On Preference-based Stochastic Linear Contextual Bandits with Knapsacks

Xin Liu
ShanghaiTech University

Abstract

This paper studies the problem of preference-based stochastic linear contextual bandits with knapsack constraints (PbLinCBwK). We propose budget-aware optimistic and randomized exploration algorithms that achieve a regret of $O((\kappa + \frac{T\nu^*}{B})\sqrt{T}\log T)$, for any total budget $B = \Omega(\sqrt{T})$. The parameters κ and $\frac{T\nu^*}{B}$ capture the effects of preference feedback and knapsack constraints, respectively. Our regret performance is near-optimal and matches the bound of LinCBwK under the mild condition $B = \Omega(\sqrt{T})$. To achieve these results, we view the process of budget consumption and stopping time as Markov processes and analyze them via the Lyapunov drift method, which is then translated into the strong regret guarantee. The experiments on a synthetic PbBwK and a simulated online content moderation via PbLinCBwK justify the theoretical results.

1 Introduction

Stochastic linear contextual bandits is a powerful framework for decision-making under uncertainty and has been applied to a broad range of practical scenarios (e.g., news recommendation Chu et al. (2011), automated algorithm configuration Schede et al. (2023); Brandt et al. (2023), online content moderation Lykouris and Weng (2024); Lee et al. (2024), network resource allocation Li et al. (2020); Guo et al. (2023); Liu and Fang (2023), online advertisement platform Lucier et al. (2024); Feng et al. (2024); Gaitonde et al. (2022)), large language model (LLM) training Ziegler et al. (2019); Ouyang et al. (2022); Dong et al. (2024)).

Proceedings of the 28[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

In a standard stochastic contextual bandit, at the beginning of each round $t \in [T]$, the learner observes a context c_t, takes an action, and receives an absolute reward feedback. The goal of the learner is to maximize the cumulative rewards over the time horizon. However, the conventional formulation has two major limitations i) it assumes the explicit feedback of reward functions; and ii) it does not take any operational constraints into account. In practice, it is often challenging (if not impossible) to obtain explicit and absolute feedback on rewards. For example, in an online recommendation system, users' preferences are inherently subjective and difficult to quantify numerically; in online content moderation, the moderators evaluate the posts in binary formats ("healthy or not") rather than absolute scores; in automated algorithm configuration, preference feedback is used for better parameter searching; in LLM model training, reinforcement learning from human feedback (RLHF) is the key technique to align LLMs, where human feedback is again in a preference form rather than absolute scores. Besides the implicit reward feedback, operational constraints play a pivotal role in practical decision-making. For example, an online recommendation system may need to adhere to budgetary constraints when displaying advertisements; an online social medium targets for efficient content moderation with minimal human intervention; LLM training is certainly constrained by computational resources, where we need to carefully schedule resource budgets to acquire high-quality feedback for post-training.

To address these challenges, this paper studies preference-based stochastic linear contextual bandits with knapsack constraints (PbLinCBwK), where the learner aims to maximize the cumulative rewards with the preference feedback while satisfying the knapsack constraints. As in a standard stochastic linear bandit, at each round $t \in [T]$, the learner observes a context c_t that is randomly sampled from a context set $\mathcal{C}$ and takes two (duel) actions (x_t, y_t) from a decision set $\mathcal{A}$. Unlike the standard bandit setup, the learner cannot directly observe the reward values but receive the preference feedback, denoted as $o_t = \mathbb{I}(x_t \succ y_t)$, in-

dicating the preferred action. Additionally, the duel actions (x_t, y_t) incur resource consumption, and the total budget is limited to B. PbLinCBwK provides an enhanced framework that captures the realistic preferential feedback and resource consumption constraints.

Main Contributions

This paper presents budget-aware optimistic and randomized exploration algorithms. The algorithms efficiently leverage preference feedback to learn latent rewards while balancing the acquisition of rewards and resource consumption to maximize cumulative rewards. The design of our algorithms is motivated by the primal-dual approach in constrained optimization. The primal modular is similar to unconstrained stochastic linear dueling bandits but with novel *symmetric* structure in designing optimistic and randomized exploration. The dual modular includes a careful design of virtual queues that updates in a gradient-descent manner to keep track of cumulative over-consumed budget in each round and mimic the behavior of the optimal Lagrangian dual multiplier. The symmetric design in primal action modular and the budget-aware design in dual modular are crucial to establish the strong theoretical performance.

Specifically, both of our algorithms achieve a regret of $O((\kappa+\frac{T\nu^*}{B})\sqrt{T}\log T))$ for any budget $B = \Omega(\sqrt{T})$. To the best of our knowledge, this is the first theoretical result in PbLinCBwK and it has three implications: i) when $B = \Theta(T)$, our regret bound is near-optimal with respect to the time horizon T, given the lower bound of regret in unconstrained stochastic linear contextual bandits is $\Omega(\sqrt{T})$ Dani et al. (2008); ii) our algorithms achieve an almost identical regret as CBwK with absolute reward feedback for $B = \Omega(T^{3/4})$ in Agrawal and Devanur (2016) where $\kappa = \Theta(1)$, capturing the possible performance loss due to implicit preference feedback; iii) our result holds under the mild condition $B = \Omega(\sqrt{T})$, which extend the budget regime of $B = \Omega(T^{3/4})$ in Agrawal and Devanur (2016).

Related Work

Preference-based bandit learning is a specialized class of bandit problems also known as dueling bandits Yue et al. (2012). This class generalizes the classical multi-armed bandits (MAB) Slivkins (2019); Lattimore and Szepesvári (2020). The traditional exploration techniques in MAB, such as Upper Confidence Bound (UCB) Auer et al. (2003); Abbasi-yadkori et al. (2011), Thompson sampling Thompson (1933); Chapelle and Li (2011), and randomized exploration Kveton et al. (2020); Vaswani et al. (2020), have been instrumental in designing efficient algorithms for dueling bandits Zoghi et al. (2014); Wu and Liu (2016). The concept of dueling bandits has been further extended to contextual dueling bandits Dudík et al. (2015), including efficient algorithms proposed by Saha (2021); Bengs et al. (2022); Saha and Krishnamurthy (2022) and Di et al. (2024). Note incorporating human feedback in the training loop of large language models and robots have been achieving huge successes Christiano et al. (2017); MacGlashan et al. (2017); Ziegler et al. (2020); Stiennon et al. (2020); Bai et al. (2022), which significantly propelled the theoretical foundation of efficient learning from preference feedback Novoseller et al. (2020); Chen et al. (2022); Wu and Sun (2023); Saha et al. (2023); Wang et al. (2023); Zhu et al. (2023).

Another line of related work is on bandits with knapsacks constraints (BwK) Agrawal and Devanur (2014); Badanidiyuru et al. (2014, 2018); Sankararaman and Slivkins (2021), where the interaction terminates if the total budget is depleted. The algorithms designed for BwK are involved than their unconstrained counterpart because it is required a careful control on the budget consumption such that the cumulative rewards are maximized. Two prominent ideas are conservative exploration technique in Amani et al. (2019); Pacchiano et al. (2021) (resembling primal method in constrained optimization) and (zero-sum) game-theoretical design with two no-regret learners on rewards and costs, respectively, in Immorlica et al. (2019, 2022) (resembling primal-dual based method in constrained optimization). However, the former approach suffers from high computational complexity as it requires constructing safe policy sets in every round. The latter approach is efficient and motivates various work Sivakumar et al. (2022); Castiglioni et al. (2022); Slivkins et al. (2023, 2024). But it usually necessitates either a dedicated module to estimate the optimal cumulative rewards or knowledge of a feasibility/safe margin with respect to the underlying oracle problem.

It is worth mentioning that a recent work Deb et al. (2024) also considered preference-based bandit learning with knapsacks. This work studied (context-free) dueling bandits with knapsacks (a special setting of PbLinCBwK) and Borda regret; it developed an Exp3-type constrained exploration algorithm, which inherits an initial warm-up phase to estimate the optimal value as in Agrawal and Devanur (2016) and achieves $\tilde{O}(\frac{T\nu^*}{B}T^{2/3})$ when the budget regime $B = \Omega(T^{3/4})$. However, we studied PbLinCBwK with a parametrized Bradley and Terry preference model and proposed adaptive algorithms (without any warm-up stage) that achieve an improved regret $\tilde{O}((\kappa+\frac{T\nu^*}{B})\sqrt{T})$ for an extended budget regime $B = \Omega(\sqrt{T})$.

2 Preference-based Stochastic Linear Contextual Bandits with Knapsacks

We study a preference-based stochastic linear contextual bandit with knapsack constraints (PbLinCBwK) defined by $\{\mathcal{C}, \mathcal{A}, \mathcal{R}, \mathcal{O}, \mathcal{W}, B\}$, where $\mathcal{C}$ is the context set (a countable set), $\mathcal{A}$ is the action set (a finite set), $\mathcal{R} : \mathcal{C} \times \mathcal{A} \to \mathbb{R}^+$ is the reward function, $\mathcal{O} : \mathcal{C} \times \mathcal{A} \times \mathcal{A} \to \{0, 1\}$ is the comparison oracle to indicate the preferred action for a given context, $\mathcal{W} : \mathcal{C} \times \mathcal{A} \to \mathbb{R}^+$ is the cost function, and B is the initial total budget. At the beginning of every round $t \in [T]$, the learner observes a context c_t that is randomly generated from the context set $\mathcal{C}$ according to a probability distribution p_c. The learner then takes two (duel) actions $x_t, y_t \in \mathcal{A}$ according to a policy defined by $\pi : \mathcal{C} \to \mathcal{A} \times \mathcal{A}$. Instead of observing individual rewards, i.e., $\mathcal{R}(c_t, x_t)$ and $\mathcal{R}(c_t, y_t)$, the learner obtains the comparison/preference feedback denoted as $o_t = \mathcal{O}(c_t, x_t, y_t)$, where $o_t = 1$ indicates a preference for x_t over y_t, and $o_t = 0$ otherwise. In PbLinCBwK, in addition to receiving rewards, costs are incurred when the duel actions (x_t, y_t) are taken denoted by $\mathcal{W}(c_t, x_t)$ and $\mathcal{W}(c_t, y_t)$. Note we only consider a single type of resource consumption for simple exposition, and this paper can be easily extended to the case where the resources are multiple dimensions. As in Agrawal and Devanur (2014); Badanidiyuru et al. (2014), we assume a null action exists, and the learner can skip the round without incurring any rewards and costs. Furthermore, we assume that $\mathcal{R}(c, x)$ and $\mathcal{W}(c, x)$ are random variables with unknown expected values $r(c, x)$ and $w(c, x)$. The context probability distribution p_c and the comparison oracle $\mathcal{O}$ are also unknown aprior.

Preference reward model: we consider the classical Bradley and Terry (1952) (BT) model to compare duel actions (x_t, y_t) as follows

$$\mathbb{P}(x_t \succ y_t | c_t) = \frac{e^{r(c_t, x_t)}}{e^{r(c_t, x_t)} + e^{r(c_t, y_t)}}, \tag{1}$$

which is the probability that the action x_t is preferred over y_t. The preference feedback $o_t = \mathcal{O}(c_t, x_t, y_t)$ is a Bernoulli random variable draw from the probability distribution in (1). We make the following assumption about the reward and cost functions.

Assumption 1. *The reward function is $\mathcal{R}(c, x) = r(c, x) + \eta$, where $r(c, x) = \langle \phi(c, x), \theta_* \rangle$ is assumed a linear function; the cost function is $\mathcal{W}(c, x) = w(c, x) + \zeta$, where $w(c, x) = \langle \psi(c, x), \omega_* \rangle$ is also a linear function. Here, $\phi(c, x) \in \mathbb{R}^d$ and $\psi(c, x) \in \mathbb{R}^d$ represent the known reward and cost feature vectors for a given context-action pair, respectively. $\theta_* \in \mathbb{R}^d$ and $\omega_* \in \mathbb{R}^d$ are the unknown parameters. Additionally, η and ζ are zero-mean random variables.*

Denote the duel actions taken by a policy π in round t as (x_t^π, y_t^π). The learner's objective is to design a (dynamic) policy π that maximizes the cumulative rewards over horizon T under *the knapsack constraints* with *only preference feedback*:

$$\max_{\pi} \quad \sum_{t=1}^{\tau} \mathcal{R}(c_t, x_t^\pi) + \mathcal{R}(c_t, y_t^\pi) \tag{2}$$

$$\text{s.t.} \quad \sum_{t=1}^{\tau} \mathcal{W}(c_t, x_t^\pi) + \mathcal{W}(c_t, y_t^\pi) \le B \tag{3}$$

The term τ represents the stopping time when the budget is exhausted. Note we can also replace the stopping time τ with T because we can continue to play the null action until T after the budget is exhausted.

It could be quite challenging to solve (2)–(3) because of the implicit reward feedback and unknown reward, cost, and context distribution. The budget constraint further complicates the problem because all actions are coupled over the time horizon. Intuitively, solving (2)–(3) requires learning the context distribution, the rewards (inferred from preference feedback), and the costs. In the following sections, we present our algorithms to address the challenges and analyze their theoretical performance.

3 Algorithm Design

We introduce a general algorithmic framework to tackle the challenges posed by *implicit preference feedback*, *unknown* cost and context distribution, as well as the budget constraint. The framework is motivated by the primal-dual approach in constrained optimization.

We begin with an oracle/offline problem by assuming the full knowledge of (explicit) rewards, costs and context distribution. Define $r(c, x, y) := r(c, x) + r(c, y)$, $w(c, x, y) := w(c, x) + w(c, y)$ and $b := B/T$. The offline problem of (2)–(3) can be formulated as follows

$$\max_{\pi} \quad \sum_{c \in \mathcal{C}, x, y \in \mathcal{A}} p_c r(c, x, y) \pi(c, x, y) \tag{4}$$

$$\text{s.t.} \quad \sum_{c \in \mathcal{C}, x, y \in \mathcal{A}} p_c w(c, x, y) \pi(c, x, y) \le b, \tag{5}$$

$$\sum_{x, y \in \mathcal{A}} \pi(c, x, y) = 1, \ \pi(c, x, y) \ge 0, \forall c \in \mathcal{C}, \tag{6}$$

where $\pi(c, x, y)$ can be viewed as the probability of taking duel actions (x, y) on context c, and note p_c is the probability that context c is sampled in every round. The optimal value of the oracle problem serves as the upper bound of its online counterpart (2)–(3) in Lemma 2. Let Π be the probability simplex defined

by (6) and its (partial) Lagrangian formulation is

$$\max_{\pi\in\Pi} \sum_c p_c(\sum_{x,y} r(c,x,y)\pi(c,x,y) - \lambda(w(c,x,y)\pi(c,x,y) - b))$$

where λ is the Lagrange multiplier associated with the knapsack constraint in (5). Assuming the (optimal) Lagrange multiplier λ^* is given, solving the optimization problem is equivalent to solving the separated subproblem for each context c in (7), because the policy π are coupled through actions (x, y) only:

$$\max_{\pi\in\Pi} \sum_{x,y} (r(c,x,y) - \lambda^* w(c,x,y))\pi(c,x,y) \quad (7)$$

This decomposition eliminates the need to learn the context distribution. Since the problem in (7) is a linear programming, an optimal solution is simply a greedy decision, $\pi(c,x,y) = 1$ for (x^*, y^*) and $\pi(c,x,y) = 0$ otherwise:

$$(x^*, y^*) \in \operatorname*{argmax}_{x,y} \; r(c,x,y) - \lambda^* w(c,x,y) \quad (8)$$

and a tie can be broken arbitrarily. Therefore, the Lagrange multiplier λ^* plays a key role to balance the reward acquisition and budget consumption. Recall $r(c,x,y)$ and $w(c,x,y)$ are additive with respect to x and y, it is observed from (8) that one of the optimal solutions has the symmetric structure of $x^* = y^*$ such that $x^* \in \arg\max_x \; r(c,x) - \lambda^* w(c,x)$. This observation motivates the "symmetric" duel actions in our algorithms as introduced below.

However, several challenges still remain to find a solution to (8): i) $r(c,x,y)$ and $w(c,x,y)$ are unknown and they can only be learned from the preference feedback and bandit feedback, respectively, and ii) it is not clear how to balance the reward acquisition and budget consumption properly. To overcome the challenges, our framework includes two key components: i) optimistic and randomized learning with maximum likelihood estimation to explore $r(c,x,y)$ and $w(c,x,y)$ based on the preference and bandit feedback and their underlying parameterized structure; ii) dual gradient-type update to dynamically track the over-consumed budget (or the Lagrange multiplier):

$$Q_{t+1} = \max\left(Q_t + w(c_t, x_t, y_t) - b, 0\right). \quad (9)$$

We carefully design a budget-aware scaling parameter $1/V$ to Q_t, i.e. Q_t/V, to approximate λ^*. Since Q_t/V is adaptive to the budget consumption, we can control the tradeoff between reward acquisition and budget consumption. Intuitively, it implicitly learns the context distribution because the duel decision in (8) would avoid spending too much for the contexts with small rewards and large costs.

Next, we formally state two algorithms guided by the intuition discussed above. The two algorithms share several similarities (e.g., parameters estimation and covariance information update) and differ in the core parts of exploration and decision, which rely on optimistic and randomized exploration, respectively. The algorithms estimate parameters, perform dedicated exploration, output the duel actions (x_t, y_t), observe preference reward feedback o_t, bandit cost feedback, make updates, and then move to the next round $t+1$. With a bit abused notation, let $\phi^+(c_t,x,y) := \phi(c_t,x) + \phi(c_t,y)$ and $\phi^-(c_t,x,y) := \phi(c_t,x) - \phi(c_t,y)$. We next introduce the details of these two algorithms.

Algorithm 1 Optimistic Duel Exploration

Initialization: $\lambda = 1$, $Q_1 = 0$, and $V = b\sqrt{T}$.
for $t = 1, \cdots, T$, **do**
Parameters Estimation: Use (regularized) MLE to estimate the parameters in (10) and (11)

$$(\hat{\theta}_t, \hat{\omega}_t) = \text{MLE}(\{\mathcal{H}_t\}, \Sigma_t, \Psi_t).$$

Optimistic Exploration: For a given context c_t, compute the optimistic estimator for duel rewards and costs

$$\widetilde{\mathcal{R}}(c_t,x,y) = \langle\hat{\theta}_t, \phi^+(c_t,x,y)\rangle + b_r(c_t,x,y),$$
$$\widetilde{\mathcal{W}}(c_t,x,y) = \langle\hat{\omega}_t, \psi^+(c_t,x,y)\rangle - b_w(c_t,x,y).$$

Duel Actions: Take the duel action such that

$$(x_t, y_t) \in \operatorname*{argmax}_{x,y\in\mathcal{A}} \; \widetilde{\mathcal{R}}(c_t,x,y) - \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t,x,y)$$

Budget Pacing: Update Q_{t+1} as in (12).
Historical Data and Covariance Matrices: Include feedback and update preference and cost covariance matrices in (13).
end for

Parameters Estimation: The first step is to learn the reward parameters via a (regularized) maximum likelihood estimation (MLE) according to the historical preference data $\{o_s, x_s, y_s, \phi(c_s,x_s), \phi(c_s,y_s)\}_{t-1}$ at beginning of time t.

$$\hat{\theta}_t = \operatorname*{argmax}_{\theta\in\Theta} \sum_{s=1}^{t-1} \ln(o_s\sigma(\langle\theta, \phi^-(c_s,x_s,y_s)\rangle) + (1-o_s)\sigma(\langle\theta, \phi^-(c_s,y_s,x_s)\rangle)) - \frac{\lambda}{2}\|\theta\|^2 \quad (10)$$

where $\sigma(z) := 1/(1+e^{-z})$ corresponds to the BT model in (1). Similarly, we estimate the cost parameter according to the value feedback for costs via the

Algorithm 2 Randomized Duel Exploration

Initialization: $\lambda = 1$, $Q_1 = 0$, and $V = b\sqrt{T}$.
for $t = 1, \cdots, T$, **do**
Parameters Estimation: as in (10) and (11).
Randomized Exploration: generate randomized exploration for the rewards and costs

$$\theta_t^0, \theta_t^1 \sim \mathcal{N}(\hat{\theta}_t, \beta_t^r \Sigma_t^{-1}) \text{ and } \omega_t \sim \mathcal{N}(\hat{\omega}_t, \beta_t^w \Psi_t^{-1})$$

Duel Actions: Take the duel action such that

$$x_t \in \underset{x \in \mathcal{A}}{\operatorname{argmax}} \ \langle \theta_t^0, \phi(c_t, x) \rangle - \frac{Q_t}{V} \langle \omega_t, \psi(c_t, x) \rangle$$
$$y_t \in \underset{y \in \mathcal{A}}{\operatorname{argmax}} \ \langle \theta_t^1, \phi(c_t, y) \rangle - \frac{Q_t}{V} \langle \omega_t, \psi(c_t, y) \rangle$$

Budget Pacing: Update Q_{t+1} as in (12).
Covariance Matrices Update: as in (13).
end for

regularized least-squares (MLE) estimator

$$\hat{\omega}_t = \underset{\omega \in \Omega}{\operatorname{argmin}} \sum_{s=1}^{t-1} (\langle \omega, \psi(c_s, x_s) \rangle - \mathcal{W}(c_s, x_s))^2 + (\langle \omega, \psi(c_s, y_s) \rangle - \mathcal{W}(c_s, y_s))^2 + \frac{\lambda}{2} \|\omega\|^2. \tag{11}$$

The estimators $\hat{\theta}_t$ and $\hat{\omega}_t$ become increasingly accurate as the learner collects more data, where the true parameters are within the confidence sets by properly setting the confidence radius. We have the following guarantee for MLE estimators $\hat{\theta}_t$ and $\hat{\omega}_t$, which are from Faury et al. (2020) and Abbasi-yadkori et al. (2011), respectively. The detailed parameters are in Appendix F.1.

Lemma 1. *Let $p \in (0, 1)$, $\beta_t^r(p)$ and $\beta_t^w(p)$ be the proper radius. Recall θ^* and ω^* be the true reward and cost parameters, the maximum likelihood estimators $\hat{\theta}_t$ and $\hat{\omega}_t$ satisfy for any $t \in [T]$ that*

$$\mathbb{P}(\|\hat{\theta}_t - \theta^*\|_{\Sigma_t} \leq \beta_t^r(p)) \geq 1 - p,$$
$$\mathbb{P}(\|\hat{\omega}_t^r - \omega^*\|_{\Psi_t} \leq \beta_t^w(p)) \geq 1 - p.$$

Based on Lemma 1, we design dedicated learning techniques to explore the duel actions space while utilizing the budget effectively guided by the core of decision making in (8). In this paper, we develop two algorithms based on the ideas of optimistic in the face of uncertainty Auer et al. (2002); Abbasi-yadkori et al. (2011) and randomized exploration Kveton et al. (2020); Vaswani et al. (2020).

Double Optimistic Exploration for Duel Action in Algorithm 1: For a given context c_t, we use $\langle \hat{\theta}_t, \phi^+(c_t, x, y) \rangle$ to approximate $r(c_t, x, y)$ and design the optimistic bonus $b_r(c_t, x, y) = \beta_t^r \|\phi^-(c_t, x, y)\|_{\Sigma_t^{-1}}$ to encourage the exploration of duel actions such that their preference difference is distinct, motivated by Di et al. (2024). Since the costs are also unknown, we use $\langle \hat{\omega}_t, \psi(c_t, x, y) \rangle$ to approximate $w(c_t, x, y)$ and design the optimistic bonus $b_w(c_t, x, y) = \beta_t^w (\|\psi(c_t, x)\|_{\Psi_t^{-1}} + \|\psi(c_t, y)\|_{\Psi_t^{-1}})$ to encourage the exploration of duel actions for budget consumption. Note the bonus terms for the rewards and costs are designed distinctly due to the different types of feedback, where we only observe the preference feedback for the rewards, but the value feedback for the costs. The factors β_t^r and β_t^w are the radius of bonus terms to control the degree of the optimism, consistent with the values in Lemma 1.

Motivated by (8), we carefully incorporate the costs induced by the duel actions such that the budget is utilized most effectively and the cumulative rewards are maximized. The key factor Q_t/V with the budget-aware quantity $V = b\sqrt{T}$ is to learn the optimal Lagrange multiplier of the baseline problem without any prior information (e.g., the distribution of contexts, rewards and costs). The duel decision in Algorithm 1 resembles (8) where we use the stochastic values to replace the accurate mean values. We call it "double optimistic exploration" because both learning and decision components favor optimistic outcomes.

Symmetric Randomized Exploration for Duel Action in Algorithm 2: For a given context c_t, we generate θ_t^0 and θ_t^1 randomly and independently around MLE estimator $\hat{\theta}_t$ to explore the reward space; generate ω_t randomly around $\hat{\omega}_t$ to explore the cost space. Algorithm 2 uses randomness to do exploration instead of imposing the optimistic bonus terms. Note the randomness terms for the rewards and costs are also designed distinctly due to the similar reason (different types of feedback for rewards and costs) as in the optimistic design. In particular, we generate "two (uncoupled) samples" (θ_t^0, θ_t^1) to explore the reward space with the preference feedback and "single (coupled) sample" (ω_t) for to learn the costs with the value feedback. The parameters $\theta_t^0, \theta_t^1 \in \mathcal{N}(\theta_t, \beta_t^r \Sigma_t^{-1})$ and $\omega_t \sim \mathcal{N}(\hat{\omega}_t, \beta_t^w \Psi_t^{-1})$ are draw from Gaussian distribution and β_t^r (β_t^w) play similar roles with the confidence radius in the optimistic exploration.

We utilize $\widetilde{\mathcal{R}}(c_t, x_t, y_t) = \langle \theta_t^0, \phi(c_t, x) \rangle + \langle \theta_t^1, \phi(c_t, y) \rangle$ to approximate $r(c_t, x, y)$ and $\widetilde{\mathcal{W}}(c_t, x_t, y_t) = \langle \omega_t, \psi(c_t, x) + \psi(c_t, y) \rangle$ to approximate $w(c_t, x, y)$, respectively. Similarly motivated by (8), we choose the duel actions (x, y) to maximize the index function

$$\langle \theta_t^0, \phi(c_t, x) \rangle + \langle \theta_t^1, \phi(c_t, y) \rangle - \frac{Q_t}{V} \langle \omega_t, \psi(c_t, x) + \psi(c_t, y) \rangle.$$

Fortunately, it is observed that the duel actions (x, y) are *symmetric* due to the decoupled randomized explo-

ration and we could take actions x_t and y_t individually as in Algorithm 2. This decoupled exploration is essential and crucial; without it, the duel actions (x_t, y_t) might stuck into the same one, i.e., $\theta_t^0 = \theta_t^1$.

Budget Pacing: Motivated by (9), we design the virtual queue that indicates how the budget is overused (per round)

$$Q_{t+1} = \max\left(Q_t + \widetilde{\mathcal{W}}(c_t, x_t, y_t) - b, 0\right), \tag{12}$$

where we use the estimator of costs $\widetilde{\mathcal{W}}(c_t, x_t, y_t)$ to replace $w(c_t, x_t, y_t)$ such that it is consistent with what used in the part of duel actions.

Historical Data and Covariance Matrices: Once the feedback $\{o_t, \mathcal{W}(c_t, x_t), \mathcal{W}(c_t, y_t)\}$ is observed, we include them in the dataset $\mathcal{H}_{t+1}$ and update the preference and cost covariance matrices

$$\begin{aligned}
\mathcal{H}_{t+1} &= \{\mathcal{H}_t, o_t, \mathcal{W}(c_t, x_t), \mathcal{W}(c_t, y_t)\},\\
\Sigma_{t+1} &= \Sigma_t + \phi^-(c_t, x_t, y_t)\phi^-(c_t, x_t, y_t)^\dagger,\\
\Psi_{t+1} &= \Psi_t + \psi(c_t, x_t)\psi(c_t, x_t)^\dagger + \psi(c_t, y_t)\psi(c_t, y_t)^\dagger.
\end{aligned} \tag{13}$$

The covariance matrices and $\mathcal{H}_{t+1}$ are used to estimate the rewards (costs) and optimize the duel decision.

We summarize optimistic and randomized exploration algorithms for PbLinCBwK in Algorithms 1 and 2 and analyze the theoretical performance as follows.

4 Main Results

We begin by defining the baseline and stating the necessary assumptions for the theoretical analysis.

Assumption 2. *The context c_t are i.i.d. across rounds. The mean reward $r(c, j) = \langle \theta^*, \phi(c, x)\rangle \in [0, 1]$. The mean cost $w(c, j) = \langle \omega^*, \psi(c, x)\rangle \in [0, 1]$; $||\psi(c, x)|| \le 1, ||\phi(c, x)|| \le 1, \forall c, x$; $||\omega^*|| \le d||\theta^*|| \le d$. The noise η_t and ζ_t are zero-mean 1-subGaussian random variables conditioned on $\{\mathcal{H}_{t-1}, x_t, y_t\}$.*

Assumption 3. *There exists a constant $\delta \in (0, 1)$ and a feasible solution π to the offline problem (4)–(6) satisfies $\sum_{c\in\mathcal{C}, x,y\in\mathcal{A}} p_c w(c, x, y)\pi(c, x, y) \le b(1 - \delta)$.*

The term δb plays a similar role with the Slater's constant in optimization. However, the term δb differs from the traditional Slater's constant because $b = B/T$ implies it is $T-$dependent. The assumption is mild and common in the literature of BwK Agrawal and Devanur (2014, 2016); Badanidiyuru et al. (2014, 2018).

The next lemma shows that the optimal value of (4)–(6) is an upper bound on the expected optimal value of (2)–(3). The proof can be found in Appendix A.

Lemma 2. *Under Assumptions 2–3, let v^* be the optimal value of the offline problem (4)–(6) and π^* is the optimal policy of the problem (2)–(3). We have*

$$\mathbb{E}[\sum_{t=1}^{T}\sum_{x,y}\mathcal{R}(c_t, x, y)\pi^*(c_t, x, y)] \le Tv^*.$$

With the baseline above, now we can define (pseudo)-regret. For a policy π, we define its (pseudo)-regret

$$\text{Regret}(T) = Tv^* - \mathbb{E}[\sum_{t=1}^{\tau}\mathcal{R}(c_t, x_t^\pi, y_t^\pi)]. \tag{14}$$

The stopping time τ in (14) is a random variable determined by our policy π. The expectation is taken w.r.t. randomness from the policy π and the environment. In the next sections, we analyze the regret and occasionally omit the superscript of $[\cdot]^\pi$ when it does not cause any confusion.

4.1 Regret Analysis

We present the theoretical results for Algorithms 1 and 2 in an order-wise sense. The detailed proof and parameters can be found in Appendices D and E.

Theorem 1. *Let $\kappa = \sup_{\theta^*,\phi} 1/\dot{\sigma}(\langle\theta, \phi\rangle)$. Under Assumptions 2–3, the double optimistic/symmetric randomized exploration for PbLinCBwK in Algorithms 1 and 2 achieve for any budget $B = \Omega(\sqrt{T})$ that*

$$\text{Regret}(T) = O((\kappa + \frac{\nu^*}{\delta b})\sqrt{T}\log T).$$

The parameters κ and $\nu^*/\delta b$ (or $T\nu^*/\delta B$) in the regret capture the effects of preference feedback and knapsack constraints, respectively, where κ could be $O(e^d)$ indicating the challenge of implicit preference feedback (note $\kappa = \Theta(1)$ when the reward is linear and observed in absolute value); To the best of our knowledge, this is the first theoretical result for PbLinCBwK. When the absolute reward feedback is available, i.e., $\kappa = \Theta(1)$, our results recover the same regret as the traditional CBwK in Agrawal and Devanur (2016) with $B = \Omega(T^{3/4})$ and extend the budget regime to $B = \Omega(\sqrt{T})$. Our regret performance is near-optimal with respect to the time horizon T when the budget $B = \Theta(T)$, given the lower bound of regret for unconstrained stochastic contextual linear bandits is $\Omega(\sqrt{T})$ Dani et al. (2008). This result improves the regret performance of $O(\frac{\nu^*}{\delta b}T^{2/3})$ in Deb et al. (2024) and implies a parameterized preference model is beneficial to achieve a better regret even in the challenging contextual setting. In particular, Deb et al. (2024) proposed Constrained D-EXP3 algorithm and proved it achieves $\tilde{O}(\frac{OPT}{B} \times T^{2/3})$ when

the budget $B = \Omega(T^{3/4})$. Their algorithm was motivated by (Agrawal and Devanur, 2016) and inherited an initial warm-up phase to estimate the optimal value and the budget assumption $B = \Omega(T^{3/4})$. Our algorithms with optimistic/randomized exploration achieve $\tilde{O}((\kappa + \frac{OPT}{B})\sqrt{T})$ when the budget $B = \Omega(\sqrt{T})$. Our theoretical results show the improvement from $O(\frac{OPT}{B} \times T^{2/3})$ to $O(\frac{OPT}{B} \times \sqrt{T})$ with an additional term $\kappa\sqrt{T}$, which are due to the preference assumption of Plackett-Luce model. It is worth emphasizing that our algorithms do not need a warm-up phase and hold under relaxed assumption $B = \Omega(\sqrt{T})$ thanks to a novel and refined analysis.

5 Proof of Theorem 1

Now we study the regret defined in (14) and decompose it according to the stopping time τ.

5.1 Regret Decomposition

Let π^* be the optimal solution to the baseline problem (4)–(6) and $(x^*, y^*) \sim \pi^*$ be the optimal duel actions sampling from it. We decompose the regret as follows

$$\begin{aligned} &\text{Regret}(T) \qquad (15)\\ &\leq 2\nu^* \mathbb{E}[T - \tau] + \mathbb{E}[\sum_{t=1}^{\tau} \mathcal{R}(c_t, x^*, y^*) - \mathcal{R}(c_t, x_t, y_t)] \end{aligned}$$

The decomposition includes two parts: "regret after stopping" and "regret before stopping". The former one is simply bounded by the remaining rounds $\times$ the optimal value ν^*; the latter one denoted by $\mathbb{E}[\text{Regret}(\tau)]$ could be further decomposed as follows (we use θ_t^0 and θ_t^1 because it is general to include both optimistic and randomized exploration where $\theta_t^0 = \theta_t^1 := \hat{\theta}_t$ in the optimistic algorithm):

$$\begin{aligned} &\mathbb{E}[\text{Regret}(\tau)]\\ &= \mathbb{E}[\sum_{t=1}^{\tau} \langle \phi^-(c_t, x^*, x_t), \theta^* \rangle + \langle \phi^-(c_t, y^*, y_t), \theta^* \rangle]\\ &= \mathbb{E}[\sum_{t=1}^{\tau} \langle \phi^-(c_t, x^*, y_t), \theta^* - \theta_t^0 \rangle + \langle \phi^-(c_t, x^*, x_t), \theta_t^0 \rangle\\ &\qquad + \langle \phi^-(c_t, y^*, x_t), \theta^* - \theta_t^1 \rangle + \langle \phi^-(c_t, y^*, y_t), \theta_t^1 \rangle\\ &\qquad + \langle \phi^-(c_t, x_t, y_t), \theta_t^0 - \theta_t^1 \rangle]\\ &= \text{Regret}(c_t, x^*) + \text{Regret}(c_t, y^*) + \mathcal{R}(c_t, \theta_t^0, \theta_t^1) \end{aligned}$$

where the first equality holds by substituting the definition of the linear reward and $\phi^-(c_t, x, y) = \phi(c_t, x) - \phi(c_t, y)$; the second equality holds by adding and subtracting corresponding terms associated with the estimated parameters θ_t^0 and θ_t^1. The regret has been decomposed into three major items: the first term is regarded as the regret induced by the duel actions (x_t, y_t) against x^* and the second term is that against y^*; the last term is on the randomized errors of θ_t^0 and θ_t^1, which is zero in optimistic exploration and is bounded in randomized exploration as follows.

Lemma 3. *Under Algorithms 1 and 2, let $p \in (0, 1)$, the ineqiality holds with a probability at least $1 - 2p$*

$$\sum_{t=1}^{\tau} \mathcal{R}(c_t, \theta_t^0, \theta_t^1) \leq \sum_{t=1}^{\tau} 4\beta_t^r \|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}}.$$

The Lemma 3 can be proved by Cauchy–Schwarz inequality and the MLE errors in Lemma 1. The details can be found in Appendix B.

5.2 Regret via Lyapunov Drift Analysis

As discussed, we provide a new perspective on analyzing the regret via Lyapunov drift analysis, which can be used to bound both "regret before stopping" and "regret after stopping". Let $L_t = Q_t^2/2$ be the Lyapunov function and $\Delta_t = L_{t+1} - L_t$ be its drift. Further let $\mathbb{E}_{\mathcal{H}_t}[\cdot] = \mathbb{E}[\cdot|\mathcal{H}_t]$ and $\text{Regret}(c_t, x, y) = \text{Regret}(c_t, x) + \text{Regret}(c_t, y)$ for a simple notation. We establish the following key lemma that bridges the one-step regret and Lyapunov drift.

Lemma 4. *Under Algorithms 1 and 2, there exists an absolute constant C_0 such that we have for any feasible policy π to (4)–(6) with $(x, y) \sim \pi$ that*

$$\begin{aligned} &\mathbb{E}_{\mathcal{H}_t}[\text{Regret}(c_t, x, y) + \frac{\Delta_t}{V} - \frac{Q_t}{V}(\mathcal{W}(c_t, x, y) - b)]\\ &\leq C_0 \mathbb{E}_{\mathcal{H}_t}[\beta_t^r \|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}} + \frac{(\widetilde{\mathcal{W}}(c_t, x_t, y_t) - b)^2}{V}]. \end{aligned}$$

Next, we analyze "regret before/after stopping".

5.3 Regret Before Stopping

Letting $(x, y) = (x^*, y^*)$ in Lemma 4, the inequality suggests that "one-step regret + Lyapunov drift" is upper bounded by three related terms: the optimal budget consumption $(\mathcal{W}(c_t, x^*, y^*) - b)$, the exploration bonus $\beta_t^r \|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}}$, the estimated consumption variance $(\widetilde{\mathcal{W}}(c_t, x_t, y_t) - b)^2$, where bounding $\mathbb{E}[\text{Regret}(\tau)]$ requires establishing their cumulative counterparts: the expected optimal consumption is always negative for every round; the cumulative exploration bonus is bounded by $O(\kappa\sqrt{T}\log T)$ with the elliptical potential lemma; most importantly, we provide a refined analysis on bounding the cumulative consumption variance by $O((Tb + Tb^2)/V)$, which is one of key component in proving the strong results in Theorem 1. The result of "regret before stopping" is summarized in the following lemma and the detailed proof is in Appendices D.2 and E.2.

Lemma 5. *Under Assumptions 2 and 3, Algorithms 1 and 2 achieve*

$$\mathbb{E}\left[\text{Regret}(\tau)\right] = O(\kappa\sqrt{T}\log T + (Tb + Tb^2)/V).$$

5.4 Regret After Stopping

In the regret decomposition, the "regret after stopping" is bounded by $2\nu^*\mathbb{E}[T-\tau]$. To minimize this regret, it is crucial to establish a "large" lower bound on the stopping time τ, ideally depleting the budget only when it is very close to T. The stopping time is the first time when the total budget is exhausted

$$\tau = \underset{\tau' \in [T]}{\text{argmax}}\{\tau' \mid \sum_{t=1}^{\tau'} \mathcal{W}(c_t, x_t, y_t) \geq B\}. \tag{16}$$

To (lower) bound the stopping time τ, we need to analyze the behavior of virtual queue because it captures the over-consumed budget against the average usage $t \times b$ for the round t. Recall the virtual queue update $Q_{t+1} = \max(Q_t + \widetilde{\mathcal{W}}(c_t, x_t, y_t) - b, 0)$. Let $M_\tau = \sum_{t=1}^{\tau}(\mathcal{W}(c_t, x_t, y_t) - \widetilde{\mathcal{W}}(c_t, x_t, y_t))$, we take the summation from $t = 1$ to τ

$$Q_{\tau+1} + b\tau + M_\tau \geq \sum_{t=1}^{\tau} \mathcal{W}(c_t, x_t, y_t). \tag{17}$$

Moreover, we show that M_τ can be bound by $\tilde{O}(\sqrt{\tau})$ with a high probability by using the martingale argument in Lemma 8 in Appendix C. Define a virtual stopping time such that $\tau_0 = \text{argmin}_{\tau' \in [T]}\{\tau' \mid Q_{\tau'+1} + b\tau' + \tilde{O}(\sqrt{\tau}) \geq B\}$, which is the first time when the upper bound of budget consumption in (17) are greater than B at τ_0. According to (16) and (17), we immediately have the true stopping time is lower bounded by the virtual stopping time lower bounded $\tau \geq \tau_0$. Now, we translate the lower bound on τ_0 into the upper bound of the virtual queue via Lyapunov drift analysis.

Lyapunov drift analysis for establishing τ_0: As discussed, we view $\{Q_t\}$ as an stochastic/Markovian process and study its convergence or upper bound via Lyapunov/potential analysis. From Lemma 4, we establish a "negative drift" of Lyapunov function, implying a high probability upper bound of virtual queue.

Lemma 6. *Under Algorithms 1 and 2, there exists a positive constant C_1 so that the Lyapunov drift satisfies*

$$\mathbb{E}\left[\Delta_t | \mathcal{H}_t = h\right] \leq -\delta b \cdot Q_t + 0.5 C_1 \beta_t^r V, \tag{18}$$

and the virtual queue satisfies

$$\mathbb{P}(Q_t \leq \frac{C_1 \beta_t^r V}{\delta b}) \geq 1 - \frac{1}{T^2}, \quad \forall t \in [T]. \tag{19}$$

Intuitively, if the virtual queue process $\{Q_t\}$ already reaches the steady state, i.e., the drift is zero $\mathbb{E}\left[\Delta_t | \mathcal{H}_t = h\right] = 0$. We immediately establish its upper bound of $O(C_1\beta_t V/\delta b)$ from (18). This intuition is formally justified by (19). Now we are ready to establish the upper bound on the remaining rounds.

Lemma 7. *Under Algorithms 1 and 2, we have*

$$T - \tau_0 \leq \frac{C_1 \beta_T^r V}{\delta b^2} + \frac{8 d \beta_T^w \sqrt{\tau \log(1+\tau)}}{b},$$

hold with a high probability $1 - 2/T^2$.

Proving Theorem 1: Combine Lemmas 5 and 7 into the regret decomposition (15), we immediately have

$$\begin{aligned}\text{Regret}(T) &\leq 2\nu^*\mathbb{E}[T-\tau] + \mathbb{E}[\text{Regret}(\tau)] \\ &= \tilde{O}(\kappa\sqrt{T} + \frac{Tb + Tb^2}{V} + \frac{V\nu^*}{\delta b^2} + \frac{\sqrt{T}\nu^*}{b})\end{aligned}$$

which proves Theorem 1 by letting $V = b\sqrt{T}$.

6 Experiments

In this section, we run two sets of experiments: 1) a synthetic scenario for preference-based stochastic MAB with knapsacks (PbBwK) and stochastic linear contextual bandits with knapsacks (PbLinCBwK) for online content moderation, respectively.

Synthetic PbBwK: In the synthetic PbBwK setting, we consider "CD-EXP3" algorithm Deb et al. (2024) as the baseline because it is the only related work in the context-free BwK setting. Besides, we construct an ideal algorithm that integrates the absolute reward value feedback into Algorithm 1 (instead of preference feedback), called "BwK-Ideal". This algorithm is the performance upper bound because it has more informative feedback. The synthetic PbBwK setting includes four arms with the expected rewards $r = [0.1, 0.2, 0.4, 0.7]$ and expected costs $w = [0.05, 0.4, 0.5, 0.7]$, respectively. At each round, the learner takes duel actions $x_t, y_t \in \{0, 1, 2, 3\}$. The learner observes the preference feedback on rewards and the (noisy) costs. The algorithms were tested with different budgets of $B = [300, 500, 700, 900]$ with $T = 2000$. The following results show the average cumulative reward values (over 200 trials) for each budget level.

Table 1 indicates that Algorithms 1 and 2 consistently outperform "CD-EXP3" across all budget levels. The "BwK-Ideal" indeed serves the upper bound for the setting due to its informative reward value feedback. Besides, we observe that Algorithm 2 with randomized exploration slightly outperformed Algorithm 1 with optimistic exploration. The possible reason could be

Algorithm/Budget	300	500	700	900
BwK-Ideal	336.0	514.7	696.2	879.1
Algorithm 1	294.2	471.0	643.9	813.6
Algorithm 2	305.0	479.2	647.2	831.2
CD-EXP3	253.7	421.6	592.1	763.6

Table 1: Cumulative rewards performance comparison for PbBwK with various budgets.

that the randomized exploration is more effective in managing the exploration-exploitation trade-off and avoids overly optimistic exploration.

Online Content Moderation: we study an online content moderation platform where the platform requires human (expert) feedback to promote the best and most appropriate reviews for displayed items. For each item, the platform decides the best two representative reviews according to its features and seeks human judgment through preference feedback. The human expert reads the two reviews and indicates their preferred review based on the underlying reward values. This system is a perfect fit for PbLinCBwK since the platform can observe the preference feedback, and the human expert must spend time and effort evaluating the reviews to make a judgment.

We study online content moderation for vehicle reviews using the car evaluation dataset from the UCI machine learning dataset. This dataset contains car reviews that are classified into one of four categories: {"acceptable", "good", "unacceptable" or "very good"}. Each review is summarized by six important features, which are {"buying", "maint", "doors", "persons", "lug boot" and "safety"}. To emulate a human expert, denoted as $\mathcal{O}$, we trained a dataset using logistic regression with the logistic function $p(c, a, \theta^*) = 1/(1 + e^{-\langle \phi(c,a), \theta^* \rangle})$. This enabled us to obtain the ground truth weight θ^*. The reward function $r(c, a) = \langle \phi(c, a), \theta^* \rangle$ represents the underlying human value on whether the review/action (a) matches the car data (c). For instance, if a car c_t is in "very good" condition, the human expert $\mathcal{O}$ would prefer the review of "very good" over "acceptable".

To simulate the human/labor costs, we assume that the costs are proportional to their underlying rewards $w(c, a) = v(a) \times r(c, a)$, where the cost factor $v(a)$ is drawn from the set of {acceptable, good, unacceptable, very good} with the probability {0.3, 0.3, 0.5, 0.5}. We have chosen these factors to simulate a scenario where giving a medium/conservative rating is relatively easier than providing a high/low rating, as these two cases require more justification.

Specifically, our experiment works as follows: at each round t, the platform randomly selects a context from the car dataset and chooses two reviews to be moderated by a human expert. The expert reads the reviews and provides feedback on their preference or comparison, which is denoted as $o_t = \mathcal{O}(c_t, x_t, y_t)$. In addition to this feedback, the platform also observes the labor costs, which are sometimes noisy, for evaluating the two reviews. The costs are represented by $\mathcal{W}(c_t, x_t)$ and $\mathcal{W}(c_t, y_t)$. Similarly, we test our algorithms with various budgets $B = [300, 500, 700, 900]$ over a time horizon of $T = 2000$. The results are the average cumulative reward values over 200 trials.

Budget B	300	500	700	900
LinCBwK-Ideal	1107.9	1542.1	2083.1	2380.9
Algorithm 1	980.4	1272.5	1941.6	2328.0
Algorithm 2	996.3	1474.0	1858.5	2203.1
CD-EXP3	406.3	844.2	1121.4	1486.3

Table 2: Reward performance comparison for online content moderation with various budgets.

The results in Table 2 demonstrate that our proposed Algorithms 1 and 2 consistently outperform the baseline "CD-EXP3" by a significant margin across all budgets. We believe that the reason for this is "CD-EXP3" does not take into account the context information, which is critical to boosting reward performance. Once again, the "LinCBwK-Ideal" algorithm serves as the upper bound for the setting due to its informative reward value feedback. Moreover, both Algorithms 1 and 2 achieve large rewards that are close to the ideal upper limit. Interestingly, we observed that the randomized exploration in Algorithm 2 outperforms the optimistic exploration in Algorithm 1 when the budget is small (300 and 500) and achieves slightly worse rewards when the budget becomes large (700 and 900). This finding might suggest that an algorithm with initial bold exploration could perform better with a sufficient budget.

7 Conclusions

In this paper, we present two budget-aware symmetric exploration algorithms for the PbLinCBwK problem, which achieve near-optimal regret performance. Our theoretical results build on a unique perspective on analyzing budget consumption using Lyapunov drift methods, along with a refined analysis of cumulative budget consumption variance. The experiments justify our algorithms outperform the baselines.

ACKNOWLEDGMENTS

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

REFERENCE

Abbasi-yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*.

Agrawal, S. and Devanur, N. (2016). Linear contextual bandits with knapsacks. In *Advances Neural Information Processing Systems (NeurIPS)*.

Agrawal, S. and Devanur, N. R. (2014). Bandits with concave rewards and convex knapsacks. In *Proceedings of the Fifteenth ACM Conference on Economics and Computation*. Association for Computing Machinery.

Amani, S., Alizadeh, M., and Thrampoulidis, C. (2019). Linear stochastic bandits under safety constraints. In *Advances Neural Information Processing Systems (NeurIPS)*, pages 9256–9266.

Auer, P., Cesa-Bianchi, N., and Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2–3):235–256.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. (2003). The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77.

Badanidiyuru, A., Kleinberg, R., and Slivkins, A. (2018). Bandits with knapsacks. *J. ACM*.

Badanidiyuru, A., Langford, J., and Slivkins, A. (2014). Resourceful contextual bandits. In *Proc. Conf. Learning Theory (COLT)*.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.

Bengs, V., Saha, A., and Hüllermeier, E. (2022). Stochastic contextual dueling bandits under linear stochastic transitivity models. In *Proceedings of the 39th International Conference on Machine Learning*, pages 1764–1786.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*.

Brandt, J., Schede, E., Haddenhorst, B., Bengs, V., Hullermeier, E., and Tierney, K. (2023). Ac-band: a combinatorial bandit-based approach to algorithm configuration. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence*.

Castiglioni, M., Celli, A., and Kroer, C. (2022). Online learning with knapsacks: the best of both worlds. In *Proceedings of the 39th International Conference on Machine Learning*, pages 2767–2783.

Cesa-Bianchi, N. and Lugosi, G. (2006). *Prediction, Learning, and Games*. Cambridge University Press.

Chapelle, O. and Li, L. (2011). An empirical evaluation of thompson sampling. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*.

Chen, X., Zhong, H., Yang, Z., Wang, Z., and Wang, L. (2022). Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *Proceedings of the 39th International Conference on Machine Learning*.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. (2017). Deep reinforcement learning from human preferences. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*.

Chu, W., Li, L., Reyzin, L., and Schapire, R. (2011). Contextual bandits with linear payoff functions. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*.

Dani, V., Hayes, T., and Kakade, S. M. (2008). Stochastic linear optimization under bandit feedback. *21st Annual Conference on Learning Theory - COLT 2008, Helsinki, Finland*.

Deb, R., Saha, A., and Banerjee, A. (2024). Think before you duel: Understanding complexities of preference learning under constrained resources. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*.

Di, Q., Jin, T., Wu, Y., Zhao, H., Farnoud, F., and Gu, Q. (2024). Variance-aware regret bounds for stochastic contextual dueling bandits. In *The Twelfth International Conference on Learning Representations*.

Dong, H., Xiong, W., Pang, B., Wang, H., Zhao, H., Zhou, Y., Jiang, N., Sahoo, D., Xiong, C., and Zhang, T. (2024). RLHF workflow: From reward modeling to online RLHF. *Transactions on Machine Learning Research*.

Dudík, M., Hofmann, K., Schapire, R. E., Slivkins, A., and Zoghi, M. (2015). Contextual dueling bandits. In *Proceedings of The 28th Conference on Learning Theory*.

Faury, L., Abeille, M., Calauzenes, C., and Fercoq, O. (2020). Improved optimistic algorithms for logistic bandits. In III, H. D. and Singh, A., editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3052–3060.

Feng, Y., Lucier, B., and Slivkins, A. (2024). Strategic budget selection in a competitive autobidding world. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, pages 213–224.

Gaitonde, J., Li, Y., Light, B., Lucier, B., and Slivkins, A. (2022). Budget pacing in repeated auctions: Regret and efficiency without convergence. *arXiv preprint arXiv:2205.08674*.

Guo, H., Cao, H., He, J., Liu, X., and Shi, Y. (2023). Pobo: Safe and optimal resource management for cloud microservices. *Performance Evaluation*, 162:102376.

Hajek, B. (1982). Hitting-time and occupation-time bounds implied by drift analysis with applications. *Ann. Appl. Prob.*, pages 502–525.

Immorlica, N., Sankararaman, K., Schapire, R., and Slivkins, A. (2019). Adversarial bandits with knapsacks. *Proc. Ann. IEEE Symp. Found. Comput. Sci.*

Immorlica, N., Sankararaman, K., Schapire, R., and Slivkins, A. (2022). Adversarial bandits with knapsacks. *Journal of the ACM*, 69(6):1–47.

Kveton, B., Zaheer, M., Szepesvari, C., Li, L., Ghavamzadeh, M., and Boutilier, C. (2020). Randomized exploration in generalized linear bandits. In Chiappa, S. and Calandra, R., editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2066–2076.

Lattimore, T. and Szepesvári, C. (2020). *Bandit Algorithms*. Cambridge University Press.

Lee, J., Namkoong, H., and Zeng, Y. (2024). Design and scheduling of an ai-based queueing system. *arXiv preprint arXiv:2406.06855*.

Li, F., Liu, J., and Ji, B. (2020). Combinatorial sleeping bandits with fairness constraints. *IEEE Transactions on Network Science and Engineering*.

Liu, Q. and Fang, Z. (2023). Learning to schedule tasks with deadline and throughput constraints. In *IEEE INFOCOM 2023 - IEEE Conference on Computer Communications*.

Liu, X., Li, B., Shi, P., and Ying, L. (2021). An efficient pessimistic-optimistic algorithm for stochastic linear bandits with general constraints. In *Advances Neural Information Processing Systems (NeurIPS)*.

Lucier, B., Pattathil, S., Slivkins, A., and Zhang, M. (2024). Autobidders with budget and roi constraints: Efficiency, regret, and pacing dynamics. In *The Thirty Seventh Annual Conference on Learning Theory*, pages 3642–3643. PMLR.

Lykouris, T. and Weng, W. (2024). Learning to defer in content moderation: The human-ai interplay. *arXiv preprint arXiv:2402.12237*.

MacGlashan, J., Ho, M. K., Loftin, R., Peng, B., Wang, G., Roberts, D. L., Taylor, M. E., and Littman, M. L. (2017). Interactive learning from policy-dependent human feedback. In *Proceedings of the 34th International Conference on Machine Learning*.

Neely, M. J. (2016). Energy-aware wireless scheduling with near-optimal backlog and convergence time tradeoffs. *IEEE/ACM Transactions on Networking*, 24(4):2223–2236.

Novoseller, E., Wei, Y., Sui, Y., Yue, Y., and Burdick, J. (2020). Dueling posterior sampling for preference-based reinforcement learning. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., and Lowe, R. (2022). Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Pacchiano, A., Ghavamzadeh, M., Bartlett, P., and Jiang, H. (2021). Stochastic bandits with linear constraints. In *Int. Conf. Artificial Intelligence and Statistics (AISTATS)*.

Saha, A. (2021). Optimal algorithms for stochastic contextual preference bandits. In *Advances in Neural Information Processing Systems*, volume 34. Curran Associates, Inc.

Saha, A. and Krishnamurthy, A. (2022). Efficient and optimal algorithms for contextual dueling bandits under realizability. In *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*.

Saha, A., Pacchiano, A., and Lee, J. (2023). Dueling rl: Reinforcement learning with trajectory preferences. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*.

Sankararaman, K. A. and Slivkins, A. (2021). Bandits with knapsacks beyond the worst case. *Advances in Neural Information Processing Systems*, 34:23191–23204.

Schede, E., Brandt, J., Tornede, A., Wever, M., Bengs, V., Hullermeier, E., and Tierney, K. (2023). A survey of methods for automated algorithm configuration. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*.

Sivakumar, V., Zuo, S., and Banerjee, A. (2022). Smoothed adversarial linear contextual bandits with knapsacks. In *Proceedings of the 39th International Conference on Machine Learning*.

Slivkins, A. (2019). Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*.

Slivkins, A., Sankararaman, K. A., and Foster, D. J. (2023). Contextual bandits with packing and covering constraints: A modular lagrangian approach via regression. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 4633–4656. PMLR.

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

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. (2020). Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Thompson, W. R. (1933). On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

Vaswani, S., Mehrabian, A., Durand, A., and Kveton, B. (2020). Old dog learns new tricks: Randomized ucb for bandit problems. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*.

Wang, Y., Liu, Q., and Jin, C. (2023). Is RLHF more difficult than standard RL? a theoretical perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Wu, H. and Liu, X. (2016). Double thompson sampling for dueling bandits. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc.

Wu, R. and Sun, W. (2023). Making rl with preference-based feedback efficient via randomization. *arXiv preprint arXiv:2310.14554*.

Yue, Y., Broder, J., Kleinberg, R., and Joachims, T. (2012). The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556.

Zhu, B., Jordan, M., and Jiao, J. (2023). Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *Proceedings of the 40th International Conference on Machine Learning*.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. (2020). Fine-tuning language models from human preferences. *arXiv preprint arXiv:2204.05862*.

Zoghi, M., Whiteson, S., Munos, R., and Rijke, M. (2014). Relative upper confidence bound for the k-armed dueling bandit problem. In *Proceedings of the 31st International Conference on Machine Learning*.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Not Applicable]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendix

A Proof of Lemma 2

Let π be the optimal policy to the following problem:

$$\max_{\pi} \ \sum_{t=1}^{\tau} \mathcal{R}(c_t, x_t^{\pi}, y_t^{\pi}) \tag{20}$$

$$\text{subject to: } \sum_{t=1}^{\tau} \mathcal{W}(c_t, x_t^{\pi}, y_t^{\pi}) \leq B \tag{21}$$

Recall $\{c_t\}_{t=1}^T$ is i.i.d. across rounds. Recall $\mathcal{R}(c,x,y)$ and $\mathcal{W}(c,x,y)$ are i.i.d. samples when duel actions (x,y) are taken given the context c. We have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\sum_{x,y}\mathcal{R}(c_t,x,y)\pi(c_t,x,y)\right] =& \sum_{t=1}^{T}\mathbb{E}\left[\mathbb{E}\left[\sum_{x,y}\mathcal{R}(c_t,x,y)\pi(c_t,x,y)|\mathcal{H}_{t-1}\right]\right] \\
=& \sum_{t=1}^{T}\mathbb{E}\left[\sum_{c}\sum_{x,y}p_c r(c,x,y)\mathbb{E}\left[\pi(c_t,x,y)|\, c_t=c,\mathcal{H}_{t-1}\right]\right] \\
=& \mathbb{E}\left[\sum_{c}\sum_{x,y}p_c r(c,x,y)\sum_{t=1}^{T}\mathbb{E}\left[\pi(c_t,x,y)|\, c_t=c,\mathcal{H}_{t-1}\right]\right] \\
=& \sum_{c}\sum_{x,y}p_c r(c,x,y)\sum_{t=1}^{T}\mathbb{E}\left[\pi(c_t,x,y)|\, c_t=c\right]
\end{aligned}$$

Similarly, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{x,y}\mathcal{W}(c_t,x,y)\pi(c_t,x,y)\right] = \sum_{c}\sum_{x,y}p_c w(c,x,y)\sum_{t=1}^{T}\mathbb{E}\left[\pi(c_t,x,y)|\, c_t=c\right].$$

Define $\hat{\pi}_T^*(c,x,y) = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\pi(c_t,x,y)|\, c_t=c\right]$, where $\hat{\pi}_T^*$ is a feasible solution to (4)–(6) because π is a feasible solution to (20)–(21). Therefore, we have

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{x,y}\mathcal{R}(c_t,x,y)\pi(c_t,x,y)\right] = T\sum_{c}\sum_{x,y}p_c r(c,x,y)\hat{\pi}_T^*(c,x,y) \leq \ T\nu^*.$$

B Proof of Lemma 3

We study the error term related to $\theta_t^0 - \theta^*$ and that related to $\theta^* - \theta_t^1$ follows the same steps.

$$\begin{aligned}
\sum_{t=1}^{T}\langle \phi^-(c_t,x_t,y_t), \theta_t^0 - \theta^*\rangle =& \sum_{t=1}^{T}\langle \phi^-(c_t,x_t,y_t), \theta_t^0 - \hat{\theta}_t + \hat{\theta}_t - \theta^*\rangle \\
\leq& \sum_{t=1}^{T}\|\phi^-(c_t,x_t,y_t)\|_{\Sigma_t^{-1}}\|\theta_t^0 - \hat{\theta}_t\|_{\Sigma_t} + \|\phi^-(c_t,x_t,y_t)\|_{\Sigma_t^{-1}}\|\hat{\theta}_t - \theta^*\|_{\Sigma_t} \\
\leq& 2\sum_{t=1}^{T}\beta_t^r\|\phi^-(c_t,x_t,y_t)\|_{\Sigma_t^{-1}}
\end{aligned}$$

where the second inequality holds because of Cachy-Shwaze inequality; the third inequality holds because $\lambda = 1$ and $\|\theta_t^0 - \hat{\theta}_t\|_{\Sigma_t} \leq \beta_t^r$ and $\|\hat{\theta}_t - \theta^*\|_{\Sigma_t} \leq \beta_t^r$ hold with a high probability $1-p$ in Lemma 1. The procedure above also holds for $\theta^* - \theta_t^1$ and the proof of Lemma 3 is completed by union bound.

C Upper Bound on Cost Mismatch

In this section, we study the upper bound on $M_\tau = \sum_{t=1}^{\tau} \mathcal{W}(c_t, x_t, y_t) - \widetilde{\mathcal{W}}(c_t, x_t, y_t)$ under Algorithms 1 and 2.

Lemma 8. *Under optimistic exploration in Algorithm 1 and randomized exploration in Algorithm 2, we have*

$$M_\tau \leq (2d + 6\beta_\tau^w(p))\sqrt{d\tau \log\left(\frac{\lambda d + \tau}{\lambda d}\right)},$$

where $\beta_\tau^w(p) = d\sqrt{\lambda} + \sqrt{2\log(1/p) + 2d\log(1 + \tau/\lambda d)}$, *holds with a high probability* $1 - p$.

Proof. We first split the term as follows

$$M_\tau = \sum_{t=1}^{\tau} \mathcal{W}(c_t, x_t, y_t) - w(c_t, x_t, y_t) + w(c_t, x_t, y_t) - \widetilde{\mathcal{W}}(c_t, x_t, y_t).$$

Combine Lemmas 9, 10 and 11, we have that

$$M_\tau \leq (2d + 6\beta_\tau^w(p))\sqrt{d\tau \log\left(\frac{\lambda d + \tau}{\lambda d}\right)}$$

holds with a high probability $1 - p$. □

Lemma 9. *Under optimistic exploration in Algorithm 1 and randomized exploration in Algorithm 2, we have*

$$\sum_{t=1}^{\tau} \mathcal{W}(c_t, x_t, y_t) - w(c_t, x_t, y_t) \leq 2d\sqrt{2\tau \log(1/p)},$$

holds with a high probability $1 - p$.

Proof. Let $\Delta(c_t, x_t, y_t) := \mathcal{W}(c_t, x_t, y_t) - w(c_t, x_t, y_t)$ and $\{\Delta(c_t, x_t, y_t)\}$ is a martingale difference sequence with bound by $2d$ by Assumptions 2. Therefore, by Azuma-Hoeffding inequality, we have

$$\mathbb{P}\left(\sum_{t=1}^{\tau} \mathcal{W}(c_t, x_t, y_t) - w(c_t, x_t, y_t) > 2d\sqrt{2\tau \log(1/p)}\right) \leq p, \quad \forall \tau \in [T].$$

□

Recall $\widetilde{\mathcal{W}}(c_t, x_t, y_t) = \langle \hat{\omega}_t, \psi^+(c_t, x_t, y_t)\rangle - b_w(c_t, x_t, y_t)$ with $b_w(c_t, x_t, y_t) = \beta_\tau^w(p)(\|\psi(c_t, x_t)\|_{\Psi_t^{-1}} + \|\psi(c_t, y_t)\|_{\Psi_t^{-1}})$ in Algorithm 1.

Lemma 10. *Under optimistic exploration in Algorithm 1, we have*

$$\sum_{t=1}^{\tau} w(c_t, x_t, y_t) - \widetilde{\mathcal{W}}(c_t, x_t, y_t) \leq 6\beta_\tau^w(p)\sqrt{d\tau \log\left(\frac{\lambda d + \tau}{\lambda d}\right)},$$

where $\beta_\tau^w(p) = d\sqrt{\lambda} + \sqrt{2\log(1/p) + 2d\log(1 + \tau/\lambda d)}$, *holds with a high probability* $1 - p$.

Proof. We study the difference of $w(c_t, x_t) - \widetilde{\mathcal{W}}(c_t, x_t)$ as follows

$$\begin{aligned}
\sum_{t=1}^{\tau} w(c_t, x_t) - \widetilde{\mathcal{W}}(c_t, x_t) &= \sum_{t=1}^{\tau} \langle \omega^* - \hat{\omega}_t, \psi(c_t, x_t)\rangle + \beta_\tau^w(p)\|\psi(c_t, x_t)\|_{\Psi_t^{-1}} \\
&\leq \sum_{t=1}^{\tau} \|\omega^* - \hat{\omega}_t\|_{\Psi_t}\|\psi(c_t, x_t)\|_{\Psi_t^{-1}} + \beta_\tau^w(p)\|\psi(c_t, x_t)\|_{\Psi_t^{-1}} \\
&\leq \sum_{t=1}^{\tau} 2\beta_\tau^w(p)\|\psi(c_t, x_t)\|_{\Psi_t^{-1}} \\
&\leq \beta_\tau^w(p)\sqrt{8d\tau \log\left(\frac{\lambda d + \tau}{\lambda d}\right)}
\end{aligned}$$

The difference of $w(c_t, y_t) - \widetilde{\mathcal{W}}(c_t, y_t)$ follows the same steps and the proof is completed by union bound. □

Recall $\widetilde{\mathcal{W}}(c_t, x_t, y_t) = \langle \omega_t, \psi^+(c_t, x_t, y_t)\rangle$ with $\omega_t \sim \mathcal{N}(\hat{\omega}_t, \beta_t^w \Psi_t^{-1})$ in Algorithm 2.

Lemma 11. *Under randomized exploration in Algorithm 2, we have*

$$\sum_{t=1}^{\tau} w(c_t, x_t, y_t) - \widetilde{\mathcal{W}}(c_t, x_t, y_t) \leq 6\sqrt{d\tau\beta_\tau^w(p)\log\left(\frac{\lambda d + \tau}{\lambda d}\right)},$$

where $\beta_\tau^w(p) = d\sqrt{\lambda} + \sqrt{2\log(1/p) + 2d\log(1+\tau/\lambda d)}$, *holds with a high probability* $1-p$.

Proof. We study the difference of $w(c_t, x_t) - \widetilde{\mathcal{W}}(c_t, x_t)$ as follows

$$\begin{aligned}
\sum_{t=1}^{\tau} w(c_t, x_t) - \widetilde{\mathcal{W}}(c_t, x_t) =& \sum_{t=1}^{\tau} \langle \omega^* - \hat{\omega}_t, \psi(c_t, x_t)\rangle + \langle \hat{\omega}_t - \omega_t, \psi(c_t, x_t)\rangle \\
\leq& \sum_{t=1}^{\tau} \|\omega^* - \hat{\omega}_t\|_{\Psi_t} \|\psi(c_t, x_t)\|_{\Psi_t^{-1}} + \|\hat{\omega}_t - \omega_t\|_{\Psi_t} \|\psi(c_t, x_t)\|_{\Psi_t^{-1}} \\
\leq& \sum_{t=1}^{\tau} 2\beta_\tau^w(p) \|\psi(c_t, x_t)\|_{\Psi_t^{-1}} \\
\leq& \beta_\tau^w(p)\sqrt{8d\tau\log\left(\frac{\lambda d + \tau}{\lambda d}\right)}
\end{aligned}$$

The difference of $w(c_t, y_t) - \widetilde{\mathcal{W}}(c_t, y_t)$ follows the same steps and the proof is completed by union bound. □

D PbLinCBwK with Optimistic Exploration

In this section, we analyze PbLinCBwK with optimistic exploration. Let $\beta_t = \beta_t^r$ and $\gamma_t = \beta_t^w$ for simple notation.

D.1 Proof of Lemma 4 under Algorithm 1

In Algorithm 1, (x_t, y_t) are the optimal solution to maximize $\widetilde{\mathcal{R}}(c_t, x, y) - \frac{Q_t}{V}(\widetilde{\mathcal{W}}(c_t, x) + \widetilde{\mathcal{W}}(c_t, y))$. We let $(x, y) = (x, y_t)$ be the baseline such that (note x could be any action in $\mathcal{A}$ including the optimal x^*)

$$\widetilde{\mathcal{R}}(c_t, x, y_t) - \frac{Q_t}{V}(\widetilde{\mathcal{W}}(c_t, x) + \widetilde{\mathcal{W}}(c_t, y_t)) \leq \widetilde{\mathcal{R}}(c_t, x_t, y_t) - \frac{Q_t}{V}(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t)).$$

Substitute the definition of $\widetilde{\mathcal{R}}(c_t, \cdot, \cdot)$, we have

$$\begin{aligned}
&\langle \hat{\theta}_t, \phi(c_t, x) + \phi(c_t, y_t)\rangle + \beta_t \|\phi^-(c_t, x, y_t)\|_{\Sigma_t^{-1}} - \frac{Q_t}{V_t}(\widetilde{\mathcal{W}}(c_t, x) + \widetilde{\mathcal{W}}(c_t, y_t)) \\
\leq& \langle \hat{\theta}_t, \phi(c_t, x_t) + \phi(c_t, y_t)\rangle + \beta_t \|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}} - \frac{Q_t}{V_t}(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t)).
\end{aligned}$$

Subtract the common terms on both sides of the inequality, we have

$$\langle \hat{\theta}_t, \phi^-(c_t, x, x_t)\rangle \leq \beta_t \|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}} - \beta_t \|\phi^-(c_t, x, y_t)\|_{\Sigma_t^{-1}} + \frac{Q_t}{V_t}(\widetilde{\mathcal{W}}(c_t, x) - \widetilde{\mathcal{W}}(c_t, x_t)). \tag{22}$$

According to Cauchy–Schwarz inequality and Lemma 1, we have

$$|\langle \theta^* - \hat{\theta}_t, \phi^-(c_t, x, y_t)\rangle| \leq \|\theta^* - \hat{\theta}\|_{\Sigma_t} \|\phi^-(c_t, x, y_t)\|_{\Sigma_t^{-1}} \leq \beta_t \|\phi^-(c_t, x, y_t)\|_{\Sigma_t^{-1}}, \tag{23}$$

holds with a high probability. Therefore, recall $\text{Regret}(c_t, x) = \langle \theta^* - \hat{\theta}_t, \phi^-(c_t, x, y_t)\rangle + \langle \hat{\theta}_t, \phi^-(c_t, x, x_t)\rangle$, we establish the regret against x (or x^*) by combining (22) and (23) as follows

$$\text{Regret}(c_t, x) \le \beta_t \|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}} + \frac{Q_t}{V_t}(\widetilde{\mathcal{W}}(c_t, x) - \widetilde{\mathcal{W}}(c_t, x_t)). \tag{24}$$

Similarly, we let $(x, y) = (x_t, y)$ be the baseline such that (note y could be any action in $\mathcal{A}$ including the optimal y^*)

$$\widetilde{\mathcal{R}}(c_t, x_t, y) - \frac{Q_t}{V}(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y)) \le \widetilde{\mathcal{R}}(c_t, x_t, y_t) - \frac{Q_t}{V}(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t)).$$

Substitute the definition of $\widetilde{\mathcal{R}}(c_t, \cdot, \cdot)$, we have

$$\begin{aligned}
&\langle \hat{\theta}_t, \phi(c_t, x_t) + \phi(c_t, y)\rangle + \beta_t \|\phi^-(c_t, x_t, y)\|_{\Sigma_t^{-1}} - \frac{Q_t}{V_t}(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y)) \\
\le &\langle \hat{\theta}_t, \phi(c_t, x_t) + \phi(c_t, y_t)\rangle + \beta_t \|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}} - \frac{Q_t}{V_t}(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t)),
\end{aligned}$$

which implies

$$\langle \hat{\theta}_t, \phi^-(c_t, y, y_t)\rangle \le \beta_t \|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}} - \beta_t \|\phi^-(c_t, x_t, y)\|_{\Sigma_t^{-1}} + \frac{Q_t}{V_t}(\widetilde{\mathcal{W}}(c_t, y) - \widetilde{\mathcal{W}}(c_t, y_t)). \tag{25}$$

According to Cauchy-Schwarz inequality and Lemma 1, we have

$$\langle \theta^* - \hat{\theta}_t, \phi^-(c_t, y, x_t)\rangle \le \|\theta^* - \hat{\theta}\|_{\Sigma_t} \|\phi^-(c_t, y, x_t)\|_{\Sigma_t^{-1}} \le \beta_t \|\phi^-(c_t, y, x_t)\|_{\Sigma_t^{-1}}, \tag{26}$$

holds with a high probability. Therefore, we establish the regret against y (or y^*) with $\text{Regret}(c_t, y) = \langle \theta^* - \hat{\theta}_t, \phi^-(c_t, y, x_t)\rangle + \langle \hat{\theta}_t, \phi^-(c_t, y, y_t)\rangle$ by combining (25) and (26) as follows

$$\text{Regret}(c_t, y) \le \beta_t \|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}} + \frac{Q_t}{V_t}(\widetilde{\mathcal{W}}(c_t, y) - \widetilde{\mathcal{W}}(c_t, y_t)). \tag{27}$$

By combining two key inequalities (24) and in (27), we have

$$\begin{aligned}
&\text{Regret}(c_t, x) + \text{Regret}(c_t, y) \\
\le &2\beta_t \|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}} + \frac{Q_t}{V_t}(\widetilde{\mathcal{W}}(c_t, x) + \widetilde{\mathcal{W}}(c_t, y) - \widetilde{\mathcal{W}}(c_t, x_t) - \widetilde{\mathcal{W}}(c_t, y_t)),
\end{aligned}$$

which implies

$$\begin{aligned}
&\text{Regret}(c_t, x) + \text{Regret}(c_t, y) + \frac{Q_t}{V_t}(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b) \\
\le\ &2\beta_t \|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}} + \frac{Q_t}{V_t}(\widetilde{\mathcal{W}}(c_t, x) + \widetilde{\mathcal{W}}(c_t, y) - b).
\end{aligned}$$

According to the virtual queue update

$$Q_{t+1} = \max\left(Q_t + \widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b, 0\right),$$

we have

$$\frac{1}{2}Q_{t+1}^2 - \frac{1}{2}Q_t^2 \le Q_t(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b) + \frac{1}{2}(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b)^2.$$

Therefore, we establish a key inequality on the "Regret + Lyapunov drift" as follows

$$\begin{aligned}
&\text{Regret}(c_t, x) + \text{Regret}(c_t, y) + \frac{Q_{t+1}^2}{2V} - \frac{Q_t^2}{2V} \\
\le &2\beta_t \|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}} + \frac{Q_t}{V}(\widetilde{\mathcal{W}}(c_t, x) + \widetilde{\mathcal{W}}(c_t, y) - b) + \frac{1}{2V}(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b)^2.
\end{aligned} \tag{28}$$

To proceed, we first define the following high probability event

$$\mathcal{E}_w = \left\{ \widetilde{\mathcal{W}}(c,x) - w(c,x) \leq 0 \quad \forall c, x, t \in [T] \right\},$$

where $p = 1/|\mathcal{C}||\mathcal{A}|T^3$ for $\gamma_t(p)$. Based on Lemma 1, the definition of $\widetilde{\mathcal{W}}(c_t, x)$, and the union bound, we have

$$\mathbb{P}\left(\mathcal{E}_w\right) \geq 1 - 1/T^2. \tag{29}$$

Now taking the conditional expectation $\mathbb{E}[\ \cdot\ |\mathcal{H}_t = h]$ on both sides of (28), we establish

$$\begin{aligned}
&\mathbb{E}\left[\text{Regret}(c_t, x) + \text{Regret}(c_t, y) | \mathcal{H}_t = h\right] + \mathbb{E}\left[\frac{Q_{t+1}^2}{2V} - \frac{Q_t^2}{2V} | \mathcal{H}_t = h\right] \\
\leq &\mathbb{E}\left[2\beta_t \|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}} + \frac{1}{2V}(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b)^2 \mid \mathcal{H}_t = h\right] \\
&+ \frac{Q}{V}\mathbb{E}\left[(\widetilde{\mathcal{W}}(c_t, x) + \widetilde{\mathcal{W}}(c_t, y) - b) \mid \mathcal{H}_t = h\right] \\
\leq &\mathbb{E}\left[2\beta_t \|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}} + \frac{1}{2V}(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b)^2 \mid \mathcal{H}_t = h\right] \\
&+ \frac{Q}{V}\mathbb{E}\left[(w(c_t, x) + w(c_t, y) - b) \mid \mathcal{H}_t = h, \mathcal{E}_w\right] + \frac{Q(2+b)}{V}\mathbb{P}\left(\mathcal{E}_w^c\right) \\
\leq &\mathbb{E}\left[2\beta_t \|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}} + \frac{1}{2V}(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b)^2 \mid \mathcal{H}_t = h\right] \\
&+ \frac{Q}{V}\mathbb{E}\left[(w(c_t, x) + w(c_t, y) - b) \mid \mathcal{H}_t = h\right] + \frac{2Q(2+b)}{V}\frac{\mathbb{P}\left(\mathcal{E}_w^c\right)}{\mathbb{P}\left(\mathcal{E}_w\right)}
\end{aligned} \tag{30}$$

where the second inequality holds under the high probability event $\mathcal{E}_w$ that

$$\widetilde{\mathcal{W}}(c_t, x) \leq w(c_t, x) \ \text{ and } \ \widetilde{\mathcal{W}}(c_t, y) \leq w(c_t, y).$$

The last inequality holds because the context is independent of $\mathcal{H}_t$ and the distribution of c_t does not change conditioned on $\mathcal{H}_t$. Therefore, we have

$$\mathbb{P}\left(c_t = c | \mathcal{H}_t = h, \mathcal{E}\right) = \mathbb{P}\left(c_t = c | \mathcal{E}\right).$$

Then we calculate that

$$\begin{aligned}
\mathbb{P}\left(c_t = c | \mathcal{E}_w\right) - \mathbb{P}\left(c_t = c\right) =& \frac{\mathbb{P}(c_t = c, \mathcal{E}_w) - \mathbb{P}\left(c_t = c\right)\mathbb{P}\left(\mathcal{E}_w\right)}{\mathbb{P}(\mathcal{E}_w)} \\
=& \frac{\mathbb{P}\left(c_t = c\right)\left(\mathbb{P}(\mathcal{E}_w | c_t = c) - \mathbb{P}\left(\mathcal{E}_w\right)\right)}{\mathbb{P}(\mathcal{E}_w)} \\
\leq& \mathbb{P}\left(c_t = c\right)\frac{1 - \mathbb{P}\left(\mathcal{E}_w\right)}{\mathbb{P}\left(\mathcal{E}_w\right)},
\end{aligned}$$

which implies that

$$\frac{Q}{V}\mathbb{E}\left[(w(c_t, x) + w(c_t, y) - b) \mid \mathcal{H}_t = h, \mathcal{E}_w\right] \leq \frac{Q(2+b)}{V}\frac{1 - \mathbb{P}\left(\mathcal{E}_w\right)}{\mathbb{P}\left(\mathcal{E}_w\right)}.$$

D.2 Proof of Lemma 5 under Algorithm 1

Since (x, y) could be any duel actions in $\mathcal{A}$, let $(x, y) = (x^*, y^*)$ in (30) that

$$\begin{aligned}
&\mathbb{E}\left[\text{Regret}(c_t, x^*) + \text{Regret}(c_t, y^*) | \mathcal{H}_t = h\right] + \mathbb{E}\left[\frac{Q_{t+1}^2}{2V} - \frac{Q_t^2}{2V} | \mathcal{H}_t = h\right] \\
\leq &\mathbb{E}\left[2\beta_t \|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}} + \frac{1}{2V}(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b)^2 \mid \mathcal{H}_t = h\right] + \frac{2(2+b)Q}{V}\frac{\mathbb{P}\left(\mathcal{E}_w^c\right)}{\mathbb{P}\left(\mathcal{E}_w\right)},
\end{aligned}$$

where the inequality holds because (x^*, y^*) satisfy the constraint in (5) such that $\mathbb{E}\left[(w(c_t, x^*) + w(c_t, y^*) - b) \mid \mathcal{H}_t = h\right] \leq 0$. We further take the expectation on both sides of the inequality and then take summation from $t = 1$ to τ that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{\tau} \text{Regret}(c_t, x^*) + \text{Regret}(c_t, y^*)\right] + \mathbb{E}\left[\frac{Q_{\tau+1}^2}{2V} - \frac{Q_1^2}{2V}\right] \\
\leq &\mathbb{E}\left[\sum_{t=1}^{\tau} 2\beta_t \|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}}\right] + \mathbb{E}\left[\sum_{t=1}^{\tau} \frac{(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b)^2}{2V}\right] + \mathbb{E}\left[\sum_{t=1}^{\tau} \frac{2(2+b)Q_t}{V} \frac{\mathbb{P}\left(\mathcal{E}_w^c\right)}{\mathbb{P}\left(\mathcal{E}_w\right)}\right].
\end{aligned}$$

Since $Q_1 = 0$ and $\mathbb{P}\left(\mathcal{E}_w^c\right) \leq 1/T^2$, we conclude that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{\tau} \text{Regret}(c_t, x^*) + \text{Regret}(c_t, y^*)\right] \\
\leq &\mathbb{E}\left[\sum_{t=1}^{\tau} 2\beta_t \|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}}\right] + 4(2+b) + \mathbb{E}\left[\sum_{t=1}^{\tau} \frac{(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b)^2}{2V}\right] \\
\leq &4\beta_T \sqrt{T \log(1+T)} + 4(2+b) + \mathbb{E}\left[\sum_{t=1}^{\tau} \frac{(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b)^2}{2V}\right]
\end{aligned}$$

where the second inequality holds by Lemma 14. Finally, we complete the proof by providing a *simple but refined analysis* on the cumulative budget consumption that

$$\mathbb{E}\left[\sum_{t=1}^{\tau} \frac{(\widetilde{\mathcal{W}}(c_t, x_t, y_t) - b)^2}{2V}\right] \leq \mathbb{E}\left[\sum_{t=1}^{\tau} \frac{\widetilde{\mathcal{W}}(c_t, x_t, y_t)^2 + b^2}{V}\right] \leq \frac{2d}{V}\mathbb{E}\left[\sum_{t=1}^{\tau} \widetilde{\mathcal{W}}(c_t, x_t, y_t)\right] + \frac{\tau b^2}{V},$$

where the second inequality holds because $\widetilde{\mathcal{W}}(c_t, x_t, y_t)$ is bounded by $2d$. Moreover, we have

$$\mathbb{E}\left[\sum_{t=1}^{\tau} \widetilde{\mathcal{W}}(c_t, x_t, y_t)\right] = \mathbb{E}\left[\sum_{t=1}^{\tau} \widetilde{\mathcal{W}}(c_t, x_t, y_t) - \mathcal{W}(c_t, x_t, y_t)\right] + \mathbb{E}\left[\sum_{t=1}^{\tau} \mathcal{W}(c_t, x_t, y_t)\right] \leq 1 + B,$$

where the last inequality holds because of $\widetilde{\mathcal{W}}(c_t, x_t, y_t)$ is an under-estimator of $\mathbb{E}[\mathcal{W}(c_t, x_t, y_t)]$ and the definition of stopping time τ. Therefore, combine with Lemma 3, we have

$$\mathbb{E}\left[\text{Regret}(\tau)\right] \leq 12\beta_T \sqrt{T \log(1+T)} + 4(2+b) + \frac{2(d + dTb + Tb^2)}{V}.$$

D.3 Proof of Lemma 6 under Algorithm 1

D.3.1 Lyapunov Drift Analysis

From (30), we have established the Lyapunov drift

$$\begin{aligned}
&\mathbb{E}\left[\frac{Q_{t+1}^2}{2} - \frac{Q_t^2}{2} | \mathcal{H}_t = h\right] \\
\leq &\mathbb{E}\left[V(\text{Regret}(c_t, x) + \text{Regret}(c_t, y)) + 2\beta_t V \|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}} + \frac{1}{2}(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b)^2 \mid \mathcal{H}_t = h\right] \\
&+ Q\mathbb{E}\left[(w(c_t, x) + w(c_t, y) - b) \mid \mathcal{H}_t = h\right] + \frac{2Q(2+b)}{V} \frac{\mathbb{P}\left(\mathcal{E}_w^c\right)}{\mathbb{P}\left(\mathcal{E}_w\right)} \\
\leq &4V + \frac{4\beta_t V}{\lambda} + Q\mathbb{E}\left[(w(c_t, x) + w(c_t, y) - b) \mid \mathcal{H}_t = h\right] + \frac{2(2+b)Q}{V} \frac{\mathbb{P}\left(\mathcal{E}_w^c\right)}{\mathbb{P}\left(\mathcal{E}_w\right)} \\
\leq &4V + 4\beta_t V + Q\mathbb{E}\left[(w(c_t, x) + w(c_t, y) - b) \mid \mathcal{H}_t = h\right] + 4(2+b) \\
\leq &-\delta b \cdot Q + 4V + 4\beta_t V + 4(2+b) \qquad (31)
\end{aligned}$$

where the second inequality holds due to the bounded rewards and costs in Assumption 2; the third inequality holds because $\lambda = 1$ and $\mathbb{P}(\mathcal{E}_w^c) \leq 1/T^2$; the last inequality holds because and the "Slater condition" in Assumption 3 that there exists a feasible policy such that $\mathbb{E}[w(c_t, x) + w(c_t, y) - b \mid \mathcal{H}_t = h] \leq -\delta b$.

Since we have the negative drift in (31), we establish the high probability bound on the virtual queue Q_t according to the Lyapunov drift lemma in Lemma 15.

D.3.2 High Probability Bound

We define the Lyapunov function $\bar{L}_t = Q_t$. To establish the high probability bound on the virtual queue, we prove conditions (i) and (ii) in Lemma 15 for $\bar{L}_t$.

Given $\mathcal{H}_t = h$ and $\bar{L}_t \geq \varphi_t = \frac{8(V+\beta_t V+(2d+b))}{\delta b}$, the conditional expected drift of $\bar{L}_t$ is

$$\begin{aligned}
&\mathbb{E}[Q_{t+1} - Q_t | \mathcal{H}_t = h] \leq \frac{1}{2Q}\mathbb{E}[Q_{t+1}^2 - Q_t^2 | \mathcal{H}_t = h] \\
\leq &\frac{-\delta b \cdot Q + 4V + 4\beta_t V + 4(2+b)}{2Q} \\
\leq &-\frac{\delta b}{2} + \frac{2V + 2\beta_t V + 2(2+b)}{Q} \\
\leq &-\frac{\delta b}{4}
\end{aligned}$$

where the first inequality holds because $2(Q_{t+1} - Q_t)Q_t \leq Q_{t+1}^2 - Q_t^2$; the second inequality holds by the "negative drift" in 31 above; and the last inequality holds given the condition $Q \geq \varphi_t = \frac{8(V+\beta_t V+(2d+b))}{\delta b}$. Moreover, for condition (ii) in Lemma 15, we have $Q_{t+1} - Q_t \leq 2d + b$.

Let $\rho = \frac{\delta b}{4}$, and $\nu_{\max} = 2d + b$. We are ready to apply Lemma 15 for $\bar{L}(t)$ and obtain

$$\mathbb{E}\left[e^{\zeta Q_t}\right] \leq 1 + \frac{2e^{\zeta(\nu_{\max}+\varphi_t)}}{\zeta\rho} \text{ with } \zeta = \frac{\rho}{\nu_{\max}^2 + \nu_{\max}\rho/3}. \tag{32}$$

We then establish the high probability bound as follows

$$\begin{aligned}
&\mathbb{P}\left(Q_t \geq \nu_{\max} + \varphi_t + \frac{\log\frac{6T}{\delta\rho}}{\zeta}\right) \leq \mathbb{P}\left(e^{\zeta Q_t} \geq e^{\zeta(\nu_{\max}+\varphi_t)+\log\frac{6T}{\delta\rho}}\right) \\
\leq &\frac{\mathbb{E}\left[e^{\zeta Q_t}\right]}{e^{\zeta(\nu_{\max}+\varphi_t)+\log\frac{6T}{\delta\rho}}} \leq \frac{1 + \frac{2e^{\zeta(\nu_{\max}+\varphi_t)}}{\zeta\rho}}{e^{\zeta(\nu_{\max}+\varphi_t)+\log\frac{6T}{\delta\rho}}} \leq 1/T^2,
\end{aligned} \tag{33}$$

where the second inequality holds by Markov inequality and the third inequality holds by (47). Therefore, there exists an absolute constant C_1 such that $\frac{C_1\beta_t V}{\delta b} \geq \nu_{\max} + \varphi_t + \frac{\log\frac{6T}{\delta\rho}}{\zeta}$ and the high probability bound holds

$$\mathbb{P}\left(Q_t > \frac{C_1\beta_t V}{\delta b}\right) \leq 1/T^2.$$

D.4 Proof of Lemma 7 under Algorithm 1

Recall the definition of virtual stopping time

$$\tau_0 = \text{argmin}_{\tau' \in [T]} \left\{\tau' \mid Q_{\tau'+1} + b\tau' + M_\tau \geq B\right\},$$

where $M_\tau = \sum_{t=1}^{\tau}(\mathcal{W}(c_t, x_t, y_t) - \widetilde{\mathcal{W}}(c_t, x_t, y_t))$. We have established both upper bounds of Q_τ and M_τ in Lemma 6 and Lemma 8, respectively. Therefore, we have

$$\frac{C_1\beta_\tau V}{\delta b} + (2d + 6\gamma_\tau)\sqrt{d\tau\log(1 + \tau/d)} + b\tau_0 \geq B.$$

Recall the definition of $b = B/T$ and by dividing b on both sides of the inequality, we establish the following upper bound on the remaining rounds

$$\frac{C_1\beta_T V}{\delta b^2} + \frac{8d\gamma_T\sqrt{\tau\log(1+\tau)}}{b} \geq T - \tau_0, \tag{34}$$

holds with a high probability.

D.5 Proof of Theorem 1 under Algorithm 1

Now we aggregate the regret after stopping and before stopping as follows

$$\begin{aligned}
\text{Regret}(T) \leq & \underbrace{2\nu^*\mathbb{E}[T-\tau]}_{\text{regret after stopping}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{\tau}\mathcal{R}(c_t,x^*,y^*)\right] - \mathbb{E}\left[\sum_{t=1}^{\tau}\mathcal{R}(c_t,x_t,y_t)\right]}_{\text{regret before stopping}} \\
\leq & 4\beta_T\sqrt{T\log(1+T)} + 4(2+b) + \frac{2(d+dTb+Tb^2)}{V} \\
& + \left(\frac{C_1\beta_T V}{\delta b^2} + \frac{8d\gamma_T\sqrt{\tau\log(1+\tau)}}{b}\right)\nu^* \\
\leq & 8\beta_T\sqrt{T\log(1+T)} + \frac{2(dTb+Tb^2)}{V} + \left(\frac{C_1\beta_T V}{\delta b^2} + \frac{8d\gamma_T\sqrt{\tau\log(1+\tau)}}{b}\right)\nu^*,
\end{aligned} \tag{35}$$

Let $V = b\sqrt{T}$ in (35), we finally have

$$\text{Regret}(T) \leq 8\beta_T\sqrt{T\log(1+T)} + 3(d+b)\sqrt{T} + \frac{C_1\beta_T\sqrt{T} + 8d\gamma_T\sqrt{\tau\log(1+\tau)}}{\delta}\frac{\nu^*}{b}.$$

Recall $\beta_T = 2\kappa d + 2\kappa\sqrt{4\log T + 2d\log(1+T/\kappa d)}$ and $\gamma_T = d + \sqrt{4\log T + 2d\log(1+T/d)}$ and the proof of Theorem 1 is completed.

E PbLinCBwK with Randomized Exploration

In this section, we analyze PbLinCBwK with randomized exploration. Let $\beta_t = \beta_t^r$ and $\gamma_t = \beta_t^w$ for a simple notation.

E.1 Proof of Lemma 4 under Algorithm 2

We first establish the upper bounds of $\text{Regret}(c_t, x)$ and $\text{Regret}(c_t, y)$ in the next two sections, respectively, and then bridge them with the Lyapunov drift.

E.1.1 Upper Bound on $\text{Regret}(c_t, x)$

We first add and subtract the term related to the cost $\widetilde{\mathcal{W}}(c_t, x_t)$ in Algorithm 2 (note x could be any action in $\mathcal{A}$ including the optimal x^*) that

$$\begin{aligned}
\text{Regret}(c_t,x) =& \langle\phi(c_t,x,y_t),\theta^* - \theta_t^0\rangle + \langle\phi(c_t,x,x_t),\theta_t^0\rangle \\
=& \langle\phi(c_t,x,y_t),\theta^* - \theta_t^0\rangle + \langle\phi(c_t,x,x_t),\theta_t^0\rangle + \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t,x_t) - \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t,x_t)
\end{aligned}$$

Recall x_t is the optimal solution to in Algorithm 2. Therefore, we have

$$\langle\phi(c_t,x),\theta_t^0\rangle - \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t,x) \leq \langle\phi(c_t,x_t),\theta_t^0\rangle - \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t,x_t), \quad \forall x\in\mathcal{A},$$

which implies that

$$\langle\phi^-(c_t,x,x_t),\theta_t^0\rangle \leq \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t,x) - \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t,x_t), \quad \forall x\in\mathcal{A}. \tag{36}$$

Therefore, $\text{Regret}(c_t, x)$ is bounded as follows

$$\begin{aligned}\text{Regret}(c_t, x) \leq & \langle \phi^-(c_t, x, y_t), \theta^* - \theta_t^0 \rangle + \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t, x) - \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t, x_t) \\ = & \langle \phi^-(c_t, x, y_t), \theta^* - \hat{\theta}_t + \hat{\theta}_t - \theta_t^0 \rangle + \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t, x) - \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t, x_t) \\ \leq & \beta_t \|\phi^-(c_t, x, y_t)\|_{\Sigma_t^{-1}} + \langle \phi^-(c_t, x, y_t), \hat{\theta}_t - \theta_t^0 \rangle + \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t, x) - \frac{Q_t}{V}\mathcal{W}(c_t, x) \\ & + \frac{Q_t}{V}\mathcal{W}(c_t, x) - \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t, x_t)\end{aligned}$$

where the first inequality holds by (36); the last inequality holds by holds because $\hat{\theta}_t$ is MLE of θ^* in Lemma (12). Since $\langle \phi^-(c_t, x, y_t), \hat{\theta}_t - \theta_t^0 \rangle \sim \mathcal{N}(0, \beta_t^2 \|\phi^-(c_t, x, y_t)\|^2_{\Sigma_t^{-1}})$ and $\langle \psi(c_t, x), \omega_t - \hat{\omega}_t \rangle \sim \mathcal{N}(0, \gamma_t^2 \|\psi(c_t, x)\|^2_{\Psi_t^{-1}})$, we conclude that the following key event $\mathcal{E}_x$ happens

$$\text{Regret}(c_t, x) - \frac{Q_t}{V}\mathcal{W}(c_t, x) + \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t, x_t) \leq 0 \tag{37}$$

with probability $p_0 = G^2(-1)/2$, where $G(\cdot)$ denotes the CDF of a standard normal distribution.

To proceed, we first define a key difference term

$$\begin{aligned} d_t^0 &= \langle \phi(c_t, x_t) - \phi(c_t, y_t), \theta_t^0 \rangle, \\ d_t^* &= \min_{\|\theta - \hat{\theta}_t\|_{\Sigma_t} \leq \sigma_t} \langle \phi(c_t, x_t) - \phi(c_t, y_t), \theta \rangle, \end{aligned}$$

which are used to define the following event $\mathcal{D}_0$ such that $\mathcal{D}_0 = \{d_t^0 \geq d_t^*\}$. Note this event occurs with a high probability

$$\mathbb{P}(\mathcal{D}_0) = \mathbb{P}(\|\theta_t^0 - \hat{\theta}_t\|_{\Sigma_t} \leq \beta_t) \geq 1 - p.$$

Moreover, define $(\theta_t^0, \theta_t^1, \omega_t) \sim \mathbb{P}_\Gamma$ and $\tilde{\mathbb{P}}_\Gamma = \mathbb{P}_\Gamma / \mathbb{P}(\mathcal{E}_x)$ if the event $\mathcal{E}_x$ happens and $\tilde{\mathbb{P}}_\Gamma = 0$, otherwise. We then study the following key term

$$\begin{aligned} & \text{Regret}(c_t, x) - \frac{Q_t}{V}\mathcal{W}(c_t, x) + \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t, x_t) \\ = & \langle \phi^-(c_t, x, y_t), \theta^* \rangle - d_t^0 - \frac{Q_t}{V}\mathcal{W}(c_t, x) + \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t, x_t) \\ \leq & \langle \phi^-(c_t, x, y_t), \theta^* \rangle - d_t^* - \frac{Q_t}{V}\mathcal{W}(c_t, x) + \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t, x_t) \\ = & \mathbb{E}_{\tilde{\mathbb{P}}_\Gamma}\left[\langle \phi(c_t, x, y_t), \theta^* \rangle - d_t^0 - \frac{Q_t}{V}\mathcal{W}(c_t, x) + \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t, x_t) + d_t^0 - d_t^*\right] \\ \leq & \mathbb{E}_{\tilde{\mathbb{P}}_\Gamma}\left[d_t^0 - d_t^*\right] \\ = & \mathbb{E}_{\mathbb{P}_\Gamma}[\mathbb{I}(\mathcal{E}_x)(d_t^0 - d_t^*)]/\mathbb{P}(\mathcal{E}_x) \\ = & \mathbb{E}_{\mathbb{P}_\Gamma}[\mathbb{I}(\mathcal{E}_x)\mathbb{I}(\mathcal{D}_0)(d_t^0 - d_t^*) + \mathbb{I}(\mathcal{E}_x)\mathbb{I}(\mathcal{D}_0^c)(d_t^0 - d_t^*)]/\mathbb{P}(\mathcal{E}_x) \\ \leq & \mathbb{E}_{\mathbb{P}_\Gamma}[\mathbb{I}(\mathcal{E}_x)\mathbb{I}(\mathcal{D}_0)(d_t^0 - d_t^*)]/\mathbb{P}(\mathcal{E}_x) + 4pT/\mathbb{P}(\mathcal{E}_x) \end{aligned} \tag{38}$$

where the first inequality holds because of the definition of d_t^0 and d_t^*; the second inequality holds because we drop the negative term as in (37) under the event $\mathcal{E}_x$; the last inequality holds because the event $\mathcal{D}_0$ holds with a high probability $1 - p$. Now we study the first major term in (38) as follows

$$\begin{aligned} \mathbb{E}_{\mathbb{P}_\Gamma}[\mathbb{I}(\mathcal{E}_x)\mathbb{I}(\mathcal{D}_0)(d_t^0 - d_t^*)]/\mathbb{P}(\mathcal{E}_x) &\leq \mathbb{E}_{\mathbb{P}_\Gamma}[\mathbb{I}(\mathcal{D}_0)(d_t^0 - d_t^*)]/\mathbb{P}(\mathcal{E}_x) \\ &\leq \mathbb{E}_{\mathbb{P}_\Gamma}[d_t^0 - d_t^*]/\mathbb{P}(\mathcal{E}_x). \end{aligned}$$

Let θ_t^* be the solution to minimizing d_t^*, we then have

$$\begin{aligned} & \text{Regret}(c_t, x) - \frac{Q_t}{V}\mathcal{W}(c_t, x) + \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t, x_t) \\ \leq & \mathbb{E}[d_t^0 - d_t^*]/\mathbb{P}(\mathcal{E}_x) + 4pT/\mathbb{P}(\mathcal{E}_x) \\ \leq & \mathbb{E}[\langle \phi^-(c_t, x_t, y_t), \theta_t^0 - \theta_t^* \rangle - \langle \phi^-(c_t, x_t, y_t), \theta_t^* - \theta^* \rangle]/\mathbb{P}(\mathcal{E}_x) + 4pT/\mathbb{P}(\mathcal{E}_x). \end{aligned}$$

For the first term above, we have

$$\begin{aligned}\langle \phi^-(c_t,x_t,y_t),\theta_t^0-\theta_t^*\rangle =&\langle \phi^-(c_t,x_t,y_t),\theta_t^0-\hat{\theta}_t+\hat{\theta}_t-\theta_t^*\rangle\\ \le&2\beta_t\|\phi^-(c_t,x_t,y_t)\|_{\Sigma_t^{-1}},\end{aligned}$$

and

$$\begin{aligned}\langle \phi^-(c_t,x_t,y_t),\theta_t^*-\theta^*\rangle =&\langle \phi^-(c_t,x_t,y_t),\theta_t^*-\hat{\theta}_t+\hat{\theta}_t-\theta^*\rangle\\ \le&2\beta_t\|\phi^-(c_t,x_t,y_t)\|_{\Sigma_t^{-1}}.\end{aligned}$$

To sum up, we conclude that

$$\begin{aligned}&\text{Regret}(c_t,x)-\frac{Q_t}{V}\mathcal{W}(c_t,x)+\frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t,x_t)\\ \le&4\beta_t\|\phi^-(c_t,x_t,y_t)\|_{\Sigma_t^{-1}}/\mathbb{P}(\mathcal{E}_x)+4pT/\mathbb{P}(\mathcal{E}_x).\end{aligned}\tag{39}$$

Due to the symmetric property, the upper bound on $\text{Regret}(c_t,y)$ follows the same steps above. For the purpose of completeness, we include it in the next section.

E.1.2 Upper Bound on Regret(c_t,y)

We first add and subtract the term related to the cost $\widetilde{\mathcal{W}}(c_t,y_t)$ in Algorithm 2 (note y could be any action in $\mathcal{A}$ including the optimal y^*) that

$$\begin{aligned}\text{Regret}(c_t,y)=&\langle\phi(c_t,y,x_t),\theta^*-\theta_t^1\rangle+\langle\phi(c_t,y,y_t),\theta_t^1\rangle\\ =&\langle\phi(c_t,y,x_t),\theta^*-\theta_t^1\rangle+\langle\phi(c_t,y,y_t),\theta_t^1\rangle+\frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t,y_t)-\frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t,y_t)\end{aligned}$$

Recall y_t is the optimal solution to in Algorithm 2. Therefore, we have

$$\langle\phi(c_t,y),\theta_t^1\rangle-\frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t,y)\le\langle\phi(c_t,y_t),\theta_t^1\rangle-\frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t,y_t),\quad\forall y\in\mathcal{A},$$

which implies that

$$\langle\phi^-(c_t,y,y_t),\theta_t^1\rangle\le\frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t,y)-\frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t,y_t),\quad\forall y\in\mathcal{A}.\tag{40}$$

Therefore, $\text{Regret}(c_t,y)$ is bounded as follows

$$\begin{aligned}\text{Regret}(c_t,y)\le&\langle\phi(c_t,y,x_t),\theta^*-\theta_t^1\rangle+\frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t,y)-\frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t,y_t)\\ =&\langle\phi(c_t,y,x_t),\theta^*-\hat{\theta}_t+\hat{\theta}_t-\theta_t^1\rangle+\frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t,y)-\frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t,y_t)\\ \le&\|\phi(c_t,y,x_t)\|^2_{\Sigma_t^{-1}}+\langle\phi(c_t,y,x_t),\hat{\theta}_t-\theta_t^0\rangle+\frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t,y)-\frac{Q_t}{V}\mathcal{W}(c_t,y)\\ &+\frac{Q_t}{V}\mathcal{W}(c_t,y)-\frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t,y_t)\end{aligned}$$

where the first inequality holds by (40); the last inequality holds by holds because $\hat{\theta}_t$ is MLE of θ^* in Lemma (12). Since $\langle\phi^-(c_t,y,x_t),\hat{\theta}_t-\theta_t^1\rangle\sim\mathcal{N}(0,\beta_t^2\|\phi^-(c_t,y,x_t)\|^2_{\Sigma_t^{-1}})$ and $\langle\psi(c_t,y),\omega_t-\hat{\omega}_t\rangle\sim\mathcal{N}(0,\gamma_t^2\|\psi(c_t,y)\|^2_{\Psi_t^{-1}})$, we conclude that the following key event $\mathcal{E}_y$ happens

$$\text{Regret}(c_t,y)-\frac{Q_t}{V}\mathcal{W}(c_t,y)+\frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t,y_t)\le0\tag{41}$$

with probability $p_0=G^2(-1)/2$, where $G(\cdot)$ denotes the CDF of a standard normal distribution.

To proceed, we first define a key difference term

$$d_t^1 = \langle \phi(c_t, x_t) - \phi(c_t, y_t), \theta_t^1 \rangle,$$
$$d_t^* = \min_{\|\theta - \hat{\theta}_t\|_{\Sigma_t} \leq \sigma_t} \langle \phi(c_t, x_t) - \phi(c_t, y_t), \theta \rangle,$$

which are used to define the following event $\mathcal{D}_1$ such that $\mathcal{D}_1 = \{d_t^1 \geq d_t^*\}$. Note this event occurs with a high probability

$$\mathbb{P}(\mathcal{D}_1) = \mathbb{P}(\|\theta_t^1 - \hat{\theta}_t\|_{\Sigma_t} \leq \beta_t) \geq 1 - p.$$

Moreover, define $(\theta_t^0, \theta_t^1, \omega_t) \sim \mathbb{P}_\Gamma$ and $\tilde{\mathbb{P}}_{\Gamma^y} = \mathbb{P}_\Gamma / \mathbb{P}(\mathcal{E}_y)$ if the event $\mathcal{E}_y$ happens and $\tilde{\mathbb{P}}_{\Gamma^y} = 0$, otherwise. We then study the following key term

$$\begin{aligned}
&\text{Regret}(c_t, y) - \frac{Q_t}{V}\mathcal{W}(c_t, y) + \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t, y_t) \\
&= \langle \phi^-(c_t, y, x_t), \theta^* \rangle - d_t^1 - \frac{Q_t}{V}\mathcal{W}(c_t, y) + \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t, y_t) \\
&\leq \langle \phi^-(c_t, y, x_t), \theta^* \rangle - d_t^* - \frac{Q_t}{V}\mathcal{W}(c_t, y) + \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t, y_t) \\
&= \mathbb{E}_{\tilde{\mathbb{P}}_{\Gamma^y}}\left[\langle \phi(c_t, y^*, x_t), \theta^* \rangle - d_t^1 - \frac{Q_t}{V}\mathcal{W}(c_t, y^*) + \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t, y_t) + d_t^1 - d_t^*\right] \\
&\leq \mathbb{E}_{\tilde{\mathbb{P}}_{\Gamma^y}}\left[d_t^1 - d_t^*\right] \\
&= \mathbb{E}_{\mathbb{P}_\Gamma}[\mathbb{I}(\mathcal{E}_y)(d_t^1 - d_t^*)]/\mathbb{P}(\mathcal{E}_y) \\
&= \mathbb{E}_{\mathbb{P}_\Gamma}[\mathbb{I}(\mathcal{E}_y)\mathbb{I}(\mathcal{D}_1)(d_t^1 - d_t^*) + \mathbb{I}(\mathcal{E}_y)\mathbb{I}(\mathcal{D}_1^c)(d_t^1 - d_t^*)]/\mathbb{P}(\mathcal{E}_y) \\
&\leq \mathbb{E}_{\mathbb{P}_\Gamma}[\mathbb{I}(\mathcal{E}_y)\mathbb{I}(\mathcal{D}_1)(d_t^1 - d_t^*)]/\mathbb{P}(\mathcal{E}_y) + 4pT/\mathbb{P}(\mathcal{E}_y)
\end{aligned} \tag{42}$$

where the first inequality holds because of the definition of d_t^1 and d_t^*; the second inequality holds because we drop the negative term as in (41) under the event $\mathcal{E}_y$; the last inequality holds because the event $\mathcal{D}_1$ holds with a high probability $1 - p$. Now we study the first major term in (42) as follows

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}_\Gamma}[\mathbb{I}(\mathcal{E}_y)\mathbb{I}(\mathcal{D}_1)(d_t^1 - d_t^*)]/\mathbb{P}(\mathcal{E}_y) &\leq \mathbb{E}_{\mathbb{P}_\Gamma}[\mathbb{I}(\mathcal{D}_1)(d_t^1 - d_t^*)]/\mathbb{P}(\mathcal{E}_y) \\
&\leq \mathbb{E}_{\mathbb{P}_\Gamma}[d_t^1 - d_t^*]/\mathbb{P}(\mathcal{E}_y).
\end{aligned}$$

Let θ_t^* be the solution to minimizing d_t^*, we then have

$$\begin{aligned}
&\text{Regret}(c_t, y) - \frac{Q_t}{V}\mathcal{W}(c_t, y) + \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t, y_t) \\
\leq &\mathbb{E}[d_t^1 - d_t^*]/\mathbb{P}(\mathcal{E}_y) + 4pT/\mathbb{P}(\mathcal{E}_y) \\
\leq &\mathbb{E}[\langle \phi^-(c_t, x_t, y_t), \theta_t^1 - \theta_t^* \rangle - \langle \phi^-(c_t, x_t, y_t), \theta_t^* - \theta^* \rangle]/\mathbb{P}(\mathcal{E}_y) + 4pT/\mathbb{P}(\mathcal{E}_y).
\end{aligned}$$

For the first term above, we have

$$\langle \phi^-(c_t, x_t, y_t), \theta_t^1 - \theta_t^* \rangle = \langle \phi^-(c_t, x_t, y_t), \theta_t^1 - \hat{\theta}_t + \hat{\theta}_t - \theta_t^* \rangle \leq 2\beta_t \|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}},$$

and

$$\langle \phi^-(c_t, x_t, y_t), \theta_t^* - \theta^* \rangle = \langle \phi^-(c_t, x_t, y_t), \theta_t^* - \hat{\theta}_t + \hat{\theta}_t - \theta^* \rangle \leq 2\beta_t \|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}}.$$

To sum up, we conclude that

$$\begin{aligned}
&\text{Regret}(c_t, y) - \frac{Q_t}{V}\mathcal{W}(c_t, y) + \frac{Q_t}{V}\widetilde{\mathcal{W}}(c_t, y_t) \\
\leq &4\beta_t \|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}}/\mathbb{P}(\mathcal{E}_y) + 4pT/\mathbb{P}(\mathcal{E}_y).
\end{aligned} \tag{43}$$

E.1.3 Proving Lemma 4 under Algorithm 2

Finally, we combine (39) and (43) and establish that

$$\begin{aligned}&\text{Regret}(c_t, x) + \text{Regret}(c_t, y)\\ \leq&\frac{8\beta_t}{p_0}\|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}} + \frac{8pT}{p_0} + \frac{Q_t}{V}(\mathcal{W}(c_t, x) + \mathcal{W}(c_t, y) - \widetilde{\mathcal{W}}(c_t, x_t) - \widetilde{\mathcal{W}}(c_t, y_t)).\end{aligned}$$

Combine these two inequality above, we immediately have

$$\begin{aligned}&\text{Regret}(c_t, x) + \text{Regret}(c_t, y) + \frac{Q_t}{V}(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b)\\ \leq&\frac{8\beta_t}{p_0}\|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}} + \frac{8pT}{p_0} + \frac{Q_t}{V}(\mathcal{W}(c_t, x) + \mathcal{W}(c_t, y) - b).\end{aligned}$$

According to the virtual queue update

$$Q_{t+1} = \max\left(Q_t + \widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b, 0\right),$$

we have

$$\frac{1}{2}Q_{t+1}^2 - \frac{1}{2}Q_t^2 \leq Q_t(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b) + \frac{1}{2}(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b)^2.$$

Therefore, we establish a key inequality on the "Regret + Lyapunov drift" as follows

$$\begin{aligned}&\text{Regret}(c_t, x) + \text{Regret}(c_t, y) + \frac{Q_{t+1}^2}{2V} - \frac{Q_t^2}{2V}\\ \leq&\frac{8\beta_t}{p_0}\|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}} + \frac{8pT}{p_0} + \frac{1}{2V}(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b)^2\\ &+ \frac{Q_t}{V}(\mathcal{W}(c_t, x) + \mathcal{W}(c_t, y) - b).\end{aligned} \tag{44}$$

The proof of Lemma 4 is completed.

E.2 Proof of Lemma 5 under Algorithm 2

Since (x, y) could be any duel action in $\mathcal{A}$, let $(x, y) = (x^*, y^*)$ in (44). Now taking the conditional expectation $\mathbb{E}[\ \cdot\ |\mathcal{H}_t = h]$ on both sides of (44), we establish

$$\begin{aligned}&\mathbb{E}\left[\text{Regret}(c_t, x^*) + \text{Regret}(c_t, y^*)|\mathcal{H}_t = h\right] + \mathbb{E}\left[\frac{Q_{t+1}^2}{2V} - \frac{Q_t^2}{2V}|\mathcal{H}_t = h\right]\\ \leq&\mathbb{E}\left[\frac{12\beta_t}{p_0}\|\phi(c_t, x_t) - \phi(c_t, y_t)\|_{\Sigma_t^{-1}} + \frac{12pT}{p_0} + \frac{1}{2V}(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b)^2 \mid \mathcal{H}_t = h\right]\\ &+ \frac{Q}{V}\mathbb{E}\left[(\mathcal{W}(c_t, x^*) + \mathcal{W}(c_t, y^*) - b) \mid \mathcal{H}_t = h\right]\\ \leq&\mathbb{E}\left[\frac{12\beta_t}{p_0}\|\phi(c_t, x_t) - \phi(c_t, y_t)\|_{\Sigma_t^{-1}} + \frac{12pT}{p_0} + \frac{1}{2V}(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b)^2 \mid \mathcal{H}_t = h\right].\end{aligned} \tag{45}$$

where the last inequality holds because (x^*, y^*) satisfy the constraint in (5). We further take the expectation on both sides of (45) and then take summation from $t = 1$ to τ that

$$\begin{aligned}&\mathbb{E}\left[\text{Regret}(c_t, x^*) + \text{Regret}(c_t, y^*)\right] + \mathbb{E}\left[\frac{Q_{\tau+1}^2}{2V} - \frac{Q_1^2}{2V}\right]\\ \leq&\mathbb{E}\left[\frac{12\beta_t}{p_0}\|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}} + \frac{12pT}{p_0} + \frac{1}{2V}(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b)^2\right].\end{aligned}$$

Since $Q_1 = 0$ and $p = 1/T^2$, we conclude that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=1}^{\tau} \text{Regret}(c_t, x^*) + \text{Regret}(c_t, y^*)\right] \\
\leq &\mathbb{E}\left[\sum_{t=1}^{\tau} \frac{12\beta_t}{p_0}\|\phi^-(c_t, x_t, y_t)\|_{\Sigma_t^{-1}}\right] + \frac{12}{Tp_0} + \mathbb{E}\left[\sum_{t=1}^{\tau} \frac{(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b)^2}{2V}\right] \\
\leq &\frac{24\beta_T}{p_0}\sqrt{T\log(1+T)} + \frac{12}{Tp_0} + \mathbb{E}\left[\sum_{t=1}^{\tau} \frac{(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b)^2}{2V}\right]
\end{aligned}$$

where the second inequality holds by Lemma 14. Finally, we complete the proof by providing a *simple but refined analysis* on the cumulative budget consumption that

$$\mathbb{E}\left[\sum_{t=1}^{\tau} \frac{(\widetilde{\mathcal{W}}(c_t, x_t, y_t) - b)^2}{2V}\right] \leq \mathbb{E}\left[\sum_{t=1}^{\tau} \frac{\widetilde{\mathcal{W}}(c_t, x_t, y_t)^2 + b^2}{V}\right] \leq \frac{2d}{V}\mathbb{E}\left[\sum_{t=1}^{\tau} \widetilde{\mathcal{W}}(c_t, x_t, y_t)\right] + \frac{\tau b^2}{V},$$

where the second inequality holds because $\widetilde{\mathcal{W}}(c_t, x_t, y_t)$ is bounded by $2d$. Moreover, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{\tau} \widetilde{\mathcal{W}}(c_t, x_t, y_t)\right] =&\mathbb{E}\left[\sum_{t=1}^{\tau} \widetilde{\mathcal{W}}(c_t, x_t, y_t) - \mathcal{W}(c_t, x_t, y_t)\right] + \mathbb{E}\left[\sum_{t=1}^{\tau} \mathcal{W}(c_t, x_t, y_t)\right] \\
\leq &(2d + 6\gamma_T)\sqrt{dT\log(1+T)} + B,
\end{aligned}$$

where the last inequality holds because of Lemma 8 and the definition of stopping time τ. Therefore, combine with Lemma 3, we have

$$\begin{aligned}
&\mathbb{E}\left[\text{Regret}(\tau)\right] \\
\leq &\frac{24\beta_T}{p_0}\sqrt{T\log(1+T)} + 12 + \frac{(4d^2 + 12\gamma_T d)\sqrt{dT\log(1+T)} + 2dTb + 2Tb^2}{V}.
\end{aligned}$$

E.3 Proof of Lemma 6 under Algorithm 2

From (44), we have established the Lyapunov drift

$$\begin{aligned}
&\mathbb{E}\left[\frac{Q_{t+1}^2}{2} - \frac{Q_t^2}{2}|\mathcal{H}_t = h\right] \\
\leq &\mathbb{E}\left[V(\text{Regret}(c_t, x) + \text{Regret}(c_t, y)) + \frac{12\beta_t V}{p_0}\|\phi(c_t, x_t) - \phi(c_t, y_t)\|_{\Sigma_t^{-1}} + \frac{12pVT}{p_0} \mid \mathcal{H}_t = h\right] \\
&+ \mathbb{E}\left[\frac{1}{2}(\widetilde{\mathcal{W}}(c_t, x_t) + \widetilde{\mathcal{W}}(c_t, y_t) - b)^2 \mid \mathcal{H}_t = h\right] + Q \cdot \mathbb{E}\left[(\mathcal{W}(c_t, x) + \mathcal{W}(c_t, y) - b) \mid \mathcal{H}_t = h\right] \\
\leq &4V + \frac{12\beta_t V}{p_0} + \frac{12V}{Tp_0} + Q \cdot \mathbb{E}\left[(\mathcal{W}(c_t, x) + \mathcal{W}(c_t, y) - b) \mid \mathcal{H}_t = h\right] \\
\leq &-\delta b \cdot Q + 4V + \frac{14\beta_t V}{p_0}
\end{aligned} \tag{46}$$

where the second inequality holds due to the bounded rewards and costs in Assumption 2 and $\lambda = 1$; the last inequality holds because and the "Slater condition" in Assumption 3 that there exists a feasible policy such that

$$\mathbb{E}\left[w(c_t, x) + w(c_t, y) - b \mid \mathcal{H}_t = h\right] \leq -\delta b.$$

Since we have the negative drift in (46), we establish the high probability bound on the virtual queue Q_t according to the Lyapunov drift lemma in Lemma 15.

E.3.1 High Probability Bound

We define the Lyapunov function $\bar{L}_t = Q_t$. To establish the high probability bound on the virtual queue, we prove conditions (i) and (ii) in Lemma 15 for $\bar{L}_t$.

Given $\mathcal{H}_t = h$ and $\bar{L}_t \geq \varphi_t = \frac{8V + \frac{28\beta_t V}{p_0}}{\delta b}$, the conditional expected drift of $\bar{L}_t$ is

$$\begin{aligned}
&\mathbb{E}[Q_{t+1} - Q_t | \mathcal{H}_t = h] \leq \frac{1}{2Q}\mathbb{E}[Q_{t+1}^2 - Q_t^2 | \mathcal{H}_t = h] \\
\leq &\frac{-\delta b \cdot Q + 4V + \frac{14\beta_t V}{p_0}}{2Q} \\
\leq &-\frac{\delta b}{2} + \frac{2V + \frac{7\beta_t V}{p_0}}{Q} \\
\leq &-\frac{\delta b}{4}
\end{aligned}$$

where the first inequality holds because $2(Q_{t+1} - Q_t)Q_t \leq Q_{t+1}^2 - Q_t^2$; the second inequality holds by the "negative drift" in 31 above; and the last inequality holds given the condition $Q \geq \varphi_t = \frac{8V + \frac{28\beta_t V}{p_0}}{\delta b}$. Moreover, for condition (ii) in Lemma 15, we have $Q_{t+1} - Q_t \leq 2d + b$.

Let $\rho = \frac{\delta b}{4}$, and $\nu_{\max} = 2d + b$. We are ready to apply Lemma 15 for $\bar{L}(t)$ and obtain

$$\mathbb{E}\left[e^{\zeta Q_t}\right] \leq 1 + \frac{2e^{\zeta(\nu_{\max} + \varphi_t)}}{\zeta\rho} \text{ with } \zeta = \frac{\rho}{\nu_{\max}^2 + \nu_{\max}\rho/3}. \tag{47}$$

We then establish the high probability bound as follows

$$\begin{aligned}
&\mathbb{P}\left(Q_t \geq \nu_{\max} + \varphi_t + \frac{\log\frac{6T}{\delta\rho}}{\zeta}\right) \leq \mathbb{P}\left(e^{\zeta Q_t} \geq e^{\zeta(\nu_{\max} + \varphi_t) + \log\frac{6T}{\delta\rho}}\right) \\
\leq &\frac{\mathbb{E}\left[e^{\zeta Q_t}\right]}{e^{\zeta(\nu_{\max} + \varphi_t) + \log\frac{6T}{\delta\rho}}} \leq \frac{1 + \frac{2e^{\zeta(\nu_{\max} + \varphi_t)}}{\zeta\rho}}{e^{\zeta(\nu_{\max} + \varphi_t) + \log\frac{6T}{\delta\rho}}} \leq 1/T^2,
\end{aligned} \tag{48}$$

where the second inequality holds by Markov inequality and the third inequality holds by (47). Therefore, there exists an absolute constant C_1 such that $\frac{C_1\beta_t V}{\delta b} \geq \nu_{\max} + \varphi_t + \frac{\log\frac{6T}{\delta\rho}}{\zeta}$ and the high probability bound holds

$$\mathbb{P}\left(Q_t > \frac{C_1\beta_t V}{\delta b}\right) \leq 1/T^2.$$

E.4 Proof of Lemma 7 under Algorithm 2

Recall the definition of virtual stopping time

$$\tau_0 = \text{argmin}_{\tau' \in [T]} \left\{\tau' \mid Q_{\tau'+1} + b\tau' + M_\tau \geq B\right\},$$

where $M_\tau = \sum_{t=1}^{\tau}(\mathcal{W}(c_t, x_t, y_t) - \widetilde{\mathcal{W}}(c_t, x_t, y_t))$. We have established both upper bounds of Q_τ and M_τ in Lemma 6 and Lemma 8, respectively. Therefore, we have

$$\frac{C_1\beta_t V}{\delta b} + (2d + 6\gamma_\tau)\sqrt{d\tau\log(1 + \tau/d)} + b\tau_0 \geq B.$$

Recall the definition of $b = B/T$ and by dividing b on both sides of the inequality, we establish the following upper bound on the remaining rounds

$$\frac{C_1\beta_T V}{\delta b^2} + \frac{8d\gamma_T\sqrt{\tau\log(1 + \tau)}}{b} \geq T - \tau_0, \tag{49}$$

holds with a high probability.

E.5 Proof of Theorem 1 under Algorithm 2

Now we aggregate the regret after stopping and before stopping as follows

$$\begin{aligned}\text{Regret}(T) \leq & \underbrace{2\nu^*\mathbb{E}[T-\tau]}_{\text{regret after stopping}} + \underbrace{\mathbb{E}\left[\sum_{t=1}^{\tau}\mathcal{R}(c_t,x^*,y^*)\right]-\mathbb{E}\left[\sum_{t=1}^{\tau}\mathcal{R}(c_t,x_t,y_t)\right]}_{\text{regret before stopping}} \\ \leq & \frac{24\beta_T}{p_0}\sqrt{T\log(1+T)}+12+\frac{(4d^2+12\gamma_T d)\sqrt{dT\log(1+T)}+2dTb+2Tb^2}{V} \\ & +\left(\frac{C_1\beta_T V}{\delta b^2}+\frac{8d\gamma_T\sqrt{\tau\log(1+\tau)}}{b}\right)\nu^* \end{aligned} \tag{50}$$

Let $V=b\sqrt{T}$ in (50), we finally have

$$\text{Regret}(T)\leq \frac{40\beta_T d}{p_0}\sqrt{T\log(1+T)}+12+2(d+b)\sqrt{T}+\frac{C_1\beta_T\sqrt{T}+8d\gamma_T\sqrt{\tau\log(1+\tau)}}{\delta}\frac{\nu^*}{b}.$$

Recall $\beta_T=2\kappa d+2\kappa\sqrt{4\log T+2d\log(1+T/\kappa d)}$ and $\gamma_T=d+\sqrt{4\log T+2d\log(1+T/d)}$ and the proof of Theorem 1 is completed.

F Supporting Lemmas

F.1 Maximum likelihood Estimation Errors

We introduce maximum likelihood estimation (MLE) errors for learning the reward and cost parameters. These can be used to define the confidence set in Algorithms 1 and 2.

Recall the key quantities

$$\begin{aligned}\kappa &= \sup_{\|\theta\|\leq d,\|\phi\|\leq 1} 1/\dot{\sigma}(\langle\theta,\phi\rangle), \\ \Sigma_t &= \kappa\lambda\mathbf{I}+\sum_{s=1}^{t-1}\phi^-(c_s,x_s,y_s)\phi^-(c_s,x_s,y_s)^\dagger, \\ \Psi_t &= \lambda\mathbf{I}+\sum_{s=1}^{t-1}\psi(c_s,x_s)\psi(c_s,x_s)^\dagger+\psi(c_s,y_s)\psi(c_t,y_s)^\dagger.\end{aligned}$$

Lemma 12 (Lemma 1 in Faury et al. (2020))**.** *For any $p>0$, with probability at least $1-p$, the following event occurs*

$$||\hat{\theta}_t-\theta_*||_{\Sigma_t}\leq\beta_t^r(p),\ \forall t\geq 1,$$

where $\beta_t^r(p)=2\kappa d\sqrt{\lambda}+2\kappa\sqrt{2\log(1/p)+2d\log(1+t/\kappa\lambda d)}$.

Lemma 13 (Theorem 2 in Abbasi-yadkori et al. (2011))**.** *For any $p>0$, with probability at least $1-p$, the following event occurs*

$$||\hat{\omega}_t-\omega_*||_{\Psi_t}\leq\beta_t^w(p),\ \forall t\geq 1,$$

where $\beta_t^w(p)=d\sqrt{\lambda}+\sqrt{2\log(1/p)+2d\log(1+t/\lambda d)}$.

Proving Lemma 1: From Lemmas 12 and 13, Lemma 1 is completed by using union bound.

F.2 Elliptical Potential Lemma

We introduce the elliptical potential lemma (Theorem 11.7 in Cesa-Bianchi and Lugosi (2006)) to bound the cumulative bonus $\sum_{s=1}^{t}\min\left(||\phi(c_s,x)||_{\Sigma_{t-1}^{-1}},1\right)$.

Lemma 14. *Let $\Sigma_0 = \lambda \mathbf{I}$ and $\phi_0, \phi_1, \cdots, \phi_{t-1} \in \mathbb{R}^d$ be a sequence of vectors with $||\phi_t|| \leq 1$ for any t and $\Sigma_t = \lambda \mathbf{I} + \sum_{s=1}^{t} \phi_s \phi_s^\dagger$. Then,*

$$\sum_{s=1}^{t} \min\left(1, ||\phi_s||^2_{\Sigma_{s-1}^{-1}}\right) \leq 2 \log\left(\frac{\det \Sigma_t}{\det \Sigma_0}\right) \leq 2d \log\left(\frac{\lambda d + t}{\lambda d}\right).$$

F.3 Lyapunov Drift Lemma

We present a lemma which will be used to derive the high probability of $\{Q_t\}$. The lemma is from Liu et al. (2021), which is a minor variation of Lemma 4.1 Neely (2016) and the results in Hajek (1982), where the radius of φ_t could be time dependent.

Lemma 15. *Let $S(t)$ be a random process, $\Phi(t)$ be its Lyapunov function with $\Phi(0) = \Phi_0$ and $\Delta(t) = \Phi(t+1) - \Phi(t)$ be the Lyapunov drift. Given an increasing sequence $\{\varphi_t\}$, ρ and $\nu_{\max}$ with $0 < \rho \leq \nu_{\max}$, if the expected drift $\mathbb{E}[\Delta(t)|S(t) = s]$ satisfies the following conditions:*

(i) There exists constants $\rho > 0$ and $\varphi_t > 0$ such that $\mathbb{E}[\Delta(t)|S(t) = s] \leq -\rho$ when $\Phi(t) \geq \varphi_t$, and

(ii) $|\Phi(t+1) - \Phi(t)| \leq \nu_{\max}$ holds with probability one;

then we have

$$\mathbb{E}[e^{\zeta \Phi(t)}] \leq e^{\zeta \Phi_0} + \frac{2e^{\zeta(\nu_{\max} + \varphi_t)}}{\zeta \rho}, \tag{51}$$

where $\zeta = \frac{\rho}{\nu_{\max}^2 + \nu_{\max}\rho/3}$.