CONTEXTUAL COMBINATORIAL BANDITS WITH POST-SERVING CONTEXTS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study contextual combinatorial bandits with post-serving contexts (C^{2}MAB-PC), where the learner selects K of m base arms using their pre-serving contexts, while additional contexts that also determine the rewards, such as follow-up engagement in slate recommendation, are revealed only after the selected arms are served. We first propose OFUC^2-PC, which is jointly optimistic over the reward parameter and the mean post-serving contexts, and analyze it with a new noisy batched truncated elliptical potential lemma (NBT-EPL). OFUC^2-PC achieves a leading regret of $\widetilde{\mathcal{O}}(m^\alpha(KT)^{1-\alpha}d_x^\alpha + d_u\sqrt{KT})$, where $\alpha \in (0, 1/2]$ is the learnability rate of the post-serving mappings and d_x and d_u are the pre-serving and augmented context dimensions, and a minimax lower bound matches this rate for additive rewards and $\alpha = 1/2$. We then propose C^{2}UCB-PC, an oracle-efficient alternative with the same leading reward-learning term, and variance-aware versions of both algorithms that improve this term from $\widetilde{\mathcal{O}}(\sqrt{KT})$ to $\widetilde{\mathcal{O}}(\sqrt{T})$ under variance-modulated smoothness. Experiments on synthetic and MovieLens environments show that our algorithms incur lower regret than the baselines and that the oracle-efficient algorithms are 18–36$\times$ faster than the joint ones.

1 INTRODUCTION

Combinatorial multi-armed bandit (CMAB) is a fundamental decision-making problem that extends the multi-armed bandit and has been extensively studied in the last two decades (Gai et al., 2012; Chen et al., 2013; Kveton et al., 2015; Combes et al., 2015; Chen et al., 2016). In each round of CMAB, the learner selects a subset of base arms (i.e., a super arm) to be pulled simultaneously. Once the super arm is pulled, each selected base arm, following some unknown distribution, will generate a random outcome as the feedback (i.e., semi-bandit feedback). Then, the learner receives a reward, which can be a general function of the pulled arms' outcomes, with the summation as the most common function. The agent's goal is to minimize the *regret*, which is the difference in cumulative rewards between always selecting the best super arm and following the agent's own policy. Motivated by large-scale applications with a huge number of base arms, recent work has developed contextual combinatorial bandits (C^{2}MAB) (Qin et al., 2014; Li et al., 2016; Takemura et al., 2021). C^{2}MAB associates each base arm with a context vector and adopts the linear bandit parameterization, so that the learner can exploit a linear structure shared across base arms instead of estimating each arm independently (Wen et al., 2015). Building on this, subsequent works enrich the model at two levels. At the base-arm level, they move beyond linear models to logistic (Liu et al., 2025) and neural (Hwang et al., 2023) parameterizations. At the super-arm level, they cover general nonlinear reward functions through smoothness conditions (Liu et al., 2023).

Despite their improved scalability and generality, the above C^{2}MAB frameworks all assume that the mean reward of a base arm is determined by contexts observed before the super arm is selected (a more detailed related work discussion in Appendix A). In many applications, however, part of the information that drives the reward is revealed only after the selected arms are served, which is referred to as *post-serving contexts* (Wang et al., 2023; Oh, 2026). Consider a streaming platform that recommends a slate of K movies to an arriving user. User demographics and movie attributes (pre-serving) are known before the slate is shown, whereas the follow-up engagement (post-serving) with each recommended movie, such as whether the user opens its detail page and how much of the movie is watched, is observed only after serving and is strongly tied to the final rating. Similarly,

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

Table 1: Comparison of the proposed algorithms with the related works.

Algorithm	Comb. action	Post-serving	Reward function	Reward regret	Context regret	Comp. efficient
poLinUCB (Wang et al., 2023)	×	✓	–	$\widetilde{\mathcal{O}}\big(d_u\sqrt{T}\big)$	$\widetilde{\mathcal{O}}(\sqrt{m}\,T^{1-\alpha}d_u^{\alpha})$	✓
C^2-UCB-T (Liu et al., 2023)	✓	×	1-norm smoothness	$\widetilde{\mathcal{O}}\big(B_1 d_x\sqrt{KT}\big)$	–	✓
VAC^2-UCB (Liu et al., 2023)	✓	×	VM smoothness	$\widetilde{\mathcal{O}}\big(B_v d_x\sqrt{T}\big)$	–	✓
OFUC^2-PC (ours)	✓	✓	1-norm smoothness	$\widetilde{\mathcal{O}}\big(B_1 d_u\sqrt{KT}\big)$	$\widetilde{\mathcal{O}}\big(B_1 m^{\alpha}(KT)^{1-\alpha}d_x^{\alpha}\big)$	×
C^{2}UCB-PC (ours)	✓	✓	1-norm smoothness	$\widetilde{\mathcal{O}}\big(B_1 d_u\sqrt{KT}\big)$	$\widetilde{\mathcal{O}}\big(B_1(1+\zeta_T/\sqrt{\lambda})\,m^{\alpha}(KT)^{1-\alpha}d_x^{\alpha}\big)$	✓
VA-OFUC^2-PC (ours)	✓	✓	VM smoothness	$\widetilde{\mathcal{O}}\big(B_v d_u\sqrt{T}\big)$	$\widetilde{\mathcal{O}}\big(B_1 m^{\alpha}(KT)^{1-\alpha}d_x^{\alpha}\big)$	×
VA-C^{2}UCB-PC (ours)	✓	✓	VM smoothness	$\widetilde{\mathcal{O}}\big(B_v d_u\sqrt{T}\big)$	$\widetilde{\mathcal{O}}\big(B_1(1+\rho_T^V/\sqrt{\lambda_V})\,m^{\alpha}(KT)^{1-\alpha}d_x^{\alpha}\big)$	✓

The reward-regret column reports the leading reward-learning term, and lower-order terms are omitted. The context regret is shown for a general learnability rate $\alpha \in (0, 1/2]$ (Assumption 1). B_1 and B_v are the coefficients of the 1-norm and the variance-modulated (VM) bounded smoothness conditions (Sec. 2.3). $\zeta_T = \widetilde{\mathcal{O}}(\sqrt{\lambda} + \sqrt{d_u})$ and $\rho_T^V = \widetilde{\mathcal{O}}(\sqrt{\lambda_V} + \sqrt{d_u})$ are reward-confidence radius parameters, so the factors $1+\zeta_T/\sqrt{\lambda}$ and $1+\rho_T^V/\sqrt{\lambda_V}$ are logarithmic for regularization parameters $\lambda, \lambda_V = \Theta(d_u)$. For poLinUCB, we report the shared-reward adaptation with arm-specific post-serving mappings.

for an LLM serving platform that routes each query to a set of K large language models (LLMs), query features and model attributes (pre-serving) are known before the routing, whereas the generation behavior (post-serving), such as the response latency, is observed only after the responses are produced and is strongly related to the quality of the response. A learner that ignores such post-serving contexts works with a misspecified reward model and can suffer linear regret even when a single arm is pulled per round (Wang et al., 2023). This motivates the following question: *Can we design a contextual Combinatorial MAB framework, together with provably efficient algorithms, for large-scale applications with post-serving contexts?*

Our main results are summarized in Table 1. First, we propose OFUC^2-PC, which is jointly optimistic over the shared reward parameter and mean post-serving contexts. Its analysis faces two challenges: (i) the design matrix is built from noisy realized contexts, whereas the regret is measured at the mean contexts, and (ii) the K arms of a round are selected before its feedback arrives. The generalized elliptical potential lemma (EPL) of Wang et al. (2023) handles (i), under a non-degenerate noise covariance condition, and the truncated EPL of Takemura et al. (2021) handles (ii). Inspired by both, we develop a noisy batched truncated EPL (NBT-EPL) that handles both without this condition. With NBT-EPL, OFUC^2-PC achieves a regret of $\widetilde{\mathcal{O}}(B_1 m^{\alpha}(KT)^{1-\alpha}d_x^{\alpha} + B_1 d_u\sqrt{KT} + B_1 d_u K)$, where $\alpha \in (0, 1/2]$ is the learnability rate of post-serving mappings, m is the number of base arms, and d_x and d_u are the pre-serving and augmented context dimensions. The leading reward-learning term matches C^{2}MAB without post-serving contexts (Liu et al., 2023), with d_u in place of d_x, so the first term is the price of post-serving contexts. For additive rewards with $\alpha = 1/2$, this bound is minimax optimal up to logarithmic factors.

Second, the joint optimism of OFUC^2-PC is a bilinear search coupled with the combinatorial action space, which is computationally expensive. We therefore propose C^{2}UCB-PC, which computes an optimistic score for each base arm and selects the super arm with one call to a combinatorial optimization oracle. C^{2}UCB-PC preserves the leading reward-learning term of OFUC^2-PC and is more than an order of magnitude faster in our experiments.

Third, under the variance-modulated smoothness condition of Liu et al. (2022), variance-aware versions of both algorithms, built on variance-weighted estimation and a variance-scaled NBT-EPL, improve the leading reward-learning term from $\widetilde{\mathcal{O}}(B_1 d_u\sqrt{KT})$ to $\widetilde{\mathcal{O}}(B_v d_u\sqrt{T})$, which matches the variance-aware C^{2}MAB baseline.

Finally, experiments on synthetic and MovieLens-based semi-synthetic environments show that our four algorithms incur lower regret than the baselines, with VA-OFUC^2-PC the lowest, and that the oracle-efficient algorithms are 18–36× faster than the joint ones.

2 Contextual Combinatorial MAB with Post-Serving Contexts

In this section, we introduce the framework of contextual combinatorial multi-armed bandits with post-serving contexts (C^{2}MAB-PC). We first present the problem formulation of C^{2}MAB-PC, then

introduce two learnability conditions for the post-serving mappings, and finally state the conditions on the combinatorial reward function.

Notation. All vectors are column vectors and are written in bold, and $[\boldsymbol{a}, \boldsymbol{b}]$ denotes the column vector that stacks $\boldsymbol{a}$ on top of $\boldsymbol{b}$. For a positive semidefinite matrix A, let $\|\boldsymbol{x}\|_A = \sqrt{\boldsymbol{x}^\top A \boldsymbol{x}}$, and let $\|\boldsymbol{x}\|_2$ denote the Euclidean norm. We use I for the identity matrix, $\mathbf{1}\{\cdot\}$ for the indicator function, $a \wedge b$ for $\min\{a, b\}$, and $A \succeq B$ for the positive semidefinite order. The notation $\widetilde{\mathcal{O}}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ hides logarithmic factors.

2.1 Problem Formulation

We specify a C^{2}MAB-PC problem by a tuple $\mathcal{I} := ([m], \mathcal{S}, d_x, d_z, \{\phi_i^\star\}_{i=1}^m, \boldsymbol{\theta}^\star, \boldsymbol{\beta}^\star, \{F_t\}_{t=1}^T)$, where $[m]$ is the set of base arms, $\mathcal{S}$ is the set of feasible combinatorial actions, d_x and d_z are the dimensions of the pre-serving and post-serving contexts, $\{\phi_i^\star\}_{i=1}^m$ are the unknown arm-specific post-serving mappings, $\boldsymbol{\theta}^\star$ and $\boldsymbol{\beta}^\star$ are the unknown shared reward parameters, and F_t is the combinatorial reward function at round t. The details of each component are described below.

Base arms. Let $[m] = \{1, \dots, m\}$ denote the set of base arms.

Combinatorial actions. Let $\mathcal{S}$ be a collection of feasible subsets of $[m]$. We use $K = \max_{S \in \mathcal{S}} |S|$ to denote the maximum number of arms that can be selected in one round.

Pre-serving and post-serving contexts. At round t, the learner first observes a pre-serving context $\boldsymbol{x}_{t,i} \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ for every arm $i \in [m]$ and selects a combinatorial action $S_t \in \mathcal{S}$. Only after the decision is made, each selected arm $i \in S_t$ reveals a post-serving context

$$\boldsymbol{z}_{t,i} = \phi_i^\star(\boldsymbol{x}_{t,i}) + \boldsymbol{\epsilon}_{t,i}, \tag{1}$$

where $\phi_i^\star : \mathbb{R}^{d_x} \to \mathbb{R}^{d_z}$ is an unknown arm-specific mapping that has to be learned and $\boldsymbol{\epsilon}_{t,i}$ is a zero-mean noise vector. For example, in movie recommendation, $\boldsymbol{x}_{t,i}$ may contain information known before serving, such as user demographics and movie attributes, whereas $\boldsymbol{z}_{t,i}$ may contain the follow-up engagement observed after the recommendation, such as how much of the movie is watched. We use $\boldsymbol{u}_{t,i} = [\boldsymbol{x}_{t,i}, \boldsymbol{z}_{t,i}]$ to denote the realized augmented context and $\bar{\boldsymbol{u}}_{t,i} = [\boldsymbol{x}_{t,i}, \phi_i^\star(\boldsymbol{x}_{t,i})]$ to denote the mean augmented context, both of dimension $d_u = d_x + d_z$.

Reward function. We assume that all arms share a common unknown parameter $\boldsymbol{w}^\star = [\boldsymbol{\theta}^\star, \boldsymbol{\beta}^\star] \in \mathbb{R}^{d_u}$. For a selected base arm $i \in S_t$, the realized and the mean base-arm rewards are

$$r_{t,i} = \boldsymbol{u}_{t,i}^\top \boldsymbol{w}^\star + \eta_{t,i}, \qquad \mu_{t,i} = \bar{\boldsymbol{u}}_{t,i}^\top \boldsymbol{w}^\star, \tag{2}$$

where $\eta_{t,i}$ is the reward noise. Let $\boldsymbol{\mu}_t = (\mu_{t,1}, \dots, \mu_{t,m})^\top$ denote the vector of mean base-arm rewards. For a combinatorial action $S \in \mathcal{S}$, let $F_t(S; \boldsymbol{\mu}_t)$ denote its expected reward, where F_t is known to the learner at decision time. For ease of exposition, we use the optimal combinatorial action $S_t^\star \in \arg\max_{S \in \mathcal{S}} F_t(S; \boldsymbol{\mu}_t)$ as the benchmark in the regret definition, which can be easily extended to an α_{o}-approximation benchmark when $S_t^\star$ is hard to compute as in Chen et al. (2013).

Observations. After selecting S_t, the learner observes the semi-bandit feedback $\{(\boldsymbol{z}_{t,i}, r_{t,i}) : i \in S_t\}$. Note that all selected-arm observations are used to estimate the shared reward parameter $\boldsymbol{w}^\star$, whereas only the observations of arm i are used to estimate its own mapping $\phi_i^\star$.

Learning objective. We define the cumulative pseudo-regret as

$$R_T = \textstyle\sum_{t=1}^T \big(F_t(S_t^\star; \boldsymbol{\mu}_t) - F_t(S_t; \boldsymbol{\mu}_t)\big). \tag{3}$$

Since S_t depends on the random feedback of earlier rounds and on any internal randomization of the learner, R_T is random, and our upper bounds hold with high probability.

Standard regularity conditions. Let $\mathcal{F}_t^{\mathrm{pre}}$ denote all observations before round t, the pre-serving contexts $\{\boldsymbol{x}_{t,i}\}_{i=1}^m$, and combinatorial action S_t, and let $\mathcal{F}_t^{\mathrm{post}}$ additionally contain the post-serving contexts $\{\boldsymbol{z}_{t,i} : i \in S_t\}$ revealed in round t. Conditioned on $\mathcal{F}_t^{\mathrm{pre}}$, the post-serving noises $\{\boldsymbol{\epsilon}_{t,i} : i \in S_t\}$ are zero-mean and independent across the selected arms. Conditioned on $\mathcal{F}_t^{\mathrm{post}}$, the reward noises $\{\eta_{t,i} : i \in S_t\}$ are zero-mean, mutually independent, and R_η-sub-Gaussian. We further assume $|\mu_{t,i}| \le B$, $\max\{\|\boldsymbol{u}_{t,i}\|_2, \|\bar{\boldsymbol{u}}_{t,i}\|_2\} \le L_u$, and $\boldsymbol{w}^\star \in \mathcal{W} := \{[\boldsymbol{\theta}, \boldsymbol{\beta}] \in \mathbb{R}^{d_u} : \|\boldsymbol{\theta}\|_2 \le 1,\ \|\boldsymbol{\beta}\|_2 \le 1\}$, where B and R_η are treated as constants and $L_u^2 = \Theta(d_u)$. We set $\lambda = \max\{1, L_u^2\} = \Theta(d_u)$. The formal definitions are given in Appendix B.

2.2 Learnability of Post-Serving Contexts

The effect of post-serving uncertainty depends on how quickly each mapping $\phi_i^\star$ can be estimated from the observed $(\boldsymbol{x}, \boldsymbol{z})$ pairs. Let $\widehat{\phi}_{t,i}$ denote the estimate of $\phi_i^\star$ computed from the pairs $\{(\boldsymbol{x}_{s,i}, \boldsymbol{z}_{s,i}) : s \le t,\ i \in S_s\}$ observed for arm i up to round t, e.g., by regularized least squares or empirical risk minimization. We consider two alternative learnability conditions on $\widehat{\phi}_{t,i}$, inspired by Wang et al. (2023) and Oh (2026), respectively. In the first, the prediction error at a context depends on how well this context is covered by the contexts previously observed for the arm. In the second, it depends only on how many observations the arm has collected.

For the first condition, let $X_{t,i} = \lambda I + \sum_{s<t} \mathbf{1}\{i \in S_s\} \boldsymbol{x}_{s,i} \boldsymbol{x}_{s,i}^\top$ denote the design matrix of the pre-serving contexts observed for arm i.

Assumption 1 (Elliptical learnability, Wang et al., 2023)**.** *There exist $\alpha \in (0, 1/2]$ and $C_0 > 0$ such that, for any arm $i \in [m]$, $\delta_\phi \in (0,1)$, and $t \ge 1$, with probability at least $1 - \delta_\phi$, the following holds simultaneously for all $\boldsymbol{x} \in \mathcal{X}$,*

$$\|\widehat{\phi}_{t-1,i}(\boldsymbol{x}) - \phi_i^\star(\boldsymbol{x})\|_2 \le C_0 \left(1 \wedge \|\boldsymbol{x}\|^2_{X_{t-1,i}^{-1}}\right)^{\alpha} \log\left(\tfrac{t}{\delta_\phi}\right). \tag{4}$$

For the second condition, let $n_{t-1}(i) = \sum_{s<t} \mathbf{1}\{i \in S_s\}$, and let $\bar{n}_{t-1}(i) = \max\{1, n_{t-1}(i)\}$, which avoids a zero sample size for arms that have not been selected yet.

Assumption 2 (Sample-size learnability, adapted from Oh, 2026)**.** *There exist $a \in (0, 1/2]$ and $C_0 > 0$ such that, for any arm $i \in [m]$, $\delta_\phi \in (0,1)$, and $t \ge 1$, with probability at least $1 - \delta_\phi$,*

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \left\|\widehat{\phi}_{t-1,i}(\boldsymbol{x}) - \phi_i^\star(\boldsymbol{x})\right\|_2 \le C_0 \sqrt{d_x}\, \bar{n}_{t-1}(i)^{-a} \log\big(m \bar{n}_{t-1}(i)/\delta_\phi\big). \tag{5}$$

Remark 1 (Comparison of learnability conditions). Assumption 1 is geometry-adaptive, since the prediction error shrinks in the directions that have been well explored in the arm-specific design matrix $X_{t,i}$. Assumption 2 abstracts away this geometry and measures learnability only through the number of observations of arm i. The two assumptions are alternatives, and neither is meant to imply the other. A canonical case covered by Assumption 1 is the arm-specific linear mapping $\phi_i^\star(\boldsymbol{x}) = M_i^\star \boldsymbol{x}$, for which the least-squares estimator satisfies equation 4 with $\alpha = 1/2$. Smaller values of α and a correspond to mappings that are harder to learn, such as nonparametric classes.

2.3 Smoothness Conditions for Reward Functions

Since the reward function F_t can be nonlinear and combinatorial, some conditions on F_t are needed to obtain meaningful regret bounds. Following standard conditions in combinatorial bandits (Chen et al., 2013; Wang & Chen, 2017; Liu et al., 2022), we assume that $F_t(S; \boldsymbol{\mu})$ satisfies the following.

Assumption 3 (Monotonicity)**.** *For any combinatorial action $S \in \mathcal{S}$ and any mean vectors $\boldsymbol{\mu}, \boldsymbol{\mu}' \in [-B, B]^m$ with $\mu_i \le \mu_i'$ for all $i \in [m]$, we have $F_t(S; \boldsymbol{\mu}) \le F_t(S; \boldsymbol{\mu}')$.*

Assumption 4 (1-norm bounded smoothness)**.** *There exists a constant $B_1 > 0$ such that, for any combinatorial action $S \in \mathcal{S}$ and any mean vectors $\boldsymbol{\mu}, \boldsymbol{\mu}' \in [-B, B]^m$, we have $|F_t(S; \boldsymbol{\mu}') - F_t(S; \boldsymbol{\mu})| \le B_1 \sum_{i \in S} |\mu_i' - \mu_i|$.*

Assumption 5 (Variance-modulated bounded smoothness, Liu et al., 2022)**.** *In the variance-aware setting, we assume $\mu_{t,i} \in [\nu, 1-\nu]$ for some $\nu \in (0, 1/2)$, and that there exist constants $B_v, B_1 > 0$ such that, for any combinatorial action $S \in \mathcal{S}$, any mean vectors $\boldsymbol{\mu}, \boldsymbol{\mu}' \in [\nu, 1-\nu]^m$, and any $\boldsymbol{\zeta}, \boldsymbol{\kappa} \in [-1, 1]^m$ with $\boldsymbol{\mu}' = \boldsymbol{\mu} + \boldsymbol{\zeta} + \boldsymbol{\kappa}$, $|F_t(S; \boldsymbol{\mu}') - F_t(S; \boldsymbol{\mu})| \le B_v \sqrt{\sum_{i \in S} \frac{\zeta_i^2}{\mu_i(1-\mu_i)}} + B_1 \sum_{i \in S} |\kappa_i|$.*

Remark 2 (Relation between the smoothness conditions). Monotonicity states that the reward of any action does not decrease when the base-arm means increase. Assumptions 4 and 5 both bound how much the reward can change when the means change. Assumption 5 implies the corresponding 1-norm bounded-smoothness bound by setting $\boldsymbol{\zeta} = \mathbf{0}$ and $\boldsymbol{\kappa} = \boldsymbol{\mu}' - \boldsymbol{\mu}$. Conversely, by the Cauchy–Schwarz inequality and $\mu_i(1-\mu_i) \le 1/4$, any reward function that satisfies Assumption 4 also satisfies Assumption 5 with $B_v = B_1\sqrt{K}/2$. The first term of Assumption 5 weights each perturbation ζ_i by the inverse of the variance proxy $\mu_i(1-\mu_i)$, allowing variance-aware analysis to replace $B_1\sqrt{K}$ with B_v in regret. Thus, the variance-modulated condition improves the leading reward-learning dependence whenever $B_v = o(B_1\sqrt{K})$, e.g., $B_v = O(1)$ for the cascading reward $F_t(S; \boldsymbol{\mu}) = 1 - \prod_{i \in S}(1 - \mu_i)$ (Liu et al., 2022).

Algorithm 1 OFUC^2-PC

Require: Set of combinatorial actions $\mathcal{S}$, regularization parameter λ, confidence levels δ_w, δ_ϕ

1: Initialize $A_0 = \lambda I$, $\boldsymbol{b}_0 = 0$, and $\widehat{\phi}_{0,i}$ for all $i \in [m]$
2: **for** $t = 1, \ldots, T$ **do**
3: Observe the pre-serving contexts $\{\boldsymbol{x}_{t,i}\}_{i=1}^m$ of all base arms
4: Construct $\mathcal{C}_{t-1}^w$ and $\{\mathcal{C}_{t-1}^\phi(i, \boldsymbol{x}_{t,i})\}_{i=1}^m$
5: Select $S_t \in \arg\max_{S \in \mathcal{S}} U_t(S)$ with $U_t(S)$ defined in equation 7
6: Play S_t and observe $\{(\boldsymbol{z}_{t,i}, r_{t,i}) : i \in S_t\}$
7: Update A_t, $\boldsymbol{b}_t$, and $\widehat{\phi}_{t,i}$ for all $i \in S_t$
8: **end for**

3 OFUC^2-PC Algorithm and Its Regret Analysis

In this section, we first introduce OFUC^2-PC (**O**ptimism in the **F**ace of **U**ncertainty for **C**^{2}MAB-**PC**) in Algorithm 1 and present its regret upper bound. We then introduce the key technical tool used in the analysis and present a matching lower bound.

3.1 OFUC^2-PC Algorithm

OFUC^2-PC is inspired by the optimistic principle of Wang et al. (2023) for a single arm and extends it to combinatorial actions, which creates two challenges. (1) All arms in S_t are selected with the same design matrix, before any post-serving feedback of round t is revealed, so their uncertainties accumulate within the round. (2) The reward parameter is estimated from the realized augmented contexts $\boldsymbol{u}_{s,i} = [\boldsymbol{x}_{s,i}, \boldsymbol{z}_{s,i}]$, which contain the post-serving noise, whereas the regret in equation 3 is evaluated at the mean augmented contexts $\bar{\boldsymbol{u}}_{t,i} = [\boldsymbol{x}_{t,i}, \phi_i^\star(\boldsymbol{x}_{t,i})]$. The analysis therefore has to control the uncertainty $\|\bar{\boldsymbol{u}}_{t,i}\|_{A_{t-1}^{-1}}$ of vectors that differ from the vectors $\boldsymbol{u}_{s,i}$ that built the design matrix A_{t-1} defined below, to which the standard elliptical potential lemma (Abbasi-Yadkori et al., 2011; Liu et al., 2023) does not apply. OFUC^2-PC handles the decision problem by joint optimism over the shared reward parameter and the post-serving contexts, and NBT-EPL (Lemma 1) combines batched truncation with noisy-context control to bound the accumulated uncertainty.

We now describe OFUC^2-PC step by step. In Line 1, we initialize the reward design matrix $A_0 = \lambda I$, the response vector $\boldsymbol{b}_0 = 0$, and the post-serving mapping estimator $\widehat{\phi}_{0,i}$ for each arm $i \in [m]$. In Line 3, at the beginning of round t, the learner observes the pre-serving contexts $\{\boldsymbol{x}_{t,i}\}_{i=1}^m$ of all base arms. In Line 4, the learner constructs the confidence sets. For the shared reward parameter, it uses the ridge regression estimate $\widehat{\boldsymbol{w}}_{t-1} = A_{t-1}^{-1}\boldsymbol{b}_{t-1}$ computed from all observations before round t, where $A_{t-1} = \lambda I + \sum_{s<t}\sum_{i\in S_s} \boldsymbol{u}_{s,i}\boldsymbol{u}_{s,i}^\top$ and $\boldsymbol{b}_{t-1} = \sum_{s<t}\sum_{i\in S_s} r_{s,i}\boldsymbol{u}_{s,i}$. Let $N_{t-1} = \sum_{s<t}|S_s|$ denote the total number of selected-arm observations before round t. Since $r_{s,i}$ is linear in the realized context $\boldsymbol{u}_{s,i}$ with noise $\eta_{s,i}$, the standard self-normalized confidence bound for linear bandits (Abbasi-Yadkori et al., 2011) applies with the confidence radius

$$\zeta_{t-1} = 2\sqrt{\lambda} + R_\eta\sqrt{2d_u \log\big(T(1 + N_{t-1}L_u^2/\lambda)/\delta_w\big)}, \tag{6}$$

where $\delta_w \in (0,1)$ is the confidence level, and the reward-parameter confidence set is $\mathcal{C}_{t-1}^w = \{\boldsymbol{w} \in \mathcal{W} : \|\boldsymbol{w} - \widehat{\boldsymbol{w}}_{t-1}\|_{A_{t-1}} \le \zeta_{t-1}\}$. We write $\zeta_T := \max_{1\le t\le T}\zeta_{t-1}$ for the largest radius. For the post-serving context of each arm i, let $e_{t-1,i}(x)$ denote the corresponding simultaneous high-probability prediction radius, using failure probability $\delta_\phi/(mT)$, i.e., the right-hand side of equation 4 under Assumption 1 or of equation 5 under Assumption 2. The post-serving confidence set is $\mathcal{C}_{t-1}^\phi(i, \boldsymbol{x}_{t,i}) = \{\boldsymbol{z} \in \mathbb{R}^{d_z} : \|\boldsymbol{z} - \widehat{\phi}_{t-1,i}(\boldsymbol{x}_{t,i})\|_2 \le e_{t-1,i}(\boldsymbol{x}_{t,i})\}$.

In Line 5, the learner selects the action by joint optimism. Given a candidate combinatorial action S, OFUC^2-PC optimizes over one shared reward parameter and one optimistic post-serving context for each arm in S. For any candidate $\boldsymbol{w}$ and $\{\boldsymbol{z}_i\}_{i\in S}$, define $\widetilde{\mu}_{t,i} = \mathrm{clip}_{[-B,B]}\big([\boldsymbol{x}_{t,i}, \boldsymbol{z}_i]^\top\boldsymbol{w}\big)$ for $i \in S$, where $\mathrm{clip}_{[-B,B]}(a) = \min\{B, \max\{-B, a\}\}$. The optimistic value of S is

$$U_t(S) = \max_{\boldsymbol{w}\in\mathcal{C}_{t-1}^w, \boldsymbol{z}_i\in\mathcal{C}_{t-1}^\phi(i,\boldsymbol{x}_{t,i}),\, i\in S} F_t\big(S; \widetilde{\boldsymbol{\mu}}_t\big), \tag{7}$$

where $\widetilde{\boldsymbol{\mu}}_t$ is the vector of the optimistic means $\widetilde{\mu}_{t,i}$, whose entries outside S do not affect $F_t(S;\cdot)$ by Assumption 4. OFUC^2-PC then selects $S_t \in \arg\max_{S\in\mathcal{S}} U_t(S)$. In Lines 6–7, the learner plays S_t,

observes $\{(\boldsymbol{z}_{t,i}, r_{t,i}) : i \in S_t\}$, and updates reward statistics as $A_t = A_{t-1} + \sum_{i\in S_t} \boldsymbol{u}_{t,i}\boldsymbol{u}_{t,i}^\top$ and $\boldsymbol{b}_t = \boldsymbol{b}_{t-1} + \sum_{i\in S_t} r_{t,i}\boldsymbol{u}_{t,i}$. For $i \in S_t$, the estimator $\widehat{\phi}_{t,i}$ is also updated with the new $(\boldsymbol{x}_{t,i}, \boldsymbol{z}_{t,i})$.

3.2 Regret Upper Bound

We first state the regret upper bound of OFUC^2-PC.

Theorem 1 (Regret of OFUC^2-PC)**.** *Suppose that the regularity conditions of Section 2 and Assumptions 3 and 4 hold, with $R_\eta = \Theta(1)$ and $\lambda = \Theta(d_u)$. For any $\delta \in (0,1)$, with probability at least $1-\delta$, the regret of OFUC^2-PC under Assumption 1 satisfies*

$$R_T = \widetilde{\mathcal{O}}\big(B_1 m^\alpha (KT)^{1-\alpha} d_x^\alpha + B_1 d_u\sqrt{KT} + B_1 d_u K\big), \tag{8}$$

and under Assumption 2, $R_T = \widetilde{\mathcal{O}}\big(B_1\sqrt{d_x}\, m^a (KT)^{1-a} + B_1 d_u\sqrt{KT} + B_1 d_u K\big)$

Discussion. The first term of Theorem 1 is the *context-learning* regret, i.e., the price of not observing the post-serving contexts at decision time. It depends on m because each arm has its own post-serving mapping. The second and third terms are the *reward-learning* regret, which is free of m since all selected-arm observations are used to estimate the shared parameter $\boldsymbol{w}^\star$. The second term is the cost of learning a d_u-dimensional linear parameter from KT noisy observations. We next discuss special cases. For the canonical rates $\alpha = a = 1/2$, both bounds of Theorem 1 become

$$R_T = \widetilde{\mathcal{O}}\big(B_1\sqrt{mKTd_x} + B_1 d_u\sqrt{KT} + B_1 d_u K\big). \tag{9}$$

In the special case without post-serving contexts, the context-learning term is absent, so the remaining terms in equation 9 recover the near-optimal regret rate for linear C^{2}MAB (Takemura et al., 2021). When $K = 1$ and $B_1 = 1$, Theorem 1 gives $\widetilde{\mathcal{O}}(m^\alpha T^{1-\alpha} d_x^\alpha + d_u\sqrt{T})$, whose reward-learning term matches the $\widetilde{\mathcal{O}}(d_u\sqrt{T})$ rate of the shared-reward variant of Wang et al. (2023). Finally, for $R_\eta = \Theta(1)$, equation 9 matches the minimax lower bound of Theorem 2 up to logarithmic factors in the non-saturated regime (Section 3.4). The full proof of Theorem 1 is given in Appendix C.

3.3 Regret Analysis and the Noisy Batched Truncated EPL (NBT-EPL)

Regret decomposition. When all confidence sets hold, optimism and Assumption 4 give $R_T \le B_1 \sum_{t=1}^T \sum_{i\in S_t} |\widetilde{\mu}_{t,i} - \mu_{t,i}|$, and each gap is at most $\min\{2B,\, 2e_{t-1,i}(\boldsymbol{x}_{t,i}) + 2\zeta_{t-1} q_{t,i}\}$, where $q_{t,i} := \|\bar{\boldsymbol{u}}_{t,i}\|_{A_{t-1}^{-1}}$ is the confidence width at the mean augmented context. Splitting the selected-arm observations (t,i) into the small-width set $\mathcal{L} = \{q_{t,i}^2 \le 1/K\}$ and the large-width set $\mathcal{H} = \{q_{t,i}^2 > 1/K\}$ as in Takemura et al. (2021), we obtain $R_T \le 2B_1 \sum_{t=1}^T \sum_{i\in S_t} e_{t-1,i}(\boldsymbol{x}_{t,i}) + 2B_1\zeta_T \sum_{(t,i)\in\mathcal{L}} q_{t,i} + 2B_1 B\,|\mathcal{H}|$, where the first term is the context-learning part R_T^ϕ and the last two form the reward-learning part R_T^w.

Reward-learning part and NBT-EPL. The main technical difficulty lies in R_T^w. The design matrix A_{t-1} is built from the realized augmented contexts $\boldsymbol{u}_{s,i}$, which contain the post-serving noise, whereas the widths $q_{t,i}$ are evaluated at the mean augmented contexts $\bar{\boldsymbol{u}}_{t,i}$. The standard elliptical potential lemma does not control these widths, because $\bar{\boldsymbol{u}}_{t,i}$ is not the vector that updates A_t, and because the K arms of a round share the same A_{t-1}. For noise-free contexts, Takemura et al. (2021) handle the second issue by truncating the squared widths at $1/K$. For a single arm per round, Wang et al. (2023) handle the first issue by a generalized EPL. Inspired by both, we bound the truncated potential $\Gamma_T = \sum_{t\le T}\sum_{i\in S_t} \min\{1/K, q_{t,i}^2\}$ of the mean augmented contexts.

Lemma 1 (Noisy batched truncated elliptical potential)**.** *Suppose that, conditioned on $\mathcal{F}_t^{\mathrm{pre}}$, the post-serving noises $\{\boldsymbol{\epsilon}_{t,i} : i \in S_t\}$ are zero-mean and independent across the selected arms, that $\|\boldsymbol{u}_{t,i}\|_2 \le L_u$, and that $\lambda \ge L_u^2$. Then, for any algorithm that selects at most K arms per round and any $\delta_E \in (0,1)$, with probability at least $1-\delta_E$,*

$$\Gamma_T \le \frac{2}{1-e^{-1}}\Big[\log\frac{\det A_T}{\det(\lambda I)} + \log\frac{1}{\delta_E}\Big] = \widetilde{\mathcal{O}}(d_u). \tag{10}$$

Lemma 1 bounds both terms of R_T^w. By the Cauchy–Schwarz inequality and $|\mathcal{L}| \le KT$, we have $\sum_{(t,i)\in\mathcal{L}} q_{t,i} \le \sqrt{KT\,\Gamma_T}$, and since every element of $\mathcal{H}$ contributes $1/K$ to Γ_T, we have $|\mathcal{H}| \le$

$K\Gamma_T$. Hence $R_T^w = \widetilde{\mathcal{O}}(B_1\zeta_T\sqrt{d_u KT} + B_1 B d_u K)$, and $\zeta_T = \widetilde{\mathcal{O}}(\sqrt{\lambda} + R_\eta\sqrt{d_u})$ gives the second and third terms of Theorem 1.

Remark 3 (Conditions of NBT-EPL). Unlike the generalized EPL of Wang et al. (2023), Lemma 1 does not require a non-degenerate noise covariance, i.e., $\mathbb{E}[\boldsymbol{\epsilon}_{t,i}\boldsymbol{\epsilon}_{t,i}^\top \mid \mathcal{F}_t^{\mathrm{pre}}] \succeq \sigma_\epsilon^2 I$, and its bound has no counterpart of the additive σ_ϵ^{-4} term of their lemma, because the noise can only inflate the expected realized width. Thus, Theorem 1 continues to hold when the post-serving noise is degenerate in some directions, e.g., when some post-serving features are deterministic given the pre-serving context. Without noise, Lemma 1 recovers the truncated EPL of Takemura et al. (2021) up to constants and the $\log(1/\delta_E)$ term. The full proof of Lemma 1 is in Appendix D.

3.4 Minimax Lower Bound

We prove the lower bound on the additive subclass $F_t(S;\boldsymbol{\mu}) = \sum_{i\in S}\mu_i$, for which $B_1 = 1$. For the lower-bound comparison, we normalize $B \geq 1$. The hard instance combines three sources of difficulty. The first is a hard family for estimating the m arm-specific post-serving mappings. The other two are inspired by the lower bound of Takemura et al. (2021) for linear C^{2}MAB, namely the stochastic uncertainty in the shared linear parameter, and the fact that all K arms of a round are committed before any feedback of that round is received.

Theorem 2 (Minimax lower bound)**.** *Let $D = \max\{d_x, d_z\} \geq 2$, and suppose that K divides m, $m/K \geq \max\{2,\, 2^{\lfloor D/2\rfloor}\}$, $KT \geq md_x$, and $R_\eta d_u\sqrt{KT} + Bd_uK \leq BKT$. Within the elliptical-learnability class with $\alpha = 1/2$, there exist a feasible family $\mathcal{S}$ and a family of instances satisfying the conditions of Section 2 such that, for any algorithm, some instance in the family yields*

$$\mathbb{E}[R_T] = \Omega\big(\sqrt{mKTd_x} + R_\eta d_u\sqrt{KT} + d_uK\big). \tag{11}$$

The lower bound matches equation 9 up to logarithmic factors in its dependence on m, K, T, d_x, and d_u, so OFUC^2-PC is minimax optimal up to logarithmic factors for additive rewards. The complete construction is deferred to Appendix G.

4 Improving Computational Efficiency with C^{2}UCB-PC

In this section, we introduce C^{2}UCB-PC (Contextual Combinatorial Upper Confidence Bound with Post-Serving Contexts), an oracle-efficient alternative to OFUC^2-PC. We first describe the algorithm and then present its regret bound and the comparison with OFUC^2-PC.

The computational bottleneck of OFUC^2-PC is the joint maximization in equation 7, where one optimistic reward parameter is coupled across all arms of a candidate action. Inspired by CUCB-style algorithms (Chen et al., 2013; Qin et al., 2014), C^{2}UCB-PC instead computes an optimistic score for each arm separately and then selects the action with a combinatorial optimization oracle. We assume access to an oracle for F_t that returns $\mathrm{Oracle}_t(\boldsymbol{v}) \in \arg\max_{S\in\mathcal{S}} F_t(S;\boldsymbol{v})$ for any mean vector $\boldsymbol{v} \in [-B, B]^m$. Concretely, in Line 4, the learner predicts the post-serving context of every arm by $\widehat{\boldsymbol{z}}_{t,i} = \widehat{\phi}_{t-1,i}(\boldsymbol{x}_{t,i})$ and forms the predicted augmented context $\widehat{\boldsymbol{u}}_{t,i} = [\boldsymbol{x}_{t,i}, \widehat{\boldsymbol{z}}_{t,i}]$. In Line 5, it computes a reward score and a post-serving score for every arm,

$$\mathrm{UCB}_{t,i}^w = \widehat{\boldsymbol{u}}_{t,i}^\top\widehat{\boldsymbol{w}}_{t-1} + \zeta_{t-1}\|\widehat{\boldsymbol{u}}_{t,i}\|_{A_{t-1}^{-1}}, \qquad \mathrm{UCB}_{t,i}^\phi = e_{t-1,i}(\boldsymbol{x}_{t,i}). \tag{12}$$

The first score is the upper confidence bound of the reward at the predicted context and the second score covers the error of this prediction. In Line 6, the learner selects $S_t = \mathrm{Oracle}_t(\widetilde{\boldsymbol{U}}_t)$ with $\widetilde{\boldsymbol{U}}_t = (\widetilde{U}_{t,1}, \ldots, \widetilde{U}_{t,m})^\top$. Thus, once the m scalar scores are computed, the action selection requires only one combinatorial oracle call.

Theorem 3 (Regret of C^{2}UCB-PC)**.** *Suppose that the regularity conditions of Section 2 and Assumptions 1, 3, and 4 hold, with $R_\eta = \Theta(1)$ and $\lambda = \Theta(d_u)$. For any $\delta \in (0,1)$, with probability at least $1-\delta$, the regret of C^{2}UCB-PC satisfies*

$$R_T = \widetilde{\mathcal{O}}\big(B_1(1 + \zeta_T/\sqrt{\lambda})\, m^\alpha (KT)^{1-\alpha} d_x^\alpha + B_1 d_u\sqrt{KT} + B_1 B d_u K\big). \tag{13}$$

Under Assumption 2, the same bound holds with $m^\alpha(KT)^{1-\alpha}d_x^\alpha$ replaced by $\sqrt{d_x}\, m^a (KT)^{1-a}$, by the argument of Appendix E.

Algorithm 2 C^{2}UCB-PC

Require: Set of combinatorial actions $\mathcal{S}$, oracle Oracle_t, regularization parameter λ, confidence levels δ_w, δ_ϕ

Initialize $A_0 = \lambda I$, $\boldsymbol{b}_0 = 0$, and $\widehat{\phi}_{0,i}$ for all $i \in [m]$
for $t = 1, \dots, T$ **do**
  Observe the pre-serving contexts $\{\boldsymbol{x}_{t,i}\}_{i=1}^m$ of all base arms
  Predict $\widehat{\boldsymbol{z}}_{t,i} = \widehat{\phi}_{t-1,i}(\boldsymbol{x}_{t,i})$ and form $\widehat{\boldsymbol{u}}_{t,i} = [\boldsymbol{x}_{t,i}, \widehat{\boldsymbol{z}}_{t,i}]$ for all $i \in [m]$
  Compute the scores $\mathrm{UCB}^w_{t,i}$ and $\mathrm{UCB}^\phi_{t,i}$ in equation 12 for all $i \in [m]$
  Select $S_t = \mathrm{Oracle}_t(\widetilde{\boldsymbol{U}}_t)$, where $\widetilde{U}_{t,i} = \mathrm{clip}_{[-B,B]}(\mathrm{UCB}^w_{t,i} + \mathrm{UCB}^\phi_{t,i})$ for all $i \in [m]$
  Play S_t and observe $\{(\boldsymbol{z}_{t,i}, r_{t,i}) : i \in S_t\}$
  Update A_t, $\boldsymbol{b}_t$, and $\widehat{\phi}_{t,i}$ for all $i \in S_t$
end for

Discussion. Compared with Theorem 1, the reward-learning terms are unchanged, and the context-learning term keeps the same exponents in m, K, T, and d_x but pays an additional factor $1+\zeta_T/\sqrt{\lambda}$. Since $\zeta_T = \widetilde{\mathcal{O}}(\sqrt{\lambda} + \sqrt{d_u})$, this factor is $\widetilde{\mathcal{O}}(1 + \sqrt{d_u/\lambda})$. C^{2}UCB-PC matches the regret of OFUC^2-PC up to logarithmic factors. In terms of computation, a round of C^{2}UCB-PC computes m scalar scores and makes one oracle call, whereas OFUC^2-PC has to solve the joint maximization in equation 7 for the candidate combinatorial actions. The full proof is given in Appendix F.

5 Improving the Regret with Variance-Aware Algorithms

In this section, we develop variance-aware versions of OFUC^2-PC and C^{2}UCB-PC and show that the leading reward-learning term improves from $\widetilde{\mathcal{O}}(B_1 d_u \sqrt{KT})$ to $\widetilde{\mathcal{O}}(B_v d_u \sqrt{T})$.

Key idea. OFUC^2-PC and C^{2}UCB-PC treat all selected-arm observations in the same way, although an observation with a lower variance is more valuable under the variance-modulated smoothness condition (Assumption 5). Inspired by the optimistic-variance approach of Liu et al. (2023), we therefore weight each observation (t, i) by $1/\bar{V}_{t,i}$ in the ridge regression of the reward parameter, where the optimistic variance $\bar{V}_{t,i}$ is the largest value of $\mu(1-\mu)$ over a confidence interval of $\mu_{t,i}$ that is computed before the feedback of round t is observed. VA-OFUC^2-PC and VA-C^{2}UCB-PC follow Algorithms 1 and 2, with the reward estimator and its confidence set replaced by their variance-weighted counterparts, whose regularization parameter and confidence radius are denoted by λ_V and ρ_T^V. In the analysis, a variance-scaled NBT-EPL controls the weighted confidence widths, and Assumption 5 turns $\sqrt{KT}$ into $\sqrt{T}$. The full algorithms and the proofs are given in Appendix H.

Theorem 4 (Regret of the variance-aware algorithms)**.** *Suppose that the regularity conditions of Section 2 and Assumptions 1, 3, and 5 hold, with $R_\eta = \Theta(1)$ and $\lambda_V = \Theta(d_u)$. For any $\delta \in (0,1)$, with probability at least $1-\delta$, the regret of VA-OFUC^2-PC satisfies*

$$R_T = \widetilde{\mathcal{O}}\big(B_1 m^\alpha (KT)^{1-\alpha} d_x^\alpha + B_v d_u \sqrt{T} + B_v d_u K\big), \tag{14}$$

and the regret of VA-C^{2}UCB-PC satisfies

$$R_T = \widetilde{\mathcal{O}}\big(B_1(1 + \rho_T^V/\sqrt{\lambda_V})\, m^\alpha (KT)^{1-\alpha} d_x^\alpha + B_v d_u \sqrt{T} + B_v d_u K\big). \tag{15}$$

Under Assumption 2, the same bounds hold with $m^\alpha (KT)^{1-\alpha} d_x^\alpha$ replaced by $\sqrt{d_x}\, m^a (KT)^{1-a}$, and the bounds with the explicit dependence on ν are given in Appendix H.

Discussion. Compared with Theorems 1 and 3, the leading reward-learning term improves from $\widetilde{\mathcal{O}}(B_1 d_u \sqrt{KT})$ to $\widetilde{\mathcal{O}}(B_v d_u \sqrt{T})$. By Remark 2, any reward function with 1-norm coefficient B_1 satisfies Assumption 5 with $B_v = B_1\sqrt{K}/2$, so the new term is never worse, and it removes the $\sqrt{K}$ factor whenever $B_v = O(B_1)$. The context-learning terms are unchanged, because these terms are governed by the learnability conditions (Assumptions 1 and 2), which carry no variance information. If the learnability conditions can be augmented with variance information, e.g., with prediction radii that scale with the variance proxy $\mu_{t,i}(1-\mu_{t,i})$, the first term could also be improved. As in Section 4, VA-C^{2}UCB-PC pays the factor $1 + \rho_T^V/\sqrt{\lambda_V}$ in the context-learning term, which is logarithmic for $\lambda_V = \Theta(d_u)$, and it keeps the separable arm-wise UCB and the computational advantage of C^{2}UCB-PC.

6 EXPERIMENTS

Setup. We evaluate our algorithms in a synthetic C^{2}MAB-PC environment and a semi-synthetic one built from MovieLens-1M (Harper & Konstan, 2015), both with the cascading reward $F(S;\boldsymbol{\mu}) = 1 - \prod_{i\in S}(1-\mu_i)$. We compare OFUC^2-PC, C^{2}UCB-PC, VA-OFUC^2-PC, and VA-C^{2}UCB-PC with three baselines: Random, X-only LinUCB (which ignores the post-serving contexts) and LinUCB($\widehat{\phi}$) of Wang et al. (2023) (which plugs the predicted post-serving context $\widehat{\phi}_i(\boldsymbol{x})$ into both action selection and reward regression). Further settings, implementation details, and results for different parameters are given in Appendices I and J.

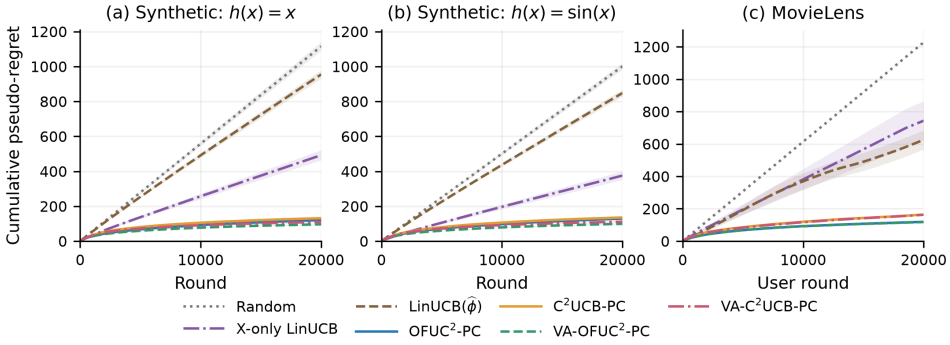

Figure 1: Regret for $K = 3$ and $T = 20000$, with shading showing the standard error over 10 seeds. **(a)** Synthetic linear mapping $h(\boldsymbol{x}) = \boldsymbol{x}$. **(b)** Synthetic sine mapping $h(\boldsymbol{x}) = \sin(\boldsymbol{x})$. **(c)** MovieLens environment. Both synthetic and MovieLens experiments use $m = 10$.

Regret. In both synthetic settings (Figure 1(a,b)), our four algorithms incur lower regret than all baselines, with VA-OFUC^2-PC the lowest, and for $K \in \{5, 7\}$ both variance-aware algorithms improve on their counterparts (Appendix I). On MovieLens (Figure 1(c)), OFUC^2-PC and VA-OFUC^2-PC reach final mean regrets of 119.14 and 117.88, about one-sixth of X-only LinUCB's, while C^{2}UCB-PC and VA-C^{2}UCB-PC achieve 163.88 and 162.58, respectively. Exploiting the post-serving contexts thus pays off, and the joint optimism of OFUC^2-PC reduces the regret further.

Table 2: Runtime in seconds at $m = 10$, $K = 3$, and $T = 20000$.

Environment	OFUC^2-PC	C^{2}UCB-PC	Speedup	VA-OFUC^2-PC	VA-C^{2}UCB-PC	Speedup
Syn., $h(\boldsymbol{x}) = \boldsymbol{x}$	48.88 ± 1.76	1.511 ± 0.007	$32.35\times$	59.34 ± 1.06	1.666 ± 0.002	$35.62\times$
Syn., $h(\boldsymbol{x}) = \sin(\boldsymbol{x})$	49.61 ± 1.17	1.519 ± 0.006	$32.66\times$	60.56 ± 0.77	1.692 ± 0.004	$35.80\times$
MovieLens	89.70 ± 3.90	4.919 ± 0.105	$18.24\times$	111.67 ± 3.12	5.049 ± 0.087	$22.11\times$

Runtime. Table 2 reports the runtime, where the joint maximization of the two joint algorithms is solved approximately by multi-start SLSQP (Kraft, 1988). The separable algorithms are 18–36$\times$ faster than their joint counterparts. Together with Figure 1, this shows the tradeoff: the joint algorithms have lower regret, and the separable ones are an order of magnitude faster.

7 CONCLUSION

We studied contextual combinatorial bandits with post-serving contexts. We proposed OFUC^2-PC and developed the noisy batched truncated elliptical potential lemma (NBT-EPL) to handle batched arm selection and noisy post-serving observations at once, whose resulting regret bound is minimax optimal up to logarithmic factors for additive rewards with $\alpha = 1/2$ under the conditions of Theorem 2. To improve the computational efficiency and the reward-learning regret, we further introduced C^{2}UCB-PC and variance-aware versions of both algorithms. Experiments on synthetic and MovieLens-based semi-synthetic environments demonstrate superior regret performance and the computational advantage of the oracle-efficient algorithms. Future directions include generalizing our framework to probabilistically triggered arms and incorporating variance information into the context learning, as discussed in Section 5.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript drafting and editing, as well as code editing and debugging for the experiments; all mathematical claims, references, proofs, code, and experimental results were reviewed by the authors, who take responsibility for the final manuscript.

REPRODUCIBILITY STATEMENT

Complete proofs are provided in the appendix. Experimental settings and implementation details are also included to facilitate reproduction.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in Neural Information Processing Systems*, 24, 2011.

Alexandra Carpentier, Claire Vernade, and Yasin Abbasi-Yadkori. The elliptical potential lemma revisited. *arXiv preprint arXiv:2010.10182*, 2020.

Wei Chen, Yajun Wang, and Yang Yuan. Combinatorial multi-armed bandit: General framework and applications. In *International Conference on Machine Learning*, pp. 151–159, 2013.

Wei Chen, Yajun Wang, Yang Yuan, and Qinshi Wang. Combinatorial multi-armed bandit and its extension to probabilistically triggered arms. *Journal of Machine Learning Research*, 17(50): 1–33, 2016.

Richard Combes, Mohammad Sadegh Talebi, Alexandre Proutiere, and Marc Lelarge. Combinatorial bandits revisited. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2116–2124, 2015.

Yi Gai, Bhaskar Krishnamachari, and Rahul Jain. Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations. *IEEE/ACM Transactions on Networking*, 20(5):1466–1478, 2012.

Nima Hamidi and Mohsen Bayati. Technical note—the elliptical potential lemma for general distributions with an application to linear thompson sampling. *Operations Research*, 71(4):1434–1439, 2022.

F. Maxwell Harper and Joseph A. Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):19:1–19:19, 2015.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

Taehyun Hwang, Kyuwook Chai, and Min-hwan Oh. Combinatorial neural bandits. In *International Conference on Machine Learning*, pp. 14203–14236. PMLR, 2023.

Dieter Kraft. A software package for sequential quadratic programming. Technical Report DFVLR-FB 88-28, DFVLR, 1988.

Branislav Kveton, Zheng Wen, Azin Ashkan, and Csaba Szepesvári. Tight regret bounds for stochastic combinatorial semi-bandits. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pp. 535–543, 2015.

Tze Leung Lai and Ching Zong Wei. Least squares estimates in stochastic regression models with applications to identification and control of dynamic systems. *The Annals of Statistics*, 10(1): 154–166, 1982.

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

Shuai Li, Baoxiang Wang, Shengyu Zhang, and Wei Chen. Contextual combinatorial cascading bandits. In *Proceedings of the 33rd International Conference on Machine Learning*, pp. 1245–1253, 2016.

Xutong Liu, Jinhang Zuo, Siwei Wang, Carlee Joe-Wong, John C. S. Lui, and Wei Chen. Batch-size independent regret bounds for combinatorial semi-bandits with probabilistically triggered arms or independent arms. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Xutong Liu, Jinhang Zuo, Siwei Wang, John C. S. Lui, Mohammad Hajiesmaili, Adam Wierman, and Wei Chen. Contextual combinatorial bandits with probabilistically triggered arms. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 22559–22593, 2023.

Xutong Liu, Xiangxiang Dai, Xuchuang Wang, Mohammad Hajiesmaili, and John CS Lui. Combinatorial logistic bandits. In *Abstracts of the 2025 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pp. 112–114, 2025.

Youngmin Oh. Robust linear dueling bandits with post-serving context under unknown delays and adversarial corruptions. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Lijing Qin, Shouyuan Chen, and Xiaoyan Zhu. Contextual combinatorial bandit and its application on diversified online recommendation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pp. 461–469, 2014.

Kei Takemura, Shinji Ito, Daisuke Hatano, Hanna Sumita, Takuro Fukunaga, Naonori Kakimura, and Ken-ichi Kawarabayashi. Near-optimal regret bounds for contextual combinatorial semi-bandits with linear payoff functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 9791–9798, 2021.

Chaoqi Wang, Ziyu Ye, Zhe Feng, Ashwinkumar Badanidiyuru, and Haifeng Xu. Follow-ups also matter: Improving contextual bandits via post-serving contexts. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Qinshi Wang and Wei Chen. Improving regret bounds for combinatorial semi-bandits with probabilistically triggered arms and its applications. In *Advances in Neural Information Processing Systems*, volume 30, pp. 1161–1171, 2017.

Zheng Wen, Branislav Kveton, and Azin Ashkan. Efficient learning in large-scale combinatorial semi-bandits. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1113–1122, 2015.

A Detailed Related Work

Contextual combinatorial bandits. Contextual combinatorial multi-armed bandits (C^2MAB) use contextual features to model the mean rewards of base arms. Qin et al. (2014) introduced the framework and proposed C^2UCB for diversified recommendation, Li et al. (2016) studied contextual combinatorial cascading bandits, and Takemura et al. (2021) established near-optimal regret bounds for linear rewards. Beyond linear models, logistic (Liu et al., 2025) and neural (Hwang et al., 2023) parameterizations have also been considered. More recently, Liu et al. (2023) extended C^2MAB to probabilistically triggered arms and developed variance-adaptive algorithms under variance-modulated smoothness conditions, inspired by the batch-size independent analysis of Liu et al. (2022). All these works assume that the contexts are fully observed before the action is selected. In contrast, we propose C^2MAB-PC, where part of the contexts is revealed only after the selected arms are served, and we develop both standard and variance-aware algorithms for it.

Bandits with post-serving contexts. Wang et al. (2023) introduced contextual bandits with post-serving contexts, where part of the contextual information is observed only after the selected arm is served. Their poLinUCB algorithm learns the mapping from pre-serving to post-serving contexts together with the reward parameters, and also covers action-dependent contexts, where each arm has its own mapping. More recently, Oh (2026) studied post-serving contexts in linear dueling bandits under unknown delays and adversarial corruptions. Our model is inspired by these works. However, they select a single arm or a pair of arms per round, and their analyses do not apply when K arms are selected before any post-serving feedback of the round is observed.

Elliptical potential lemmas. The elliptical potential lemma (EPL), which dates back to Lai & Wei (1982), controls the accumulated uncertainty of sequentially selected feature vectors and is central to linear bandit analysis (Abbasi-Yadkori et al., 2011). Two extensions are closest to ours. For contextual combinatorial semi-bandits, Takemura et al. (2021) truncate the potential to handle multiple arms that are selected in the same round with the same design matrix, but their contexts are noise-free. For post-serving contexts, Wang et al. (2023) develop a generalized EPL for noisy feature observations, but it covers one arm per round and requires a non-degenerate noise covariance. Other generalizations consider different powers of the inverse design matrix (Carpentier et al., 2020) and general prior and noise distributions(Hamidi & Bayati, 2022). Inspired by the first two, our noisy batched truncated EPL (NBT-EPL) handles batched selection and noisy observations jointly, without the covariance condition.

B Technical Conditions

We collect here the filtration, noise, and boundedness conditions used throughout the analysis. All other notation follows the main text.

Filtration Let $\mathcal{H}_{t-1}$ denote the history before round t. Define

$$\mathcal{F}_t^{\mathrm{pre}} = \sigma(\mathcal{H}_{t-1}, \{\boldsymbol{x}_{t,i}\}_{i=1}^m, S_t)\,,$$

which contains the information available after observing the current pre-serving contexts and selecting S_t, but before observing the current post-serving contexts. After $\{\boldsymbol{z}_{t,i} : i \in S_t\}$ are observed, define

$$\mathcal{F}_t^{\mathrm{post}} = \sigma(\mathcal{F}_t^{\mathrm{pre}}, \{\boldsymbol{z}_{t,i} : i \in S_t\})\,.$$

Noise conditions Conditional on $\mathcal{F}_t^{\mathrm{pre}}$, the post-serving innovations $\{\boldsymbol{\epsilon}_{t,i} : i \in S_t\}$ are independent across the selected arms and satisfy

$$\mathbb{E}[\boldsymbol{\epsilon}_{t,i} \mid \mathcal{F}_t^{\mathrm{pre}}] = 0, \qquad \|\boldsymbol{\epsilon}_{t,i}\|_2 \le L_\epsilon.$$

No lower bound on their conditional covariance is required.

Conditional on $\mathcal{F}_t^{\mathrm{post}}$, the reward noises $\{\eta_{t,i} : i \in S_t\}$ are mutually independent and each is R_η-sub-Gaussian:

$$\mathbb{E}\left[e^{s\eta_{t,i}} \mid \mathcal{F}_t^{\mathrm{post}}\right] \le \exp\left(\frac{s^2 R_\eta^2}{2}\right), \qquad \forall s \in \mathbb{R}.$$

Boundedness We assume

$$\|\boldsymbol{\theta}^\star\|_2 \le 1, \qquad \|\boldsymbol{\beta}^\star\|_2 \le 1, \qquad \|\boldsymbol{x}_{t,i}\|_2 \le L_x,$$

and

$$\max\{\|\boldsymbol{u}_{t,i}\|_2, \|\bar{\boldsymbol{u}}_{t,i}\|_2\} \le L_u, \qquad |\mu_{t,i}| \le B.$$

Define

$$\mathcal{W} = \left\{\boldsymbol{w} = \begin{bmatrix}\boldsymbol{\theta}\\ \boldsymbol{\beta}\end{bmatrix} : \|\boldsymbol{\theta}\|_2 \le 1,\ \|\boldsymbol{\beta}\|_2 \le 1\right\}, \qquad \boldsymbol{w}^\star \in \mathcal{W},$$

and assume $L_u^2 = \Theta(d_u)$. We take

$$\lambda = \max\{1, L_u^2\} = \Theta(d_u).$$

Confidence events We split the failure probability as

$$\delta_\phi = \delta_w = \delta_E = \frac{\delta}{3}.$$

Applying Assumption 1 with failure probability $\delta_\phi/(mT)$ to each arm-round query and taking a union bound gives a simultaneous context-confidence event $\mathcal{E}_\phi$ with probability at least $1 - \delta_\phi$. By the corresponding confidence guarantees and a union bound,

$$\mathbb{P}\left(\mathcal{E}_\phi \cap \mathcal{E}_w \cap \mathcal{E}_E\right) \ge 1 - \delta.$$

C Proof of Theorem 1

We prove the result under Assumption 1. The sample-size-based case follows by replacing the post-serving context term with the bound established in Appendix E.

Work on the joint confidence event

$$\mathcal{E} = \mathcal{E}_\phi \cap \mathcal{E}_w \cap \mathcal{E}_E,$$

which holds with probability at least $1 - \delta$ by Appendix B.

For each selected arm, define

$$q_{t,i} = \|\bar{\boldsymbol{u}}_{t,i}\|_{A_{t-1}^{-1}}$$

and

$$\ell_{t,i} = |\widetilde{\mu}_{t,i} - \mu_{t,i}|,$$

where $(\widetilde{\boldsymbol{w}}_t, \{\widetilde{\boldsymbol{z}}_{t,i}\}_{i \in S_t})$ are the optimizers in equation 7.

Since the true pair

$$(\boldsymbol{w}^\star, \{\phi_i^\star(\boldsymbol{x}_{t,i})\}_{i \in S})$$

is feasible on $\mathcal{E}$, optimism gives

$$F_t(S_t^\star; \boldsymbol{\mu}_t) \le U_t(S_t^\star) \le U_t(S_t) = F_t(S_t; \widetilde{\boldsymbol{\mu}}_t).$$

Therefore, by the 1-norm bounded-smoothness condition,

$$R_T \le B_1 \sum_{t=1}^{T} \sum_{i \in S_t} \ell_{t,i}. \tag{16}$$

Because both $\widetilde{\boldsymbol{z}}_{t,i}$ and $\phi_i^\star(\boldsymbol{x}_{t,i})$ belong to the same post-serving confidence set,

$$\|\widetilde{\boldsymbol{z}}_{t,i} - \phi_i^\star(\boldsymbol{x}_{t,i})\|_2 \le 2e_{t-1,i}(\boldsymbol{x}_{t,i}).$$

Since $\widetilde{\boldsymbol{w}}_t \in \mathcal{C}_{t-1}^w \subseteq \mathcal{W}$, the definition of $\mathcal{W}$ implies $\|\widetilde{\boldsymbol{\beta}}_t\|_2 \le 1$. Therefore,

$$\left|(\widetilde{\boldsymbol{u}}_{t,i} - \bar{\boldsymbol{u}}_{t,i})^\top \widetilde{\boldsymbol{w}}_t\right| \le \|\widetilde{\boldsymbol{z}}_{t,i} - \phi_i^\star(\boldsymbol{x}_{t,i})\|_2 \|\widetilde{\boldsymbol{\beta}}_t\|_2 \le 2e_{t-1,i}(\boldsymbol{x}_{t,i})$$

Moreover, since both $\widetilde{\boldsymbol{w}}_t$ and $\boldsymbol{w}^\star$ belong to $\mathcal{C}_{t-1}^w$,

$$\|\widetilde{\boldsymbol{w}}_t - \boldsymbol{w}^\star\|_{A_{t-1}} \le 2\zeta_{t-1}.$$

Hence, by Cauchy–Schwarz,

$$\left|\widetilde{\boldsymbol{u}}_{t,i}^\top \widetilde{\boldsymbol{w}}_t - \bar{\boldsymbol{u}}_{t,i}^\top \boldsymbol{w}^\star\right| \le 2e_{t-1,i}(\boldsymbol{x}_{t,i}) + 2\zeta_{t-1} q_{t,i}.$$

Since $\mu_{t,i} \in [-B, B]$ and clipping is nonexpansive,

$$\ell_{t,i} = |\widetilde{\mu}_{t,i} - \mu_{t,i}| \le 2e_{t-1,i}(\boldsymbol{x}_{t,i}) + 2\zeta_{t-1} q_{t,i}.$$

Clipping also gives

$$\ell_{t,i} \le 2B.$$

C.1 Post-Serving Context Term

Define the post-serving context contribution by

$$R_T^\phi = 2B_1 \sum_{t=1}^{T} \sum_{i \in S_t} e_{t-1,i}(\boldsymbol{x}_{t,i}).$$

Under Assumption 1,

$$e_{t-1,i}(\boldsymbol{x}_{t,i}) \le C_0 \left(1 \wedge \|\boldsymbol{x}_{t,i}\|_{X_{t-1,i}^{-1}}^2\right)^\alpha \log\left(\frac{mT^2}{\delta_\phi}\right).$$

For arm i, let

$$\mathcal{T}_i = \{t \in [T] : i \in S_t\}, \qquad n_i = |\mathcal{T}_i|.$$

The standard elliptical-potential bound (Abbasi-Yadkori et al., 2011) gives

$$\sum_{t\in\mathcal{T}_i} \left(1 \wedge \|\boldsymbol{x}_{t,i}\|^2_{X^{-1}_{t-1,i}}\right) \le 2\log\frac{\det X_{T,i}}{\det(\lambda I)}.$$

By concavity,

$$\begin{aligned}\sum_{t\in\mathcal{T}_i} \left(1 \wedge \|\boldsymbol{x}_{t,i}\|^2_{X^{-1}_{t-1,i}}\right)^{\alpha} &\le n_i^{1-\alpha}\left[2\log\frac{\det X_{T,i}}{\det(\lambda I)}\right]^{\alpha} \\ &= \widetilde{\mathcal{O}}\left(n_i^{1-\alpha} d_x^{\alpha}\right).\end{aligned}$$

Therefore,

$$R_T^{\phi} = \widetilde{\mathcal{O}}\left(B_1 d_x^{\alpha} \sum_{i=1}^{m} n_i^{1-\alpha}\right).$$

Since

$$\sum_{i=1}^{m} n_i = \sum_{t=1}^{T} |S_t| \le KT,$$

concavity implies

$$\sum_{i=1}^{m} n_i^{1-\alpha} \le m^{\alpha}(KT)^{1-\alpha}.$$

Thus,

$$R_T^{\phi} = \widetilde{O}\left(B_1 m^{\alpha}(KT)^{1-\alpha} d_x^{\alpha}\right). \tag{17}$$

C.2 Reward-Parameter Term

Split the selected-arm occurrences into

$$\mathcal{L} = \left\{(t,i) : i \in S_t,\ q_{t,i}^2 \le \frac{1}{K}\right\}$$

and

$$\mathcal{H} = \left\{(t,i) : i \in S_t,\ q_{t,i}^2 > \frac{1}{K}\right\}.$$

Using equation 16, the bound on $\ell_{t,i}$, and the clipping bound,

$$R_T \le 2B_1 \sum_{t=1}^{T}\sum_{i\in S_t} e_{t-1,i}(\boldsymbol{x}_{t,i}) + 2B_1\zeta_T \sum_{(t,i)\in\mathcal{L}} q_{t,i} + 2B_1 B|\mathcal{H}|.$$

Recall the NBT-EPL potential

$$\Gamma_T = \sum_{t=1}^{T}\sum_{i\in S_t} \min\left\{\frac{1}{K}, q_{t,i}^2\right\}.$$

By Lemma 1,

$$\Gamma_T = \widetilde{\mathcal{O}}(d_u).$$

For the low-width occurrences, Cauchy–Schwarz gives

$$\sum_{(t,i)\in\mathcal{L}} q_{t,i} \le \sqrt{KT \sum_{(t,i)\in\mathcal{L}} q_{t,i}^2} \le \sqrt{KT\Gamma_T},$$

and hence

$$R_{T,\text{low}}^{w} = \widetilde{O}\left(B_1\zeta_T\sqrt{KTd_u}\right)$$

For the high-width occurrences,

$$\frac{|\mathcal{H}|}{K} \leq \Gamma_T,$$

so

$$|\mathcal{H}| = \widetilde{\mathcal{O}}(Kd_u).$$

Therefore,

$$R^w_{T,\text{high}} = \widetilde{\mathcal{O}}(B_1 BKd_u).$$

Combining the two parts,

$$R^w_T = \widetilde{\mathcal{O}}\left(B_1\zeta_T\sqrt{KTd_u} + B_1BKd_u\right).$$

Using

$$\zeta_T = \widetilde{\mathcal{O}}\left(\sqrt{\lambda} + R_\eta\sqrt{d_u}\right),$$

we obtain

$$R^w_T = \widetilde{\mathcal{O}}\left(B_1R_\eta d_u\sqrt{KT} + B_1\sqrt{\lambda d_u KT} + B_1Bd_uK\right). \tag{18}$$

C.3 Combination

Combining equation 17 and equation 18 yields

$$R_T = \widetilde{O}\left(B_1 m^\alpha (KT)^{1-\alpha} d_x^\alpha + B_1R_\eta d_u\sqrt{KT} + B_1\sqrt{\lambda d_u KT} + B_1Bd_uK\right),$$

which proves Theorem 1 under Assumption 1.

For $\alpha = 1/2$,

$$R_T = \widetilde{O}\left(B_1\sqrt{mKTd_x} + B_1R_\eta d_u\sqrt{KT} + B_1\sqrt{\lambda d_u KT} + B_1Bd_uK\right),$$

D Proof of Lemma 1

Proof. For each round t, let

$$P_t = A_{t-1}^{-1}, \qquad b_{t,i} = \boldsymbol{u}_{t,i}^\top P_t \boldsymbol{u}_{t,i}.$$

$$q_{t,i}^2 = \|\bar{\boldsymbol{u}}_{t,i}\|^2_{A_{t-1}^{-1}}.$$

Let

$$\boldsymbol{\xi}_{t,i} = \begin{bmatrix} 0 \\ \boldsymbol{\epsilon}_{t,i} \end{bmatrix},$$

so

$$\boldsymbol{u}_{t,i} = \bar{\boldsymbol{u}}_{t,i} + \boldsymbol{\xi}_{t,i}.$$

we have

$$b_{t,i} = q_{t,i}^2 + 2\bar{\boldsymbol{u}}_{t,i}^\top P_t \boldsymbol{\xi}_{t,i} + \boldsymbol{\xi}_{t,i}^\top P_t \boldsymbol{\xi}_{t,i}.$$

By $\mathbb{E}[\boldsymbol{\xi}_{t,i} \mid \mathcal{F}_t^{\text{pre}}] = 0$,

$$\mathbb{E}\left[b_{t,i} \mid \mathcal{F}_t^{\text{pre}}\right] = q_{t,i}^2 + \mathbb{E}\left[\boldsymbol{\xi}_{t,i}^\top P_t \boldsymbol{\xi}_{t,i} \mid \mathcal{F}_t^{\text{pre}}\right] \geq q_{t,i}^2.$$

Define

$$\rho_{t,i} = \min\left\{1, \frac{1}{Kq_{t,i}^2}\right\},$$

with $\rho_{t,i} = 1$ when $q_{t,i} = 0$

Set

$$Z_{t,i} = \rho_{t,i} b_{t,i}.$$

Then

$$\rho_{t,i} q_{t,i}^2 = \min\left\{\frac{1}{K}, q_{t,i}^2\right\},$$

$$\mathbb{E}\left[Z_{t,i} \mid \mathcal{F}_t^{\mathrm{pre}}\right] \geq \min\left\{\frac{1}{K}, q_{t,i}^2\right\}.$$

Since $A_{t-1}^{-1} \preceq \lambda^{-1} I$ and $\lambda \geq L_u^2$,

$$0 \leq Z_{t,i} \leq b_{t,i} \leq \frac{\|\boldsymbol{u}_{t,i}\|_2^2}{\lambda} \leq 1.$$

For each round define

$$X_t = \sum_{i \in S_t} \min\left\{\frac{1}{K}, q_{t,i}^2\right\}, \qquad Y_t = \sum_{i \in S_t} Z_{t,i}, \qquad W_t = 1 \wedge Y_t.$$

Since $|S_t| \leq K$,

$$0 \leq X_t \leq 1.$$

Because $0 \leq Z_{t,i} \leq 1$,

$$W_t \geq 1 - \prod_{i \in S_t} (1 - Z_{t,i}).$$

$$\begin{aligned}
\mathbb{E}\left[W_t \mid \mathcal{F}_t^{\mathrm{pre}}\right] &\geq 1 - \prod_{i \in S_t} \left(1 - \mathbb{E}[Z_{t,i} \mid \mathcal{F}_t^{\mathrm{pre}}]\right) \\
&\geq 1 - \prod_{i \in S_t} \left(1 - \min\left\{\frac{1}{K}, q_{t,i}^2\right\}\right) \\
&\geq 1 - e^{-X_t} \\
&\geq (1 - e^{-1}) X_t,
\end{aligned}$$

where the final inequality uses $X_t \in [0, 1]$.

Next define

$$\boldsymbol{v}_{t,i} = A_{t-1}^{-1/2} \boldsymbol{u}_{t,i}$$

$$D_t = \log\det A_t - \log\det A_{t-1}.$$

Then

$$\begin{aligned}
D_t &= \log\det\left(I + \sum_{i \in S_t} v_{t,i} \boldsymbol{v}_{t,i}^\top\right) \\
&\geq \log\left(1 + \sum_{i \in S_t} \|\boldsymbol{v}_{t,i}\|_2^2\right) \\
&= \log\left(1 + \sum_{i \in S_t} b_{t,i}\right) \\
&\geq \log(1 + Y_t) \\
&\geq (\log 2) W_t.
\end{aligned}$$

Let

$$\mu_t = \mathbb{E}[W_t \mid \mathcal{F}_t^{\mathrm{pre}}].$$

For any $\gamma > 0$, define

$$c_\gamma := 1 - e^{-\gamma}$$

and

$$M_n = \exp\left(c_\gamma \sum_{t=1}^{n} \mu_t - \gamma \sum_{t=1}^{n} W_t\right), \qquad M_0 = 1.$$

Since $0 \le W_t \le 1$ and $w \mapsto e^{-\gamma w}$ is convex on $[0,1]$,

$$e^{-\gamma W_t} \le (1 - W_t) + W_t e^{-\gamma} = 1 - c_\gamma W_t.$$

Taking conditional expectation with respect to $\mathcal{F}_t^{\mathrm{pre}}$ gives

$$\mathbb{E}\big[e^{-\gamma W_t} \mid \mathcal{F}_t^{\mathrm{pre}}\big] \le 1 - c_\gamma \mu_t \le e^{-c_\gamma \mu_t},$$

where the last inequality uses $1 - x \le e^{-x}$. Moreover,

$$M_t = M_{t-1} \exp\left(c_\gamma \mu_t - \gamma W_t\right).$$

Since M_{t-1} and μ_t are $\mathcal{F}_t^{\mathrm{pre}}$-measurable,

$$\begin{aligned}
\mathbb{E}[M_t \mid \mathcal{F}_t^{\mathrm{pre}}] &= M_{t-1} e^{c_\gamma \mu_t} \mathbb{E}\big[e^{-\gamma W_t} \mid \mathcal{F}_t^{\mathrm{pre}}\big] \\
&\le M_{t-1} e^{c_\gamma \mu_t} e^{-c_\gamma \mu_t} \\
&= M_{t-1}.
\end{aligned}$$

Thus $\{M_t\}_{t \ge 0}$ is a nonnegative supermartingale.

Taking $\gamma = \log 2$ and applying Ville's inequality (Howard et al., 2020), with probability at least $1 - \delta_E$,

$$\sum_{t=1}^{T} \mu_t \le 2 \log 2 \sum_{t=1}^{T} W_t + 2 \log \frac{1}{\delta_E}.$$

Therefore,

$$\begin{aligned}
(1 - e^{-1})\Gamma_T = (1 - e^{-1}) \sum_{t=1}^{T} X_t \\
&\le \sum_{t=1}^{T} \mu_t \\
&\le 2 \log 2 \sum_{t=1}^{T} W_t + 2 \log \frac{1}{\delta_E} \\
&\le 2 \sum_{t=1}^{T} D_t + 2 \log \frac{1}{\delta_E} \\
&= 2 \log \frac{\det A_T}{\det(\lambda I)} + 2 \log \frac{1}{\delta_E}.
\end{aligned}$$

Hence

$$\Gamma_T \le \frac{2}{1 - e^{-1}} \left[\log \frac{\det A_T}{\det(\lambda I)} + \log \frac{1}{\delta_E}\right].$$

Finally,

$$\operatorname{tr}(A_T - \lambda I) = \sum_{t=1}^{T} \sum_{i \in S_t} \|\boldsymbol{u}_{t,i}\|_2^2 \le K T L_u^2.$$

By AM–GM,

$$\log \frac{\det A_T}{\det(\lambda I)} \le d_u \log\left(1 + \frac{K T L_u^2}{\lambda d_u}\right).$$

Consequently,

$$\Gamma_T = \widetilde{\mathcal{O}}(d_u),$$

□

E SAMPLE-SIZE LEARNABILITY ANALYSIS

We show how Assumption 2 changes the post-serving context-learning term in Theorem 1. The reward-parameter analysis is unchanged.

For every arm $i \in [m]$, recall

$$n_{t-1}(i) = \sum_{s<t} \mathbf{1}\{i \in S_s\}, \qquad \bar{n}_{t-1}(i) = \max\{1, n_{t-1}(i)\}.$$

Under Assumption 2,

$$\sup_{\boldsymbol{x}\in\mathcal{X}} \left\| \widehat{\phi}_{t-1,i}(\boldsymbol{x}) - \phi_i^\star(\boldsymbol{x}) \right\|_2 \le C_0 \sqrt{d_x}\, \bar{n}_{t-1}(i)^{-a} \log\left(\frac{m^2 T \bar{n}_{t-1}(i)}{\delta_\phi} \right).$$

Since $\bar{n}_{t-1}(i) \le T$, let

$$\Lambda_\phi = \log\left(\frac{m^2 T^2}{\delta_\phi} \right).$$

Then the prediction radius can be bounded by

$$e_{t-1,i} \le C_0 \sqrt{d_x}\, \bar{n}_{t-1}(i)^{-a} \Lambda_\phi.$$

As in Appendix C, the post-serving context contribution satisfies

$$R_T^\phi \le 2B_1 \sum_{t=1}^{T} \sum_{i\in S_t} e_{t-1,i}.$$

Therefore,

$$R_T^\phi \le 2B_1 C_0 \sqrt{d_x} \Lambda_\phi \sum_{t=1}^{T} \sum_{i\in S_t} \bar{n}_{t-1}(i)^{-a}.$$

For each arm, define

$$\mathcal{T}_i = \{t \in [T] : i \in S_t\}, \qquad n_i = |\mathcal{T}_i|.$$

For any arm with $n_i \ge 1$, At the j-th selection of arm i, it has been selected $j-1$ times previously. Hence

$$\sum_{t\in\mathcal{T}_i} \bar{n}_{t-1}(i)^{-a} = 1 + \sum_{r=1}^{n_i-1} r^{-a}.$$

Since $a \in (0,1)$, an integral comparison gives

$$1 + \sum_{r=1}^{n_i-1} r^{-a} \le \frac{2}{1-a} n_i^{1-a}.$$

Thus,

$$R_T^\phi \le \frac{4B_1 C_0 \sqrt{d_x} \Lambda_\phi}{1-a} \sum_{i=1}^{m} n_i^{1-a}.$$

Let

$$N_T = \sum_{i=1}^{m} n_i = \sum_{t=1}^{T} |S_t| \le KT.$$

Since x^{1-a} is concave,

$$\sum_{i=1}^{m} n_i^{1-a} \le m^a \left(\sum_{i=1}^{m} n_i \right)^{1-a} = m^a N_T^{1-a}.$$

Therefore,

$$R_T^\phi \le \frac{4B_1 C_0 \sqrt{d_x} \Lambda_\phi}{1-a} m^a (KT)^{1-a}.$$

Equivalently,

$$R_T^{\phi} = \widetilde{\mathcal{O}}\left(B_1\sqrt{d_x}\, m^a (KT)^{1-a}\right).$$

Replacing the context-learning term in Appendix C by this bound yields the second statement of Theorem 1. In particular, for $a = 1/2$,

$$R_T^{\phi} = \widetilde{\mathcal{O}}\left(B_1\sqrt{mKTd_x}\right).$$

F Proof of Theorem 3

Work on the joint confidence event. For each arm, write

$$\widehat{\boldsymbol{z}}_{t,i} = \widehat{\phi}_{t-1,i}(\boldsymbol{x}_{t,i}), \qquad \widehat{\boldsymbol{u}}_{t,i} = \begin{bmatrix} \boldsymbol{x}_{t,i} \\ \widehat{\boldsymbol{z}}_{t,i} \end{bmatrix},$$

and let

$$e_{t,i} = e_{t-1,i}^{\delta_\phi/(mT)}(\boldsymbol{x}_{t,i}).$$

Then

$$\|\widehat{\boldsymbol{z}}_{t,i} - \phi_i^\star(\boldsymbol{x}_{t,i})\|_2 \le e_{t,i}.$$

Optimism. The reward-parameter confidence bound gives

$$\widehat{\boldsymbol{u}}_{t,i}^\top \boldsymbol{w}^\star \le \widehat{\boldsymbol{u}}_{t,i}^\top \widehat{\boldsymbol{w}}_{t-1} + \zeta_{t-1}\|\widehat{\boldsymbol{u}}_{t,i}\|_{A_{t-1}^{-1}}.$$

Moreover,

$$\begin{aligned} \mu_{t,i} - \widehat{\boldsymbol{u}}_{t,i}^\top \boldsymbol{w}^\star &= \langle \phi_i^\star(\boldsymbol{x}_{t,i}) - \widehat{\boldsymbol{z}}_{t,i}, \boldsymbol{\beta}^\star \rangle \\ &\le e_{t,i}. \end{aligned}$$

Hence

$$\mu_{t,i} \le U_{t,i}^{\mathrm{raw}}.$$

Since $\mu_{t,i} \in [-B, B]$,

$$U_{t,i} = \widetilde{U}_{t,i} = \mathrm{clip}_{[-B,B]}(U_{t,i}^{\mathrm{raw}}) \ge \mu_{t,i}.$$

By monotonicity, for every feasible set S,

$$F_t(S; \boldsymbol{\mu}_t) \le F_t(S; \widetilde{\boldsymbol{U}}_t).$$

By the C^{2}UCB-PC selection rule,

$$F_t(S_t^\star; \boldsymbol{\mu}_t) \le F_t(S_t^\star; \widetilde{\boldsymbol{U}}_t) \le F_t(S_t; \widetilde{\boldsymbol{U}}_t).$$

Hence, by the 1-norm bounded-smoothness condition,

$$R_T^{\mathrm{C^2UCB\text{-}PC}} \le B_1 \sum_{t=1}^{T} \sum_{i \in S_t} (U_{t,i} - \mu_{t,i}).$$

Regret decomposition. Define

$$q_{t,i} = \|\bar{\boldsymbol{u}}_{t,i}\|_{A_{t-1}^{-1}}, \qquad \boldsymbol{d}_{t,i}^{\phi} = \widehat{\boldsymbol{u}}_{t,i} - \bar{\boldsymbol{u}}_{t,i}.$$

Then

$$\|\boldsymbol{d}_{t,i}^{\phi}\|_2 \le e_{t,i}.$$

Since $A_{t-1} \succeq \lambda I$,

$$\begin{aligned} \|\widehat{\boldsymbol{u}}_{t,i}\|_{A_{t-1}^{-1}} &\le \|\bar{\boldsymbol{u}}_{t,i}\|_{A_{t-1}^{-1}} + \|\boldsymbol{d}_{t,i}^{\phi}\|_{A_{t-1}^{-1}} \\ &\le q_{t,i} + \frac{e_{t,i}}{\sqrt{\lambda}}. \end{aligned}$$

Using the reward confidence bound and $\|(\boldsymbol{d}_{t,i}^{\phi})\|_2 \le e_{t,i}$,

$$\begin{aligned} U_{t,i}^{\mathrm{raw}} - \mu_{t,i} &= \widehat{\boldsymbol{u}}_{t,i}^{\top}(\widehat{\boldsymbol{w}}_{t-1} - \boldsymbol{w}^{\star}) + (\boldsymbol{d}_{t,i}^{\phi})^{\top}\boldsymbol{w}^{\star} \\ &\quad + \zeta_{t-1}\|\widehat{\boldsymbol{u}}_{t,i}\|_{A_{t-1}^{-1}} + e_{t,i} \\ &\le 2\zeta_{t-1}q_{t,i} + 2\left(1 + \frac{\zeta_{t-1}}{\sqrt{\lambda}}\right)e_{t,i}. \end{aligned}$$

Clipping also gives

$$U_{t,i} - \mu_{t,i} \le 2B.$$

Define

$$\mathcal{L} = \left\{(t,i) : i \in S_t,\ q_{t,i}^2 \le \frac{1}{K}\right\},$$

and

$$\mathcal{H} = \left\{(t,i) : i \in S_t,\ q_{t,i}^2 > \frac{1}{K}\right\}.$$

Therefore,

$$\begin{aligned} R_T^{\mathrm{C^2UCB\text{-}PC}} &\le 2B_1\zeta_T \sum_{(t,i)\in\mathcal{L}} q_{t,i} + 2B_1 B|\mathcal{H}| \\ &\quad + 2B_1\left(1 + \frac{\zeta_T}{\sqrt{\lambda}}\right)\sum_{t=1}^{T}\sum_{i\in S_t} e_{t,i}. \end{aligned} \tag{19}$$

Reward-parameter contribution. By Lemma 1,

$$\Gamma_T = \sum_{t=1}^{T}\sum_{i\in S_t} \min\left\{\frac{1}{K}, q_{t,i}^2\right\} = \widetilde{\mathcal{O}}(d_u).$$

Thus,

$$\sum_{(t,i)\in\mathcal{L}} q_{t,i} \le \sqrt{KT\Gamma_T} = \widetilde{\mathcal{O}}\left(\sqrt{KTd_u}\right),$$

while

$$|\mathcal{H}| \le K\Gamma_T = \widetilde{\mathcal{O}}(Kd_u).$$

Hence the first two terms in equation 19 are

$$\widetilde{\mathcal{O}}\left(B_1\zeta_T\sqrt{KTd_u} + B_1BKd_u\right).$$

Post-serving context contribution. The same arm-wise generalized elliptical-potential argument as in Appendix C gives

$$\sum_{t=1}^{T}\sum_{i\in S_t} e_{t,i} = \widetilde{\mathcal{O}}\left(m^{\alpha}(KT)^{1-\alpha}d_x^{\alpha}\right).$$

Substituting into equation 19 yields

$$R_T^{\mathrm{C^2UCB\text{-}PC}} = \widetilde{\mathcal{O}}\left(B_1\zeta_T\sqrt{KTd_u} + B_1BKd_u + B_1\left(1 + \frac{\zeta_T}{\sqrt{\lambda}}\right)m^{\alpha}(KT)^{1-\alpha}d_x^{\alpha}\right).$$

This proves theorem 3.

G Proof of Theorem 2

We prove the context-learning and reward-parameter lower bounds separately on two subclasses of our problem.

For simplicity, let

$$q = \frac{m}{K},$$

and partition the arms into K groups $G_1, \ldots, G_K$, each containing q arms. Consider the feasible family

$$\mathcal{S} = \{S \subseteq [m] : |S \cap G_g| = 1,\ \forall g \in [K]\}.$$

G.1 Post-Serving Context Hardness

Partition the horizon into d_x blocks of length

$$n_C = \left\lfloor \frac{T}{d_x} \right\rfloor.$$

We ignore the remaining at most $d_x - 1$ rounds. Let $s = \sqrt{\lambda/2}$. During block j, all arms use

$$\boldsymbol{x}_{t,g,a} = s\boldsymbol{e}_j.$$

For each pair (g, j), independently choose

$$J_{g,j} \sim \mathrm{Unif}(G_g)$$

and define

$$p_{g,a,j} = \frac{1}{2} + \Delta_C \mathbf{1}\{a = J_{g,j}\}, \qquad \Delta_C = c \min\left\{1, \sqrt{\frac{q}{n_C}}\right\},$$

for a sufficiently small universal constant $c > 0$.

Let $\boldsymbol{e}_1^{(z)}$ be the first standard basis vector of $\mathbb{R}^{d_z}$ and define

$$\phi^\star_{g,a}(s\boldsymbol{e}_j) = p_{g,a,j}\boldsymbol{e}_1^{(z)}.$$

Generate

$$Y_{t,g,a} \sim \mathrm{Bernoulli}(p_{g,a,j}), \qquad \boldsymbol{z}_{t,g,a} = Y_{t,g,a}\boldsymbol{e}_1^{(z)},$$

and set

$$\boldsymbol{\theta}^\star = 0, \qquad \boldsymbol{\beta}^\star = \boldsymbol{e}_1^{(z)}, \qquad \eta_{t,g,a} = 0.$$

Then

$$r_{t,g,a} = \boldsymbol{z}_{t,g,a}^\top \boldsymbol{\beta}^\star = Y_{t,g,a}.$$

For each fixed (g, j), the learner faces a standard q-armed stochastic bandit problem of length n_C. Hence the standard minimax lower bound for stochastic multi-armed bandits (Lattimore & Szepesvári, 2020) gives

$$\mathbb{E}[R_{g,j}] = \Omega\left(\min\{n_C, \sqrt{qn_C}\}\right).$$

Summing over the Kd_x independent subproblems yields

$$\mathbb{E}[R_T^C] = \Omega\left(\min\{KT, \sqrt{mKTd_x}\}\right).$$

It remains to verify that the construction satisfies the canonical elliptical-learnability condition. If $s\boldsymbol{e}_j$ has been observed $N_{t,i,j}$ times for arm i, then

$$\|s\boldsymbol{e}_j\|^2_{X_{t,i}^{-1}} = \frac{s^2}{\lambda + N_{t,i,j}s^2} = \frac{1}{N_{t,i,j} + 2}.$$

Since Bernoulli mean estimation has error $\widetilde{\mathcal{O}}((1 + N_{t,i,j})^{-1/2})$, the construction satisfies Assumption 1 with $\alpha = 1/2$.

G.2 Reward-Parameter Hardness

We next isolate the difficulty of learning the shared reward parameter. Let

$$D = \max\{d_x, d_z\}.$$

Since

$$D \le d_u = d_x + d_z \le 2D,$$

we have

$$D = \Theta(d_u).$$

We embed the standard reward hard instances of Takemura et al. (2021) into the larger feature block.

If $d_x \geq d_z$, we use the pre-serving coordinates, set

$$\boldsymbol{z}_{t,i} = 0, \qquad \boldsymbol{\beta}^\star = 0,$$

and identify the unknown parameter in the Takemura et al. construction with the corresponding coordinates of $\boldsymbol{\theta}^\star$. Thus

$$r_{t,i} = \boldsymbol{x}_{t,i}^\top \boldsymbol{\theta}^\star + \eta_{t,i}.$$

If $d_z > d_x$, we instead set

$$\boldsymbol{\theta}^\star = 0, \qquad \boldsymbol{\epsilon}_{t,i} = 0,$$

and embed the Takemura et al. feature vectors in the post-serving coordinates. Let ℓ index the feature stages. For each stage, choose a distinct pre-serving context $\boldsymbol{x}^{(\ell)}$ and define the fixed mapping by

$$\phi_i^\star(\boldsymbol{x}^{(\ell)}) = \boldsymbol{v}_{\ell,i},$$

For a round t in stage ℓ, write $\boldsymbol{v}_{t,i} = \boldsymbol{v}_{\ell,i}$. We may reveal these deterministic mappings to the learner, so the post-serving features are effectively known before action selection. The unknown reward parameter is now $\boldsymbol{\beta}^\star$, and

$$\boldsymbol{z}_{t,i} = \boldsymbol{v}_{t,i}, \qquad r_{t,i} = \boldsymbol{v}_{t,i}^\top \boldsymbol{\beta}^\star + \eta_{t,i}.$$

Hence, in either case, the learner faces the same linear reward-learning problem as in Takemura et al. (2021) on $D = \Theta(d_u)$ effective dimensions. For the stochastic reward construction, use

$$d_1 = \left\lfloor \frac{D}{2} \right\rfloor$$

coordinates. By the assumption of Theorem 2,

$$q = \frac{m}{K} \geq 2^{d_1},$$

so each group contains enough arms to realize all hypercube feature vectors required by the construction of Takemura et al. (2021, Theorem 5). The remaining $\Theta(D)$ coordinates are used for the same-round batch construction of Takemura et al. (2021, Theorem 6), discarding at most a constant number of coordinates when necessary.

The hypercube construction of Takemura et al. (2021, Theorem 5) therefore gives

$$R_1 = \Omega\left(R_\eta d_u \sqrt{KT}\right).$$

The same-round batch construction of Takemura et al. (2021, Theorem 6), applied on the second subspace, gives

$$R_2 = \Omega\left(\min\{Bd_uK, BKT\}\right).$$

The two constructions define two admissible subclasses of our problem. Hence, the minimax regret over the full class is at least the maximum of their respective lower bounds. Since

$$\max\{R_1, R_2\} \geq \frac{R_1 + R_2}{2},$$

we obtain

$$\mathbb{E}[R_T^R] = \Omega\left(R_\eta d_u \sqrt{KT} + \min\{Bd_uK, BKT\}\right).$$

Moreover, the condition

$$KT = \Omega\left(\left(\frac{R_\eta d_u}{B}\right)^2\right)$$

used by Takemura et al. (2021, Theorem 7) is automatically satisfied in the non-saturated regime considered below.

G.3 COMBINATION

Define

$$A_T = \min\{KT, \sqrt{mKTd_x}\}$$

and

$$W_T = \min\left\{R_\eta d_u\sqrt{KT} + Bd_uK,\ BKT\right\}.$$

The context-hard construction gives

$$\inf_\pi \sup_{\nu\in\mathcal{P}} \mathbb{E}_\nu^\pi[R_T] = \Omega(A_T),$$

while the reward-hard construction gives

$$\inf_\pi \sup_{\nu\in\mathcal{P}} \mathbb{E}_\nu^\pi[R_T] = \Omega(W_T).$$

Therefore,

$$\inf_\pi \sup_{\nu\in\mathcal{P}} \mathbb{E}_\nu^\pi[R_T] = \Omega\left(\max\{A_T, W_T\}\right).$$

Since

$$\max\{A_T, W_T\} \geq \frac{A_T + W_T}{2},$$

we obtain

$$\inf_\pi \sup_{\nu\in\mathcal{P}} \mathbb{E}_\nu^\pi[R_T] = \Omega\left(\min\{KT, \sqrt{mKTd_x}\} + \min\left\{R_\eta d_u\sqrt{KT} + Bd_uK,\ BKT\right\}\right).$$

In the non-saturated regime

$$\sqrt{mKTd_x} \leq KT$$

and

$$R_\eta d_u\sqrt{KT} + Bd_uK \leq BKT,$$

the lower bound becomes

$$\inf_\pi \sup_{\nu\in\mathcal{P}} \mathbb{E}_\nu^\pi[R_T] = \Omega\left(\sqrt{mKTd_x} + R_\eta d_u\sqrt{KT} + Bd_uK\right).$$

This proves Theorem 2.

H VARIANCE-AWARE ALGORITHMS AND THEIR REGRET ANALYSIS

We give the details of the variance-aware extension described in Section 5. The post-serving context model remains unchanged.

Assume that the feasible arm means lie in

$$[\nu, 1-\nu], \qquad \nu \in (0, 1/2),$$

and define

$$V_{\min} = \nu(1-\nu), \qquad V_{\max} = \frac{1}{4}.$$

We further assume

$$\mathcal{W} \subseteq \left\{\boldsymbol{w} \in \mathbb{R}^{d_u} : \|\boldsymbol{w}\|_2 \leq S_w\right\}, \qquad \|\boldsymbol{\beta}^\star\|_2 \leq 1.$$

For every round t, let $F_t(S;\boldsymbol{\mu})$ denote the expected reward of action S under the arm-mean vector $\boldsymbol{\mu}$. Following Liu et al. (2022), we assume that F_t is monotone and satisfies the variance-modulated (B_v, B_1)-bounded smoothness condition: for any $\boldsymbol{\mu}, \boldsymbol{\mu}' \in [\nu, 1-\nu]^m$ and $\boldsymbol{\zeta}, \boldsymbol{\kappa} \in [-1, 1]^m$ such that $\boldsymbol{\mu}' = \boldsymbol{\mu} + \boldsymbol{\zeta} + \boldsymbol{\kappa}$,

$$|F_t(S;\boldsymbol{\mu}') - F_t(S;\boldsymbol{\mu})| \leq B_v\sqrt{\sum_{i\in S}\frac{\zeta_i^2}{\mu_i(1-\mu_i)}} + B_1\sum_{i\in S}|\kappa_i|, \tag{20}$$

H.1 VARIANCE-AWARE ALGORITHMS

This subsection gives the full description of VA-OFUC^2-PC (Algorithm 3) and VA-C^{2}UCB-PC (Algorithm 4). Both follow their counterparts in the main text (Algorithms 1 and 2) and change only how the shared reward parameter is estimated. We first explain the idea, then describe the two components that the algorithms share, and finally go through each algorithm step by step.

Idea. OFUC^2-PC and C^{2}UCB-PC estimate the reward parameter by an unweighted ridge regression, which treats all selected-arm observations in the same way. Under the variance-modulated smoothness condition equation 20, however, the estimation error ζ_i of arm i enters the regret through $\zeta_i^2/(\mu_{t,i}(1-\mu_{t,i}))$, i.e., relative to the variance proxy of this arm, so the observations of arms with a smaller variance proxy are more valuable. Inspired by the optimistic-variance approach of Liu et al. (2023), the variance-aware algorithms therefore weight observation (t,i) by $1/\bar{V}_{t,i}$ in the ridge regression, where $\bar{V}_{t,i}$ is an optimistic estimate of the unknown variance proxy $\mu_{t,i}(1-\mu_{t,i})$. With these weights, the confidence width of arm i becomes $s_{t,i} = \|\bar{\boldsymbol{u}}_{t,i}\|^2_{(G_t^V)^{-1}}/\bar{V}_{t,i}$, which is the quantity that equation 20 produces and that the variance-scaled NBT-EPL (Lemma 2) controls. Two requirements shape the construction. First, $\bar{V}_{t,i}$ has to be computed before the feedback of round t is observed, so that the weights are $\mathcal{F}_t^{\mathrm{pre}}$-measurable and the self-normalized concentration of Appendix H.2 still applies. Second, $\bar{V}_{t,i}$ has to stay in $[V_{\min}, 1/4]$, which is why all mean estimates are restricted to $[\nu, 1-\nu]$.

Shared component 1: variance-weighted reward estimator. Let $\lambda_V \geq \max\{1, L_u^2/V_{\min}\}$ be the regularization parameter, $S_w = \sup_{\boldsymbol{w}\in\mathcal{W}} \|\boldsymbol{w}\|_2$, and $R_V = R_\eta/\sqrt{V_{\min}}$. Both algorithms maintain the weighted statistics of the rounds before t,

$$G_t^V = \lambda_V I + \sum_{s<t}\sum_{i\in S_s} \frac{\boldsymbol{u}_{s,i}\boldsymbol{u}_{s,i}^\top}{\bar{V}_{s,i}}, \qquad \boldsymbol{c}_t^V = \sum_{s<t}\sum_{i\in S_s} \frac{r_{s,i}\boldsymbol{u}_{s,i}}{\bar{V}_{s,i}}, \tag{21}$$

the weighted ridge estimator $\widehat{\boldsymbol{w}}_t^V = (G_t^V)^{-1}\boldsymbol{c}_t^V$, and the confidence set

$$\mathcal{C}_t^{w,V} = \left\{\boldsymbol{w}\in\mathcal{W} : \|\boldsymbol{w} - \widehat{\boldsymbol{w}}_t^V\|_{G_t^V} \leq \rho_t^V\right\}, \qquad \rho_t^V = \sqrt{\lambda_V}S_w + R_V\sqrt{2\log\frac{1}{\delta_w} + \log\frac{\det G_t^V}{\det(\lambda_V I)}}. \tag{22}$$

The statistics start from $G_1^V = \lambda_V I$ and $\boldsymbol{c}_1^V = 0$ and are updated after each round by

$$G_{t+1}^V = G_t^V + \sum_{i\in S_t}\frac{\boldsymbol{u}_{t,i}\boldsymbol{u}_{t,i}^\top}{\bar{V}_{t,i}}, \qquad \boldsymbol{c}_{t+1}^V = \boldsymbol{c}_t^V + \sum_{i\in S_t}\frac{r_{t,i}\boldsymbol{u}_{t,i}}{\bar{V}_{t,i}}. \tag{23}$$

Appendix H.2 shows that $\boldsymbol{w}^\star \in \mathcal{C}_t^{w,V}$ for all t with probability at least $1-\delta_w$.

Shared component 2: optimistic variance. Suppose that an interval $[\underline{\mu}_{t,i}, \overline{\mu}_{t,i}] \subseteq [\nu, 1-\nu]$ contains the mean reward $\mu_{t,i}$ when all confidence sets hold. The optimistic variance is the largest variance proxy over this interval,

$$\bar{V}_{t,i} = \max_{v\in[\underline{\mu}_{t,i},\overline{\mu}_{t,i}]} v(1-v) = \check{\mu}_{t,i}(1-\check{\mu}_{t,i}), \qquad \check{\mu}_{t,i} = \mathrm{clip}_{[\underline{\mu}_{t,i},\overline{\mu}_{t,i}]}(1/2), \tag{24}$$

where the closed form holds because $v(1-v)$ is concave with its maximum at $1/2$. By construction, $V_{\min} \leq \bar{V}_{t,i} \leq 1/4$, and $\mu_{t,i}(1-\mu_{t,i}) \leq \bar{V}_{t,i}$ when all confidence sets hold. If the interval is empty, which can happen only when some confidence set fails, we set $\bar{V}_{t,i} = 1/4$. The two algorithms differ in how they compute the interval and in how they select the action.

VA-OFUC^2-PC step by step. Lines 1 and 3 of Algorithm 3 are the same as in Algorithm 1, with the weighted statistics in place of A_0 and $\boldsymbol{b}_0$. In Line 4, the learner computes the weighted estimator and constructs the variance-weighted reward confidence set, together with the same post-serving confidence sets $\mathcal{C}_{t-1}^{\phi}(i, \boldsymbol{x}_{t,i})$ as in OFUC^2-PC. In Line 5, it selects the action by joint optimism as

Algorithm 3 VA-OFUC^2-PC

Require: Set of combinatorial actions $\mathcal{S}$, regularization parameter λ_V, range parameter ν, confidence levels δ_w, δ_ϕ

1: Initialize $G_1^V = \lambda_V I$, $\boldsymbol{c}_1^V = 0$, and $\widehat{\phi}_{0,i}$ for all $i \in [m]$
2: **for** $t = 1, \dots, T$ **do**
3: Observe the pre-serving contexts $\{\boldsymbol{x}_{t,i}\}_{i=1}^m$ of all base arms
4: Compute $\widehat{\boldsymbol{w}}_t^V = (G_t^V)^{-1}\boldsymbol{c}_t^V$ and construct $\mathcal{C}_t^{w,V}$ in equation 22 and $\{\mathcal{C}_{t-1}^{\phi}(i, \boldsymbol{x}_{t,i})\}_{i=1}^m$
5: Select $S_t \in \arg\max_{S\in\mathcal{S}} U_t^V(S)$ with $U_t^V(S)$ defined in equation 25
6: Compute $[\underline{\mu}_{t,i}, \overline{\mu}_{t,i}]$ by equation 26 and $\bar{V}_{t,i}$ by equation 24 for all $i \in S_t$
7: Play S_t and observe $\{(\boldsymbol{z}_{t,i}, r_{t,i}) : i \in S_t\}$
8: Update G_{t+1}^V and $\boldsymbol{c}_{t+1}^V$ by equation 23, and $\widehat{\phi}_{t,i}$ for all $i \in S_t$
9: **end for**

Algorithm 4 VA-C^{2}UCB-PC

Require: Set of combinatorial actions $\mathcal{S}$, oracle Oracle_t, regularization parameter λ_V, range parameter ν, confidence levels δ_w, δ_ϕ

1: Initialize $G_1^V = \lambda_V I$, $\boldsymbol{c}_1^V = 0$, and $\widehat{\phi}_{0,i}$ for all $i \in [m]$
2: **for** $t = 1, \dots, T$ **do**
3: Observe the pre-serving contexts $\{\boldsymbol{x}_{t,i}\}_{i=1}^m$ of all base arms
4: Compute $\widehat{\boldsymbol{w}}_t^V = (G_t^V)^{-1}\boldsymbol{c}_t^V$ and the radius ρ_t^V in equation 22
5: Predict $\widehat{\boldsymbol{z}}_{t,i} = \widehat{\phi}_{t-1,i}(\boldsymbol{x}_{t,i})$ and form $\widehat{\boldsymbol{u}}_{t,i} = [\boldsymbol{x}_{t,i}, \widehat{\boldsymbol{z}}_{t,i}]$ for all $i \in [m]$
6: Compute the upper confidence bound $\overline{\mu}_{t,i}$ in equation 27 for all $i \in [m]$
7: Select $S_t = \mathrm{Oracle}_t(\overline{\boldsymbol{\mu}}_t)$
8: Compute the lower confidence bound $\underline{\mu}_{t,i}$ in equation 27 and $\bar{V}_{t,i}$ by equation 24 for all $i \in S_t$
9: Play S_t and observe $\{(\boldsymbol{z}_{t,i}, r_{t,i}) : i \in S_t\}$
10: Update G_{t+1}^V and $\boldsymbol{c}_{t+1}^V$ by equation 23, and $\widehat{\phi}_{t,i}$ for all $i \in S_t$
11: **end for**

in equation 7, with two changes. The reward confidence set is $\mathcal{C}_t^{w,V}$, and the optimistic means are constrained to $[\nu, 1-\nu]$ instead of being clipped to $[-B, B]$, i.e.,

$$U_t^V(S) = \max_{\substack{\boldsymbol{w}\in\mathcal{C}_t^{w,V},\ \boldsymbol{z}_i\in\mathcal{C}_{t-1}^{\phi}(i,\boldsymbol{x}_{t,i}),\\ [\boldsymbol{x}_{t,i},\boldsymbol{z}_i]^\top\boldsymbol{w}\in[\nu,1-\nu],\ i\in S}} F_t\big(S; \widetilde{\boldsymbol{\mu}}_t\big), \qquad \widetilde{\mu}_{t,i} = [\boldsymbol{x}_{t,i}, \boldsymbol{z}_i]^\top \boldsymbol{w}. \tag{25}$$

When all confidence sets hold, the true pair $(\boldsymbol{w}^\star, \{\phi_i^\star(\boldsymbol{x}_{t,i})\}_{i\in S})$ satisfies these constraints because $\mu_{t,i} \in [\nu, 1-\nu]$, so the maximization is feasible and $U_t^V(S) \geq F_t(S; \boldsymbol{\mu}_t)$. In Line 6, the learner computes the optimistic variances of the selected arms, which are the only ones that enter the update. The interval of arm i is obtained from the two confidence sets,

$$\overline{\mu}_{t,i} = \min\Big\{1-\nu,\ \max_{\boldsymbol{z},\boldsymbol{w}} [\boldsymbol{x}_{t,i}, \boldsymbol{z}]^\top \boldsymbol{w}\Big\}, \qquad \underline{\mu}_{t,i} = \max\Big\{\nu,\ \min_{\boldsymbol{z},\boldsymbol{w}} [\boldsymbol{x}_{t,i}, \boldsymbol{z}]^\top \boldsymbol{w}\Big\}, \tag{26}$$

where the maximum and the minimum are over $\boldsymbol{z} \in \mathcal{C}_{t-1}^{\phi}(i, \boldsymbol{x}_{t,i})$ and $\boldsymbol{w} \in \mathcal{C}_t^{w,V}$, and $\bar{V}_{t,i}$ then follows from equation 24. Note that this line uses only information that is available before the feedback of round t, so $\bar{V}_{t,i}$ is $\mathcal{F}_t^{\mathrm{pre}}$-measurable. In Lines 7–8, the learner plays S_t, observes the post-serving contexts and rewards of the selected arms, and updates the weighted statistics with the weights $1/\bar{V}_{t,i}$ and the mapping estimators of the selected arms.

VA-C^{2}UCB-PC step by step. Algorithm 4 follows Algorithm 2 and avoids the joint maximization. In Line 4, the learner computes the weighted estimator and its radius. In Line 5, it predicts the post-serving context of every arm and forms the predicted augmented context, as in C^{2}UCB-PC. In Line 6, it computes an arm-wise confidence interval from a reward bonus and a post-serving bonus,

$$\begin{aligned} \overline{\mu}_{t,i} &= \min\big\{1-\nu,\ \widehat{\boldsymbol{u}}_{t,i}^\top \widehat{\boldsymbol{w}}_t^V + \rho_t^V \|\widehat{\boldsymbol{u}}_{t,i}\|_{(G_t^V)^{-1}} + e_{t-1,i}(\boldsymbol{x}_{t,i})\big\}, \\ \underline{\mu}_{t,i} &= \max\big\{\nu,\ \widehat{\boldsymbol{u}}_{t,i}^\top \widehat{\boldsymbol{w}}_t^V - \rho_t^V \|\widehat{\boldsymbol{u}}_{t,i}\|_{(G_t^V)^{-1}} - e_{t-1,i}(\boldsymbol{x}_{t,i})\big\}. \end{aligned} \tag{27}$$

The reward bonus $\rho_t^V \|\widehat{\boldsymbol{u}}_{t,i}\|_{(G_t^V)^{-1}}$ accounts for the error of $\widehat{\boldsymbol{w}}_t^V$ at the predicted context, and the post-serving bonus $e_{t-1,i}(\boldsymbol{x}_{t,i})$ accounts for the prediction error of the context, whose effect on the

mean reward is at most $\|\boldsymbol{\beta}^\star\|_2\, e_{t-1,i}(\boldsymbol{x}_{t,i}) \le e_{t-1,i}(\boldsymbol{x}_{t,i})$. The upper confidence bounds $\overline{\mu}_{t,i}$ of all arms are the scores of the action selection. In Line 7, the learner selects $S_t = \mathrm{Oracle}_t(\overline{\boldsymbol{\mu}}_t)$ with $\overline{\boldsymbol{\mu}}_t = (\overline{\mu}_{t,1}, \ldots, \overline{\mu}_{t,m})^\top$, which uses the same oracle as C^{2}UCB-PC. In Line 8, it computes the lower confidence bounds and the optimistic variances of the selected arms, again before any feedback of round t is observed. Lines 9–10 are the same as in Algorithm 3.

Comparison of the two algorithms. VA-OFUC^2-PC keeps the joint optimism of OFUC^2-PC and additionally solves two bilinear problems for every selected arm to obtain the interval in equation 26, so it is computationally expensive. All quantities of VA-C^{2}UCB-PC are available in closed form, and its action selection needs one oracle call per round. The price is the factor $1 + \rho_T^V/\sqrt{\lambda_V}$ in the context-learning term, because the reward bonus in equation 27 is evaluated at the predicted context $\widehat{\boldsymbol{u}}_{t,i}$, whereas the variance-scaled NBT-EPL controls the mean augmented context $\bar{\boldsymbol{u}}_{t,i}$.

H.2 Variance-Weighted Confidence Set

Let $\bar{V}_{t,i}$ denote an optimistic variance computed before observing the current post-serving context and reward. Define

$$G_t^V = \lambda_V I + \sum_{s<t} \sum_{i \in S_s} \frac{\boldsymbol{u}_{s,i} \boldsymbol{u}_{s,i}^\top}{\bar{V}_{s,i}},$$

and

$$\boldsymbol{c}_t^V = \sum_{s<t} \sum_{i \in S_s} \frac{r_{s,i} \boldsymbol{u}_{s,i}}{\bar{V}_{s,i}}.$$

The variance-weighted ridge estimator is

$$\widehat{\boldsymbol{w}}_t^V = (G_t^V)^{-1} \boldsymbol{c}_t^V.$$

Since $\eta_{t,i}$ is conditionally R_η-sub-Gaussian and $\bar{V}_{t,i} \ge V_{\min}$, the normalized reward noise is conditionally R_V-sub-Gaussian with

$$R_V = \frac{R_\eta}{\sqrt{V_{\min}}}.$$

$$\mathbb{E}\left[\exp\left(a \frac{\eta_{t,i}}{\sqrt{\bar{V}_{t,i}}} \right) \middle| \mathcal{F}_t^{\mathrm{post}} \right] \le \exp\left(\frac{a^2 R_V^2}{2} \right).$$

Since $\bar{V}_{s,i}$ is $\mathcal{F}_s^{\mathrm{pre}}$-measurable, the standard self-normalized martingale inequality (Abbasi-Yadkori et al., 2011) gives

$$\|\widehat{\boldsymbol{w}}_t^V - \boldsymbol{w}^\star\|_{G_t^V} \le \rho_t^V,$$

where

$$\rho_t^V = \sqrt{\lambda_V} S_w + R_V \sqrt{2 \log \frac{1}{\delta_w} + \log \frac{\det G_t^V}{\det(\lambda_V I)}}.$$

Hence

$$\mathcal{C}_t^{w,V} = \left\{ \boldsymbol{w} \in \mathcal{W} : \|\boldsymbol{w} - \widehat{\boldsymbol{w}}_t^V\|_{G_t^V} \le \rho_t^V \right\}.$$

Since $\bar{V}_{t,i} \ge V_{\min}$,

$$\rho_T^V := \max_{t \le T} \rho_t^V = \widetilde{\mathcal{O}}\left(\sqrt{\lambda_V} S_w + R_V \sqrt{d_u} \right).$$

H.3 Optimistic Variance

Using the current context and reward-parameter confidence sets, define

$$\overline{\mu}_{t,i} = \max_{\substack{\boldsymbol{z} \in \mathcal{C}_{t-1}^{\phi}(i, \boldsymbol{x}_{t,i}) \\ \boldsymbol{w} \in \mathcal{C}_t^{w,V}}} \begin{bmatrix} \boldsymbol{x}_{t,i} \\ \boldsymbol{z} \end{bmatrix}^\top \boldsymbol{w}$$

and similarly

$$\underline{\mu}_{t,i} = \min_{\substack{\boldsymbol{z}\in\mathcal{C}_{t-1}^{\phi}(i,\boldsymbol{x}_{t,i}) \\ \boldsymbol{w}\in\mathcal{C}_t^{w,V}}} \begin{bmatrix} \boldsymbol{x}_{t,i} \\ \boldsymbol{z} \end{bmatrix}^\top \boldsymbol{w}.$$

On the confidence event,

$$\underline{\mu}_{t,i} \le \mu_{t,i} \le \overline{\mu}_{t,i}.$$

Define the clipped interval

$$I_{t,i} = [\underline{\mu}_{t,i}, \overline{\mu}_{t,i}] \cap [\nu, 1-\nu].$$

Define

$$\check{\mu}_{t,i} \in \arg\max_{v\in I_{t,i}} v(1-v).$$

Then

$$V_{\min} \le \bar{V}_{t,i} \le \frac{1}{4}.$$

Moreover, with

$$c_V = \frac{1}{4V_{\min}},$$

we have

$$\frac{1}{\mu_{t,i}(1-\mu_{t,i})} \le \frac{c_V}{\bar{V}_{t,i}}. \tag{28}$$

H.4 VA-OFUC^2-PC

Let $(\widetilde{\boldsymbol{w}}_t, \{\widetilde{\boldsymbol{z}}_{t,i}\})$ denote the optimistic parameters selected by the variance-weighted analogue of OFUC^2-PC, with the additional constraints

$$\nu \le \begin{bmatrix} \boldsymbol{x}_{t,i} \\ \boldsymbol{z}_i \end{bmatrix}^\top \boldsymbol{w} \le 1-\nu, \qquad i \in S.$$

Define

$$\widetilde{\mu}_{t,i} = \begin{bmatrix} \boldsymbol{x}_{t,i} \\ \widetilde{\boldsymbol{z}}_{t,i} \end{bmatrix}^\top \widetilde{\boldsymbol{w}}_t.$$

By optimism,

$$\Delta_t \le F_t(S_t; \widetilde{\boldsymbol{\mu}}_t) - F_t(S_t; \boldsymbol{\mu}_t).$$

For every selected arm, define the clipped intermediate mean

$$\mu_{t,i}^{\mathrm{mid}} = \mathrm{clip}_{[\nu,1-\nu]}\left(\bar{\boldsymbol{u}}_{t,i}^\top \widetilde{\boldsymbol{w}}_t\right).$$

We then decompose

$$\widetilde{\mu}_{t,i} - \mu_{t,i} = \zeta_{t,i} + \kappa_{t,i},$$

where

$$\zeta_{t,i} = \mu_{t,i}^{\mathrm{mid}} - \mu_{t,i}, \qquad \kappa_{t,i} = \widetilde{\mu}_{t,i} - \mu_{t,i}^{\mathrm{mid}}.$$

Since

$$\mu_{t,i}, \mu_{t,i}^{\mathrm{mid}}, \widetilde{\mu}_{t,i} \in [\nu, 1-\nu],$$

we have

$$|\zeta_{t,i}| \le 1-2\nu < 1, \qquad |\kappa_{t,i}| \le 1-2\nu < 1.$$

Moreover, since clipping onto $[\nu, 1-\nu]$ cannot increase the distance to a point in $[\nu, 1-\nu]$,

$$|\zeta_{t,i}| \le \left|\bar{\boldsymbol{u}}_{t,i}^\top(\widetilde{\boldsymbol{w}}_t - \boldsymbol{w}^\star)\right| \le 2\rho_T^V \|\bar{\boldsymbol{u}}_{t,i}\|_{(G_t^V)^{-1}},$$

and

$$|\kappa_{t,i}| \le \left|(\widetilde{\boldsymbol{u}}_{t,i} - \bar{\boldsymbol{u}}_{t,i})^\top \widetilde{\boldsymbol{w}}_t\right| \le 2e_{t-1,i}(\boldsymbol{x}_{t,i}).$$

Applying equation 20 together with equation 28 gives

$$\Delta_t \le 2B_v\sqrt{c_V}\rho_T^V \sqrt{\sum_{i\in S_t} \frac{\|\bar{\boldsymbol{u}}_{t,i}\|^2_{(G_t^V)^{-1}}}{\bar{V}_{t,i}}} + 2B_1 \sum_{i\in S_t} e_{t-1,i}(\boldsymbol{x}_{t,i}).$$

Variance-Scaled NBT-EPL

Define

$$s_{t,i} = \frac{\|\bar{\boldsymbol{u}}_{t,i}\|^2_{(G_t^V)^{-1}}}{\bar{V}_{t,i}}, \qquad \Gamma_T^V = \sum_{t=1}^{T} \sum_{i \in S_t} \min\left\{\frac{1}{K}, s_{t,i}\right\}.$$

Lemma 2 (Variance-scaled NBT-EPL)**.** *If*

$$\lambda_V \geq \max\left\{1, \frac{L_u^2}{V_{\min}}\right\},$$

then, with probability at least $1 - \delta_E$*,*

$$\Gamma_T^V = \widetilde{\mathcal{O}}(d_u).$$

Proof. Set

$$\boldsymbol{a}_{t,i} = \frac{\bar{\boldsymbol{u}}_{t,i}}{\sqrt{\bar{V}_{t,i}}}, \qquad \boldsymbol{y}_{t,i} = \frac{\boldsymbol{u}_{t,i}}{\sqrt{\bar{V}_{t,i}}}.$$

Because $\bar{V}_{t,i}$ is $\mathcal{F}_t^{\mathrm{pre}}$-measurable,

$$\mathbb{E}\left[\boldsymbol{y}_{t,i} - \boldsymbol{a}_{t,i} \mid \mathcal{F}_t^{\mathrm{pre}}\right] = 0,$$

and

$$\|\boldsymbol{y}_{t,i}\|_2 \leq \frac{L_u}{\sqrt{V_{\min}}}.$$

Thus Lemma 1 applies directly to the scaled batches and gives

$$\Gamma_T^V \leq \frac{2}{1 - e^{-1}} \left[\log \frac{\det G_{T+1}^V}{\det(\lambda_V I)} + \log \frac{1}{\delta_E}\right] = \widetilde{\mathcal{O}}(d_u).$$

□

Final Regret Bound For the low-width occurrences, Lemma 2 and Cauchy–Schwarz give

$$R_{\mathrm{low}}^{w,V} = \widetilde{\mathcal{O}}\left(B_v \sqrt{c_V} \rho_T^V \sqrt{T d_u}\right).$$

For the high-width occurrences,

$$N_H^V \leq K \Gamma_T^V = \widetilde{\mathcal{O}}(K d_u),$$

and using

$$|\zeta_{t,i}| \leq 1 - 2\nu, \qquad \mu_{t,i}(1 - \mu_{t,i}) \geq V_{\min},$$

we obtain

$$R_{\mathrm{high}}^{w,V} = \widetilde{\mathcal{O}}\left(\frac{B_v(1 - 2\nu)}{\sqrt{V_{\min}}} K d_u\right).$$

Therefore,

$$R_T^{w,V} = \widetilde{\mathcal{O}}\left(B_v \sqrt{c_V} \rho_T^V \sqrt{T d_u} + \frac{B_v(1 - 2\nu)}{\sqrt{V_{\min}}} K d_u\right).$$

Finally, the context-learning contribution is bounded as in Appendix C:

$$R_T^{\phi} = \widetilde{\mathcal{O}}\left(B_1 m^{\alpha} (KT)^{1-\alpha} d_x^{\alpha}\right).$$

Combining the two contributions gives

$$R_T = \widetilde{\mathcal{O}}\left(B_1 m^{\alpha} (KT)^{1-\alpha} d_x^{\alpha} + B_v \sqrt{c_V} \rho_T^V \sqrt{T d_u} + \frac{B_v(1 - 2\nu)}{\sqrt{V_{\min}}} K d_u\right).$$

H.5 VA-C^{2}UCB-PC

We next extend the separable C^{2}UCB-PC policy of Section 4 to the variance-aware setting. We reuse the variance-weighted reward estimator, confidence set, and variance-scaled potential introduced in Appendix H.2 and H.3.

The main difference from the VA-OFUC^2-PC policy is that VA-C^{2}UCB-PC does not jointly optimize over $\boldsymbol{w} \in \mathcal{C}_t^{w,V}$ and $\boldsymbol{z} \in \mathcal{C}_{t-1}^{\phi}(i, \boldsymbol{x}_{t,i})$. Instead, it constructs an arm optimistic mean by separately adding a reward parameter bonus and a post-serving prediction bonus.

Arm optimistic mean and variance Recall

$$\widehat{\boldsymbol{z}}_{t,i} = \widehat{\phi}_{t-1,i}(\boldsymbol{x}_{t,i}), \qquad \widehat{\boldsymbol{u}}_{t,i} = \begin{bmatrix} \boldsymbol{x}_{t,i} \\ \widehat{\boldsymbol{z}}_{t,i} \end{bmatrix}$$

For each arm i, define

$$\overline{\mu}_{t,i} = \min\left\{1-\nu,\, \widehat{\boldsymbol{u}}_{t,i}^\top \widehat{\boldsymbol{w}}_t^V + \rho_t^V \|\widehat{\boldsymbol{u}}_{t,i}\|_{(G_t^V)^{-1}} + e_{t-1,i}(\boldsymbol{x}_{t,i})\right\}$$

and

$$\underline{\mu}_{t,i} = \max\left\{\nu,\, \widehat{\boldsymbol{u}}_{t,i}^\top \widehat{\boldsymbol{w}}_t^V - \rho_t^V \|\widehat{\boldsymbol{u}}_{t,i}\|_{(G_t^V)^{-1}} - e_{t-1,i}(\boldsymbol{x}_{t,i})\right\}$$

On the joint confidence event,

$$\underline{\mu}_{t,i} \le \mu_{t,i} \le \overline{\mu}_{t,i}$$

Indeed,

$$\begin{aligned} \mu_{t,i} - \widehat{\boldsymbol{u}}_{t,i}^\top \widehat{\boldsymbol{w}}_t^V &= \bar{\boldsymbol{u}}_{t,i}^\top \boldsymbol{w}^\star - \widehat{\boldsymbol{u}}_{t,i}^\top \widehat{\boldsymbol{w}}_t^V \\ &= (\bar{\boldsymbol{u}}_{t,i} - \widehat{\boldsymbol{u}}_{t,i})^\top \boldsymbol{w}^\star + \widehat{\boldsymbol{u}}_{t,i}^\top (\boldsymbol{w}^\star - \widehat{\boldsymbol{w}}_t^V) \end{aligned}$$

Since the pre-serving components of $\bar{\boldsymbol{u}}_{t,i}$ and $\widehat{\boldsymbol{u}}_{t,i}$ coincide,

$$\begin{aligned} \left|(\bar{\boldsymbol{u}}_{t,i} - \widehat{\boldsymbol{u}}_{t,i})^\top \boldsymbol{w}^\star\right| &= \left|\left\langle \phi_i^\star(\boldsymbol{x}_{t,i}) - \widehat{\phi}_{t-1,i}(\boldsymbol{x}_{t,i}), \boldsymbol{\beta}^\star \right\rangle\right| \\ &\le e_{t-1,i}(\boldsymbol{x}_{t,i}) \end{aligned}$$

Moreover, the variance-weighted reward confidence set gives

$$\begin{aligned} \left|\widehat{\boldsymbol{u}}_{t,i}^\top (\boldsymbol{w}^\star - \widehat{\boldsymbol{w}}_t^V)\right| &\le \|\widehat{\boldsymbol{u}}_{t,i}\|_{(G_t^V)^{-1}} \|\boldsymbol{w}^\star - \widehat{\boldsymbol{w}}_t^V\|_{G_t^V} \\ &\le \rho_t^V \|\widehat{\boldsymbol{u}}_{t,i}\|_{(G_t^V)^{-1}} \end{aligned}$$

Combining the two inequalities gives the claimed confidence interval.

Unlike the VA-OFUC^2-PC policy, which obtains the optimistic variance by jointly optimizing over the reward and post-serving confidence sets, VA-C^{2}UCB-PC uses the arm-wise interval $[\underline{\mu}_{t,i}, \overline{\mu}_{t,i}]$ above. Define

$$\check{\mu}_{t,i} \in \arg\max_{v \in [\underline{\mu}_{t,i}, \overline{\mu}_{t,i}]} v(1-v)$$

and

$$\bar{V}_{t,i} = \check{\mu}_{t,i}(1-\check{\mu}_{t,i})$$

Since

$$\mu_{t,i} \in [\underline{\mu}_{t,i}, \overline{\mu}_{t,i}] \subseteq [\nu, 1-\nu]$$

we again have

$$V_{\min} \le \bar{V}_{t,i} \le \frac{1}{4}$$

and

$$\frac{1}{\mu_{t,i}(1-\mu_{t,i})} \le \frac{c_V}{\bar{V}_{t,i}}$$

The VA-C^{2}UCB-PC policy selects

$$S_t \in \arg\max_{S \in \mathcal{S}} F_t(S; \overline{\boldsymbol{\mu}}_t)$$

where

$$\overline{\boldsymbol{\mu}}_t = (\overline{\mu}_{t,1}, \ldots, \overline{\mu}_{t,m})$$

This uses the same combinatorial optimization oracle for F_t as C^{2}UCB-PC in Section 4.

Regret decomposition. By monotonicity of F_t and optimism,

$$F_t(S_t^\star; \boldsymbol{\mu}_t) \le F_t(S_t^\star; \overline{\boldsymbol{\mu}}_t) \le F_t(S_t; \overline{\boldsymbol{\mu}}_t)$$

and hence

$$\Delta_t \le F_t(S_t; \overline{\boldsymbol{\mu}}_t) - F_t(S_t; \boldsymbol{\mu}_t)$$

The key difference from the VA-OFUC^2-PC policy appears when bounding the arm optimistic gap. Since clipping at $1-\nu$ can only decrease this gap,

$$\begin{aligned}
\overline{\mu}_{t,i} - \mu_{t,i} &\le \widehat{\boldsymbol{u}}_{t,i}^\top \widehat{\boldsymbol{w}}_t^V - \bar{\boldsymbol{u}}_{t,i}^\top \boldsymbol{w}^\star + \rho_t^V \|\widehat{\boldsymbol{u}}_{t,i}\|_{(G_t^V)^{-1}} + e_{t-1,i}(\boldsymbol{x}_{t,i}) \\
&= \widehat{\boldsymbol{u}}_{t,i}^\top (\widehat{\boldsymbol{w}}_t^V - \boldsymbol{w}^\star) + (\widehat{\boldsymbol{u}}_{t,i} - \bar{\boldsymbol{u}}_{t,i})^\top \boldsymbol{w}^\star \\
&\quad + \rho_t^V \|\widehat{\boldsymbol{u}}_{t,i}\|_{(G_t^V)^{-1}} + e_{t-1,i}(\boldsymbol{x}_{t,i}) \\
&\le 2\rho_t^V \|\widehat{\boldsymbol{u}}_{t,i}\|_{(G_t^V)^{-1}} + 2e_{t-1,i}(\boldsymbol{x}_{t,i})
\end{aligned}$$

The variance-scaled NBT-EPL, however, controls the clean context $\bar{\boldsymbol{u}}_{t,i}$ rather than $\widehat{\boldsymbol{u}}_{t,i}$. Since

$$G_t^V \succeq \lambda_V I$$

we have

$$(G_t^V)^{-1} \preceq \frac{1}{\lambda_V} I$$

and therefore

$$\begin{aligned}
\|\widehat{\boldsymbol{u}}_{t,i} - \bar{\boldsymbol{u}}_{t,i}\|_{(G_t^V)^{-1}} &\le \frac{\|\widehat{\boldsymbol{u}}_{t,i} - \bar{\boldsymbol{u}}_{t,i}\|_2}{\sqrt{\lambda_V}} \\
&\le \frac{e_{t-1,i}(\boldsymbol{x}_{t,i})}{\sqrt{\lambda_V}}
\end{aligned}$$

Consequently,

$$\begin{aligned}
\|\widehat{\boldsymbol{u}}_{t,i}\|_{(G_t^V)^{-1}} &\le \|\bar{\boldsymbol{u}}_{t,i}\|_{(G_t^V)^{-1}} + \|\widehat{\boldsymbol{u}}_{t,i} - \bar{\boldsymbol{u}}_{t,i}\|_{(G_t^V)^{-1}} \\
&\le \|\bar{\boldsymbol{u}}_{t,i}\|_{(G_t^V)^{-1}} + \frac{e_{t-1,i}(\boldsymbol{x}_{t,i})}{\sqrt{\lambda_V}}
\end{aligned}$$

Using $\rho_t^V \le \rho_T^V$,

$$\begin{aligned}
\overline{\mu}_{t,i} - \mu_{t,i} &\le 2\rho_T^V \|\bar{\boldsymbol{u}}_{t,i}\|_{(G_t^V)^{-1}} \\
&\quad + 2\left(1 + \frac{\rho_T^V}{\sqrt{\lambda_V}}\right) e_{t-1,i}(\boldsymbol{x}_{t,i})
\end{aligned}$$

This is the main statistical difference from the VA-OFUC^2-PC policy. For the VA-OFUC^2-PC policy, the corresponding post-serving term is $2e_{t-1,i}(\boldsymbol{x}_{t,i})$. In VA-C^{2}UCB-PC, evaluating the reward bonus at $\widehat{\boldsymbol{u}}_{t,i}$ introduces the additional $\frac{\rho_T^V}{\sqrt{\lambda_V}}$ factor.

To apply the variance-modulated bounded-smoothness condition, write

$$\overline{\mu}_{t,i} - \mu_{t,i} = \zeta_{t,i} + \kappa_{t,i}$$

where

$$\zeta_{t,i} = \min\left\{\overline{\mu}_{t,i} - \mu_{t,i},\ 2\rho_T^V \|\bar{\boldsymbol{u}}_{t,i}\|_{(G_t^V)^{-1}}\right\}$$

and

$$\kappa_{t,i} = \overline{\mu}_{t,i} - \mu_{t,i} - \zeta_{t,i}$$

Then

$$0 \le \zeta_{t,i} \le 2\rho_T^V \|\bar{\boldsymbol{u}}_{t,i}\|_{(G_t^V)^{-1}}$$

and

$$0 \le \kappa_{t,i} \le 2\left(1 + \frac{\rho_T^V}{\sqrt{\lambda_V}}\right) e_{t-1,i}(\boldsymbol{x}_{t,i})$$

Applying the variance-modulated bounded-smoothness condition gives

$$\Delta_t \le B_v \sqrt{\sum_{i \in S_t} \frac{\zeta_{t,i}^2}{\mu_{t,i}(1-\mu_{t,i})}} + B_1 \sum_{i \in S_t} |\kappa_{t,i}|$$

We bound these two contributions separately.

Reward-learning contribution. Recall from Appendix H.4 that

$$s_{t,i} = \frac{\|\bar{\boldsymbol{u}}_{t,i}\|^2_{(G_t^V)^{-1}}}{\bar{V}_{t,i}}$$

and

$$\Gamma_T^V = \sum_{t=1}^{T} \sum_{i \in S_t} \min\left\{\frac{1}{K}, s_{t,i}\right\}$$

By Lemma 2,

$$\Gamma_T^V = \widetilde{O}(d_u)$$

We separate the reward-learning term into low- and high-width occurrences according to whether $s_{t,i} \le 1/K$ or $s_{t,i} > 1/K$. Using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$,

$$R_T^{w,V} \le R_{\text{low}}^{w,V} + R_{\text{high}}^{w,V}$$

For the low-width occurrences,

$$\begin{aligned}
\frac{\zeta_{t,i}^2}{\mu_{t,i}(1-\mu_{t,i})} &\le 4(\rho_T^V)^2 \frac{\|\bar{\boldsymbol{u}}_{t,i}\|^2_{(G_t^V)^{-1}}}{\mu_{t,i}(1-\mu_{t,i})} \\
&\le 4c_V(\rho_T^V)^2 \frac{\|\bar{\boldsymbol{u}}_{t,i}\|^2_{(G_t^V)^{-1}}}{\bar{V}_{t,i}} \\
&= 4c_V(\rho_T^V)^2 s_{t,i}
\end{aligned}$$

Hence

$$\begin{aligned}
R_{\text{low}}^{w,V} &\le 2B_v\sqrt{c_V}\rho_T^V \sum_{t=1}^{T} \sqrt{\sum_{\substack{i \in S_t: \\ s_{t,i} \le 1/K}} s_{t,i}} \\
&\le 2B_v\sqrt{c_V}\rho_T^V \sqrt{T \sum_{t=1}^{T} \sum_{\substack{i \in S_t: \\ s_{t,i} \le 1/K}} s_{t,i}} \\
&\le 2B_v\sqrt{c_V}\rho_T^V \sqrt{T\Gamma_T^V}
\end{aligned}$$

Therefore,

$$R_{\text{low}}^{w,V} = \widetilde{O}\left(B_v\sqrt{c_V}\rho_T^V\sqrt{Td_u}\right)$$

This low term is the same as for the VA-OFUC^2-PC policy: the variance-modulated smoothness condition aggregates the selected arms within each round before Cauchy–Schwarz is applied across the T rounds, producing $\sqrt{T}$ rather than $\sqrt{KT}$.

We next consider the high-width occurrences. Let N_H^V denote the total number of selected-arm occurrences satisfying

$$s_{t,i} > \frac{1}{K}$$

Each such occurrence contributes $1/K$ to Γ_T^V, so

$$\frac{N_H^V}{K} \le \Gamma_T^V$$

and therefore

$$N_H^V \le K\Gamma_T^V = \widetilde{O}(Kd_u)$$

Since

$$\mu_{t,i}, \overline{\mu}_{t,i} \in [\nu, 1-\nu]$$

and

$$0 \le \zeta_{t,i} \le \overline{\mu}_{t,i} - \mu_{t,i}$$

we have

$$|\zeta_{t,i}| \le 1 - 2\nu$$

Moreover,

$$\mu_{t,i}(1-\mu_{t,i}) \ge V_{\min}$$

Let $N_{t,H}^V$ denote the number of high-width selected arms in round t. Then

$$\begin{aligned} B_v \sqrt{\sum_{\substack{i \in S_t: \\ s_{t,i} > 1/K}} \frac{\zeta_{t,i}^2}{\mu_{t,i}(1-\mu_{t,i})}} &\le \frac{B_v(1-2\nu)}{\sqrt{V_{\min}}} \sqrt{N_{t,H}^V} \\ &\le \frac{B_v(1-2\nu)}{\sqrt{V_{\min}}} N_{t,H}^V \end{aligned}$$

Summing over t gives

$$\begin{aligned} R_{\text{high}}^{w,V} &\le \frac{B_v(1-2\nu)}{\sqrt{V_{\min}}} N_H^V \\ &= \widetilde{O}\left(\frac{B_v(1-2\nu)}{\sqrt{V_{\min}}} Kd_u \right) \end{aligned}$$

Combining the low- and high-width terms,

$$R_T^{w,V} = \widetilde{O}\left(B_v \sqrt{c_V} \rho_T^V \sqrt{Td_u} + \frac{B_v(1-2\nu)}{\sqrt{V_{\min}}} Kd_u \right)$$

Post-serving context contribution. The second part of the bounded-smoothness decomposition satisfies

$$R_T^{\phi} \le B_1 \sum_{t=1}^{T} \sum_{i \in S_t} |\kappa_{t,i}|$$

Using the bound on $\kappa_{t,i}$,

$$R_T^{\phi} \le 2B_1 \left(1 + \frac{\rho_T^V}{\sqrt{\lambda_V}} \right) \sum_{t=1}^{T} \sum_{i \in S_t} e_{t-1,i}(\boldsymbol{x}_{t,i})$$

The same arm-wise generalized elliptical-potential argument as in Appendix C gives

$$\sum_{t=1}^{T}\sum_{i\in S_t} e_{t-1,i}(\boldsymbol{x}_{t,i}) = \widetilde{O}\left(m^{\alpha}(KT)^{1-\alpha}d_x^{\alpha}\right)$$

Therefore,

$$R_T^{\phi} = \widetilde{O}\left(B_1\left(1+\frac{\rho_T^V}{\sqrt{\lambda_V}}\right)m^{\alpha}(KT)^{1-\alpha}d_x^{\alpha}\right)$$

Compared with the VA-OFUC^2-PC policy, the dependence on m, K, T, and d_x is unchanged, but the separable policy pays the additional multiplicative factor $\frac{\rho_T^V}{\sqrt{\lambda_V}}$ in the post-serving context term.

Final regret bound. Combining the reward-learning and post-serving context contributions gives

$$R_T^{\text{VA-C}^2\text{UCB-PC}} = \widetilde{O}\left(B_1\left(1+\frac{\rho_T^V}{\sqrt{\lambda_V}}\right)m^{\alpha}(KT)^{1-\alpha}d_x^{\alpha} + B_v\sqrt{c_V}\rho_T^V\sqrt{Td_u} + \frac{B_v(1-2\nu)}{\sqrt{V_{\min}}}Kd_u\right)$$

In particular, for $\alpha = 1/2$,

$$R_T^{\text{VA-C}^2\text{UCB-PC}} = \widetilde{O}\left(B_1\left(1+\frac{\rho_T^V}{\sqrt{\lambda_V}}\right)\sqrt{mKTd_x} + B_v\sqrt{c_V}\rho_T^V\sqrt{Td_u} + \frac{B_v(1-2\nu)}{\sqrt{V_{\min}}}Kd_u\right)$$

I Additional Synthetic Experiment Details

Experimental settings. We use $m = 10$, $T = 20000$, and $K \in \{3, 5, 7\}$. Contexts are $\boldsymbol{x}_{t,i} = (1, a_{t,i}, b_{t,i})^{\top}$, with $a_{t,i}, b_{t,i} \sim \text{Unif}[-0.25, 0.25]$ independently. Arm-specific mapping coefficients are sampled once per seed: $\gamma_{i,1}^{\star} \sim \text{Unif}[-0.05, 0.05]$ and $\gamma_{i,2}^{\star}, \gamma_{i,3}^{\star} \sim \text{Unif}[-0.04, 0.04]$. Post-serving contexts follow

$$\boldsymbol{z}_{t,i} = (\boldsymbol{\gamma}_i^{\star})^{\top}h(\boldsymbol{x}_{t,i}) + \boldsymbol{\epsilon}_{t,i}, \qquad h(\boldsymbol{x}) \in \{\boldsymbol{x}, \sin(\boldsymbol{x})\},$$

where $\boldsymbol{\epsilon}_{t,i}$ equals ± 0.03 with equal probability. Rewards are Bernoulli with mean $\boldsymbol{x}_{t,i}^{\top}\boldsymbol{\theta}^{\star} + \boldsymbol{\beta}^{\star}\boldsymbol{z}_{t,i}$, where $\boldsymbol{\theta}^{\star} = (0.12, 0.08, -0.04)^{\top}$ and $\boldsymbol{\beta}^{\star} = 0.6$. Mapping estimators use the corresponding representation $h(\boldsymbol{x})$. We reuse the underlying random draws across mappings and report means with one standard error over ten seeds. Pseudo-regret uses decision-time means and the exact top-K oracle for $F(S;\boldsymbol{\mu}) = 1 - \prod_{i\in S}(1-\mu_i)$.

Implementation. Mapping ridge regularization is 1; reward ridge regularization is 1.5 for ordinary PC methods and 6 for VA methods. We fix $c_w = 0.15$, $\delta_{\phi} = \delta_w = 0.05/3$, and $\nu = 0.02$ across settings. The joint methods use approximate multi-start SLSQP with up to six starting subsets per sign of $\boldsymbol{\beta}$. The separable methods select the top K optimistic arm scores.

Plug-in baseline. LinUCB ($\widehat{\phi}$), adapted from Wang et al. (2023), uses predicted features $[\boldsymbol{x}; \widehat{\phi}_i(\boldsymbol{x})]$ for both action selection and reward regression. After observing selected-arm feedback, it first updates the mapping estimator and then uses the updated predictions in the reward-regression update. The four PC methods instead update reward regression using realized features $[\boldsymbol{x}; \boldsymbol{z}]$. The plug-in baseline also uses horizon-dependent regularization $\lambda_{\text{plug}} = L_u\sqrt{d_u T}\log(T/\delta_w)$, where $d_u = 4$ and $L_u = \sqrt{1 + 0.63^2}$, and includes accumulated mapping uncertainty in its confidence bonus. Its mapping estimator uses the matching representation $h(\boldsymbol{x})$ in all settings.

Additional regret results. Figures 2 and 3 show results for $K = 5$ and $K = 7$. Both variance-aware algorithms achieve lower final mean regret than their ordinary counterparts in the linear and sine settings. VA-OFUC^2-PC achieves the lowest final mean regret in both settings.

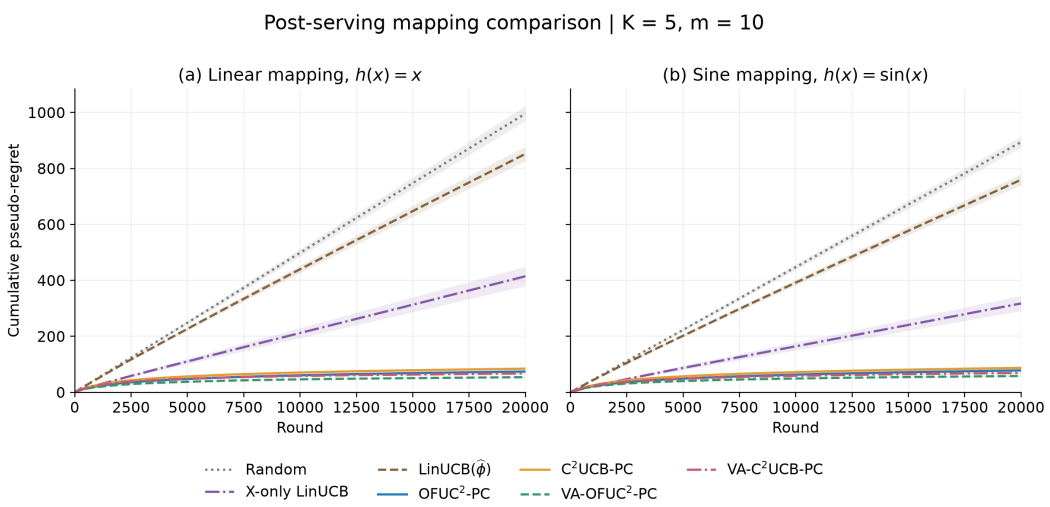

Figure 2: Cumulative pseudo-regret for $K = 5$. Panels correspond to linear and sine mappings; shading indicates one standard error over ten seeds.

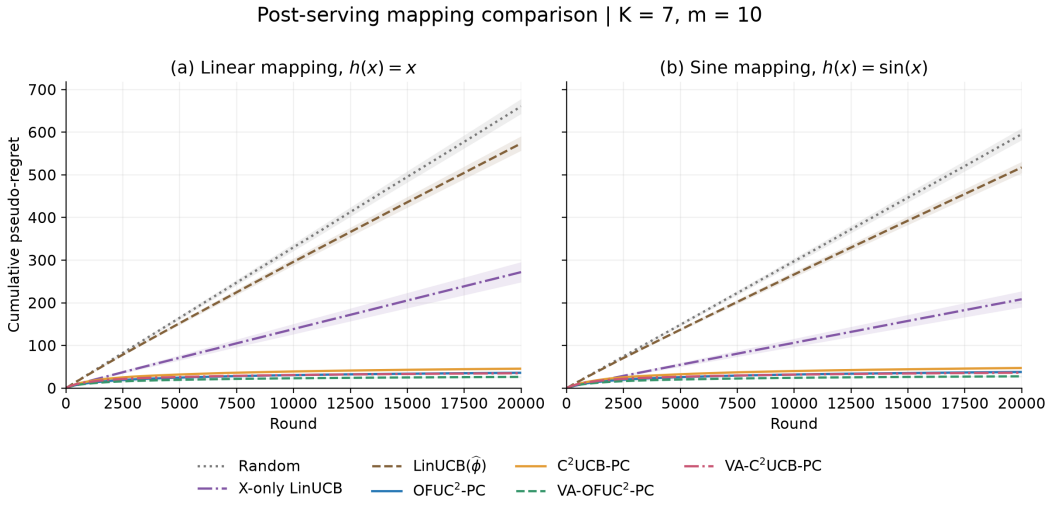

Figure 3: Cumulative pseudo-regret for $K = 7$, using the same settings and uncertainty convention as Figure 2.

Runtime protocol. For the runtime comparison, we report the linear and sine settings with $K = 3$, $T = 20000$, and $m \in \{10, 12, 14\}$. Runs are sequential with one BLAS thread, following an untimed 50-round warmup; each measured run starts with a fresh policy. Timing includes initialization, action selection, feedback, updates, validation, and regret evaluation. Method and mapping orders rotate across runs. All remaining environment and algorithm parameters match the regret experiments. Figure 4 reports the mean runtime and one standard error over three seeds.

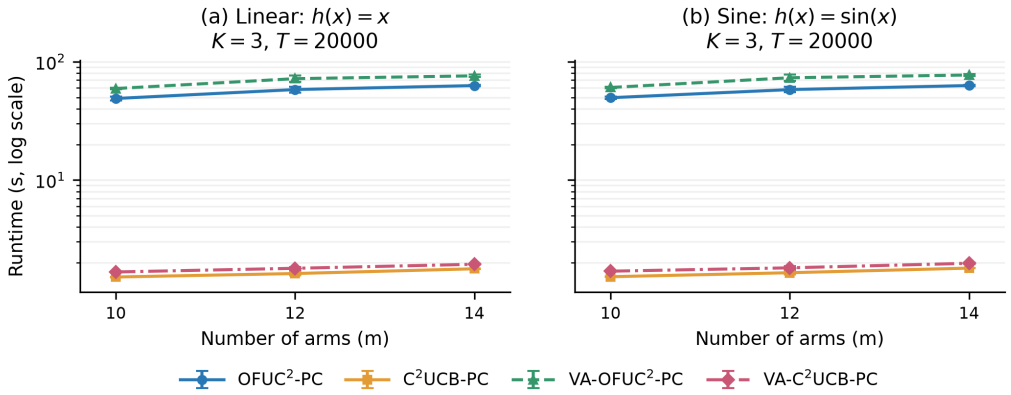

Figure 4: Synthetic runtime under linear and sine post-serving mappings, with $K = 3$, $T = 20000$, and $m \in \{10, 12, 14\}$. Points and error bars show the mean and one standard error over three seeds. The runtime axis is logarithmic.

J Additional MovieLens Semi-Synthetic Experiment Details

We use MovieLens-1M, containing 1,000,209 ratings from 6040 users. Users are split into 3040 representation-training, 2000 construction, and 1000 evaluation users. A frozen neural rating predictor and history encoder provide 16-dimensional user representations, and we use the corresponding 16-dimensional movie directions for $m = 10$ fixed candidate movies.

Let $\boldsymbol{e}_u, \boldsymbol{v}_i \in \mathbb{R}^{16}$ denote the normalized user and movie representations. The pre-serving context is

$$\boldsymbol{x}_{u,i} = [1; \boldsymbol{e}_u; \boldsymbol{v}_i]/\sqrt{3},$$

so $d_x = 33$, $d_z = 1$, and $d_u = 34$.

The environment is calibrated using only the construction users. Let $y^{\mathrm{NN}}_{u,i}$ be the prediction of the frozen neural model, and let $\boldsymbol{\theta}_{\mathrm{fit}}$ be the shared ridge fit of $y^{\mathrm{NN}}_{u,i}$ on $\boldsymbol{x}_{u,i}$, with regularization 1 and an unregularized intercept. Define

$$A_{u,i} = \boldsymbol{x}_{u,i}^\top \boldsymbol{\theta}_{\mathrm{fit}}, \qquad G_{u,i} = \boldsymbol{e}_u^\top \boldsymbol{v}_i - m_i,$$

where m_i is the construction-set mean of $\boldsymbol{e}_u^\top \boldsymbol{v}_i$ for movie i. After standardizing $A_{u,i}$ and $G_{u,i}$ on the construction set, the decision-time mean is

$$\mu_{u,i} = 0.15 + \kappa(0.025)\frac{A_{u,i} - \bar{A}}{\sigma_A} + \kappa(0.025)\frac{G_{u,i}}{\sigma_G}, \qquad \kappa = 0.884602.$$

For the post-serving context, we use

$$z_{t,i} = 0.1 + 0.115390\, G_{u_t,i} + \epsilon_{t,i}, \qquad \epsilon_{t,i} \sim \mathrm{Unif}[-0.0191522, 0.0191522].$$

For each movie, this is a linear post-serving mapping of $\boldsymbol{x}_{u_t,i}$. Rewards are generated as

$$r_{t,i} \sim \mathrm{Bernoulli}\left(\boldsymbol{x}_{u_t,i}^\top \boldsymbol{\theta}^\star + 0.5 z_{t,i}\right),$$

where $\boldsymbol{\theta}^\star$ is shared across movies and is chosen to give the mean $\mu_{u,i}$ above. All calibration quantities are fixed before the evaluation runs.

We compare Random, X-only LinUCB, LinUCB($\widehat{\phi}$), OFUC^2-PC, C^{2}UCB-PC, and their variance-aware versions. The PC methods use separate linear post-serving estimators for each movie and a shared reward estimator. We use $\lambda_\phi = 1$, $\lambda_{\mathrm{base}} = 1.5$, $\lambda_{\mathrm{VA}} = 6$, $c_w = 0.15$, and $\nu = 0.02$. The joint methods use multistart SLSQP.

For each of ten seeds, we run $T = 20000$ rounds with a common sampled user sequence and common potential outcomes across policies. At each round, $K = 3$ movies are selected, and pseudo-regret uses

$$F(S; \boldsymbol{\mu}_t) = 1 - \prod_{i \in S} (1 - \mu_{t,i}),$$

with the oracle selecting the three largest decision-time means.

Runtime protocol. We benchmark the four PC methods with $m = 10$, $K = 3$, and $T = 20000$ over seeds $0, \ldots, 9$. Methods run sequentially with one BLAS thread after 50 untimed warmup rounds, with a fresh policy for each measured trajectory and the method order rotated across seeds. Figure 5 reports the mean runtime and one standard error.

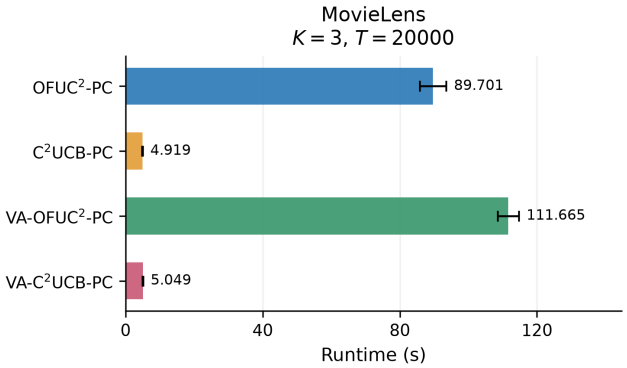

Figure 5: MovieLens runtime for the four PC methods, with $m = 10$, $K = 3$, and $T = 20000$. Bars and error bars show the mean and one standard error over ten seeds.