Best-of-Both Worlds for linear contextual bandits with paid observations

Anonymous Authors[1]

Abstract

We study the problem of linear contextual bandits with paid observations, where at each round the learner selects an action in order to minimize its loss in a given context, and can then decide to pay a fixed cost to observe the loss of any arm. Building on the Follow-the-Regularized-Leader framework with efficient estimators via Matrix Geometric Resampling, we introduce a computationally efficient Best-of-Both-Worlds (BOBW) algorithm for this problem. We show that it achieves the minimax-optimal regret of $\Theta(T^{2/3})$ in adversarial settings, while guaranteeing poly-logarithmic regret in (corrupted) stochastic regimes. Our approach builds on the framework from (Tsuchiya & Ito, 2024) to design BOBW algorithms for "hard problem", using analysis techniques tailored for the setting that we consider.

1. Introduction

Multi-armed bandits (MAB) have emerged as one of the most popular models for sequential decision-making under uncertainty (Lattimore & Szepesvári, 2020; Bubeck & Cesa-Bianchi, 2012). In this framework, a learning agent repeatedly chooses among a finite set of actions (called "arms") and observes a noisy reward for the chosen arm, with the goal of maximizing cumulative reward over time. The appeal of the bandit model lies in its ability to capture the fundamental exploration-exploitation trade-off, that can be encountered in many sequential decision-making scenarios. Nevertheless, the classical bandit framework does not adequately capture two aspects that arise naturally in modern interactive learning systems: the dependence of rewards on user-specific contexts, and the potential cost of acquiring feedback.

An illustrative example is online content recommendation. Indeed, the quality of a recommendation depends crucially on the user who receives it: a video, news article, or product may be highly relevant to one user but uninteresting to another. This motivates the use of *contextual* bandit models (Abe & Long, 1999; Beygelzimer et al., 2011), where the expected reward depends on a context vector that describes the user or environment. A widely studied and practically successful instance is the linear contextual bandit model (Langford & Zhang, 2007; Li et al., 2010). In this setting, the reward is modeled as the dot product between the observed context vector and an unknown arm-specific parameter. Linear contextual bandits offer a useful balance: they are expressive enough to capture heterogeneity in user preferences, while permitting efficient learning through regularized least-squares estimation.

A second challenge is that, in practice, feedback may not be observed automatically. While in standard bandits the learner always receives the reward of the chosen arm, in recommendation systems feedback often comes only if the user provides it (*e.g.*, through ratings or explicit reviews). Actively requesting feedback at every round is undesirable, as it may burden or annoy users. A natural abstraction is therefore to associate a cost with each observation, so that the learner must strategically decide when feedback is worth acquiring. This leads to the framework of bandits with paid observations, first formalized by Seldin et al. (2014).

A third, orthogonal challenge is the nature of the reward-generating process. In some cases, user behavior is well modeled by a stochastic distribution, while in others it may be adversarial. Designing Best-of-Both-Worlds (BoBW) algorithms, that are versatile enough to perform optimally under both regimes, has become a central theme in bandit research (Bubeck & Slivkins, 2012; Zimmert & Seldin, 2022; Dann et al., 2023; Tsuchiya & Ito, 2024).

Motivated by these observations, in this work we introduce the setting of linear contextual bandits with paid observations, which simultaneously incorporates the challenges of contextual modeling, costly feedback acquisition, and uncertainty about the reward generation process. We design a new algorithm within the Follow-the-Regularized-Leader (FTRL) framework, extending ideas from recent advances in best-of-both-worlds algorithms for bandits (Kuroki et al., 2024; Tsuchiya & Ito, 2024). Our algorithm achieves regret guarantees in both stochastic and adversarial regimes,

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

thereby solving the main challenges of the setting that we consider.

Achieving Best-of-Both-Worlds (BoBW) performance in hard problems, *i.e.* problems that incur a minimax regret of $\Theta\big(T^{2/3}\big)$ in the adversarial regime, is a significant challenge, as highlighted in (Tsuchiya & Ito, 2024). The standard approaches used in other settings often fail without substantial modifications. Fortunately, (Tsuchiya & Ito, 2024) introduced a dedicated framework designed to facilitate the design and analysis of BoBW algorithms for such problems. #DB: the following will likely change after fixing the proof. # While this framework forms the basis of our analysis, several challenges arise in adapting it to our setting. First, the general formulation assumes the existence of a single optimal arm throughout the learning process, which does not hold in the contextual linear setting where the optimal action varies with the context. Second, our setting introduces a new key parameter, the smallest non-negative eigenvalue of the context distribution ($\lambda_{\min}$), introduced in Section 2, which necessitates specific tuning of several algorithmic parameters. Third, we identify and resolve an inconsistency in prior applications of the BoBW framework to bandits with paid observations, thereby obtaining tighter regret guarantees; we elaborate on this point in Section 4. Structural differences in our setting require various other adjustments to the technical proofs.

1.1. Detailed literature review

In this section we detail existing results related to the different components of the settings that we consider.

Linear Contextual Bandits Contextual bandits extend classical multi-armed bandits by allowing the reward distribution to depend on an observed context, which can vary across rounds. To enable efficient decision-making, one must adopt a suitable model to capture how the context influences the rewards. In this work we consider the *linear contextual bandit* model (Langford & Zhang, 2007; Li et al., 2010), that we formally describe in Section 2. This model is closely-related to the well-studied *stochastic linear bandit* framework, since in both settings the average reward of each arm is given by the inner product of an arm feature vector and a parameter vector. The two formulations differ in the source of uncertainty: in stochastic linear bandits the arm features are known and the underlying parameter is unknown, whereas in (stochastic) linear contextual bandits the arm-specific features are fixed but unknown, while the context vector is revealed at the beginning of each round.

Most approaches used in linear contextual bandits are borrowed from the stochastic linear bandit literature, in which algorithms follow general principles such as *Optimism in Face of Uncertainty* (Abe & Long, 1999; Dani et al., 2008; Abbasi-Yadkori et al., 2011; Flynn et al., 2023), *Thompson Sampling* (Agrawal & Goyal, 2013; Abeille & Lazaric, 2017; Abeille et al., 2025), *Information Directed Sampling* (Kirschner et al., 2020), or (asymptotic) lower bound matching (Lattimore & Szepesvári, 2017; Degenne et al., 2020). Nonetheless, linear contextual bandits exhibit specific properties compared to standard linear bandits. In particular, Bastani et al. (2021) showed that under suitable assumptions on *context diversity*, even a simple greedy strategy can achieve logarithmic regret.

While the above works assume stochastic rewards, this assumption can be restrictive in practice. To address this, Neu & Olkhovskaya (2020) introduced an adversarial formulation of linear contextual bandits, in which arm parameters are fixed by an oblivious adversary. They derived a $\widetilde{\mathcal{O}}(\sqrt{KdT})$ regret bound for an exponential-weights algorithm (Auer et al., 2002), where d is the parameter dimension, K is the number of arms, and T is the horizon. Building on this, Olkhovskaya et al. (2023) obtained refined first and second-order bounds. In parallel, Kuroki et al. (2024) established the first *Best-of-Both-Worlds* guarantees in this setting, showing that one can achieve simultaneously polylogarithmic regret in the stochastic regime and $\widetilde{\mathcal{O}}(Kd\sqrt{T})$ regret in the adversarial case.

Bandits with Paid Observations. This framework was introduced by Seldin & Slivkins (2014) to capture a feedback structure lying between the standard multi-armed bandit and full-information settings. In this model, the learner may choose to observe the reward of *any* arm at a fixed cost. They established that the minimax regret in this setting is $\Theta((cK)^{1/3}T^{2/3} + \sqrt{T})$, and proposed an algorithm matching this lower bound.

Prior to this, several related models were proposed to account for the possibility of observing additional feedback beyond the chosen arm (Mannor & Shamir, 2011; Avner et al., 2012; Alon et al., 2013), though these formulations do not explicitly capture the cost of information acquisition. An alternative approach is to impose a *budget* on the total observation cost, as in (Yun et al., 2018; Efroni et al., 2021). However, this formulation requires the decision-maker to know both the acquisition cost of each arm and an overall budget, thereby placing regret minimization and acquisition costs on different scales. By contrast, the bandits-with-paid-observations framework integrates both aspects under a unified metric by directly subtracting observation costs from the rewards.

Best-of-Both-Worlds (BoBW). The design of algorithms that perform well simultaneously in stochastic and adversarial regimes has become a central theme in the bandit literature. The foundational work of Bubeck & Slivkins (2012); Seldin & Slivkins (2014) initiated this line of research by asking whether one can achieve logarithmic regret in the stochastic setting while retaining $\tilde{O}(\sqrt{T})$ regret in the adversarial case. Their results provided only partial success, either with suboptimal bounds or with algorithms of limited practicality. Later, Zimmert & Seldin (2022) first obtained the optimal best-of-both-worlds guarantees in the K-armed bandit setting. This breakthrough has since inspired the development of BoBW algorithms across a variety of bandit problems (Amir et al., 2022; Rouyer et al., 2022; Saha & Gaillard, 2022; Tsuchiya et al., 2023; Jin et al., 2023; Zimmert & Marinov, 2024; Kato & Ito, 2025).

Of particular relevance to our work, Kuroki et al. (2024) studied linear contextual bandits through the black-box reduction framework of Dann et al. (2023), which can be used to design BoBW algorithms for problems whose minimax regret scales as $\sqrt{T}$. More recently, Tsuchiya & Ito (2024) proposed a general recipe for constructing BoBW algorithms in so-called "hard" online learning problems, namely those with minimax regret of order $\Theta(T^{2/3})$. They further show that several known bandit models, including multi-armed bandits with paid observations, fall within this framework. Our work is inspired by their approach, however, a direct application of their method does not yield optimal bounds in our setting (see Section 4). This motivates the need for a careful adaptation of their ideas, which we develop in the remainder of the paper.

2. Problem Definition

In this section we formalize the setting of *linear bandits with paid observations*, and state the main assumptions used in the analysis presented in Section 4.

Interaction protocol. The interaction between the learning agent and the environment has a total duration of $T \in \mathbb{N}$ time steps, where T is unknown to the learner. Context vectors are drawn independently from a fixed distribution $\mathcal{D}$ supported on a compact, full-dimensional subset $\mathcal{X} \subseteq \mathbb{R}^d$. At each round t, the following steps occur:

1. For each action $a \in [K] := \{1, \ldots, K\}$, the environment selects a loss parameter $\theta_{t,a} \in \mathbb{R}^d$.
2. A context $X_t \in \mathcal{X}$ is drawn from $\mathcal{D}$.
3. The learner observes X_t, chooses an action $A_t \in [K]$, and an observation set $O_t \subseteq [K]$.
4. The learner incurs loss $\ell_t(X_t, A_t) + c|O_t|$, where ℓ_t is a loss function that depends on the environment parameters $(\theta_{t,a})_{a\in[K]}$, $c > 0$ is the known unit cost of observation, and $|O_t|$ is the cardinality of the observation set. It then observes the losses $\{\ell_t(X_t, o) : o \in O_t\}$.

Following Seldin & Slivkins (2014), the learner may query multiple arms in each round, paying cost c per queried arm. When $c = 0$, the learner is incentivized to query all arms, recovering the *full-information* (or "experts") setting.

Assumptions. To enable algorithm design and analysis, we adopt standard assumptions from the linear contextual bandit literature (Kuroki et al., 2024):

1. For $X \sim \mathcal{D}$, $\|X\|_2 \le X_{\max}$ almost surely.
2. For any $t \in [T]$, $a \in [K]$, $\|\theta_{t,a}\|_2 \le \Theta_{\max}$.
3. For any $t \in [T]$, $x \in \mathcal{X}$, $a \in [K]$, $\ell_t(x, a) \in [-1, 1]$. [Antoine: inconsistent with the noise model]

We denote by $\Sigma = \mathbb{E}_{X\sim\mathcal{D}}[XX^\intercal] \succ 0$ the covariance matrix of the context distribution, and by $\lambda_{\min} > 0$ its minimum non zero eigenvalue, assumed to be known to the learner. While the learner does not know $\mathcal{D}$ in full, we assume access to independent samples from $\mathcal{D}$ between rounds, for instance through a simulator.

We now define how the loss $\ell_t(x, a)$ is constructed in each of the regimes considered in this work, for a given step $t \in [T]$, context $x \in \mathcal{X}$ and arm $a \in [K]$.

Adversarial regime. The loss satisfies $\ell_t(x, a) := \langle x, \theta_{t,a}\rangle$, where $\theta_{t,a}$ is chosen by an *oblivious* adversary: the entire sequence $(\theta_{t,a})_{t\in[T],a\in[K]}$ can be arbitrary, but is fixed before the interaction starts.

Stochastic regime. The loss is defined by $\ell_t(x, a) := \langle x, \theta_a\rangle + \varepsilon_{t,a}$ where θ_a is a fixed, unknown parameter for each arm a, and $\varepsilon_{t,a}$ is a zero-mean random noise bounded, independent across rounds and arms.

Corrupted stochastic regime. The loss satisfies $\ell_t(x, a) := \langle x, \theta_{t,a}\rangle + \varepsilon_{t,a}$, , where $\varepsilon_{t,a}$ is again a zero-mean random noise bounded in $[-1, 1]$. In this regime, the adversary may corrupt the parameters over time, but only within a limited budget: there exists fixed but unknown vectors $(\theta_a)_{a\in[K]}$ and a constant $C > 0$ such that $\sum_{t=1}^T \max_{a\in[K]} \|\theta_{t,a} - \theta_a\|_2 \le C$. The extreme cases $C = 0$ and $C = T$ recover, respectively, the stochastic regime and the adversarial regime (up to the presence of random noise).

Let Π denote the set of deterministic policies $\pi : \mathcal{X} \mapsto [K]$. We define the best policy in hindsight $\pi_T^\star$ by

$$\pi_T^\star : x \in \mathcal{X} \mapsto \operatorname*{arg\,min}_{a\in[K]} \mathbb{E}\left[\sum_{t=1}^T \ell_t(x, a)\right],$$

where potential randomness of the loss distribution. The learners' ojective is to minimize the expected cumulative regret against $\pi_T^\star$,

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T}(\ell_t(X_t, A_t) - \ell_t(X_t, \pi_T^\star(X_t)))\right] \quad (1)$$
$$+ \mathbb{E}\left[\sum_{t=1}^{T} c \cdot |O_t|\right],$$

where the expectation here additionally includes the learner's internal randomization.

#DB: If we re-write the proof with ghost sample, explain here#

Additional definitions. In the (corrupted) stochastic regime, we further define, for any context $x \in \mathcal{X}$,

$$\Delta_{\text{min}}(x) := \min_{a \neq \pi_T^\star(x)} \left\langle x, \theta_a - \theta_{\pi_T^\star(x)} \right\rangle$$

and the minimum sub-optimality gap

$$\Delta_{\text{min}} := \min_{x \in \mathcal{X}} \Delta_{\text{min}}(x).$$

If the distribution $\mathcal{D}$ over contexts is discrete, then Δ_{min} is always strictly positive if all arms have distinct parameters. However, in the case where $\mathcal{D}$ is continuous, it is possible that $\Delta_{\text{min}} = 0$. In such cases, stochastic regret guarantees depending on Δ_{min}^{-1} become vacuous. Nonetheless, the adversarial regret bounds remain valid regardless of the value of Δ_{min}.

We denote $\mathcal{H}_t = \sigma\big(X_s, A_s, O_s, \{l_s(X_s, o)\}_{o \in O_s}, s \leq t\big)$ the filtration generated by all past contexts, actions, and observed losses. Finally, we use equivalently the notation $a = \mathcal{O}(b)$ or $a \lesssim b$ when there exists a constant $\omega > 0$ such that $a \leq \omega b$, where ω is independent of the following problem-dependent quantities: $T, d, K, \Sigma, \mathcal{D}, C, \Delta_{\text{min}}$.

3. ALGORITHM

As is standard in the best-of-both-worlds literature, our algorithm builds on the *Follow-the-Regularized-Leader* (FTRL) framework (see, *e.g.*, Shalev-Shwartz, 2012, Sec. 2.3). This general principle is characterized by three key design choices: a *loss estimator*, a *learning-rate schedule*, and an appropriate *regularizer*.

To obtain loss estimates adapted to the linear contextual setting, we follow the approach of Kuroki et al. (2024), constructing importance-weighted regression estimates of the losses. For computational efficiency, we employ the *Matrix Geometric Resampling (MGR)* method (Neu & Bartók, 2013; Bartók et al., 2014; Kuroki et al., 2024), which guarantees tractability while controlling both the bias and variance of the estimates (see also Neu, 2015).

The other components of our algorithm are more directly inspired by Algorithm 2 of Tsuchiya & Ito (2024), which addresses the best-of-both-worlds problem for multi-armed bandits with paid observations. In particular, we adopt their use of a Tsallis entropy regularizer [Antoine: why?] , an adaptive learning-rate schedule, and the computation of an *observation probability* that is uniform across arms. This probability is derived from the sampling probability vector produced by FTRL. This idea to use distinct observation and sampling probabilities originates from the initial work of Seldin & Slivkins (2014).

In the following, we detail the components of our algorithm for linear contextual bandits with paid observations. The pseudo-code can be found in Algorithm 1.

Sampling distribution (FTRL). We recall that, at each round $t \geq 1$, the learner observes a context vector X_t, and must choose an action $A_t \in [K]$. As a first step, our algorithm computes a sampling distribution $q_t(\cdot \,|\, X_t) \in \Delta_K$, where Δ_K denotes the $(K-1)$-dimensional probability simplex. Following Tsuchiya & Ito (2024), given a context x, this distribution is obtained through the *Follow-the-Regularized-Leader* (FTRL) principle, by solving the optimization problem [Antoine: minus missing in front of the second entropy?]

$$q_t(\cdot \,|\, x) \in \underset{q \in \Delta_K}{\arg\min} \left\{ \sum_{s=1}^{t-1} \left\langle q, \widetilde{\ell}_s(x) \right\rangle + \psi_t(q) + \bar{\beta} H_{\bar{\alpha}}(q) \right\}. \quad (2)$$

Note that $x \mapsto q_t(\cdot \,|\, x)$ is $\mathcal{H}_{t-1}$-measurable. This formulation involves the following components:

- **Loss estimates.** For each round $s \leq t-1$,

$$\widetilde{\ell}_s(x) := \left(\left\langle x, \widetilde{\theta}_{s,1} \right\rangle, \ldots, \left\langle x, \widetilde{\theta}_{s,K} \right\rangle\right)^\top, \quad (3)$$

  where $\widetilde{\theta}_{s,a}$ is an estimator of the linear loss parameter $\theta_{s,a} \in \mathbb{R}^d$ (see Eq. (5)).

- **Regularizer.** We use the Tsallis entropy, with

$$\psi_t(q) := -\frac{H_\alpha(q)}{\eta_t}, \text{ for } H_\alpha(q) := \frac{1}{\alpha - 1} \sum_{a=1}^{K} (q_a^\alpha - q_a),$$

  #DB: Should be $\frac{1}{\alpha-1}$, I see the error is propagated from (Tsuchiya & Ito, 2024) (below Eq. (10)), but I guess in their derivations they then use the right one. #
  where $\eta_t > 0$ is the learning rate at time t, and we fix $\alpha := 1 - (\log K)^{-1}$. For convenience, we also define $\beta_t := 1/\eta_t$.

- **Additional parameters.** We set $\bar{\alpha} := 1 - \alpha$ and

$$\bar{\beta} := \frac{32 K d \sqrt{c}}{(1-\alpha)^2 \sqrt{\beta_1} \min(1, \lambda_{\text{min}})},$$

where c, K, and $\lambda_{\min}$ are as introduced in Section 2. The term $\beta_1 = \eta_1^{-1}$ is introduced here in order to simplify some parts of the analysis, since we will define the learning rate such that $\beta_t \geq \beta_1$ holds for all time steps $t \geq 1$.

The definition of the FTRL distribution in Eq. (2) follows Algorithm 2 of Tsuchiya & Ito (2024), with two key modifications. The first, as previously discussed, is the use of loss estimates specifically adapted to the linear contextual structure of our setting.

The second is the value of $\bar{\beta}$ before the second regularization term, which we use in the analysis to control the evolution of $H_\alpha(q_t)$ between rounds (see Lemma **??**), in particular at the beginning of the interaction (since this term does not scale up with t). This value is adjusted by the parameter $\lambda_{\min}$ to account for the impact of the context distribution in the analysis.

Estimation of the linear losses. We rely on a standard importance-weighted estimator, adapted from Kuroki et al. (2024). The key modification is that, instead of using the sampled action, we use the actions that are *observed* (if any) at round t. Specifically, for $t \geq 1$ and $a \in [K]$, we could estimate $\theta_{t,a}$ by

$$\widehat{\theta}_{t,a} := \Sigma_{t,a}^{-1} X_t \ell_t(X_t, a) \mathbb{1}_{\{a \in O_t\}}, \tag{4}$$

where $\Sigma_{t,a} := \mathbb{E}[\mathbb{1}_{a \in O_t} X_t X_t^\top \mid \mathcal{H}_{t-1}]$ [Antoine: maybe just define it independently of a?] . However, computing $\Sigma_{t,a}^{-1}$ exactly is computationally impractical for two reasons. First, matrix inversion at every round costs $\mathcal{O}(d^3)$ operations, which becomes prohibitive in high dimensions. Second, evaluating $\Sigma_{t,a}$ itself may be extremely costly: even in the discrete-context case, it requires computing observation probabilities for all possible contexts, with complexity at least $\mathcal{O}(|\mathcal{X}|)$, and moreover presupposes full knowledge of the context distribution.

To circumvent this issue, we approximate $\Sigma_{t,a}^{-1}$ using the *Matrix Geometric Resampling* (MGR) procedure, described in Algorithm **??** (Appendix). Computationally, MGR only requires sampling M_t contexts independently from $\mathcal{D}$, evaluating their observation probabilities (*i.e.*, those the algorithm would assign if the context were observed at round t), and performing basic algebraic operations. This reduces the dependence of the cost from $|\mathcal{X}|$ to $\mathcal{O}(\log(T))$, while only requesting access to a sampler of $\mathcal{D}$.

Accordingly, the estimator used in our algorithm is

$$\widetilde{\theta}_{t,a} := \Sigma_{t,a}^{+} X_t \, \ell_t(X_t, a) \mathbb{1}_{\{a \in O_t\}}, \tag{5}$$

where $\Sigma_{t,a}^{+}$ is the approximation of $\Sigma_{t,a}^{-1}$ returned by the MGR routine. Denote $p_{t,\min} = \min p_t$. Guided by our analysis, we set the number of MGR iterations to

$$M_t := \left\lceil \frac{4K}{p_{t,\min}\lambda_{\min}} \ln(t) \right\rceil, \tag{6}$$

[Antoine: Can probably replace $p_{t,\min}$ by γ_t with uniform mixing.] which ensures sufficiently accurate approximation of $\Sigma_{t,a}^{+}$. Compared to Kuroki et al. (2024), where the bias of the estimator is controlled via a forced exploration rate, in our setting this role is played by the observation probability p_t.

Observation probability. Since observing each arm incurs a fixed cost c, the observation probability p_t must balance variance reduction with cost. For any context x, we define

$$\begin{aligned} z_t(x) &:= \frac{4cKd^2}{(1-\alpha)\lambda_{\min}^2} \left(q_{t*}(x)^{2-\alpha} + \sum_{i \neq I_t} q_t(i \mid x)^{2-\alpha} \right), \\ u_t(x) &:= \frac{8d \max(c,1)}{(1-\alpha)\lambda_{\min}} q_{t*}(x)^{1-\alpha}, \text{ where} \\ I_t(x) &:= \arg\max_{i \in [K]} q_t(i \mid x), \text{ and} \\ q_{t*}(x) &:= \min\{q_t(I_t(x) \mid x), 1 - q_t(I_t(x) \mid x)\}. \end{aligned} \tag{7}$$

#DB: d here introduced artificially, redo later.#

Compared to Algorithm 2 in Tsuchiya & Ito (2024), we have modified the definitions of the quantities z_t and u_t to include the $\lambda_{\min}$ and d terms, which becomes necessary to appropriately control the variance of importance-weighted losses. For a learning rate η_t, we then define the observation probability as

$$p_t(x) := \min\left\{ \frac{\sqrt{z_t(x)\eta_t} + u_t(x)\eta_t}{cK}, 1 \right\}. \tag{8}$$

This tuning seems to differ from the one proposed in Eq. 93 of Tsuchiya & Ito (2024) for their BoBW algorithm in the MAB with paid observations setting. As we explain in Section 4, our choice avoids a factor $(\frac{1}{cK} + cK)$ in the regret bound, which would otherwise render the guarantee vacuous when c is very small. Moreover, Eq. (7) shows that without this inverse scaling in c, the observation probability would converge to zero for small c under a fixed sampling probability, which is an unintuitive and undesirable behavior.

The fact that the probability p_t is uniform across arms has two important consequences for the MGR scheme. First, it removes the need for the forced exploration mechanism used in Kuroki et al. (2024) to control the bias (see their Lemma 9), and instead leads to a different result, formalized in our Lemma **??**. Second, since $\Sigma_{t,a}$ is identical for all arms, we only need to compute a single pseudo-inverse Σ_t^{+} per round. As a result, MGR only needs to be executed once at each time step, significantly reducing the overall computational cost.

Learning rate. The learning rate η_t balances stability and adaptivity of FTRL, and is chosen to ensure optimal regret in both regimes. We follow Rule 2 of the framework presented in Tsuchiya & Ito (2024) and use the update rule

$$\frac{1}{\eta_{t+1}} = \frac{1}{\eta_t} + \frac{1}{h_t(X_t)}\Big(2\sqrt{z_t(X_t)\eta_t} + u_t(X_t)\eta_t\Big), \quad (9)$$

where $h_t(X_t)$ denotes the entropy $H(q_t(\cdot \,|\, X_t))$. For notational convenience we set $\gamma_t(x) = cKp_t(x)$ [Antoine: where is it used?] . We also choose η_1 to ensure that $p_t \leq \frac{1}{2}$ for all time steps,

$$\eta_1 = \frac{(1-\alpha)\lambda_{\min}^2}{64\max(c,1)K}. \quad (10)$$

Algorithm 1 FTRL for linear contextual bandits with paid observations

```
Input: $K$ arms, cost $c$, minimum eigenvalue $\lambda_{\min}$.
Initialize $\eta_1$ as in Eq. (10), and for any arm $a \in [K]$, set $\widetilde{\theta}_{0,a} = 0$.
for $t = 1, 2, \ldots, T$ do
    Observe $X_t$ and compute $q_t(\cdot \,|\, X_t)$ as in Eq. (2).
    Sample $A_t \sim q_t(\cdot \,|\, X_t)$.
    Compute $p_t(X_t)$ as in Eq. (8).
    For any $a$, observe $\ell_t(X_t, a)$ with prob. $p_t(X_t)$.
    Suffer the loss $\ell_t(X_t, A_t) + c|O_t|$.
    Update $\eta_t$ to $\eta_{t+1}$ according to Eq. (9).
    For any $a$, compute and store $\widetilde{\theta}_{t,a}$ via Alg. ??.
    Compute and store $\Sigma_t^+$ via MGR (see Alg. ??) with $M_t$ iterations.
end for
```

Computation time and memory. The total space and time complexity of Algorithm 1 are respectively $\mathcal{O}\big(Td^2\big)$ and $\mathcal{O}\big(K^2T^2d^2\log T\big)$. Details can be found in Appendix **??**.

4. REGRET ANALYSIS

We now introduce the main theoretical result of this work, which is that Algorithm 1 achieves Best-of-Both-Worlds regret guarantees in the setting of linear bandits with paid observations, under the assumptions introduced in Section 2.

#DB: We have to change the dimension dependency after the fix. #

Theorem 4.1. *In the adversarial regime, the regret of Algorithm 1 satisfies*

$$R_T \lesssim \left(\frac{cKd^2\log K}{\lambda_{\min}^2}\right)^{1/3} T^{2/3} + \sqrt{\frac{\max(c,1)d\log K \cdot T}{\lambda_{\min}}} + \kappa$$

with [Antoine: $\log KT$ or $T\log K$?]

$$\kappa = \sqrt{\frac{cKd^2\log K}{\lambda_{\min}^2}} + \frac{\max(c,1)d\log K}{\lambda_{\min}} + \frac{\max(c,1)K\log K}{\lambda_{\min}^2} + \frac{32Kd\sqrt{c}}{(1-\alpha)^2\sqrt{\beta_1}\min(1,\lambda_{\min})}.$$

while in the corrupted stochastic regime with corruption level C, it satisfies

$$R_T \lesssim \frac{d\sqrt{\max(c,1)K\log K}}{\lambda_{\min}\Delta_{min}^2} \cdot \log\big(T\Delta_{min}^3\big) + \left(\frac{C^2d\sqrt{\max(c,1)K\log K}}{\lambda_{\min}\Delta_{min}^2} \cdot \log\left(\frac{T\Delta_{min}}{C}\right)\right)^{1/3} + \kappa + \kappa',$$

where we further define

$$\kappa' = \left(\left(\frac{cKd^2\log K}{\lambda_{\min}^2}\right)^{1/3} + \sqrt{\frac{\max(c,1)d\log K}{\lambda_{\min}}}\right) \times \left(\frac{1}{\Delta_{min}^3} + \frac{C}{\Delta_{min}}\right)^{2/3}.$$

This result shows that Algorithm 1 achieves the minimax-optimal $\mathcal{O}\big(T^{2/3}\big)$ regret in the adversarial regime, while smoothly adapting to the (possibly corrupted) stochastic regime with logarithmic dependence on T when $C = 0$. These bounds match the known lower bounds from Seldin et al. (2014), which applies to our setting since it encompasses the standard multi-armed bandit (by taking $d = 1$ and $X_t = 1$ a.s.), and extend the Best-of-Both-Worlds (BoBW) framework of Tsuchiya & Ito (2024) to the setting of linear bandits.

While the dependence in T is thus known to be optimal, the optimal dependence in other problem-specific parameters remains unknown, as this is the first work to address this setting. However, since our algorithm builds upon and generalizes both Algorithm 2 from Kuroki et al. (2024) and Algorithm 2 from Tsuchiya & Ito (2024), we can compare our regret bounds to theirs, even if the settings do not perfectly align. We consider first the limiting case where $c \to 0$, corresponding to the full-information setting, in which all losses are observed. In this regime, the first term of the adversarial regret bound vanishes, and we have

$$R_T \lesssim \sqrt{\frac{dT\log(K)}{\lambda_{\min}}}.$$

This matches, up to logarithmic factors, the adversarial regret bound established for Algorithm 2 in Kuroki et al. (2024), namely

$$R_T \lesssim \sqrt{T\left(d + \frac{\log T}{\lambda_{\min}}\right)K\log K\log T}.$$

In our case, the factor K is replaced by $\log K$, which reflects the full-information nature of our setting, a standard improvement in such regimes. However, in the stochastic regime, our regret exhibits an additional $\frac{1}{\Delta_{\text{min}}}$ factor compared to the full-information bounds in Kuroki et al. (2024). But on the countrary, our algorithm has a better $\log T$ dependence, thus our bound is better if T is significantly larger than $\frac{1}{\Delta_{\text{min}}}$. However, we do not know whether our improved $\log T$ dependency stems from being in the full-information setting or from other factors. We can at least observe that the dependence on the setting-specific parameters d and $\lambda_{\min}$ in our bounds matches that of their Algorithm 2.

Another useful comparison is to consider the special case $d = 1, \mathcal{X} = \{1\}$, in which case we recover the setting of Seldin & Slivkins (2014). From their Corollary 17, Algorithm 2 of Tsuchiya & Ito (2024) obtain an adversarial regret bound of

$$R_T \lesssim \left((cK)^{1/3} T^{2/3} (\log K)^{1/3}\right),$$

which is exactly the scaling that we obtain with Theorem 4.1 in this setting. This observation furthermore still holds in the stochastic setting.

These comparisons suggest that, while we can not establish optimality in general due to the lack of known lower bounds, our algorithm can be viewed as a strict generalization of the approach in Tsuchiya & Ito (2024) for bandits with paid observations, since we recover their guarantees in this setting. Moreover, since the dependencies in d and $\lambda_{\min}$ are known to be optimal compared to previous approaches when $c = 0$, this further supports the relevance of our design beyond prior approaches.

A detailed proof of the theorem can be found in Appendix **??**. In the following, we present the main steps of the proofs, highlighting the technical arguments that required to be adapted from the existing frameworks.

Proof sketch. As a preliminary step of the analysis, we isolate the difficulty induced by the use of (biased) MGR estimates (Eq. (5)) instead of using the unbiased estimators from Eq. (3). Following the proof technique of Kuroki et al. (2024), we introduce an auxiliary game where these estimators are treated as unbiased, and for which the regret would thus become

$$\tilde{R}_T := \mathbb{E}\left[\sum_{t=1}^{T} \left\langle X_t, \tilde{\theta}_{t,A_t} \right\rangle - \left\langle X_t, \tilde{\theta}_{t,\pi^\star(X_t)} \right\rangle\right].$$

We can verify that the actual regret of our algorithm thus satisfies

$$R_T \leq \tilde{R}_T + 2 \sum_{t=1}^{T} \max_{a\in[K]} \left| \mathbb{E}\left[\left\langle X_t, \tilde{\theta}_{t,a} - \theta_{t,a} \right\rangle\right] \right|.$$

Then, in Lemma **??** we prove that the second term of this upper bound can be upper bounded by a constant, independent of all problem parameters. In the following, we thus focus on upper bounding $\tilde{R}_T$. We write the following proof steps with the notation R_T, with an abuse of notation, since previous result showed that both terms have the same scaling in T.

The remainder of the analysis builds on the general framework introduced by Tsuchiya & Ito (2024) to build Best-of-Both-Worlds algorithms for problems with minimax regret scaling with $T^{2/3}$, and in particular their instantiation of this framework to tackle standard multi-armed bandit with paid observations (without the linear contextual structure). Our first contribution is an adaptation of their Theorem 7 to accommodate the linear contextual structure, that we introduce below.

Lemma 4.2 (Adaptation of Theorem 7 of Tsuchiya & Ito, 2024)**.** *Suppose that Algorithm 1 satisfies the following conditions in the adversarial regime:*

(i) $R_T \leq \sum_{t=1}^{T} \mathbb{E}\Big[\Big(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\Big) h_t + \frac{z_t \eta_t}{\gamma_t} + \gamma_t\Big] + \bar{\beta}\bar{h},$

(ii) $\mathbb{E}[h_{t+1} \mid \mathcal{H}_t] \leq 2\mathbb{E}[h_t \mid \mathcal{H}_{t-1}]$ *for all* $t \geq 1$.

Then the regret can be bounded as

$$R_T \lesssim (z_{\max} h_1)^{1/3} T^{2/3} + \sqrt{u_{\max} h_1 T} + \kappa,$$

where

$$z_{\max} = \max_{t\in[T]} z_t \leq 4cK \log K \frac{1}{\lambda_{\min}^2},$$
$$u_{\max} = \max_{t\in[T]} u_t \leq 4 \max(c, 1) \log K \frac{1}{\lambda_{\min}},$$

and

$$\kappa := \sqrt{z_{\max}\eta_1} + u_{\max}\eta_1 + \tfrac{h_1}{\eta_1} + \bar{\beta} h_{\max}.$$

Moreover, if Algorithm 1 satisfies the following conditions in the stochastic regime: there exists a constant $\rho > 0$ such that, for any $t \geq 1$,

(iii) $\sqrt{z_t h_t} \leq \sqrt{\rho}(1 - q_t(\pi_T^\star(X_t) \mid X_t))$*, and*

(iv) $u_t h_t \leq \rho\,(1 - q_t(\pi_T^\star(X_t) \mid X_t))$,

then, for $T \geq \tau := \frac{1}{\Delta_{min}^3} + \frac{C}{\Delta_{min}}$ *it holds that*

$$R_T \lesssim \frac{\rho}{\Delta_{min}^2} \log\big(T\Delta_{min}^3\big) + \Big(\tfrac{C^2\rho}{\Delta_{min}^2} \log\big(\tfrac{T\Delta_{min}}{C}\big)\Big)^{1/3} + \kappa'$$

with

$$\kappa' := \kappa + \Big((z_{\max}h_1)^{1/3} + \sqrt{u_{\max}h_1}\Big)\Big(\tfrac{1}{\Delta_{min}^3} + \tfrac{C}{\Delta_{min}}\Big)^{2/3}.$$

While Lemma 4.2 adapts Theorem 7 from Tsuchiya & Ito (2024), it differs in several significant aspects. First, condition (i) is new and replaces conditions (i)–(ii) in the original theorem, and both lead to a similar proof structure, our condition better adjust the framework to our setting. Second, condition (ii) is a relaxed reformulation of condition (iii) in Tsuchiya & Ito (2024), which is necessary to handle the stochasticity of contexts in our setting. With careful use of the tower rule, we show that this weaker assumption is sufficient for the regret analysis. Finally, conditions (iii) and (iv) are reformulations of conditions (iv) and (v) from Tsuchiya & Ito (2024), and the corresponding proof techniques carry over with only little modifications. The detailed proof of this lemma is deferred to Appendix **??**.

To establish Theorem 4.1, it then suffices to verify that Algorithm 1 satisfies each of the four conditions.

Condition (i) follows from the standard FTRL regret decomposition: the stability term bound is direct to obtain, while the penalty term is controlled using Lemma **??** (in Appendix), which is similarly to the proof of Tsuchiya & Ito (2024, Theorem 8).

We prove condition (ii) in Lemma **??**. The proof consists in applying Lemma 15 from Tsuchiya & Ito (2024) (restated as Lemma **??**) for each fixed context, and to conclude via linearity of expectation. A key challenge arises from the fact that, in our setting, we have the bound $\mathbb{E}\left[\langle X_t, \hat{\theta}_{t,a}\rangle^2\right] \leq \frac{1}{\lambda_{\min}^2 p_t}$, which contrasts with the original bound $\mathbb{E}\left[\ell_t^2\right] \leq \frac{1}{p_t}$ in the non-contextual case. Since Lemma **??** only accommodates a constant upper bound, this discrepancy required a careful adjustment of several parameters, specifically u_t and $\bar{\beta}$, which represents a slight modification in the precise behavior of the algorithm.

Finally, conditions (iii) and (iv) are verified by combining entropy bounds from Tsuchiya & Ito (2024) with direct control of the variance-like quantities z_t and u_t, thereby linking them to the optimal action probability.

Together, these arguments ensure that Algorithm 1 satisfies the assumptions of Lemma 4.2, which directly yields the regret guarantees stated in Theorem 4.1.

The full derivations and supporting lemmas are deferred to Appendix **??**, where we carefully establish that each condition of the lemma holds in our setting. □

While the definition of p_t in Tsuchiya & Ito (2024) differs from ours by a factor $(cK)^{-1}$, this appears to be a simple typo in their presentation. Indeed, their analysis assumes $p_t = \frac{1}{cK}\left(\sqrt{z_t\eta_t} + u_t\eta_t\right)$, even though the statement of their Algorithm 2 defines $p_t := \sqrt{z_t\eta_t} + u_t\eta_t$. We can use this observation to comment on the optimality of the tuning of p_t with respect to the analysis used to derive BoBW regret bounds for our algorithm.

Indeed, a step in the analysis (see Eq. (**??**)) involves the quantity $\gamma_t' := \gamma_t - \frac{u_t}{\beta_t}$. With our definition, this yields $\gamma_t' = \sqrt{z_t/\beta_t}$, while using the unnormalized p_t (without $1/(cK)$) gives

$$\gamma_t' = cK\sqrt{z_t\eta_t} + (cK-1)u_t\eta_t \geq cK\sqrt{z_t\eta_t},$$

assuming $cK \geq 1$. This leads to the bound

$$\sum_{t=1}^{T}\mathbb{E}\left[\frac{z_t\eta_t}{\gamma_t'} + \gamma_t\right] \leq \sum_{t=1}^{T}\mathbb{E}\left[\frac{1}{cK}\sqrt{\frac{z_t}{\beta_t}} + cK\left(\sqrt{\frac{z_t}{\beta_t}} + \frac{u_t}{\beta_t}\right)\right]$$

$$\leq \left(\frac{1}{cK} + cK\right)\sum_{t=1}^{T}\mathbb{E}\left[2\sqrt{\frac{z_t}{\beta_t}} + \frac{u_t}{\beta_t}\right].$$

The factor $(cK)^{-1} + cK$ then propagates through the analysis and degrades the regret bound. More generally, an overestimation of p_t by a multiplicative factor ω leads to a regret that is worsened by a factor proportional to $\omega + \omega^{-1}$, so $\omega = 1$ (our tuning) is optimal.

5. DISCUSSION

We proposed an algorithm achieving BoBW regret guarantees in the setting of *linear contextual bandits with paid observations*, with explicit scaling in problem dimensions (d, K) and parameters ($\lambda_{\min}$, $\Delta_{\min}$, c).

However, an important limitation, shared with the analysis of Algorithm 2 from Kuroki et al. (2024), arises in the stochastic setting when the context space is continuous. In such cases, the quantity $\Delta_{\min}$ is often zero, which implies that the regret bound remains at $\Theta\left(T^{2/3}\right)$, even though the environment is stochastic and should, in principle, allow for better rates. This issue also affects discrete but finely spaced context spaces, where $\Delta_{\min} > 0$ but can be arbitrarily small, leading to overly pessimistic bounds in practice. Nevertheless, Bastani et al. (2021) demonstrates that under suitable regularity conditions on the context distribution, it is possible to achieve logarithmic regret in continuous settings without any dependence on $\Delta_{\min}$. Extending such ideas to our setting, and combining them with BoBW-style guarantees, could lead to improved regret bounds, potentially polylogarithmic or polynomially better than $\sqrt{T}$ or $T^{2/3}$. We believe this is a promising direction for future work.

Finally, as previously discussed, since this setting is novel, there are currently no lower bounds specifically tailored to it. Existing lower bounds only apply to simplified or special cases of our setting. Developing minimax and stochastic lower bounds that are adapted to this setting, precisely capturing all dimensions and parameters, would therefore be an interesting contribution to improve the understanding of this setting.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems 24*, pp. 2312–2320, 2011.

Abe, N. and Long, P. M. Associative reinforcement learning using linear probabilistic concepts. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pp. 3–11. Morgan Kaufmann, 1999.

Abeille, M. and Lazaric, A. Linear thompson sampling revisited. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pp. 176–184. PMLR, 2017.

Abeille, M., Janz, D., and Pike-Burke, C. When and why randomised exploration works (in linear bandits). In *Proceedings of the International Conference on Algorithmic Learning Theory*, volume 272, pp. 4–22. PMLR, 2025.

Agrawal, S. and Goyal, N. Thompson sampling for contextual bandits with linear payoffs. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 127–135. JMLR.org, 2013.

Alon, N., Cesa-Bianchi, N., Gentile, C., and Mansour, Y. From bandits to experts: A tale of domination and independence. In *Advances in Neural Information Processing Systems 26*, pp. 1610–1618, 2013.

Amir, I., Azov, G., Koren, T., and Livni, R. Better best of both worlds bounds for bandits with switching costs. *Advances in Neural Information Processing Systems*, 35: 15800–15810, 2022.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. E. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

Avner, O., Mannor, S., and Shamir, O. Decoupling exploration and exploitation in multi-armed bandits. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Bartók, G., Foster, D., Pál, D., Rakhlin, A., and Szepesvári, C. Partial monitoring—classification, regret bounds, and algorithms. *Mathematics of Operations Research*, 39(4): 967–997, 2014. doi: 10.1287/moor.2014.0663.

Bastani, H., Bayati, M., and Khosravi, K. Mostly exploration-free algorithms for contextual bandits. *Management Science*, 67(3):1329–1349, 2021. doi: 10.1287/mnsc.2020.3605.

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. E. Contextual bandit algorithms with supervised learning guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15, pp. 19–26. JMLR.org, 2011.

Bubeck, S. and Cesa-Bianchi, N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5(1):1–122, 2012. doi: 10.1561/2200000024.

Bubeck, S. and Slivkins, A. The best of both worlds: Stochastic and adversarial bandits. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23, pp. 42.1–42.23. PMLR, 2012.

Dani, V., Hayes, T. P., and Kakade, S. M. Stochastic linear optimization under bandit feedback. In *Proceedings of the 21st Annual Conference on Learning Theory*, pp. 355–366. Omnipress, 2008.

Dann, C., Wei, C.-Y., and Zimmert, J. A blackbox approach to best of both worlds in bandits and beyond. In *Proceedings of the 36th Annual Conference on Learning Theory*, volume 195, pp. 5503–5570. PMLR, 2023.

Degenne, R., Shao, H., and Koolen, W. M. Structure adaptive algorithms for stochastic bandits. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 2443–2452. PMLR, 2020.

Efroni, Y., Merlis, N., Saha, A., and Mannor, S. Confidence-budget matching for sequential budgeted learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 2937–2947. PMLR, 2021.

Flynn, H., Reeb, D., Kandemir, M., and Peters, J. R. Improved algorithms for stochastic linear bandits using tail bounds for martingale mixtures. In *Advances in Neural Information Processing Systems 36*, 2023.

Jin, T., Liu, J., and Luo, H. Improved best-of-both-worlds guarantees for multi-armed bandits: Ftrl with general regularizers and multiple optimal arms. *Advances in Neural Information Processing Systems*, 36:30918–30978, 2023.

Kato, M. and Ito, S. Lc-tsallis-inf: Generalized best-of-both-worlds linear contextual bandits. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, volume 258, pp. 3655–3663. PMLR, 2025.

Kirschner, J., Lattimore, T., and Krause, A. Information directed sampling for linear partial monitoring. In *Proceedings of the Conference on Learning Theory*, volume 125, pp. 2328–2369. PMLR, 2020.

Kuroki, Y., Rumi, A., Tsuchiya, T., Vitale, F., and Cesa-Bianchi, N. Best-of-both-worlds algorithms for linear contextual bandits. In Dasgupta, S., Mandt, S., and Li, Y. (eds.), *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1216–1224. PMLR, 2024.

Langford, J. and Zhang, T. The epoch-greedy algorithm for multi-armed bandits with side information. In *Advances in Neural Information Processing Systems 20*, pp. 817–824, 2007.

Lattimore, T. and Szepesvári, C. The end of optimism? an asymptotic analysis of finite-armed linear bandits. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pp. 728–737. PMLR, 2017.

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

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 661–670. ACM, 2010. doi: 10.1145/1772690.1772758.

Mannor, S. and Shamir, O. From bandits to experts: On the value of side-observations. In *Advances in Neural Information Processing Systems 24*, pp. 684–692, 2011.

Neu, G. Explore no more: Improved high-probability regret bounds for non-stochastic bandits. *arXiv preprint arXiv:1506.03271*, 2015.

Neu, G. and Bartók, G. An efficient algorithm for learning with semi-bandit feedback. *arXiv preprint arXiv:1305.2732*, 2013.

Neu, G. and Olkhovskaya, J. Efficient and robust algorithms for adversarial linear contextual bandits. In *Proceedings of the Conference on Learning Theory*, volume 125, pp. 3049–3068. PMLR, 2020.

Olkhovskaya, J., Mayo, J. J., van Erven, T., Neu, G., and Wei, C.-Y. First- and second-order bounds for adversarial linear contextual bandits. In *Advances in Neural Information Processing Systems 36*, 2023.

Rouyer, C., van der Hoeven, D., Cesa-Bianchi, N., and Seldin, Y. A near-optimal best-of-both-worlds algorithm for online learning with feedback graphs. In *Advances in Neural Information Processing Systems 35*, 2022.

Saha, A. and Gaillard, P. Versatile dueling bandits: Best-of-both world analyses for learning from relative preferences. In *Proceedings of the International Conference on Machine Learning*, pp. 19011–19026. PMLR, 2022.

Seldin, Y. and Slivkins, A. One practical algorithm for both stochastic and adversarial bandits. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 1287–1295. JMLR.org, 2014.

Seldin, Y., Bartlett, P., Crammer, K., and Abbasi-Yadkori, Y. Prediction with limited advice and multiarmed bandits with paid observations. In Xing, E. and Jebara, T. (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 280–287, Beijing, China, 2014. PMLR.

Shalev-Shwartz, S. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

Tsuchiya, T. and Ito, S. A simple and adaptive learning rate for ftrl in online learning with minimax regret of $theta(t^{2/3})$ and its application to best-of-both-worlds. *NeurIPS*, 2024.

Tsuchiya, T., Ito, S., and Honda, J. Further adaptive best-of-both-worlds algorithm for combinatorial semi-bandits. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 8117–8144. PMLR, 2023.

Yun, D., Proutiére, A., Ahn, S., Shin, J., and Yi, Y. Multi-armed bandit with additional observations. *Proceedings of ACM Measurement and Analysis of Computing Systems*, 2(1):13:1–13:22, 2018. doi: 10.1145/3179416.

Zimmert, J. and Marinov, T. V. Productive bandits: Importance weighting no more. In *Advances in Neural Information Processing Systems 37*, pp. 85360–85388, 2024.

Zimmert, J. and Seldin, Y. Tsallis-inf: An optimal algorithm for stochastic and adversarial bandits, 2022.

.1. Technical tools

Lemma .1 (Neu & Olkhovskaya, 2020, Lemma 3)**.** *Let $\pi^\star$ be a fixed stochastic policy and let X_0 be a sample from the context distribution $\mathcal{D}$ independent from $\mathcal{H}_T$. For any $t \in [T]$, any action $a \in [K]$, suppose that π_t is $\mathcal{H}_{t-1}$-measurable and that $\mathbb{E}\left[\widehat{\theta}_{t,a} \,\middle|\, \mathcal{H}_{t-1}\right] = \theta_{t,a}$. Then, it holds that*

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K}(\pi_t(a \,|\, X_t) - \pi^\star(a \,|\, X_t))\langle X_t, \theta_{t,a}\rangle\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K}(\pi_t(a \,|\, X_0) - \pi^\star(a \,|\, X_0))\left\langle X_0, \widehat{\theta}_{t,a}\right\rangle\right].$$

Lemma .2. *[Antoine: Exercise 28.12 from Lattimore & Szepesvári (2020) as a lemma + proof]*

Lemma .3. *[Antoine: Convexity of the negative Tsallis entropy, corresponding Bregman divergence]*

A. Temp Antoine

A.1. Regret analysis

Recall the definition of the regret

$$\begin{aligned} R_T &= \mathbb{E}\left[\sum_{t=1}^{T}(\ell_t(X_t, A_t) - \ell_t(X_t, \pi^\star_T(X_t)))\right] + c\,\mathbb{E}\left[\sum_{t=1}^{T}|O_t|\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K}(q_t(a \,|\, X_t) - \pi^\star_T(a \,|\, X_t))\langle X_t, \theta_{t,a}\rangle\right] + cK\,\mathbb{E}\left[\sum_{t=1}^{T}p_t(X_t)\right], \end{aligned}$$

where we used the fact that $\mathbb{E}[\ell_t(X_t, a) \,|\, \mathcal{H}_{t-1}] = \langle X_t, \theta_{t,a}\rangle$ for any $a \in [K]$ for the first term, and that $\mathbb{E}[|O_t| \,|\, \mathcal{H}_{t-1}, X_t] = Kp_t(X_t)$ for the second term, which holds because at time t, each arm a is observed with probability $p_t(X_t)$.

We introduce a ghost sample $X_0 \sim \mathcal{D}$ independent from $\mathcal{H}_T$. Conditional on $\mathcal{H}_{t-1}$, both X_t and X_0 are i.i.d. from $\mathcal{D}$, and p_t is $\mathcal{H}_{t-1}$-measurable, hence, we have

$$\mathbb{E}[p_t(X_t) \,|\, \mathcal{H}_{t-1}] = \mathbb{E}[p_t(X_0) \,|\, \mathcal{H}_{t-1}].$$

Recall the definition of the importance-weighted estimator $\widehat{\theta}_{t,a} = \Sigma_{t,a}^{-1} X_t \ell_t(X_t, a)\mathbb{1}_{\{a \in O_t\}}$ and note that $\mathbb{E}\left[\widehat{\theta}_{t,a} \,\middle|\, \mathcal{H}_{t-1}\right] = \theta_{t,a}$. By Lemma .1, we can further rewrite the regret as

$$\begin{aligned} R_T &= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K}(q_t(a \,|\, X_0) - \pi^\star_T(a \,|\, X_0))\left\langle X_0, \widehat{\theta}_{t,a}\right\rangle\right] + cK\,\mathbb{E}\left[\sum_{t=1}^{T}p_t(X_0)\right] \\ &= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K}(q_t(a \,|\, X_0) - \pi^\star_T(a \,|\, X_0))\left\langle X_0, \widetilde{\theta}_{t,a}\right\rangle\right] + cK\,\mathbb{E}\left[\sum_{t=1}^{T}p_t(X_0)\right] \\ &\quad + \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K}(q_t(a \,|\, X_0) - \pi^\star_T(a \,|\, X_0))\left\langle X_0, \widehat{\theta}_{t,a} - \widetilde{\theta}_{t,a}\right\rangle\right] \\ &\le \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a=1}^{K}(q_t(a \,|\, X_0) - \pi^\star_T(a \,|\, X_0))\left\langle X_0, \widetilde{\theta}_{t,a}\right\rangle\right] + cK\,\mathbb{E}\left[\sum_{t=1}^{T}p_t(X_0)\right] \\ &\quad + 2\sum_{t=1}^{T}\max_{a \in [K]}\left|\mathbb{E}\left[\left\langle X_0, \widehat{\theta}_{t,a} - \widetilde{\theta}_{t,a}\right\rangle\right]\right|. \end{aligned}$$

where $\widetilde{\theta}_{t,a}$ are is MGR estimate defined as $\widetilde{\theta}_{t,a} = \Sigma_t^{+} X_t \ell_t(X_t, a)\mathbb{1}_{\{a \in O_t\}}$ (see Eq. 5). For any context $x \in \mathcal{X}$, we define the auxiliary regret

$$\widetilde{R}_T(x) \coloneqq \sum_{t=1}^{T}\sum_{a=1}^{K}(q_t(a \,|\, x) - \pi^\star_T(a \,|\, x))\left\langle x, \widetilde{\theta}_{t,a}\right\rangle + cK\sum_{t=1}^{T}p_t(x)$$

and the bias induced by MGR

$$\texttt{bias}_{\texttt{MGR}} \coloneqq \sum_{t=1}^{T} \max_{a\in[K]} \left|\mathbb{E}\left[\left\langle X_0, \widehat{\theta}_{t,a} - \widetilde{\theta}_{t,a}\right\rangle\right]\right|.$$

With this notation, we can write the inequality above as

$$R_T \leq \mathbb{E}_{X_0\sim\mathcal{D}}\left[\widetilde{R}_T(X_0)\right] + 2\,\texttt{bias}_{\texttt{MGR}}.$$

We bound the two terms separately. In Lemma **??** we prove that the bias induced by MGR is upper bounded as

$$2\texttt{bias}_{\texttt{MGR}} \leq \frac{\pi^2}{3}\,.$$

Thus, it remains to upper bound $\mathbb{E}_{X_0\sim\mathcal{D}}\left[\widetilde{R}_T(X_0)\right]$. Let us fix a context $x \in \mathcal{X}$ and recall that at time t, the distribution $q_t(\cdot\,|\,x)$ is

$$q_t(\cdot\,|\,x) = \operatorname*{arg\,min}_{q\in\Delta_K}\left\{\sum_{s=1}^{t-1}\sum_{a=1}^{K} q(a)\left\langle x, \widetilde{\theta}_{s,a}\right\rangle + \mathcal{R}_t(q)\right\},$$

where we denoted $\mathcal{R}_t(q) = \frac{1}{\eta_t}(-H_\alpha(q)) + \bar{\beta}(-H_{\bar{\alpha}}(q))$. By the standard FTRL analysis (Lattimore & Szepesvári, 2020, Exercise 28.12), we have

$$\begin{aligned}\widetilde{R}_T(x) \leq{}& \sum_{t=1}^{T}\left(\sum_{a=1}^{K}(q_t(a\,|\,x) - q_{t+1}(a\,|\,x))\left\langle x, \widetilde{\theta}_{t,a}\right\rangle - D_{\mathcal{R}_t}(q_{t+1}(\cdot\,|\,x), q_t(\cdot\,|\,x))\right)\\ &+ \mathcal{R}_{T+1}(\pi^\star_T(\cdot\,|\,x)) - \mathcal{R}_1(q_1(\cdot\,|\,x)) + \sum_{t=1}^{T}\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t}\right)H_\alpha(q_{t+1}(\cdot\,|\,x))\\ &+ cK\sum_{t=1}^{T} p_t(x).\end{aligned}$$