(s(\mathcal{I}), \mathcal{I})$. The denominator of $\rho(\mathcal{I})$ acts as a scaling factor that reflects the geometry of the candidate set $\mathcal{I}$ via the problem's sensitivity to perturbations and plays a technical role in the subsequent analysis. Combining feasibility and performance considerations, we define the overall sub-optimality gap as follows.

Definition 4.4 (**Sub-optimality Gap**)**.** *For a candidate set $\mathcal{I} \in \mathfrak{B}$, the* sub-optimality gap *is defined as $\Delta(\mathcal{I}) := \max\big(s(\mathcal{I}),\, \rho(\mathcal{I})\big)$, with the corresponding worst-case sub-optimality given by $\Delta^\star := \min_{\mathcal{I}\in\mathfrak{B}:\,\mathcal{I}\neq\mathcal{I}^\star} \Delta(\mathcal{I})$.*

Since the family $\mathfrak{B}$ is finite, $\Delta^\star$ is well defined. This characterization enables us to distinguish the optimal basis from all suboptimal ones. Lemma 4.1 formalizes this, showing that $\Delta(\mathcal{I})$ plays a role analogous to the gap in classical multi-armed bandit problems; the proof is provided in Section C.2.

Lemma 4.1 (Suboptimality Characterization). *Under Assumption 3.2, $\Delta(\mathcal{I}^\star) = 0$, and $\Delta^\star > 0$.*

Example 1. *We provide a simple constrained toy example to illustrate these notions. We consider a single context (suppressing the dependence on c) and two arms with rewards and costs: $r_1 = a_1$, $\psi_1 = b$ and $r_2 = a_2$, $\psi_2 = -b$, where $a_1 > a_2$ and $b \in (0, 1]$. The offline problem reads*

$$\max_{\pi \in \mathbb{R}^2} \quad a_1\pi_1 + a_2\pi_2 \text{ s.t.} \qquad (1):\ b\pi_1 - b\pi_2 \le 0, \quad (2):\ \pi_1 + \pi_2 = 1, (3):\ \pi_1 \ge 0, \qquad (4):\ \pi_2 \ge 0.$$

Here $|\mathcal{C}|\,(|\mathcal{A}| - 1) = 1$, so candidate sets are singletons: $\mathfrak{B} = \{\mathcal{I}_1, \mathcal{I}_2, \mathcal{I}_3\}$ with $\mathcal{I}_1 = \{(1)\}$, $\mathcal{I}_2 = \{(3)\}$, and $\mathcal{I}_3 = \{(4)\}$, and enforcing the simplex equality (2) together with any single constraint among (1), (3), (4) at equality pins down a unique point, namely $\pi_{\mathcal{I}_1} = (1/2, 1/2)$, $\pi_{\mathcal{I}_2} = (0, 1)$, and $\pi_{\mathcal{I}_3} = (1, 0)$. The optimal policy is $\pi^\star = (1/2, 1/2)$, with $\mathcal{I}^\star = \mathcal{I}_1$ and $v^\star = (a_1 + a_2)/2$.

- *For $\mathcal{I}_2$, we have $\Gamma(s, \mathcal{I}_2) = \{\pi : b\pi_1 - b\pi_2 \le s,\ \pi_1 = 0,\ \pi_1 + \pi_2 = 1\}$. Clearly, $s(\mathcal{I}_2) = 0$ and the only feasible policy is $\pi = (0, 1)$, yielding $\mathcal{L}(\mathcal{I}_2) = 0$. Hence $\rho(\mathcal{I}_2) = \frac{(a_1+a_2)/2 - a_2}{1+0} = (a_1 - a_2)/2$ and $\Delta(\mathcal{I}_2) = \max\big(0,\ (a_1 - a_2)/2\big) = (a_1 - a_2)/2$, reflecting performance suboptimality.*
- *For $\mathcal{I}_3$, we have $\Gamma(s, \mathcal{I}_3) = \{\pi : b\pi_1 - b\pi_2 \le s,\ \pi_2 = 0,\ \pi_1 + \pi_2 = 1\}$. Clearly, $s(\mathcal{I}_3) = b$ and the only feasible policy is $\pi = (1, 0)$, yielding $\mathcal{L}(\mathcal{I}_3) = 0$. Hence $\rho(\mathcal{I}_3) = \frac{(a_1+a_2)/2 - a_1}{1+0} = (a_2 - a_1)/2 < 0$ and $\Delta(\mathcal{I}_3) = \max\big(b,\ (a_2 - a_1)/2\big) = b$, reflecting infeasibility.*

Therefore, $\Delta^\star = \min\big(\Delta(\mathcal{I}_2),\ \Delta(\mathcal{I}_3)\big) = \min\big((a_1 - a_2)/2,\ b\big)$ captures both performance and infeasibility effects. As either the performance gap $a_1 - a_2$ or the infeasibility gap b shrinks, $\Delta^\star \to 0$, making it inherently difficult to distinguish optimal from suboptimal active sets. In this regime, logarithmic regret can no longer be expected, and the problem transitions from a problem-dependent (logarithmic) regime to a worst-case regime.

5 Algorithm and Theoretical Results

In contrast to prior work (Slivkins et al., 2024; Guo & Liu, 2024; Guo et al., 2025), which relies on primal–dual methods, our approach is purely primal, solving at each round an approximation of the oracle (2).

OPALP: OPtimistic Approximation of Linear Programming

Inputs: Action set $\mathcal{A}$, context set $\mathcal{C}$, distribution $\mathcal{P}_\mathcal{C}$.
for $t = 1$ **to** T **do**
  Observe context c_t.
  Estimation: Construct estimators $\{\hat{r}_t, \check{\psi}_t\}$ from the learning oracles $\{\Omega_r, \Omega_\psi\}$.
  Policy Optimization: Compute the policy π_t as follows:

$$\pi_t = \text{argmax } LP(\hat{r}_t, \check{\psi}_t). \tag{4}$$

  Sampling: Play an arm $a_t \sim \boldsymbol{\pi}_t(c_t)$.
  Feedback Observation: Observe reward $r_t(c_t, a_t)$ and cost $\psi_t(c_t, a_t)$.
  History Update: $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1} \cup \{c_t, a_t, r_t(c_t, a_t), \psi_t(c_t, a_t)\}$.

Upon observing a context c_t and leveraging the estimators $\{\hat{r}_t, \check{\psi}_t\}$ provided by the estimation procedure (see Assumption 3.3), the **OPALP** algorithm constructs and solves a linear program (4) that approximates the offline oracle problem (2). This approximation is *optimistic* in nature, employing an optimistic reward estimator together with a cost estimator that makes the constraint easier to satisfy. The resulting policy is then used to select the action a_t. Finally, the history is updated

with the new observations, which are subsequently leveraged by the estimation procedure in future rounds. Because the algorithm relies on estimated quantities, it is essential to verify that the linear program (4), which forms its algorithmic core, remains feasible. This property is established in the following lemma, proved in Section C.1.

Lemma 5.1. *Under Assumptions 3.1, 3.2, and 3.3, with probability at least $1-\delta$, the optimistic linear program in* (4) *is feasible for all $t \in [T]$.*

With the above in hand, we now evaluate the performance of Algorithm **OPALP** in terms of regret and constraint violation. The result is stated in the following theorem, whose proof is deferred to Section C.3.

Theorem 5.1. *Under Assumptions 3.1, 3.2, and 3.3, the regret and cumulative constraint violation of **OPALP** satisfy*

$$\mathcal{R}_T = \mathcal{O}\left(\min\left\{\frac{\log(T)}{\Delta^\star},\ \sqrt{T\log(T)}\right\}\right), \quad \mathcal{V}_T = \tilde{\mathcal{O}}\left(\sqrt{T}\right).$$

*In particular, **OPALP** attains logarithmic regret $\mathcal{O}\big(\log(T)/\Delta^\star\big)$ whenever the gap $\Delta^\star$ is bounded away from zero, while never exceeding the previous work rate $\tilde{\mathcal{O}}(\sqrt{T})$, irrespective of $\Delta^\star$.*

As shown in Theorem 5.1, the cumulative constraint violation of our algorithm scales as $\widetilde{\mathcal{O}}(\sqrt{T})$, matching the best-known rates in prior work (Guo et al., 2025). The primary contribution lies instead in the regret analysis: to the best of our knowledge, this is the first work to establish problem-dependent regret bounds for stochastic contextual bandits with general constraints. The primal approach adopted by **OPALP** enables the introduction of a tailored gap parameter $\Delta^\star$, which characterizes the optimal active set $\mathcal{I}^\star$. In practice, **OPALP** solves an approximation of the oracle problem based on estimated rewards and costs; consequently, estimation errors may give rise to two issues: (i) the algorithm may activate an incorrect set, that is, $\mathcal{I}_t \neq \mathcal{I}^\star$; and (ii) even when the correct set is activated, that is, $\mathcal{I}_t = \mathcal{I}^\star$, the resulting policy is computed from the estimated problem and may therefore still incur regret. Notably, **OPALP** addresses both issues effectively: we show that the event $I_t \neq I^\star$ occurs only when the global estimation error (defined in Definition C.1) exceeds $\Delta^\star$ (see Proposition C.4), and that the cumulative regret contribution of such rounds is at most $\mathcal{O}\big(\log(T)/\Delta^\star\big)$; whenever $\mathcal{I}_t = \mathcal{I}^\star$, we establish that the per-round regret vanishes (see Proposition C.5).

A natural question is why the active-set characterization based on $\Delta^\star$ does not also lead to improved rates for the constraint violation. The reason lies in the use of optimistic estimates in Assumption 3.3. Whenever **OPALP** saturates the optimal active set $\mathcal{I}^\star$, the resulting policy π_t may intentionally relax the constraint, estimated in a lower-confidence-bound fashion that is inherently easier to satisfy than the true constraint, in order to achieve no round-wise regret and thus stronger performance guarantees. Equivalently, π_t belongs to the class of policies that saturate $\mathcal{I}^\star$ while improving reward at the expense of constraint satisfaction. Consequently, under the estimation framework of Assumption 3.3, refined problem-dependent guarantees for the constraint violation cannot be expected. One might instead consider pessimistic cost estimates of the form $0 \leq \hat{\psi}_t(c,a) - \psi(c,a) \leq \varepsilon_t(c,a,\delta)$. Such conservative estimates would place greater emphasis on constraint satisfaction. However, this approach raises fundamental difficulties. First, without additional strong assumptions, there is no guarantee that the approximate problem $\mathrm{LP}(\hat{r}_t, \hat{\psi}_t)$ remains feasible; in particular, an analogue of Lemma 5.1 cannot be established in general. Second, even when feasibility holds, there is no guarantee that the optimal policy $\pi^\star$ belongs to the plausibly safe set: if $\pi^\star$ saturates the global constraint **GC**, pessimistic cost estimates may incorrectly deem it infeasible and exclude it from consideration.

Lazy-OPALP: as shown in (4), **OPALP** repeatedly solves a linear program, which is more computationally demanding than the gradient-based updates used by primal–dual methods (Guo et al., 2025). While the latter are more efficient computationally, they typically offer weaker theoretical guarantees, resulting in a trade-off between regret performance and runtime. The appropriate choice therefore depends on the priorities of the application. To reduce computational cost, a lazy update scheme can be used (as in Abbasi-Yadkori et al. (2011)): the linear program is solved only when estimation errors change significantly, and the current policy is reused otherwise. This limits the number of LP solves to logarithmic in T, while preserving the original guarantees up to constant factors. Implementation details of **Lazy-OPALP**, and its theoretical justification are provided in Section D.

6 OPALP AND CONTEXTUAL BANDITS WITH KNAPSACKS (CBWK)

For budget or resource constraints, it may be appealing from a practitioner's perspective to enforce the budget strictly, as in the CBwK paradigm—that is, without allowing any overconsumption. To align with the CBwK setting, we define $\psi_t(c, a) = w_t(c, a) - b$, where $w_t(c, a) \in [0, 1]$ denotes the random budget consumption of action a in context c at time t (so that $\psi_t \in [-1, 1]$, consistently with the setting of Section 3), and $b = B/T \in (0, 1]$, with B the total initial budget. Once the cumulative consumption exceeds B at some stopping time τ, the interaction is terminated. As a result, the total budget constraint is always satisfied and the constraint violation is identically zero, i.e., $\mathcal{V}_T^{\mathrm{CBwK}} \equiv 0$. The corresponding regret is defined as $\mathcal{R}_T^{\mathrm{CBwK}} := \mathbb{E}\left[\sum_{t=1}^{\tau}\left(v^\star - r(c_t, a_t)\right) + (T - \tau)\, v^\star\right]$, where $v^\star$ denotes the optimal value in the standard CBwK setting (see Agrawal & Devanur (2016)). This naturally raises the question: what regret guarantees can still be established for **OPALP** in this setting, where the interaction terminates once the budget is exhausted? The answer is provided by Corollary 6.1, whose proof is deferred to Section C.4.

Corollary 6.1. *Under Assumptions 3.1 to 3.3, within the CBwK framework, **OPALP** achieves*

$$\mathcal{R}_T^{\mathrm{CBwK}} = \widetilde{\mathcal{O}}\left(\left(1 + \frac{v^\star}{b}\right)\sqrt{T}\right).$$

The resulting regret bound mirrors the guarantee established in Agrawal & Devanur (2016) and is meaningful in the regime $B = \Omega(v^\star\sqrt{T})$. Note that the standard knapsack setting (Badanidiyuru et al., 2018) involves multiple resource consumption costs subject to multiple budget constraints; this poses no difficulty, since our approach readily generalizes to multiple constraints.

7 EXPERIMENTS

To corroborate our theoretical results, we conduct several numerical experiments. Owing to space constraints, the detailed experimental setup and results are deferred to Section F. First, to confirm the theoretical bounds and the claimed improvement over prior work, we benchmark **OPALP** against three primal–dual baselines, **Optimistic**^{3} (Guo et al., 2025), **LOE2D** (Guo & Liu, 2024), and **LagrangeCBwLC** (Slivkins et al., 2024), on synthetic data, and against **Optimistic**^{3} on a real-world dataset (Section F.1). Second, to better understand the sub-optimality gap $\Delta^\star$, we evaluate the regret on a sequence of instances with decreasing $\Delta^\star$ (Section F.2). Third, to illustrate that the known-context-distribution assumption is not necessary, we compare **OPALP** with a variant that estimates the context distribution from data (Section F.3). Fourth, to evaluate the computational burden of **OPALP** relative to **Optimistic**^{3}, we measure their average per-round runtimes on instances of increasing size, and verify that **Lazy-OPALP** matches the empirical performance of **OPALP** at a reduced computational cost (Section F.4). Finally, to validate our CBwK result, we benchmark **OPALP** against the method of Agrawal & Devanur (2016) (Section F.5).

CONCLUSION

In this paper, we study stochastic constrained contextual bandits with general reward and cost function classes, thereby capturing a broad range of practical problems. We propose the **OPALP** algorithm, which leverages reward and cost estimates supplied by external online regression oracles to solve a noisy approximation of the oracle LP and significantly improves the theoretical guarantees by achieving logarithmic, problem-dependent regret. This improvement stems from a refined analysis exploiting the structure of the problem, together with a geometric duality between the optimal solution of the associated linear program and its active set. We further corroborate our theoretical results through several numerical experiments.

In doing so, we highlight a trade-off between improved theoretical guarantees (our method) and lighter computational overhead (prior primal–dual approaches), the choice between which is left to the practitioner depending on their priorities and available computational resources.

AI USE STATEMENT

In this work, we used generative AI tools to edit the manuscript for readability and to check spelling, grammar, and the consistency of notation and citations. We have not used generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, provide ingredients for or write proofs, propose or refine hypotheses, design or provide feedback on the methodology or experiments, implement methods, or interpret results; the remaining tasks with required disclosure (generating synthetic data, translation, dataset cleaning, and qualitative data analysis) are not applicable to this work. We have reviewed all AI-assisted work: every suggested edit was checked by the authors against the original text and the mathematical content. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

This work is primarily theoretical. All assumptions required to derive our results are precisely stated and discussed in the main text (Assumptions 3.1 to 3.3), and are further examined in Sections A and E. Proof sketches are provided in the main text, and complete proofs are included in the appendices. The full protocol of the simulations (instance, parameters, number of replications) is specified in Section F.

REFERENCES

Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. Online least squares estimation with self-normalized processes: An application to bandit problems, 2011. (Cited on 1, 8)

Shipra Agrawal and Nikhil R. Devanur. Linear contextual bandits with knapsacks, 2016. URL https://arxiv.org/abs/1507.06738. (Cited on 9, 32)

Shipra Agrawal, Nikhil R. Devanur, and Lihong Li. An efficient algorithm for contextual bandits with knapsacks, and an extension to concave objectives. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir (eds.), *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 4–18, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL https://proceedings.mlr.press/v49/agrawal16.html. (Cited on 1)

Sanae Amani, Mahnoosh Alizadeh, and Christos Thrampoulidis. Linear stochastic bandits under safety constraints. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/09a8a8976abcdfdee15128b4cc02f33a-Paper.pdf. (Cited on 1, 2)

Ashwinkumar Badanidiyuru, John Langford, and Aleksandrs Slivkins. Resourceful contextual bandits. In Maria Florina Balcan, Vitaly Feldman, and Csaba Szepesvári (eds.), *Proceedings of The 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pp. 1109–1134, Barcelona, Spain, 13–15 Jun 2014. PMLR. URL https://proceedings.mlr.press/v35/badanidiyuru14.html. (Cited on 1)

Ashwinkumar Badanidiyuru, Robert Kleinberg, and Aleksandrs Slivkins. Bandits with knapsacks. *J. ACM*, 65(3), March 2018. ISSN 0004-5411. doi: 10.1145/3164539. URL https://doi.org/10.1145/3164539. (Cited on 1, 4, 9)

Santiago R. Balseiro, Omar Besbes, and Dana Pizarro. Survey of dynamic resource-constrained reward collection problems: Unified model and analysis. *Operations Research*, 72(5):2168–2189, September 2024. doi: 10.1287/opre.2023.2441. URL https://ideas.repec.org/a/inm/oropre/v72y2024i5p2168-2189.html. (Cited on 2)

Dorian Baudry, Nadav Merlis, Mathieu Benjamin Molina, Hugo Richard, and Vianney Perchet. Multi-armed bandits with guaranteed revenue per arm. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 379–387. PMLR,

02–04 May 2024. URL https://proceedings.mlr.press/v238/baudry24a.html. (Cited on 2)

Ahmed Ben Yahmed, Hafedh El Ferchichi, Marc Abeille, and Vianney Perchet. Contextual multi-armed bandits with minimum aggregated revenue constraints. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qbhPULMkuQ. (Cited on 1, 2)

Dimitris Bertsimas and John N. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, Belmont, MA, 1997. (Cited on 18)

Guanting Chen, Xiaocheng Li, and Yinyu Ye. An improved analysis of lp-based control for revenue management, 2022a. URL https://arxiv.org/abs/2101.11092. (Cited on 2)

Tianrui Chen, Aditya Gangrade, and Venkatesh Saligrama. Strategies for safe multi-armed bandits with logarithmic regret and risk. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3123–3148. PMLR, 17–23 Jul 2022b. URL https://proceedings.mlr.press/v162/chen22e.html. (Cited on 1, 2)

Zhaohua Chen, Rui Ai, Mingwei Yang, Yuqi Pan, Chang Wang, and Xiaotie Deng. Contextual decision-making with knapsacks beyond the worst case. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 88147–88193. Curran Associates, Inc., 2024. doi: 10.52202/079017-2798. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/a0e1c2c40fc245b5fe7251ea33fbb045-Paper-Conference.pdf. (Cited on 2, 27)

Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In Geoffrey Gordon, David Dunson, and Miroslav Dudík (eds.), *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 208–214, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/chu11a.html. (Cited on 1)

Evgenii E Chzhen, Christophe Giraud, Zhen LI, and Gilles Stoltz. Small total-cost constraints in contextual bandits with knapsacks, with application to fairness. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=uZvG0HLkOB. (Cited on 1, 3, 5, 16)

Lorenzo Croissant, Marc Abeille, and Bruno Bouchard. Near-continuous time reinforcement learning for continuous state-action spaces, 2023. URL https://arxiv.org/abs/2309.02815. (Cited on 16)

Rohan Deb and Aadirupa Saha. Think before you duel: Understanding complexities of preference learning under constrained resources, 2023. URL https://arxiv.org/abs/2312.17229. (Cited on 1)

Rohan Deb, Mohammad Ghavamzadeh, and Arindam Banerjee. Conservative contextual bandits: Beyond linear representations. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=SThJXvucjQ. (Cited on 2)

Rohan Deb, Mohammad Ghavamzadeh, and Arindam Banerjee. Thompson sampling for constrained bandits. In *Reinforcement Learning Conference*, 2025b. URL https://openreview.net/forum?id=5HL4QaNQcf. (Cited on 2)

Arthur Flajolet and Patrick Jaillet. Logarithmic regret bounds for bandits with knapsacks, 2017. URL https://arxiv.org/abs/1510.01800. (Cited on 2)

Dylan Foster and Alexander Rakhlin. Beyond UCB: Optimal and efficient contextual bandits with regression oracles. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3199–3210. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/foster20a.html. (Cited on 5, 16)

Dylan J. Foster and Akshay Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination, 2021. URL https://arxiv.org/abs/2107.02237. (Cited on 5)

Dylan J. Foster and Alexander Rakhlin. Foundations of reinforcement learning and interactive decision making, 2023. URL https://arxiv.org/abs/2312.16730. (Cited on 16)

{Dylan J.} Foster, Alekh Agarwal, Miroslav Dudik, Luo Haipeng, and {Robert E.} Schapire. Practical contextual bandits with regression oracles. In Jennifer Dy and Andreas Krause (eds.), *35th International Conference on Machine Learning, ICML 2018*, 35th International Conference on Machine Learning, ICML 2018, pp. 2482–2517. International Machine Learning Society (IMLS), 2018. Publisher Copyright: © 2018 35th International Conference on Machine Learning, ICML 2018. All rights reserved.; 35th International Conference on Machine Learning, ICML 2018 ; Conference date: 10-07-2018 Through 15-07-2018. (Cited on 5, 16)

Aditya Gangrade, Tianrui Chen, and Venkatesh Saligrama. Safe linear bandits over unknown polytopes. In Shipra Agrawal and Aaron Roth (eds.), *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 1755–1795. PMLR, 30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/gangrade24a.html. (Cited on 2)

Hengquan Guo and Xin Liu. Stochastic constrained contextual bandits via lyapunov optimization based estimation to decision framework. In Shipra Agrawal and Aaron Roth (eds.), *Proceedings of Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 2204–2231. PMLR, 30 Jun–03 Jul 2024. URL https://proceedings.mlr.press/v247/guo24a.html. (Cited on 2, 3, 5, 7, 9, 29, 30)

Hengquan Guo and Xin Liu. On stochastic contextual bandits with knapsacks in small budget regime. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FCMpUOZkxi. (Cited on 1, 5)

Hengquan Guo, Lingkai Zu, and Xin Liu. Triple-optimistic learning for stochastic contextual bandits with general constraints. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=NhJ4cCifqF. (Cited on 2, 3, 4, 5, 7, 8, 9, 27, 29, 30, 31)

Yuxuan Han, Jialin Zeng, Yang Wang, Yang Xiang, and Jiheng Zhang. Optimal contextual bandits with knapsacks under realizability via regression oracles. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 5011–5035. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/han23b.html. (Cited on 1, 3)

Abbas Kazerouni, Mohammad Ghavamzadeh, Yasin Abbasi-Yadkori, and Benjamin Van Roy. Conservative contextual linear bandits. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 3913–3922, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964. (Cited on 2)

Raunak Kumar and Robert D. Kleinberg. Non-monotonic resource utilization in the bandits with knapsacks problem. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088. (Cited on 1, 2)

T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020. ISBN 9781108486828. URL https://books.google.fr/books?id=bydXzAEACAAJ. (Cited on 1)

Fengjiao Li, Jia Liu, and Bo Ji. Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering*, 7(3):1799–1813, 2020. doi: 10.1109/TNSE.2019.2954310. (Cited on 1)

Lihong Li, Yu Lu, and Dengyong Zhou. Provably optimal algorithms for generalized linear contextual bandits. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2071–2080. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/li17c.html. (Cited on 1)

Xin Liu. On preference-based stochastic linear contextual bandits with knapsacks. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 2890–2898. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/liu25g.html. (Cited on 1)

Aldo Pacchiano, Mohammad Ghavamzadeh, Peter Bartlett, and Heinrich Jiang. Stochastic bandits with linear constraints. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 2827–2835. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/pacchiano21a.html. (Cited on 1, 2)

Aldo Pacchiano, Mohammad Ghavamzadeh, and Peter Bartlett. Contextual bandits with stage-wise constraints. *Journal of Machine Learning Research*, 26(170):1–57, 2025. URL http://jmlr.org/papers/v26/24-0267.html. (Cited on 2)

Tao Qin and Tie-Yan Liu. Introducing letor 4.0 datasets, 2013. URL https://arxiv.org/abs/1306.2597. (Cited on 29)

Daniel Russo and Benjamin Van Roy. Eluder dimension and the sample complexity of optimistic exploration. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/41bfd20a38bb1b0bec75acf0845530a7-Paper.pdf. (Cited on 5)

Karthik Abinav Sankararaman and Aleksandrs Slivkins. Bandits with knapsacks beyond the worst case. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 23191–23204. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/c3395dd46c34fa7fd8d729d8cf88b7a8-Paper.pdf. (Cited on 2)

Alexander Schrijver. *Theory of Linear and Integer Programming*. John Wiley & Sons, Chichester, 1986. (Cited on 18)

Vidyashankar Sivakumar, Shiliang Zuo, and Arindam Banerjee. Smoothed adversarial linear contextual bandits with knapsacks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20253–20277. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/sivakumar22a.html. (Cited on 1)

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. URL http://jmlr.org/papers/v25/24-1220.html. (Cited on 2, 3, 5, 7, 9, 29, 30)

Long Tran-Thanh, Archie Chapman, Alex Rogers, and Nicholas R. Jennings. Knapsack based optimal policies for budget-limited multi-armed bandits, 2012. URL https://arxiv.org/abs/1204.1909. (Cited on 1)

Lequn Wang, Yiwei Bai, Wen Sun, and Thorsten Joachims. Fairness of exposure in stochastic bandits. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10686–10696. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/wang21b.html. (Cited on 1)

Larry Wasserman. Density estimation. Lecture notes for 36-708: Statistical Methods for Machine Learning, Carnegie Mellon University, 2019. URL https://www.stat.cmu.edu/~larry/=sml/densityestimation.pdf. (Cited on 27)

Huasen Wu, R. Srikant, Xin Liu, and Chong Jiang. Algorithms with logarithmic or sublinear regret for constrained contextual bandits, 2015. URL https://arxiv.org/abs/1504.06937. (Cited on 2)

APPENDIX CONTENTS

A A Closer Look at The Assumptions **16**

B Oracle Behavior **17**

B.1 Proof of Lemma 3.1 17

C Theoretical Results **17**

C.1 Feasibility of the Optimistic LP 17

C.2 Sub-Optimality Characterisation 18

C.3 OPALP Theoretical Guarantees 19

C.4 OPALP and Contextual Bandits with Knapsack (CBwK) 24

D Computational Complexity **25**

E Relaxed Assumptions **27**

E.1 From Known to Unknown Context Distribution 27

E.2 Extension to Multiple Global Constraints 28

F Details of Experiments **29**

F.1 Comparison with the State of the Art 29

F.2 The Effect of $\Delta^\star$ on the Regret 30

F.3 Unknown Context Distribution 31

F.4 Computational Complexity 31

F.5 OPALP and Contextual Bandits with Knapsacks (CBwK) 32

A A Closer Look at The Assumptions

Assumption 3.1: This assumption is standard and states that, although rewards and costs are modeled using general function classes, the observed realizations are noisy versions of underlying mean functions. Specifically, both reward and cost observations are assumed to be generated by adding centered (zero-mean) noise to nominal functions, which captures stochastic variability while preserving a well-defined expected structure.

Assumption 3.2. This assumption guarantees that the offline oracle problem used as a benchmark is feasible. If this condition is violated, infeasibility can in principle be detected through the infeasibility of the optimization problem solved by **OPALP**. In such situations, there is no canonical way to proceed; depending on the application, a natural approach is to relax the constraints by introducing a positive slack variable and to redefine the problem accordingly. The non-degeneracy condition is standard in parametric linear programming: it requires the optimal vertex to be uniquely determined by linearly independent active constraints. It holds for generic problem data—degenerate instances form a Lebesgue-null set of parameters (r, ψ, p)—and is what renders the optimal basis $\mathcal{I}^\star$ identifiable; at degenerate optima, ties between bases make the active set non-identifiable and the gap $\Delta^\star$ vanishes. In general, non-degeneracy can be enforced by a preprocessing step that slightly perturbs the problem data.

Assumption 3.3: Under the noise model described in Assumption 3.1 , it is well established in the literature Foster & Rakhlin (2023) that online regression oracles can be used to construct accurate estimators of the underlying reward and cost function. Such oracles leverage historical observations to produce estimators that control the cumulative estimation error under an appropriate loss function. Moreover, online regression oracles are often computationally efficient and enjoy strong theoretical guarantees Foster & Rakhlin (2020). Broadly speaking, when dealing with general function classes $\mathcal{F}$ under an additive noise model (realizability assumption), the standard construction (Foster et al., 2018)—further discussed in Chzhen et al. (2023)—proceeds as follows. One builds a sequence of confidence sets $\mathcal{F}_t \subset \mathcal{F}$ such that

$$\mathbb{P}\big(\forall t,\ f^\star \in \mathcal{F}_t\big) \geq 1 - \delta,$$

where $f^\star$ denotes the true unknown function. Then, for any estimate $\hat{f}_t \in \mathcal{F}_t$,

$$\forall\, t, c, a: \quad \big|f^\star(c,a) - \hat{f}_t(c,a)\big| \leq \omega_t(c,a,\delta) := \sup_{f,f' \in \mathcal{F}_t} \big(f(c,a) - f'(c,a)\big).$$

By Lemma 21 of Croissant et al. (2023), this yields $\sum_{t=1}^T \omega_t^2(c_t, a_t, \delta) \lesssim \log\Big(\frac{|\mathcal{F}|}{\delta}\Big)$. We apply this construction separately to the reward and cost classes:

- For the reward class, choose, for all (c,a), $\hat{r}_t(c,a) \in \operatorname{argmax}_{r' \in \mathcal{R}_t} r'(c,a)$, and define $\omega_t^r(c,a,\delta)$ accordingly.
- For the cost class, choose, for all (c,a), $\check{\psi}_t(c,a) \in \operatorname{argmin}_{\psi' \in \Psi_t} \psi'(c,a)$, and define $\omega_t^\psi(c,a,\delta)$ accordingly.

Letting $\varepsilon_t(c,a,\delta) := \max\big(\omega_t^r(c,a,\delta),\, \omega_t^\psi(c,a,\delta)\big)$, we obtain

$$\sum_{t=1}^T \varepsilon_t^2(c_t,a_t,\delta) \leq \sum_{t=1}^T \big[\big(\omega_t^r(c_t,a_t,\delta)\big)^2 + \big(\omega_t^\psi(c_t,a_t,\delta)\big)^2\big] \lesssim \log\left(\frac{\max\{|\mathcal{R}|, |\Psi|\}}{\delta}\right).$$

This confirms that Assumption 3.3 is standard in the literature: we have provided references for both how such estimates are computed in practice and the theoretical foundations underpinning the required per-round, one-sided, and uniform guarantees under realizability.

B Oracle Behavior

B.1 Proof of Lemma 3.1

Lemma 3.1. *Under Assumptions 3.1 and 3.2, let π denote the optimal policy of problem* (1)*. Then,*

$$\mathbb{E}\left[\sum_{t=1}^{T}\langle \boldsymbol{\pi}(c_t), \boldsymbol{r}(c_t)\rangle\right] \leq T\, v^\star.$$

Proof of Lemma 3.1. Let $\pi = (\pi_t)_{t\in[T]}$ be the optimal policy of the planning problem (1); it is history-adapted, i.e., each $\pi_t : \mathcal{C} \to \Delta_{\mathcal{A}}$ is $\mathcal{H}_{t-1}$-measurable. Fix $t \in [T]$. Since $c_t \sim \mathcal{P}_{\mathcal{C}}$ is independent of $\mathcal{H}_{t-1}$, the tower rule and Assumption 3.1 give

$$\mathbb{E}\big[r_t(c_t, a_t)\big] = \mathbb{E}\left[\sum_{a\in\mathcal{A}} \pi_t(c_t, a)\, r(c_t, a)\right] = \sum_{c\in\mathcal{C}} p_c \sum_{a\in\mathcal{A}} \mathbb{E}\big[\pi_t(c, a)\big]\, r(c, a),$$

where the last step uses the independence of c_t and π_t. Define the deterministic policy $\bar{\pi}(c, a) := \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\big[\pi_t(c, a)\big]$; each $\bar{\pi}(c)$ is a probability vector, as a convex combination of probability vectors. Summing the display over t, and arguing identically for the cost,

$$\mathbb{E}\left[\sum_{t=1}^{T} r_t(c_t, a_t)\right] = T\sum_{c\in\mathcal{C}} p_c\, \langle \bar{\boldsymbol{\pi}}(c), \boldsymbol{r}(c)\rangle\,, \qquad \mathbb{E}\left[\sum_{t=1}^{T} \psi_t(c_t, a_t)\right] = T\sum_{c\in\mathcal{C}} p_c\, \langle \bar{\boldsymbol{\pi}}(c), \boldsymbol{\psi}(c)\rangle \leq 0,$$

where the inequality holds since π is feasible for (1) (almost-sure feasibility implies feasibility in expectation). Thus $\bar{\pi}$ is feasible for the offline problem (2), whence $\sum_{c\in\mathcal{C}} p_c \langle \bar{\boldsymbol{\pi}}(c), \boldsymbol{r}(c)\rangle \leq v^\star$, which is the claim.

□

C Theoretical Results

C.1 Feasibility of the Optimistic LP

Lemma 5.1. *Under Assumptions 3.1, 3.2, and 3.3, with probability at least $1-\delta$, the optimistic linear program in* (4) *is feasible for all $t \in [T]$.*

Proof of Lemma 5.1. We recall that for all $t \in [T]$, the constraint of the approximated $LP(\hat{r}_t, \check{\psi}_t)$ is:

$$\begin{aligned} &\sum_{c\in\mathcal{C}} p_c \left\langle \boldsymbol{\pi}(c), \check{\boldsymbol{\psi}}_t(c)\right\rangle \leq 0 \\ \implies &\sum_{c\in\mathcal{C}}\sum_{a\in\mathcal{A}} p_c \pi(c, a)\check{\psi}_t(c, a) \leq 0. \end{aligned}$$

On the other hand, given Assumption 3.2 and the definition of the oracle problem (2), we have:

$$\sum_{c\in\mathcal{C}}\sum_{a\in\mathcal{A}} p_c \pi^\star(c, a)\psi(c, a) \leq 0.$$

In addition, Assumption 3.3 implies with probability at least $1-\delta$:

$$\begin{aligned} \forall t \in [T],\ \forall c \in \mathcal{C}, \forall a \in \mathcal{A}:\ &\check{\psi}_t(c, a) \leq \psi(c, a) \\ \implies &\sum_{c\in\mathcal{C}}\sum_{a\in\mathcal{A}} p_c \pi^\star(c, a)\check{\psi}_t(c, a) \leq \sum_{c\in\mathcal{C}}\sum_{a\in\mathcal{A}} p_c \pi^\star(c, a)\psi(c, a) \\ \implies &\sum_{c\in\mathcal{C}}\sum_{a\in\mathcal{A}} p_c \pi^\star(c, a)\check{\psi}_t(c, a) \leq 0. \end{aligned}$$

This shows that $\pi^\star$ is a feasible solution to $LP(\hat{r}_t, \check{\psi}_t)$, thereby establishing the feasibility of the latter and concluding the proof. □

C.2 Sub-Optimality Characterisation

Proposition C.1 ($\mathcal{L}(\mathcal{I})$ **Property**). *For any candidate set $\mathcal{I} \in \mathfrak{B}$, the coefficient $\mathcal{L}(\mathcal{I})$ given in Definition 4.3 is nonnegative and finite, and the minimum in its definition is attained. More precisely, the value function*

$$z_{\mathcal{I}}(s) \;:=\; \max_{\pi \in \Gamma(s,\mathcal{I})} \sum_{c \in \mathcal{C}} p_c \, \langle \boldsymbol{\pi}(c), \boldsymbol{r}(c) \rangle \,, \qquad s \geq s(\mathcal{I}),$$

is concave and piecewise linear with finitely many pieces on $[s(\mathcal{I}), \infty)$, and $\mathcal{L}(\mathcal{I})$ equals its right derivative at $s(\mathcal{I})$.

Proof of Proposition C.1. Fix $\mathcal{I} \in \mathfrak{B}$ and write $\tilde{r}, \tilde{\psi} \in \mathbb{R}^{|\mathcal{A}||\mathcal{C}|}$ for the vectors with entries $\tilde{r}_{(c,a)} := p_c \, r(c,a)$ and $\tilde{\psi}_{(c,a)} := p_c \, \psi(c,a)$, so that $z_{\mathcal{I}}(s) = \max_{\pi \in \Gamma(s,\mathcal{I})} \tilde{r}^\top \pi$.

Step 0 (well-posedness). Let $\Pi_{\mathcal{I}} := \{\pi \in \Pi : \pi(c,a) = 0 \; \forall (c,a) \in \mathcal{J} \cap \mathcal{I}\}$. First, $\Pi_{\mathcal{I}} \neq \emptyset$: if some context c had all its actions in $\mathcal{J} \cap \mathcal{I}$, its $|\mathcal{A}|$ zero constraints together with its simplex equality would be $|\mathcal{A}| + 1$ vectors supported on the $|\mathcal{A}|$-dimensional coordinate block of context c – indeed, the simplex row is the sum of the zero rows – hence linearly dependent, contradicting that $\mathcal{I}$ forms a basic index set (Definition 4.1). Therefore every context retains at least one free action, and placing unit mass on one free action per context exhibits a point of $\Pi_{\mathcal{I}}$. Second, define the continuous function $h : \Pi_{\mathcal{I}} \to \mathbb{R}_{\geq 0}$ by $h(\pi) := \left| \tilde{\psi}^\top \pi \right|$ if $\mathbf{GC} \in \mathcal{I}$ and $h(\pi) := (\tilde{\psi}^\top \pi)_+$ otherwise, so that for every $s \geq 0$,

$$\Gamma(s, \mathcal{I}) \;=\; \{\pi \in \Pi_{\mathcal{I}} : h(\pi) \leq s\}.$$

Since $\Pi_{\mathcal{I}}$ is nonempty and compact and h is continuous, $s(\mathcal{I}) = \min_{\pi \in \Pi_{\mathcal{I}}} h(\pi)$ is attained; consequently, for every $s \geq s(\mathcal{I})$ the set $\Gamma(s, \mathcal{I})$ is a nonempty compact polytope, $z_{\mathcal{I}}(s)$ is attained, and $z_{\mathcal{I}}(s) \in [0,1]$ since $r \in [0,1]$ and each $\boldsymbol{\pi}(c)$ is a probability vector. Moreover $\Gamma(s_1, \mathcal{I}) \subseteq \Gamma(s_2, \mathcal{I})$ for $s(\mathcal{I}) \leq s_1 \leq s_2$, so $z_{\mathcal{I}}$ is nondecreasing.

Step 1 (dual representation with an s-independent feasible set). For each $s \geq s(\mathcal{I})$, the program defining $z_{\mathcal{I}}(s)$ is a linear program whose constraint matrix A (encoding the constraint $\tilde{\psi}^\top \pi \leq s$, the constraint $-\tilde{\psi}^\top \pi \leq s$ when $\mathbf{GC} \in \mathcal{I}$, the equalities $\pi(c,a) = 0$ on $\mathcal{J} \cap \mathcal{I}$, the simplex equalities, and $\pi \geq 0$) does not depend on s; only the right-hand side does, and it is affine in s: $b(s) = b_0 + s\, d$ with $d \in \{0,1\}^m$ supported on the one or two rows involving s. By Step 0 the primal is feasible with finite value, so LP strong duality (Bertsimas & Tsitsiklis, 1997, Theorem 4.4) gives

$$z_{\mathcal{I}}(s) \;=\; \min_{y \in D} \, b(s)^\top y, \qquad s \geq s(\mathcal{I}),$$

where the dual feasible set $D \subseteq \mathbb{R}^m$ is a nonempty polyhedron determined by A and $\tilde{r}$ alone, hence *independent of s*.

Step 2 (finite minimum of affine functions). By the Minkowski–Weyl decomposition (Schrijver, 1986, Corollary 7.1b), there exist finite sets $V, R \subseteq \mathbb{R}^m$ such that $D = \mathrm{conv}(V) + \mathrm{cone}(R)$; no pointedness of D is required (this matters here, as D contains free dual variables associated with the equality constraints). Fix $s \geq s(\mathcal{I})$. Since $\min_{y \in D} b(s)^\top y = z_{\mathcal{I}}(s)$ is finite, we must have $b(s)^\top \rho \geq 0$ for every $\rho \in R$ (a recession direction with $b(s)^\top \rho < 0$ would drive the infimum to $-\infty$). Consequently the minimum over D is achieved on $\mathrm{conv}(V)$, and by linearity on V itself:

$$z_{\mathcal{I}}(s) \;=\; \min_{v \in V} \, \big(b_0 + s\, d\big)^\top v, \qquad s \geq s(\mathcal{I}).$$

As a pointwise minimum of finitely many affine functions of s, $z_{\mathcal{I}}$ is concave and piecewise linear with at most $|V|$ pieces on the closed half-line $[s(\mathcal{I}), \infty)$; in particular it is continuous at $s(\mathcal{I})$ and admits there a finite right derivative

$$(z_{\mathcal{I}})'_+(s(\mathcal{I})) \;=\; \min \big\{\, d^\top v \;:\; v \in V \text{ active at } s(\mathcal{I}) \,\big\} \;<\; \infty.$$

Step 3 (identification with $\mathcal{L}(\mathcal{I})$). For $s(\mathcal{I}) \leq s_1 < s_2$, let $g(s_1, s_2) := \frac{z_{\mathcal{I}}(s_2) - z_{\mathcal{I}}(s_1)}{s_2 - s_1}$ denote the chord slope. By concavity, g is nonincreasing in each of its arguments, so

$$\sup_{s(\mathcal{I}) \leq s_1 < s_2} g(s_1, s_2) \;=\; \lim_{h \downarrow 0} \frac{z_{\mathcal{I}}(s(\mathcal{I}) + h) - z_{\mathcal{I}}(s(\mathcal{I}))}{h} \;=\; (z_{\mathcal{I}})'_+(s(\mathcal{I})),$$

which is finite by Step 2 and nonnegative since $z_{\mathcal{I}}$ is nondecreasing (Step 0). By Definition 4.3, a constant $L \geq 0$ is admissible if and only if $L \geq g(s_1, s_2)$ for all $s(\mathcal{I}) \leq s_1 < s_2$, i.e., if and only if $L \geq (z_{\mathcal{I}})'_+(s(\mathcal{I}))$. The admissible set is therefore the closed half-line $[(z_{\mathcal{I}})'_+(s(\mathcal{I})), \infty)$, whose minimum is attained and equals $(z_{\mathcal{I}})'_+(s(\mathcal{I}))$. This proves $\mathcal{L}(\mathcal{I}) = (z_{\mathcal{I}})'_+(s(\mathcal{I})) \in [0, \infty)$. □

Lemma 4.1 (Suboptimality Characterization). *Under Assumption 3.2,* $\Delta(\mathcal{I}^\star) = 0$, *and* $\Delta^\star > 0$.

Proof of Lemma 4.1. By Assumption 3.2, $\pi^\star$ is the unique optimal solution of (3) and a non-degenerate vertex: the constraints active at $\pi^\star$, jointly with the simplex equalities, are linearly independent and determine $\pi^\star$ uniquely, hence number exactly $|\mathcal{C}|\,|\mathcal{A}|$; the active inequality constraints therefore number exactly $|\mathcal{C}|\,(|\mathcal{A}| - 1)$ and form a candidate set. It is moreover the only candidate set contained in the active set, since any candidate set is a subset of the same cardinality and must coincide with it. This establishes existence and uniqueness of $\mathcal{I}^\star \in \mathfrak{B}$. Moreover, $\pi^\star = \pi_{\mathcal{I}^\star} \in \Gamma(0, \mathcal{I}^\star)$, so $s(\mathcal{I}^\star) = 0$ and $v^\star = \max_{\pi \in \Gamma(0,\mathcal{I}^\star)} f(\pi, r)$, giving $\Delta(\mathcal{I}^\star) = 0$.

Let $\mathcal{I} \in \mathfrak{B}$ with $\mathcal{I} \neq \mathcal{I}^\star$. If $\Gamma(0, \mathcal{I}) = \emptyset$, then $s(\mathcal{I}) > 0$ and thus $\Delta(\mathcal{I}) > 0$. Otherwise $s(\mathcal{I}) = 0$, and we claim $\Gamma(0, \mathcal{I}) = \{\pi_{\mathcal{I}}\}$: every $\pi \in \Gamma(0, \mathcal{I})$ satisfies the defining equalities of $\mathcal{I}$ (when $\mathbf{GC} \in \mathcal{I}$, the two inequalities of $\Gamma(0, \mathcal{I})$ pin the generic constraint to equality), and this system has the unique solution $\pi_{\mathcal{I}}$ by Definition 4.1. The policy $\pi_{\mathcal{I}}$ is feasible for (2). Furthermore, $\pi_{\mathcal{I}} \neq \pi^\star$: otherwise all constraints of $\mathcal{I}$ would be active at $\pi^\star$, i.e., $\mathcal{I}$ would be a candidate set contained in the active set of $\pi^\star$, which by the uniqueness argument above forces $\mathcal{I} = \mathcal{I}^\star$, a contradiction. Uniqueness of the optimal solution then yields $f(\pi_{\mathcal{I}}, \boldsymbol{r}) < v^\star$, hence $\rho(\mathcal{I}) > 0$ and $\Delta(\mathcal{I}) > 0$. Since $\mathfrak{B}$ is finite, $\Delta^\star > 0$ follows. □

C.3 OPALP Theoretical Guarantees

Before proceeding to the proof of Theorem 5.1, we require several preliminary results that will be used in the subsequent analysis.

Definition C.1. *For all* $t \in [T]$, *we define the averaged global estimation error:*

$$\xi_t(\pi_t, \delta) = \sum_{c \in \mathcal{C}} \sum_{a \in \mathcal{A}} p_c \pi_t(c, a) \varepsilon_t(c, a, \delta) = \mathbb{E}_{c \sim \mathcal{P}_{\mathcal{C}}, a \sim \pi_t} \left[\varepsilon_t(c, a, \delta) \right].$$

Throughout this section, we denote by $\mathcal{GE}$ the event of Assumption 3.3 under which the one-sided estimation guarantees hold uniformly over $t \in [T]$, $c \in \mathcal{C}$, and $a \in \mathcal{A}$; it satisfies $\mathbb{P}(\mathcal{GE}) \geq 1 - \delta$. Propositions C.2, C.3, C.4 and C.5 are all consequences of this single event, so no union bound is required when they are invoked jointly.

Proposition C.2. *With probability at least* $1 - \delta$, *the following holds:*

$$\forall t \in [T], \qquad \sum_{c \in \mathcal{C}} p_c \langle \boldsymbol{\pi}_t(c), \boldsymbol{\psi}(c) \rangle \;\leq\; \xi_t(\pi_t, \delta).$$

Proof of Proposition C.2. Given Assumption 3.3, with probability at least $1 - \delta$, we have:

$$\begin{aligned}
\forall t \in [T], \forall c \in \mathcal{C}, \forall a \in \mathcal{A}, \quad & \psi(c, a) \leq \check{\psi}_t(c, a) + \varepsilon_t(c, a, \delta) \\
\implies & \pi_t(c, a)\psi(c, a) \leq \pi_t(c, a)\check{\psi}_t(c, a) + \pi_t(c, a)\varepsilon_t(c, a, \delta) \\
\implies & \sum_{c,a} p_c \pi_t(c, a)\psi(c, a) \leq \sum_{c,a} p_c \pi_t(c, a)\check{\psi}_t(c, a) + \sum_{c,a} p_c \pi_t(c, a)\varepsilon_t(c, a, \delta) \\
\implies & \sum_{c \in \mathcal{C}} p_c \left\langle \boldsymbol{\pi}_t(c), \boldsymbol{\psi}(c) \right\rangle \leq \sum_{c \in \mathcal{C}} p_c \left\langle \boldsymbol{\pi}_t(c), \check{\boldsymbol{\psi}}_t(c) \right\rangle + \xi_t(\pi_t, \delta) \\
\implies & \sum_{c \in \mathcal{C}} p_c \left\langle \boldsymbol{\pi}_t(c), \boldsymbol{\psi}(c) \right\rangle \leq \xi_t(\pi_t, \delta).
\end{aligned}$$

The last inequality is due to $\sum_{c \in \mathcal{C}} p_c \left\langle \boldsymbol{\pi}_t(c), \check{\boldsymbol{\psi}}_t(c) \right\rangle \leq 0$. □

Proposition C.3. *With probability at least* $1 - \delta$, *the following holds:*

$$\forall t \in [T], \qquad \sum_{c \in \mathcal{C}} p_c \left\langle \boldsymbol{\pi}^\star(c) - \boldsymbol{\pi}_t(c), \boldsymbol{r}(c) \right\rangle \;\leq\; \xi_t(\pi_t, \delta).$$

Proof of Proposition C.3. Given Assumption 3.3, with probability at least $1-\delta$, we have:

$$\begin{aligned}
\forall t \in [T], \forall c \in \mathcal{C}, \forall a \in \mathcal{A}, \quad & \hat{r}_t(c,a) \leq r(c,a) + \varepsilon_t(c,a,\delta) \\
\implies & p_c \pi_t(c,a) \hat{r}_t(c,a) \leq p_c \pi_t(c,a) r(c,a) + p_c \pi_t(c,a) \varepsilon_t(c,a,\delta) \\
\implies & \sum_{c,a} p_c \pi_t(c,a) \hat{r}_t(c,a) \leq \sum_{c,a} p_c \pi_t(c,a) r(c,a) + \sum_{c,a} p_c \pi_t(c,a) \varepsilon_t(c,a,\delta) \\
\implies & \sum_{c \in \mathcal{C}} p_c \left\langle \boldsymbol{\pi}_t(c), \hat{\boldsymbol{r}}_t(c) \right\rangle \leq \sum_{c \in \mathcal{C}} p_c \left\langle \boldsymbol{\pi}_t(c), \boldsymbol{r}(c) \right\rangle + \xi_t(\pi_t, \delta).
\end{aligned}$$

In the proof of Lemma 5.1, we show that $\pi^\star$ is a feasible policy for the problem $LP(\hat{r}_t, \check{\psi}_t)$, hence we directly obtain:

$$\sum_{c \in \mathcal{C}} p_c \left\langle \boldsymbol{\pi}^\star(c), \hat{\boldsymbol{r}}_t(c) \right\rangle \leq \sum_{c \in \mathcal{C}} p_c \left\langle \boldsymbol{\pi}_t(c), \hat{\boldsymbol{r}}_t(c) \right\rangle . \tag{5}$$

And noticing by Assumption 3.3, that:

$$\sum_{c \in \mathcal{C}} p_c \left\langle \boldsymbol{\pi}^\star(c), \boldsymbol{r}(c) \right\rangle \leq \sum_{c \in \mathcal{C}} p_c \left\langle \boldsymbol{\pi}^\star(c), \hat{\boldsymbol{r}}_t(c) \right\rangle ,$$

completes the proof. □

Knowing $\mathcal{I}^\star$ effectively localizes and defines the optimal allocation $\pi^\star$. Hence, having a very tight estimate is strongly tied to activating $\mathcal{I}^\star$ and achieving the optimal policy. Conversely, activating a suboptimal set $\mathcal{I}_t$ is closely linked to the looseness of estimation.

Proposition C.4. *For any round t in which **OPALP** saturates a suboptimal active set, i.e., $\mathcal{I}_t \neq \mathcal{I}^\star$, the following holds with probability at least $1-\delta$:*

$$\xi_t(\pi_t, \delta) \ \geq \ \Delta^\star.$$

Proof of Proposition C.4. We rely for the proof on the impact of saturating the wrong set $\mathcal{I}_t \neq \mathcal{I}^\star$ on feasibility and performance.

Infeasibility. Recall the set $\Gamma(s, \mathcal{I})$ given in Definition 4.2. At round t, this set effectively links the chosen allocation π_t, the set of activated constraints $\mathcal{I}_t$, and $\xi_t(\pi_t, \delta)$.

Lemma C.1. *If π_t is the policy chosen by **OPALP** at time t, then with probability at least $1-\delta$:*

$$\pi_t \in \Gamma(\xi_t(\pi_t, \delta), \mathcal{I}_t).$$

Proof of Lemma C.1. By definition π_t is feasible to the problem $LP(\hat{r}_t, \check{\psi}_t)$ and activates $\mathcal{I}_t$, hence we have:

$$\sum_{c \in \mathcal{C}} p_c \left\langle \boldsymbol{\pi}_t(c), \check{\boldsymbol{\psi}}_t(c) \right\rangle \leq 0 \tag{6}$$

$$\text{if } \mathbf{GC} \in \mathcal{I}_t, \ \sum_{c \in \mathcal{C}} p_c \left\langle \boldsymbol{\pi}_t(c), \check{\boldsymbol{\psi}}_t(c) \right\rangle = 0, \tag{7}$$

$$\pi_t \in \Pi. \tag{8}$$

Given Assumption 3.3, with probability at least $1-\delta$, we have:

$$\begin{aligned}
\forall t \in [T], \forall c \in \mathcal{C}, \forall a \in \mathcal{A}, \quad & \check{\psi}_t(c,a) - \varepsilon_t(c,a,\delta) \leq \psi(c,a) \leq \check{\psi}_t(c,a) + \varepsilon_t(c,a,\delta) \\
\implies & \pi_t(c,a) \check{\psi}_t(c,a) - \pi_t(c,a) \varepsilon_t(c,a,\delta) \leq \pi_t(c,a) \psi(c,a) \leq \pi_t(c,a) \check{\psi}_t(c,a) + \pi_t(c,a) \varepsilon_t(c,a,\delta) \\
\implies & \sum_{c,a} p_c \pi_t(c,a) \check{\psi}_t(c,a) - \sum_{c,a} p_c \pi_t(c,a) \varepsilon_t(c,a,\delta) \leq \sum_{c,a} p_c \pi_t(c,a) \psi(c,a) \leq \sum_{c,a} p_c \pi_t(c,a) \check{\psi}_t(c,a) + \sum_{c,a} p_c \pi_t(c,a) \varepsilon_t(c \\
\implies & \sum_{c \in \mathcal{C}} p_c \left\langle \boldsymbol{\pi}_t(c), \check{\boldsymbol{\psi}}_t(c) \right\rangle - \xi_t(\pi_t, \delta) \leq \sum_{c \in \mathcal{C}} p_c \left\langle \boldsymbol{\pi}_t(c), \boldsymbol{\psi}(c) \right\rangle \leq \sum_{c \in \mathcal{C}} p_c \left\langle \boldsymbol{\pi}_t(c), \check{\boldsymbol{\psi}}_t(c) \right\rangle + \xi_t(\pi_t, \delta).
\end{aligned} \tag{9}$$

From (6) and (9), we conclude:

$$\sum_{c\in\mathcal{C}} p_c \left\langle \boldsymbol{\pi}_t(c), \boldsymbol{\psi}(c) \right\rangle \leq \xi_t(\pi_t, \delta).$$

Similarly, from (7) and (9):

$$\text{if } \mathbf{GC} \in \mathcal{I}_t, \quad -\xi_t(\pi_t, \delta) \leq \sum_{c\in\mathcal{C}} p_c \left\langle \boldsymbol{\pi}_t(c), \boldsymbol{\psi}(c) \right\rangle.$$

Finally, by the definition of $\mathcal{I}_t$, we also have:

$$\forall (c, a) \in \mathcal{J} \cap \mathcal{I}_t, \quad \pi_t(c, a) = 0,$$

which completes the proof. □

The inclusion $\pi_t \in \Gamma(\xi_t(\pi_t, \delta), \mathcal{I}_t)$ established in Lemma C.1 implies that the set $\Gamma(\xi_t(\pi_t, \delta), \mathcal{I}_t)$ is nonempty. Since $s(\mathcal{I}_t)$ is the minimal such slack (Definition 4.2), it must be the case that:

$$\xi_t(\pi_t, \delta) \geq s(\mathcal{I}_t), \tag{10}$$

where $s(\mathcal{I}_t)$ is given in Definition 4.2.

Performance Gap. Another tension arises due to performance. We consider the following problem:

$$z_s(r, \mathcal{I}) = \max_{\pi\in\Gamma(s,\mathcal{I})} \sum_{c\in\mathcal{C}} p_c \left\langle \boldsymbol{\pi}(c), \boldsymbol{r}(c) \right\rangle.$$

Given Definition 4.3, we have:

$$z_{\xi_t(\pi_t,\delta)}(r, \mathcal{I}_t) - z_{s(\mathcal{I}_t)}(r, \mathcal{I}_t) \leq \mathcal{L}(\mathcal{I}_t) \left(\xi_t(\pi_t, \delta) - s(\mathcal{I}_t)\right).$$

On the other hand, it holds that:

$$\sum_{c\in\mathcal{C}} p_c \left\langle \boldsymbol{\pi}_t(c), \boldsymbol{r}(c) \right\rangle \leq z_{\xi_t(\pi_t,\delta)}(r, \mathcal{I}_t),$$

and by Proposition C.3, we have:

$$\sum_{c\in\mathcal{C}} p_c \left\langle \boldsymbol{\pi}^\star(c), \boldsymbol{r}(c) \right\rangle - \xi_t(\pi_t, \delta) \leq \sum_{c\in\mathcal{C}} p_c \left\langle \boldsymbol{\pi}_t(c), \boldsymbol{r}(c) \right\rangle.$$

Hence, we obtain:

$$\underbrace{\sum_{c\in\mathcal{C}} p_c \left\langle \boldsymbol{\pi}^\star(c), \boldsymbol{r}(c) \right\rangle}_{v^\star} - \xi_t(\pi_t, \delta) - z_{s(\mathcal{I}_t)}(r, \mathcal{I}_t) \leq \mathcal{L}(\mathcal{I}_t) \left(\xi_t(\pi_t, \delta) - s(\mathcal{I}_t)\right).$$

This implies:

$$\begin{aligned} \frac{v^\star - z_{s(\mathcal{I}_t)}(r, \mathcal{I}_t) + \mathcal{L}(\mathcal{I}_t) s(\mathcal{I}_t)}{1 + \mathcal{L}(\mathcal{I}_t)} &\leq \xi_t(\pi_t, \delta) \\ \implies \frac{v^\star - z_{s(\mathcal{I}_t)}(r, \mathcal{I}_t)}{1 + \mathcal{L}(\mathcal{I}_t)} &\leq \xi_t(\pi_t, \delta) \\ \implies \rho(\mathcal{I}_t) &\leq \xi_t(\pi_t, \delta). \end{aligned} \tag{11}$$

Combining equations (10) and (11), we obtain that: $\xi_t(\pi_t, \delta) \geq \Delta(\mathcal{I}_t)$. Hence, if $\mathcal{I}_t$ is sub-optimal then $\xi_t(\pi_t, \delta) \geq \Delta^\star$, which completes the proof. □

Proposition C.5. *On the good event $\mathcal{GE}$ of Assumption 3.3 (which holds with probability at least $1-\delta$): for any round t in which **OPALP** saturates the optimal active set, i.e., $\mathcal{I}_t = \mathcal{I}^\star$, the following holds:*

$$\sum_{c\in\mathcal{C}} p_c \,\langle \boldsymbol{\pi}^\star(c) - \boldsymbol{\pi}_t(c),\, \boldsymbol{r}(c)\rangle \le 0.$$

Proof of Proposition C.5. Associate to the generic constraint $\sum_c p_c \,\langle \boldsymbol{\pi}(c), \boldsymbol{\psi}(c)\rangle \le 0$ of the fluid program (3) the Lagrangian, defined for $\lambda \ge 0$ by

$$\mathcal{L}_\lambda(\boldsymbol{\pi}) := \sum_{c\in\mathcal{C}} p_c \,\langle \boldsymbol{\pi}(c),\, \boldsymbol{r}(c) - \lambda\, \boldsymbol{\psi}(c)\rangle\,, \qquad \mathcal{A}^\star_\lambda(c) := \operatorname*{argmax}_a \big\{\, r(c,a) - \lambda\, \psi(c,a)\,\big\},$$

where $\mathcal{A}^\star_\lambda(c)$ collects the Lagrange-optimal actions in context c. By Assumption 3.2, the oracle LP is non-degenerate and admits a unique optimal basis $\mathcal{I}^\star$ together with an associated optimal dual multiplier $\lambda^\star$. Strong LP duality and complementary slackness then give

$$v^\star = \max_{\boldsymbol{\pi}} \mathcal{L}_{\lambda^\star}(\boldsymbol{\pi}) = \mathcal{L}_{\lambda^\star}(\boldsymbol{\pi}^\star), \qquad \lambda^\star \sum_{c\in\mathcal{C}} p_c \,\langle \boldsymbol{\pi}^\star(c), \boldsymbol{\psi}(c)\rangle = 0, \tag{12}$$

the maximum being taken over all $\boldsymbol{\pi}$ with each $\boldsymbol{\pi}(c)$ a distribution over actions. As $\mathcal{L}_{\lambda^\star}$ separates across contexts, its maximizers are exactly the policies whose per-context support is contained in $\mathcal{A}^\star_{\lambda^\star}(c)$.

Step 1 (the active set pins the support). By non-degeneracy (Assumption 3.2), the number of constraints active at $\boldsymbol{\pi}^\star$ (jointly with the simplex equalities) equals the dimension of the ambient space, so $\boldsymbol{\pi}^\star$ has no zero coordinate outside $\mathcal{J}\cap\mathcal{I}^\star$: if $\pi^\star(c,a) = 0$ for some $(c,a) \notin \mathcal{I}^\star$, the active set at $\boldsymbol{\pi}^\star$ would strictly contain the linearly independent system determined by $\mathcal{I}^\star$, a contradiction. Hence $\operatorname{supp}(\boldsymbol{\pi}^\star) = \mathcal{J}\setminus\mathcal{I}^\star$ exactly, and complementary slackness in the true LP yields $\mathcal{J}\setminus\mathcal{I}^\star \subseteq \{(c,a) : a \in \mathcal{A}^\star_{\lambda^\star}(c)\}$. Since $\mathcal{I}_t = \mathcal{I}^\star$ forces $\pi_t(c,a) = 0$ on $\mathcal{J}\cap\mathcal{I}^\star$, we conclude $\operatorname{supp}(\boldsymbol{\pi}_t) \subseteq \mathcal{J}\setminus\mathcal{I}^\star$, so $\boldsymbol{\pi}_t$ maximizes $\mathcal{L}_{\lambda^\star}$ context-wise, that is,

$$\mathcal{L}_{\lambda^\star}(\boldsymbol{\pi}_t) = \max_{\boldsymbol{\pi}} \mathcal{L}_{\lambda^\star}(\boldsymbol{\pi}) = v^\star. \tag{13}$$

Step 2 (optimism makes the true cost non-negative). Let $\check{\boldsymbol{\psi}}_t$ be the optimistic (lower-confidence) cost estimate with which **OPALP** defines its round-t program; on $\mathcal{GE}$ we have $\check{\boldsymbol{\psi}}_t \le \boldsymbol{\psi}$ componentwise. If $\lambda^\star = 0$ the second term in (15) below vanishes. If $\lambda^\star > 0$, the generic constraint is active and therefore belongs to $\mathcal{I}^\star = \mathcal{I}_t$, so $\boldsymbol{\pi}_t$ saturates it *in the optimistic program*, $\sum_c p_c\langle \boldsymbol{\pi}_t(c), \check{\boldsymbol{\psi}}_t(c)\rangle = 0$. Since $\check{\boldsymbol{\psi}}_t \le \boldsymbol{\psi}$ and $\boldsymbol{\pi}_t \ge 0$,

$$\sum_{c\in\mathcal{C}} p_c \,\langle \boldsymbol{\pi}_t(c), \boldsymbol{\psi}(c)\rangle \;\ge\; \sum_{c\in\mathcal{C}} p_c \,\big\langle \boldsymbol{\pi}_t(c), \check{\boldsymbol{\psi}}_t(c)\big\rangle = 0. \tag{14}$$

Step 3 (conclusion). By linearity of the value and constraint functionals,

$$\sum_{c\in\mathcal{C}} p_c \,\langle \boldsymbol{\pi}_t(c), \boldsymbol{r}(c)\rangle = \mathcal{L}_{\lambda^\star}(\boldsymbol{\pi}_t) + \lambda^\star \sum_{c\in\mathcal{C}} p_c \,\langle \boldsymbol{\pi}_t(c), \boldsymbol{\psi}(c)\rangle\,. \tag{15}$$

Substituting (13) and (14) and using $\lambda^\star \ge 0$,

$$\sum_{c\in\mathcal{C}} p_c \,\langle \boldsymbol{\pi}_t(c), \boldsymbol{r}(c)\rangle = v^\star + \lambda^\star \underbrace{\sum_{c\in\mathcal{C}} p_c \,\langle \boldsymbol{\pi}_t(c), \boldsymbol{\psi}(c)\rangle}_{\ge 0} \;\ge\; v^\star = \sum_{c\in\mathcal{C}} p_c \,\langle \boldsymbol{\pi}^\star(c), \boldsymbol{r}(c)\rangle\,,$$

where the last equality is (12) (recall $\mathcal{L}_{\lambda^\star}(\boldsymbol{\pi}^\star) = v^\star$ and $\lambda^\star \sum_{c\in\mathcal{C}} p_c\langle \boldsymbol{\pi}^\star(c), \boldsymbol{\psi}(c)\rangle = 0$). Rearranging yields $\sum_{c\in\mathcal{C}} p_c \,\langle \boldsymbol{\pi}^\star(c) - \boldsymbol{\pi}_t(c), \boldsymbol{r}(c)\rangle \le 0$, which is the claim. □

With all of this in hand we are ready to prove the main theorem.

Theorem 5.1. *Under Assumptions 3.1, 3.2, and 3.3, the regret and cumulative constraint violation of **OPALP** satisfy*

$$\mathcal{R}_T = \mathcal{O}\Big(\min\Big\{\frac{\log(T)}{\Delta^\star},\ \sqrt{T\log(T)}\Big\}\Big), \quad \mathcal{V}_T = \tilde{\mathcal{O}}\Big(\sqrt{T}\Big).$$

*In particular, **OPALP** attains logarithmic regret $\mathcal{O}\big(\log(T)/\Delta^\star\big)$ whenever the gap $\Delta^\star$ is bounded away from zero, while never exceeding the previous work rate $\tilde{\mathcal{O}}(\sqrt{T})$, irrespective of $\Delta^\star$.*

Proof of Theorem 5.1. We assess each metric separately.

Regret. Recall the good event $\mathcal{GE}$ of Assumption 3.3. Then

$$\begin{aligned}
\mathcal{R}_T &= T\,v^\star - \mathbb{E}\left[\sum_{t=1}^{T} r_t(c_t, a_t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{c\in\mathcal{C}} p_c\Big(\langle \boldsymbol{\pi}^\star(c), \boldsymbol{r}(c)\rangle - \langle \boldsymbol{\pi}_t(c), \boldsymbol{r}(c)\rangle\Big)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{c\in\mathcal{C}} p_c\Big(\langle \boldsymbol{\pi}^\star(c), \boldsymbol{r}(c)\rangle - \langle \boldsymbol{\pi}_t(c), \boldsymbol{r}(c)\rangle\Big)\,\Big|\,\mathcal{GE}\right]\mathbb{P}(\mathcal{GE}) \\
&\quad + \mathbb{E}\left[\sum_{t=1}^{T}\sum_{c\in\mathcal{C}} p_c\Big(\langle \boldsymbol{\pi}^\star(c), \boldsymbol{r}(c)\rangle - \langle \boldsymbol{\pi}_t(c), \boldsymbol{r}(c)\rangle\Big)\,\Big|\,\mathcal{GE}^c\right]\mathbb{P}(\mathcal{GE}^c) \\
&\le \mathbb{E}\left[\sum_{t=1}^{T}\sum_{c\in\mathcal{C}} p_c\Big(\langle \boldsymbol{\pi}^\star(c), \boldsymbol{r}(c)\rangle - \langle \boldsymbol{\pi}_t(c), \boldsymbol{r}(c)\rangle\Big)\,\Big|\,\mathcal{GE}\right] + T\delta \\
&\overset{(a)}{\le} \mathbb{E}\left[\sum_{t=1}^{T} \xi_t(\pi_t,\delta)\,\mathbb{1}_{[\xi_t(\pi_t,\delta)\ge\Delta^\star]}\right] + \frac{1}{T},
\end{aligned}$$

where (a) follows from the choice $\delta = 1/T^2$, from Proposition C.3, from Proposition C.4, and from Proposition C.5. We bound the last expression in two complementary ways.

(i) Gap-dependent bound. Whenever $\xi_t(\pi_t,\delta) \ge \Delta^\star$ we have $\xi_t(\pi_t,\delta) \le \xi_t(\pi_t,\delta)^2/\Delta^\star$; moreover, by Jensen's inequality,

$$\xi_t(\pi_t,\delta)^2 = \left(\mathbb{E}_{\substack{c_t\sim\mathcal{P}_c,\\ a_t\sim\pi_t}}\big[\varepsilon_t(c_t,a_t)\big]\right)^2 \le \mathbb{E}_{\substack{c_t\sim\mathcal{P}_c,\\ a_t\sim\pi_t}}\big[\varepsilon_t(c_t,a_t)^2\big].$$

Therefore

$$\mathcal{R}_T \le \mathbb{E}\left[\sum_{t=1}^{T}\frac{\xi_t(\pi_t,\delta)^2}{\Delta^\star}\right] + \frac{1}{T} \le \frac{1}{\Delta^\star}\,\mathbb{E}\left[\sum_{t=1}^{T}\varepsilon_t(c_t,a_t)^2\right] + \frac{1}{T} \le \mathcal{O}\Big(\frac{\log(T)}{\Delta^\star}\Big),$$

where the last inequality uses Assumption 3.3.

(ii) Worst-case bound (independent of $\Delta^\star$). Dropping the indicator, $\xi_t(\pi_t,\delta)\,\mathbb{1}_{[\xi_t(\pi_t,\delta)\ge\Delta^\star]} \le \xi_t(\pi_t,\delta)$, and applying the Cauchy–Schwarz inequality exactly as in the constraint-violation bound below,

$$\mathcal{R}_T \le \mathbb{E}\left[\sum_{t=1}^{T}\xi_t(\pi_t,\delta)\right] + \frac{1}{T} \le \mathbb{E}\left[\sqrt{T\sum_{t=1}^{T}\xi_t(\pi_t,\delta)^2}\right] + \frac{1}{T} \le \tilde{\mathcal{O}}\Big(\sqrt{T}\Big),$$

again by Assumption 3.3. This bound holds for every value of $\Delta^\star$.

Combining (i) and (ii) yields the stated regret bound,

$$\mathcal{R}_T \le \mathcal{O}\Big(\min\Big\{\frac{\log(T)}{\Delta^\star},\ \sqrt{T\log(T)}\Big\}\Big).$$

Constraint violation. By Proposition C.2, the per-round expected cost of the policy played by **OPALP** is upper bounded by the global estimation error $\xi_t(\pi_t, \delta)$. The constraint-violation bound then follows:

$$\begin{aligned}
\mathcal{V}_T &= \mathbb{E}\left[\sum_{t=1}^{T} \psi(c_t, a_t)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T} \sum_{c\in\mathcal{C}} p_c \left\langle \boldsymbol{\pi}_t(c), \boldsymbol{\psi}(c)\right\rangle\right] \\
&\le \mathbb{E}\left[\sum_{t=1}^{T} \xi_t(\pi_t, \delta)\right] && \text{(Proposition C.2)} \\
&\le \mathbb{E}\left[\sqrt{T\sum_{t=1}^{T} \xi_t(\pi_t, \delta)^2}\right] && \text{(Cauchy–Schwarz)} \\
&\le \tilde{\mathcal{O}}\left(\sqrt{T}\right) && \text{(Assumption 3.3).}
\end{aligned}$$

□

C.4 OPALP and Contextual Bandits with Knapsack (CBwK)

Corollary 6.1. *Under Assumptions 3.1 to 3.3, within the CBwK framework,* ***OPALP*** *achieves*

$$\mathcal{R}_T^{\mathrm{CBwK}} = \widetilde{\mathcal{O}}\left(\left(1 + \frac{v^\star}{b}\right)\sqrt{T}\right).$$

Proof of Corollary 6.1. Let τ be the round at which the budget is exhausted, capped at the horizon:

$$\tau := \min\left\{t \in [T] : \sum_{s=1}^{t} w_s(c_s, a_s) \ge B\right\} \wedge T.$$

The regret decomposes as

$$\mathcal{R}_T^{\mathrm{CBwK}} = \overbrace{\mathbb{E}\left[\sum_{t=1}^{\tau}\left(v^\star - r(c_t, a_t)\right)\right]}^{\mathcal{R}^{\mathrm{pre}}} + \overbrace{\mathbb{E}\left[(T-\tau)\, v^\star\right]}^{\mathcal{R}^{\mathrm{post}}}.$$

Throughout, write $v_t := \sum_{c\in\mathcal{C}} p_c \langle \boldsymbol{\pi}_t(c), \boldsymbol{\psi}(c)\rangle$ for the expected cost of the policy played at round t, and let $\mathcal{GE}$ denote the good event of Assumption 3.3.

A martingale correction. Since π_t is $\mathcal{H}_{t-1}$-measurable and $c_t \sim \mathcal{P}_\mathcal{C}$ is independent of $\mathcal{H}_{t-1}$, we have $\mathbb{E}\left[\psi_t(c_t, a_t) \mid \mathcal{H}_{t-1}\right] = v_t$. Hence $D_t := \psi_t(c_t, a_t) - v_t$ is a martingale difference sequence with $|D_t| \le 2$ (as $\psi_t \in [-1, 1]$), and $M_t := \sum_{s\le t} D_s$ is a martingale. Since $\tau \le T$ is a bounded stopping time, optional stopping yields $\mathbb{E}\left[M_\tau^2\right] = \mathbb{E}\left[\sum_{s\le\tau} D_s^2\right] \le 4T$, whence

$$\mathbb{E}\,|M_\tau| \;\le\; \sqrt{\mathbb{E}\left[M_\tau^2\right]} \;\le\; 2\sqrt{T}. \tag{16}$$

This step is needed because feasibility of π_t in $LP(\hat{r}_t, \check{\psi}_t)$ controls only the expected cost v_t, and not the realized sum $\sum_{s\le\tau} \psi_s(c_s, a_s)$; the two differ exactly by M_τ.

For the $\mathcal{R}^{\mathrm{post}}$ term: On the event $\{\tau < T\}$ the budget is exhausted, so

$$\sum_{s=1}^{\tau} \psi_s(c_s, a_s) = \sum_{s=1}^{\tau} w_s(c_s, a_s) - b\tau \;\ge\; B - b\tau \;=\; b(T-\tau),$$

while on $\{\tau = T\}$ we trivially have $b(T-\tau) = 0$. Hence, pathwise,

$$b(T-\tau) \;\le\; \left(\sum_{s=1}^{\tau} \psi_s(c_s, a_s)\right)_+ \;\le\; |M_\tau| + \left(\sum_{s=1}^{\tau} v_s\right)_+ ,$$

using $(x+y)_+ \le |x| + (y)_+$. On the event $\mathcal{GE}$, Proposition C.2 gives $v_s \le \xi_s(\pi_s, \delta)$ with $\xi_s(\pi_s, \delta) \ge 0$, so that $\left(\sum_{s\le\tau} v_s\right)_+ \le \sum_{s\le T} \xi_s(\pi_s, \delta)$; on $\mathcal{GE}^c$ we bound the sum crudely by T, contributing $T\delta = 1/T$ for $\delta = 1/T^2$. Moreover,

$$\begin{aligned}
\mathbb{E}\left[\sum_{s=1}^{T} \xi_s(\pi_s, \delta)\right] &\le \sqrt{T\,\mathbb{E}\left[\sum_{s=1}^{T} \xi_s(\pi_s, \delta)^2\right]} && \text{(Cauchy–Schwarz)} \\
&\le \sqrt{T\,\mathbb{E}\left[\sum_{s=1}^{T} \varepsilon_s(c_s, a_s, \delta)^2\right]} && \text{(Jensen, Definition C.1)} \\
&\le \tilde{\mathcal{O}}(\sqrt{T}) && \text{(Assumption 3.3).}
\end{aligned}$$

Combining the last three displays with (16),

$$b\,\mathbb{E}[T-\tau] \;\le\; \tilde{\mathcal{O}}(\sqrt{T}) \qquad \Longrightarrow \qquad \mathcal{R}^{\mathrm{post}} = \mathbb{E}\left[(T-\tau)\,v^\star\right] \;\le\; \frac{v^\star}{b}\,\tilde{\mathcal{O}}(\sqrt{T}).$$

For the $\mathcal{R}^{\mathrm{pre}}$ term: Write

$$v^\star - r(c_t, a_t) = \Big(v^\star - \sum_{c\in\mathcal{C}} p_c \langle \boldsymbol{\pi}_t(c), \boldsymbol{r}(c)\rangle\Big) + \Big(\sum_{c\in\mathcal{C}} p_c \langle \boldsymbol{\pi}_t(c), \boldsymbol{r}(c)\rangle - r(c_t, a_t)\Big).$$

The second bracket is a bounded martingale difference (again since π_t is $\mathcal{H}_{t-1}$-measurable and $c_t \perp \mathcal{H}_{t-1}$), so by optional stopping its sum up to the bounded stopping time τ has zero expectation. For the first bracket, we bound its sum up to τ by the sum of its positive parts up to T and proceed exactly as in the proof of Theorem 5.1: on $\mathcal{GE}$, Proposition C.5 gives a zero contribution whenever $\mathcal{I}_t = \mathcal{I}^\star$, while Propositions C.3 and C.4 give the bound $\xi_t(\pi_t, \delta)\,\mathbb{1}_{[\xi_t(\pi_t,\delta)\ge\Delta^\star]}$ otherwise. This yields

$$\mathcal{R}^{\mathrm{pre}} = \mathcal{O}\left(\min\left\{\frac{\log(T)}{\Delta^\star},\ \sqrt{T\log(T)}\right\}\right) = \tilde{\mathcal{O}}(\sqrt{T}).$$

As a conclusion,

$$\mathcal{R}_T^{\mathrm{CBwK}} = \tilde{\mathcal{O}}\left(\left(1 + \frac{v^\star}{b}\right)\sqrt{T}\right).$$

□

D Computational Complexity

Lazy-OPALP employs a reduced update frequency compared to its vanilla counterpart, **OPALP**. Define the *clipped* estimation errors

$$\bar{\varepsilon}_t(c, a, \delta) := \max\left\{\varepsilon_t(c, a, \delta),\ \frac{1}{T}\right\}, \qquad \forall (c, a) \in \mathcal{C}\times\mathcal{A}.$$

The policy optimization step is triggered only when the clipped estimation error at some context–action pair has shrunk by a factor of $\sqrt{2}$ since the last update. Between such events, the policy and estimators $\{\hat{r}_t, \check{\psi}_t\}$ from the most recent update are retained. Since $\bar{\varepsilon}_t(c, a, \delta) \in [1/T,\, 2]$, each pair can trigger at most $\log_{\sqrt{2}}(2T)$ updates, so this scheme reduces the number of LP solves to at most $\mathcal{O}(\log T)$ over the horizon T.

Lazy-OPALP: Lazy OPtimistic Approximation of Linear Programming

Inputs: action set $\mathcal{A}$, context set $\mathcal{C}$, distribution $\mathcal{P}_{\mathcal{C}}$.
Initialize $\tau \leftarrow 1$ `// last round at which the policy was updated`
for $t = 1$ **to** T **do**
  Observe the context c_t.
  Estimation: Construct estimators $\{\hat{r}_t, \check{\psi}_t\}$ via the regression oracles $\{\Omega_r, \Omega_\psi\}$, and set $\bar{\varepsilon}_t(c, a, \delta) \leftarrow \max\{\varepsilon_t(c, a, \delta),\ 1/T\}$ for all $(c, a) \in \mathcal{C} \times \mathcal{A}$.
  Policy Optimization: Compute the policy π_t as follows:
  if $\exists\, (c, a) \in \mathcal{C} \times \mathcal{A}$ *such that* $\bar{\varepsilon}_t(c, a, \delta) < \dfrac{\bar{\varepsilon}_\tau(c, a, \delta)}{\sqrt{2}}$ **then**
    $\pi_t \leftarrow \operatorname{argmax}\ \mathrm{LP}(\hat{r}_t, \check{\psi}_t)$
    $\tau \leftarrow t$
  else
    $\pi_t \leftarrow \pi_\tau$
  Sampling: Play an action $a_t \sim \boldsymbol{\pi}_t(c_t)$.
  Feedback: Observe reward $r_t(c_t, a_t)$ and cost $\psi_t(c_t, a_t)$.
  History Update: $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1} \cup \{c_t, a_t, r_t(c_t, a_t), \psi_t(c_t, a_t)\}$.

Update count. Fix a pair (c, a). Each update triggered by (c, a) requires $\bar{\varepsilon}_t(c, a, \delta) < \bar{\varepsilon}_\tau(c, a, \delta)/\sqrt{2}$, i.e., the clipped width of that pair shrinks by a factor of at least $\sqrt{2}$ relative to its value at the previous update. Since the clipped width takes values in $[1/T, 2]$, the pair (c, a) can trigger at most $\log_{\sqrt{2}}(2T) = \mathcal{O}(\log T)$ updates over the horizon. Summing over the (finitely many) pairs, the total number of LP solves is $\mathcal{O}(\log T)$.

Preservation of the guarantees. Notably, this reduction in update frequency does not deteriorate the theoretical guarantees: the analysis and the characterization of the saturating constraint set remain valid exactly as in the non-lazy case. Consequently, the logarithmic and $\widetilde{\mathcal{O}}(\sqrt{T})$ bounds are intrinsically tied to the respective bounds of

$$\mathbb{E}\left[\sum_{t=1}^{T} \xi_t(\pi_t, \delta)^2\right] \quad \text{and} \quad \mathbb{E}\left[\sum_{t=1}^{T} \xi_t(\pi_t, \delta)\right],$$

where

$$\xi_t(\pi_t, \delta) := \sum_{c \in \mathcal{C}} \sum_{a \in \mathcal{A}} p_c\, \pi_t(c, a)\, \varepsilon_{\tau_t}(c, a, \delta),$$

and τ_t denotes the most recent round prior to t at which a policy update occurred (so that $\pi_t = \pi_{\tau_t}$ is computed from the estimators of round τ_t, whose widths govern its guarantees). By negation of the trigger condition at every non-update round between τ_t and t, we have

$$\bar{\varepsilon}_{\tau_t}(c, a, \delta) \ \le\ \sqrt{2}\, \bar{\varepsilon}_t(c, a, \delta), \qquad \forall (c, a) \in \mathcal{C} \times \mathcal{A},$$

and moreover $\varepsilon_{\tau_t} \le \bar{\varepsilon}_{\tau_t}$ and $\bar{\varepsilon}_t \le \varepsilon_t + 1/T$ by the definition of the clipping. Under **Lazy-OPALP**, using $(x+y)^2 \le 2x^2 + 2y^2$, we obtain

$$
\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T} \xi_t(\pi_t,\delta)^2\right] &\le \mathbb{E}\left[\sum_{t=1}^{T}\sum_{c\in\mathcal{C}}\sum_{a\in\mathcal{A}} p_c\,\pi_t(c,a)\,\varepsilon_{\tau_t}(c,a,\delta)^2\right] && \text{(Jensen's inequality)}\\
&\le \mathbb{E}\left[2\sum_{t=1}^{T}\sum_{c\in\mathcal{C}}\sum_{a\in\mathcal{A}} p_c\,\pi_t(c,a)\,\bar{\varepsilon}_t(c,a,\delta)^2\right] && (\varepsilon_{\tau_t}^2 \le \bar{\varepsilon}_{\tau_t}^2 \le 2\bar{\varepsilon}_t^2)\\
&\le \mathbb{E}\left[4\sum_{t=1}^{T}\sum_{c\in\mathcal{C}}\sum_{a\in\mathcal{A}} p_c\,\pi_t(c,a)\,\varepsilon_t(c,a,\delta)^2\right] + \frac{4}{T} && (\bar{\varepsilon}_t^2 \le 2\varepsilon_t^2 + 2/T^2)\\
&= \mathbb{E}\left[4\sum_{t=1}^{T}\varepsilon_t(c_t,a_t,\delta)^2\right] + \frac{4}{T} && \text{(tower rule, } c_t \perp \mathcal{H}_{t-1},\ a_t\sim\pi_t(c_t))\\
&\le \mathcal{O}(\log T) && (Assumption\ 3.3),\\
\mathbb{E}\left[\sum_{t=1}^{T} \xi_t(\pi_t,\delta)\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{c\in\mathcal{C}}\sum_{a\in\mathcal{A}} p_c\,\pi_t(c,a)\,\varepsilon_{\tau_t}(c,a,\delta)\right]\\
&\le \mathbb{E}\left[\sqrt{2}\sum_{t=1}^{T}\sum_{c\in\mathcal{C}}\sum_{a\in\mathcal{A}} p_c\,\pi_t(c,a)\,\varepsilon_t(c,a,\delta)\right] + \sqrt{2} && (\varepsilon_{\tau_t} \le \sqrt{2}\,\bar{\varepsilon}_t \le \sqrt{2}\,\varepsilon_t + \sqrt{2}/T)\\
&= \mathbb{E}\left[\sqrt{2}\sum_{t=1}^{T}\varepsilon_t(c_t,a_t,\delta)\right] + \sqrt{2} && \text{(tower rule, } c_t \perp \mathcal{H}_{t-1},\ a_t\sim\pi_t(c_t))\\
&\le \mathbb{E}\left[\sqrt{2}\sqrt{T\sum_{t=1}^{T}\varepsilon_t(c_t,a_t,\delta)^2}\right] + \sqrt{2} && \text{(Cauchy–Schwarz)}\\
&\le \widetilde{\mathcal{O}}(\sqrt{T}) && (Assumption\ 3.3).
\end{aligned}
$$

As a result, the lazy version of the algorithm preserves the same theoretical guarantees as its vanilla counterpart, up to constant factors.

Practical Deployment: from a practical perspective, the computational burden can be further reduced using standard optimization techniques. Modern LP solvers leverage presolve routines to simplify instances and warm-starting to reuse solutions from similar problems. Moreover, solver tolerances can be relaxed to match estimation accuracy, avoiding unnecessary numerical precision without affecting empirical performance.

E Relaxed Assumptions

E.1 From Known to Unknown Context Distribution

In the main text, we adopt the known-context-distribution assumption, which is also imposed by our baseline (Guo et al., 2025), primarily for clarity of presentation; it is not a prerequisite of our approach. Indeed, our method operates entirely on estimated quantities, and the context distribution can be handled in the same manner, which is why this requirement is mild. As noted in Chen et al. (2024), when the context distribution satisfies suitable smoothness properties—for instance, when it belongs to the β_p-order L_p-Hölder smoothness (Wasserman, 2019)— it can be efficiently estimated via empirical averaging. Concretely, Assumption 3.3 would then concern estimators of the products $r'(c,a) = p_c\, r(c,a)$ and $\psi'(c,a) = p_c\,\psi(c,a)$, rather than of $r(c,a)$ and $\psi(c,a)$ alone. The remainder of the analysis proceeds in exactly the same spirit, both in the construction of the relevant

quantities and in the regret decomposition. Namely, the offline problem reads as follows:

$$\mathrm{LP}(r', \psi') := \max_{\pi \in \mathbb{R}^{|\mathcal{A}| \times |\mathcal{C}|}} \sum_{c \in \mathcal{C}} \langle \boldsymbol{\pi}(c), \boldsymbol{r}'(c) \rangle \quad \text{s.t.} \quad \begin{aligned} &\sum_{c \in \mathcal{C}} \langle \boldsymbol{\pi}(c), \boldsymbol{\psi}'(c) \rangle \leq 0, \\ &\sum_{a \in \mathcal{A}} \pi(c, a) = 1, \quad \forall\, c \in \mathcal{C}, \\ &\pi(c, a) \geq 0, \quad \forall\, c \in \mathcal{C},\ \forall\, a \in \mathcal{A}. \end{aligned} \tag{17}$$

We then assume access to online regression oracles $\{\Omega_{r'}, \Omega_{\psi'}\}$ for the weighted reward and cost functions, producing estimators $\hat{r}'_t$ and $\check{\psi}'_t$ such that, with probability at least $1 - \delta$, the event

$$\xi := \left\{ \begin{aligned} &\forall\, t \in [T],\ \forall\, c \in \mathcal{C},\ \forall\, a \in \mathcal{A} : \\ &\quad 0 \leq \hat{r}'_t(c, a) - r'(c, a) \leq \varepsilon_t(c, a, \delta), \\ &\quad 0 \leq \psi'(c, a) - \check{\psi}'_t(c, a) \leq \varepsilon_t(c, a, \delta) \end{aligned} \right\}$$

holds. We further adapt the key set used in the definitions of the infeasibility and sub-optimality gaps:

$$\Gamma'(s, \mathcal{I}) := \left\{ \pi \in \Pi : \begin{aligned} &\sum_{c \in \mathcal{C}} \langle \boldsymbol{\pi}(c), \boldsymbol{\psi}'(c) \rangle \leq s, \\ &\sum_{c \in \mathcal{C}} \langle \boldsymbol{\pi}(c), \boldsymbol{\psi}'(c) \rangle \geq -s \text{ if } \mathbf{GC} \in \mathcal{I}, \\ &\pi(c, a) = 0 \quad \forall\, (c, a) \in \mathcal{J} \cap \mathcal{I} \end{aligned} \right\},$$

and the remainder of the analysis carries over unchanged. In particular, since context arrivals are fully observed, the products r' and ψ' are directly estimable from data without ever estimating p_c; the linear program written in the primed quantities is identical to (3), so the candidate sets (Definition 4.1), the gaps $s(\mathcal{I})$, $\rho(\mathcal{I})$, and $\Delta^\star$ (Definitions 4.2 to 4.4), as well as Lemmas 4.1 and 5.1, Proposition C.1, and Propositions C.2 to C.5, hold verbatim with (r', ψ', Γ') in place of (r, ψ, Γ), and the averaged global estimation error $\xi_t(\pi_t, \delta)$ of Definition C.1 is unchanged. Consequently, Theorem 5.1 holds with the same rates: the context distribution is handled implicitly, agnostically, and at no cost in the guarantees. Hence, the known-context-distribution assumption adopted in the main text does not simplify the learning problem; it is made purely for expository purposes.

E.2 Extension to Multiple Global Constraints

Our approach, and in particular the construction underlying the definition of the sub-optimality gap $\Delta^\star$, extends to the case of multiple constraints. Specifically, consider d distinct generic constraints, indexed by $\{\mathbf{GC}_1, \ldots, \mathbf{GC}_d\}$, so that the offline problem reads:

$$\mathrm{LP}(r, \psi) := \max_{\pi \in \mathbb{R}^{|\mathcal{A}| \times |\mathcal{C}|}} \sum_{c \in \mathcal{C}} p_c \, \langle \boldsymbol{\pi}(c), \boldsymbol{r}(c) \rangle \quad \text{s.t.} \quad \begin{aligned} &\sum_{c \in \mathcal{C}} p_c \, \langle \boldsymbol{\pi}(c), \boldsymbol{\psi}_i(c) \rangle \leq 0, \quad \forall\, i \in [d], \\ &\sum_{a \in \mathcal{A}} \pi(c, a) = 1, \quad \forall\, c \in \mathcal{C}, \\ &\pi(c, a) \geq 0, \quad \forall\, c \in \mathcal{C},\ \forall\, a \in \mathcal{A}, \end{aligned}$$

where $\boldsymbol{\psi}_i(c)$ denotes the mean cost vector of the i-th constraint under context c. We impose Assumption 3.2 on the linear program above, namely that it is feasible and non-degenerate. The estimation oracles now comprise one cost oracle per constraint, $\{\Omega_r, \Omega_{\psi_1}, \ldots, \Omega_{\psi_d}\}$, producing estimators $\hat{r}_t$ and $\{\check{\psi}_{i,t}\}_{i \in [d]}$ such that, with probability at least $1 - \delta$, the event

$$\xi := \left\{ \begin{aligned} &\forall\, t \in [T],\ \forall\, c \in \mathcal{C},\ \forall\, a \in \mathcal{A} : \\ &\quad 0 \leq \hat{r}_t(c, a) - r(c, a) \leq \varepsilon_t(c, a, \delta), \\ &\quad 0 \leq \psi_i(c, a) - \check{\psi}_{i,t}(c, a) \leq \varepsilon_t(c, a, \delta), \quad \forall\, i \in [d] \end{aligned} \right\}$$

holds, with $\sum_{t=1}^{T} \varepsilon_t(c_t, a_t, \delta)^2 = \mathcal{O}\big(\log\big(\max\{|\mathcal{R}|, |\Psi_1|, \ldots, |\Psi_d|\}/\delta \big) \big)$, where Ψ_i denotes the function class of the i-th cost. At each round, **OPALP** solves $\mathrm{LP}(\hat{r}_t, \check{\psi}_{1,t}, \ldots, \check{\psi}_{d,t})$; the feasibility argument of Lemma 5.1 applies constraint-wise, since $\check{\psi}_{i,t} \leq \psi_i$ componentwise on ξ implies that $\pi^\star$ remains feasible for the optimistic program.

The combinatorial objects generalize as follows. The inequality constraints of the LP are now indexed by $\{\mathbf{GC}_1, \ldots, \mathbf{GC}_d\} \cup \mathcal{J}$, and a candidate set $\mathcal{I}$ is any subset of $\{\mathbf{GC}_1, \ldots, \mathbf{GC}_d\} \cup \mathcal{J}$ that, jointly with the $|\mathcal{C}|$ simplex equalities, forms a basic index set (Definition 4.1); in particular, $|\mathcal{I}| = |\mathcal{C}|\,|\mathcal{A}| - |\mathcal{C}|$, of which up to d indices may correspond to generic constraints. Under the non-degeneracy assumption, the optimal policy $\pi^\star$ remains a non-degenerate vertex characterized by a unique optimal basis $\mathcal{I}^\star$, where $\mathbf{GC}_i \in \mathcal{I}^\star$ indicates that $\pi^\star$ saturates the i-th constraint, and $(c, a) \in \mathcal{I}^\star \cap \mathcal{J}$ indicates that $\pi^\star(c, a) = 0$. The set of pseudo-feasible policies becomes:

$$\Gamma(s, \mathcal{I}) := \left\{ \pi \in \Pi : \begin{array}{l} \displaystyle\sum_{c \in \mathcal{C}} p_c \, \langle \boldsymbol{\pi}(c), \boldsymbol{\psi}_i(c) \rangle \le s, \quad \forall\, i \in [d], \\ \displaystyle\sum_{c \in \mathcal{C}} p_c \, \langle \boldsymbol{\pi}(c), \boldsymbol{\psi}_i(c) \rangle \ge -s \quad \text{if } \mathbf{GC}_i \in \mathcal{I}, \\ \pi(c, a) = 0 \quad \forall\, (c, a) \in \mathcal{J} \cap \mathcal{I} \end{array} \right\},$$

so that the slack s now relaxes all d constraints simultaneously, and saturation is required only for those constraints indexed in $\mathcal{I}$. The infeasibility gap $s(\mathcal{I})$, the sensitivity $\mathcal{L}(\mathcal{I})$, the performance gap $\rho(\mathcal{I})$, and the sub-optimality gap $\Delta(\mathcal{I}) = \max\{s(\mathcal{I}), \rho(\mathcal{I})\}$ are then defined verbatim from this modified $\Gamma(s, \mathcal{I})$, and $\Delta^\star := \min_{\mathcal{I} \neq \mathcal{I}^\star} \Delta(\mathcal{I})$ remains positive by the same argument as in Lemma 4.1, since the family of candidate sets is finite and non-degeneracy ensures uniqueness of $\mathcal{I}^\star$.

The analysis carries over with only notational changes. Lemma C.1 holds with the multi-constraint Γ: the played policy π_t satisfies each estimated constraint, and the one-sided estimation guarantees transfer this, constraint by constraint, to a $\xi_t(\pi_t, \delta)$-relaxation of each true constraint, yielding $\pi_t \in \Gamma(\xi_t(\pi_t, \delta), \mathcal{I}_t)$ and hence Proposition C.4. For Proposition C.5, the scalar multiplier $\lambda^\star$ is replaced by the optimal dual vector $\boldsymbol{\lambda}^\star \in \mathbb{R}^d_{\ge 0}$, the Lagrangian becomes $\mathcal{L}_{\boldsymbol{\lambda}^\star}(\boldsymbol{\pi}) = \sum_c p_c \langle \boldsymbol{\pi}(c), \boldsymbol{r}(c) - \sum_{i \in [d]} \lambda_i^\star \boldsymbol{\psi}_i(c) \rangle$, and complementary slackness applies coordinate-wise: for each i with $\lambda_i^\star > 0$, the constraint $\mathbf{GC}_i$ is active, hence belongs to $\mathcal{I}^\star = \mathcal{I}_t$, and π_t saturates it in the optimistic program, so that $\sum_c p_c \langle \boldsymbol{\pi}_t(c), \boldsymbol{\psi}_i(c) \rangle \ge 0$ term by term. The regret bound of Theorem 5.1 therefore holds unchanged, while the constraint-violation bound, applied to each of the d constraints via Proposition C.2, yields $\mathcal{V}_T^{(i)} = \tilde{\mathcal{O}}(\sqrt{T})$ for every $i \in [d]$, i.e., an $\tilde{\mathcal{O}}(d\sqrt{T})$ bound on the aggregate violation. The dependence on d is thus confined to the number of oracles, the cardinality of the candidate family, and this linear factor in the aggregate violation; the rates in T are unaffected.

F Details of Experiments

To corroborate our theoretical results, we conduct several numerical experiments evaluating the performance of **OPALP**.

F.1 Comparison with the State of the Art

We compare **OPALP** against three primal–dual baselines: **Optimistic**^{3} (Guo et al., 2025), **LOE2D** (Guo & Liu, 2024), and **LagrangeCBwLC** (Slivkins et al., 2024), the latter granted the *true* Slater constant of the instance. All methods share the same estimators (empirical means with standard confidence widths); each baseline is tuned in its favor on held-out seeds, selecting the configuration with the smallest final regret among those whose final violation is within a fixed band around that of **OPALP**.

Synthetic Data. We consider a setting with 20 contexts and 10 arms, with a randomly generated context distribution. For each context–arm pair (c, a), we associate a Gaussian reward distribution with mean $r(c, a) \in [0, 1]$ and a Gaussian cost distribution with mean $\psi(c, a) \in [-1, 1]$, both generated uniformly at random (noise level $\sigma = 0.5$); instances are screened so that the general constraint is active at the optimum. Figure 1(a) shows that the regret of **OPALP** flattens after an initial learning phase — the logarithmic regime of Theorem 5.1 — whereas that of all three baselines keeps growing at a $\sqrt{T}$ rate. Figure 1(b) reports the cumulative violation; all methods incur violations of the same order, with **OPALP** attaining the smallest.

Real-World Dataset. On real-world data we benchmark **OPALP** against **Optimistic**^{3}, using the large-scale learning-to-rank dataset Microsoft MSLR-WEB30k (Qin & Liu, 2013). Following Guo

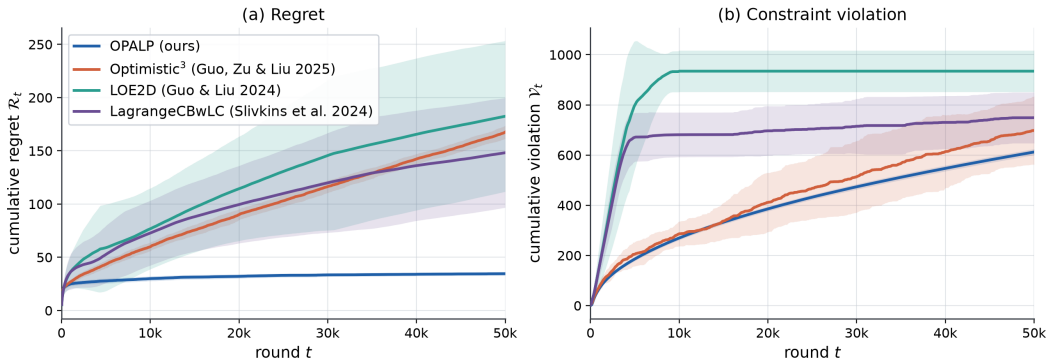
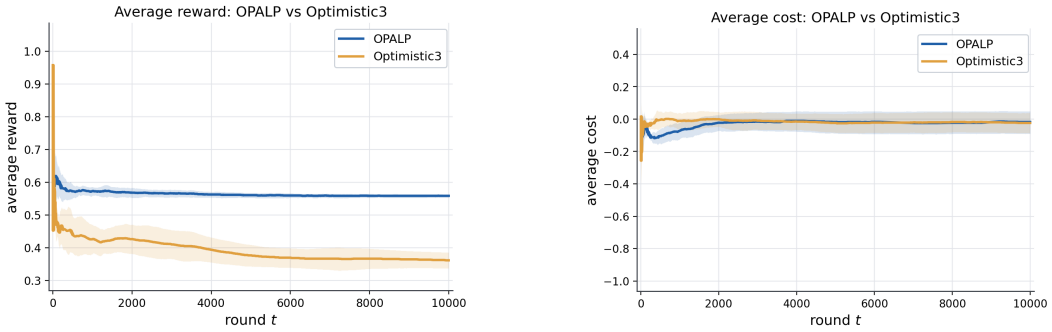


Figure 1: Cumulative regret (a) and cumulative constraint violation (b) of **OPALP** and three primal–dual baselines (Guo et al., 2025; Guo & Liu, 2024; Slivkins et al., 2024) on synthetic data (20 contexts, 10 arms; mean $\pm$ one standard deviation over 10 runs).

et al. (2025), we adopt a similar general-constraint setting, in which the reward $r(x, a)$ is the relevance score (normalized to $[0, 1]$) assigned to the recommended document for the incoming user, and the cost $\psi(x, a)$ of each arm is drawn uniformly at random from $[-0.5, 1]$ and remains fixed throughout each trial. We employ gradient-boosted tree regression and the empirical mean to estimate the reward and cost functions, respectively. Observations of rewards and costs are perturbed by Gaussian noise $\mathcal{N}(0, 0.05)$, and the reported results are averaged over 20 independent trials. As shown in Figure 2, **OPALP** attains a higher average reward than **Optimistic**^{3} at comparable average cost; the early fluctuations correspond to the estimation phase, during which both algorithms occasionally play infeasible policies.



(a) Average reward.

(b) Average cost.

Figure 2: Average reward (a) and average cost (b) under **OPALP** and **Optimistic**^{3} on MSLR-WEB30k (mean $\pm$ one standard deviation over 20 trials).

F.2 The Effect of $\Delta^\star$ on the Regret

We assess the effect of $\Delta^\star$ on the regret using Example 1, averaged over 30 runs, varying the gap such that $\Delta_1^\star > \Delta_2^\star \gg \Delta_3^\star$. As shown in Figure 3, the regret scales inversely with $\Delta^\star$, at a rate consistent with the $\mathcal{O}(\log(T)/\Delta^\star)$ dependence of Theorem 5.1, and the flattening of the regret curve — marking the identification of $\mathcal{I}^\star$ — occurs later on smaller-gap instances. As $\Delta^\star \to 0$, distinguishing optimal from sub-optimal active sets becomes inherently difficult, and the regret behavior transitions from the problem-dependent (logarithmic) regime to a worst-case regime.

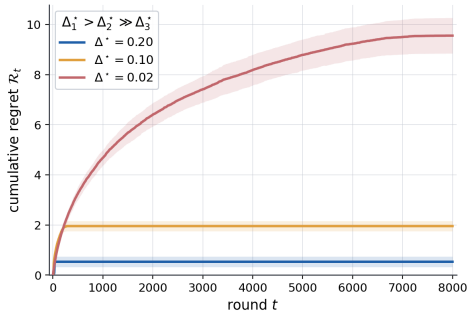


Figure 3: Cumulative regret of **OPALP** on instances of Example 1 with decreasing sub-optimality gap $\Delta^\star$ (mean $\pm$ one standard deviation over 30 runs).

F.3 Unknown Context Distribution

We provide empirical evidence that the assumption of a known context distribution can be relaxed without affecting the regret. Instead of plugging p_c into the fluid LP, we estimate the weighted coefficients $p_c\, r(c,a)$ and $p_c\, \psi(c,a)$ directly, forming a UCB on the former and an LCB on the latter by propagating both the empirical context-frequency uncertainty $\hat{p}_c \pm \delta_p$ and the reward/cost uncertainty through the product; p_c thus never appears explicitly in the program. As shown in Figure 4, the estimated-p_c variant attains a regret statistically indistinguishable from **OPALP** with known p_c, while preserving the $\widetilde{\mathcal{O}}(\sqrt{T})$ scaling of the constraint violation; the additional uncertainty in p_c surfaces only as a slightly larger, but identically scaling, violation constant.

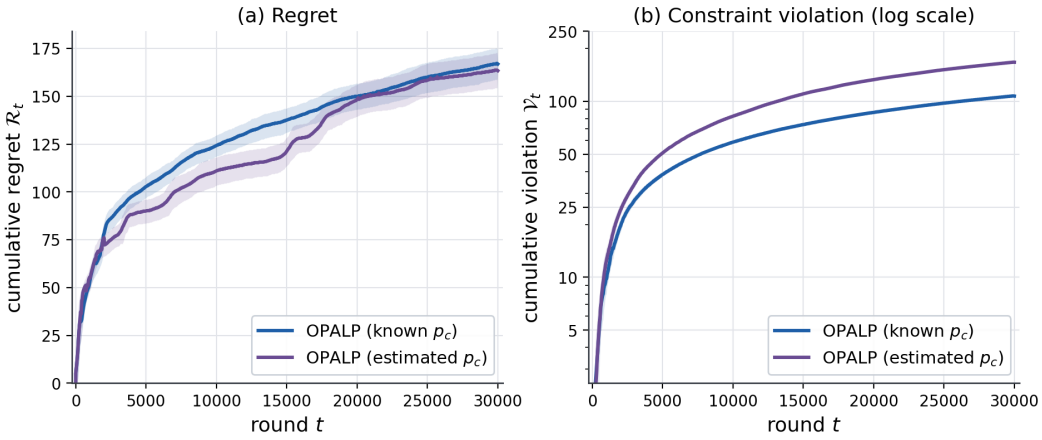


Figure 4: **Relaxing the known context distribution.** Cumulative regret (left) and constraint violation (right) for **OPALP** with known p_c versus a variant that estimates p_c directly through the weighted LP coefficients, on an instance with 10 contexts and 5 arms, averaged over 16 runs (shaded: ± 1 std). The regret curves coincide, while estimating p_c incurs a modestly larger, but identically scaling, constraint violation.

F.4 Computational Complexity

Computational burden. We empirically compare the computational complexity of **OPALP** with that of the recent method **Optimistic**^{3} of Guo et al. (2025). We fix the number of arms at $|\mathcal{A}| = 10$ and vary the number of contexts $|\mathcal{C}| \in \{10, 100, 500, 1000, 5000, 10000\}$, reporting the average per-round runtime, in seconds, in Table 2. As the decision space grows, the per-round LP solve makes **OPALP** significantly more expensive than the gradient-based updates of **Optimistic**^{3}, motivating the lazy variant evaluated next.

Algorithm \ $\lvert\mathcal{C}\rvert$	10	100	500	1000	5000	10000
OPALP	$5 \cdot 10^{-4}$	$1.3 \cdot 10^{-3}$	$1.1 \cdot 10^{-2}$	$2.5 \cdot 10^{-2}$	0.5	4
Optimistic^3	$5 \cdot 10^{-4}$	$7 \cdot 10^{-4}$	$3 \cdot 10^{-3}$	$8 \cdot 10^{-3}$	0.08	0.2

Table 2: Average per-round runtime (in seconds) for **OPALP** and **Optimistic**^{3} across instances of varying size.

Vanilla OPALP vs. Lazy-OPALP. As discussed in Section D, **Lazy-OPALP** offers an attractive alternative to **OPALP** for mitigating the computational burden: it re-solves the linear program only when the estimation errors have shrunk significantly, reducing the per-round cost while preserving performance. We compare the two algorithms on an instance with 20 contexts and 10 arms, using random Gaussian distributions and averaging over 20 runs. As shown in Figure 5, we observe no significant performance deterioration, consistent with the theoretical insights.

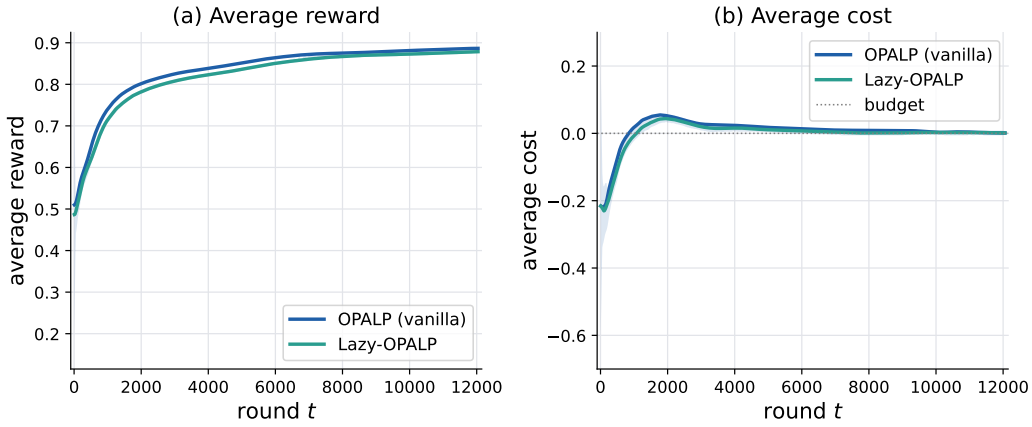


Figure 5: Average reward (a) and average cost (b) of **OPALP** and its lazy counterpart **Lazy-OPALP** (20 contexts, 10 arms; mean $\pm$ one standard deviation over 20 runs).

F.5 OPALP and Contextual Bandits with Knapsacks (CBwK)

Following Section 6, we instantiate **OPALP** in the CBwK setting as described in the main text — using a stopping time and incorporating the budget into the cost samples — with 20 arms and 10 contexts. We consider budgets $B \in \{300, 500, 700, 900\}$ with $T = 2000$ and average over 20 runs. As reported in Table 3, **OPALP** achieves final cumulative reward within 1–3% of that of the algorithm of Agrawal & Devanur (2016), which is specifically designed for the CBwK setting — despite **OPALP** being a general-constraint method not tailored to knapsacks.

Algorithm \ Budget	300	500	700	900
OPALP	298.2	484.7	660.3	838.9
Agrawal & Devanur (2016)	307.8	489.6	666.2	840.0

Table 3: Final cumulative reward of **OPALP** and Agrawal & Devanur (2016) under the vanilla CBwK setting for various initial budgets (average over 20 runs, $T = 2000$).