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

Anonymous Authors¹

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

¹Anonymous 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.

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,

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

057 Achieving Best-of-Both-Worlds (BoBW) performance in
 058 hard problems, *i.e.* problems that incur a minimax regret
 059 of $\Theta(T^{2/3})$ in the adversarial regime, is a significant chal-
 060 lenge, as highlighted in (Tsuchiya & Ito, 2024). The stan-
 061 dard approaches used in other settings often fail without
 062 substantial modifications. Fortunately, (Tsuchiya & Ito,
 063 2024) introduced a dedicated framework designed to facil-
 064 itate the design and analysis of BoBW algorithms for such
 065 problems. #DB: the following will likely change after fixing
 066 the proof. # While this framework forms the basis of our
 067 analysis, several challenges arise in adapting it to our setting.
 068 First, the general formulation assumes the existence of a
 069 single optimal arm throughout the learning process, which
 070 does not hold in the contextual linear setting where the
 071 optimal action varies with the context. Second, our setting
 072 introduces a new key parameter, the smallest non-negative
 073 eigenvalue of the context distribution (λ_{\min}), introduced
 074 in Section 2, which necessitates specific tuning of several
 075 algorithmic parameters. Third, we identify and resolve an
 076 inconsistency in prior applications of the BoBW framework
 077 to bandits with paid observations, thereby obtaining tighter
 078 regret guarantees; we elaborate on this point in Section 4.
 079 Structural differences in our setting require various other
 080 adjustments to the technical proofs.

081 1.1. Detailed literature review

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

084 **Linear Contextual Bandits** Contextual bandits extend
 085 classical multi-armed bandits by allowing the reward distri-
 086 bution to depend on an observed context, which can vary
 087 across rounds. To enable efficient decision-making, one
 088 must adopt a suitable model to capture how the context in-
 089 fluences the rewards. In this work we consider the *linear*
 090 *contextual bandit* model (Langford & Zhang, 2007; Li et al.,
 091 2010), that we formally describe in Section 2. This model
 092 is closely-related to the well-studied *stochastic linear bandit*
 093 framework, since in both settings the average reward of
 094 each arm is given by the inner product of an arm feature
 095 vector and a parameter vector. The two formulations differ
 096 in the source of uncertainty: in stochastic linear bandits the
 097 arm features are known and the underlying parameter is
 098 unknown, whereas in (stochastic) linear contextual bandits
 099 the arm-specific features are fixed but unknown, while the
 100 context vector is revealed at the beginning of each round.

101 Most approaches used in linear contextual bandits are bor-
 102 rowed from the stochastic linear bandit literature, in which
 103 algorithms follow general principles such as *Optimism in
 104 Face of Uncertainty* (Abe & Long, 1999; Dani et al., 2008;

105 Abbasi-Yadkori et al., 2011; Flynn et al., 2023), *Thompson
 106 Sampling* (Agrawal & Goyal, 2013; Abeille & Lazaric,
 107 2017; Abeille et al., 2025), *Information Directed Sampling*
 108 (Kirschner et al., 2020), or (asymptotic) lower bound match-
 109 ing (Lattimore & Szepesvári, 2017; Degenne et al., 2020).
 110 Nonetheless, linear contextual bandits exhibit specific prop-
 111 erties compared to standard linear bandits. In particular,
 112 Bastani et al. (2021) showed that under suitable assump-
 113 tions on *context diversity*, even a simple greedy strategy can
 114 achieve logarithmic regret.

115 While the above works assume stochastic rewards, this as-
 116 sumption can be restrictive in practice. To address this,
 117 Neu & Olkhovskaya (2020) introduced an adversarial for-
 118 mulation of linear contextual bandits, in which arm param-
 119 eters are fixed by an oblivious adversary. They derived a
 120 $\tilde{\mathcal{O}}(\sqrt{KdT})$ regret bound for an exponential-weights algo-
 121 rithm (Auer et al., 2002), where d is the parameter dimen-
 122 sion, K is the number of arms, and T is the horizon. Build-
 123 ing on this, Olkhovskaya et al. (2023) obtained refined first
 124 and second-order bounds. In parallel, Kuroki et al. (2024)
 125 established the first *Best-of-Both-Worlds* guarantees in this
 126 setting, showing that one can achieve simultaneously poly-
 127 logarithmic regret in the stochastic regime and $\tilde{\mathcal{O}}(Kd\sqrt{T})$
 128 regret in the adversarial case.

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

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

150 **Best-of-Both-Worlds (BoBW).** The design of algorithms
 151 that perform well simultaneously in stochastic and adver-
 152 sarial regimes has become a central theme in the bandit
 153 literature. The foundational work of Bubeck & Slivkins
 154 (2012); Seldin & Slivkins (2014) initiated this line of re-

110 search by asking whether one can achieve logarithmic regret
 111 in the stochastic setting while retaining $\tilde{O}(\sqrt{T})$ regret in the
 112 adversarial case. Their results provided only partial success,
 113 either with suboptimal bounds or with algorithms of limited
 114 practicality. Later, Zimmert & Seldin (2022) first obtained
 115 the optimal best-of-both-worlds guarantees in the K -armed
 116 bandit setting. This breakthrough has since inspired the
 117 development of BoBW algorithms across a variety of ban-
 118 dit problems (Amir et al., 2022; Rouyer et al., 2022; Saha
 119 & Gaillard, 2022; Tsuchiya et al., 2023; Jin et al., 2023;
 120 Zimmert & Marinov, 2024; Kato & Ito, 2025).

121 Of particular relevance to our work, Kuroki et al. (2024)
 122 studied linear contextual bandits through the black-box re-
 123 duction framework of Dann et al. (2023), which can be
 124 used to design BoBW algorithms for problems whose min-
 125 imax regret scales as \sqrt{T} . More recently, Tsuchiya & Ito
 126 (2024) proposed a general recipe for constructing BoBW
 127 algorithms in so-called “hard” online learning problems,
 128 namely those with minimax regret of order $\Theta(T^{2/3})$. They
 129 further show that several known bandit models, including
 130 multi-armed bandits with paid observations, fall within this
 131 framework. Our work is inspired by their approach, how-
 132 ever, a direct application of their method does not yield
 133 optimal bounds in our setting (see Section 4). This moti-
 134 vates the need for a careful adaptation of their ideas, which
 135 we develop in the remainder of the paper.

2. Problem Definition

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

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

- 150 1. For each action $a \in [K] := \{1, \dots, K\}$, the environ-
 151 ment selects a loss parameter $\theta_{t,a} \in \mathbb{R}^d$.
- 152 2. A context $X_t \in \mathcal{X}$ is drawn from \mathcal{D} .
- 153 3. The learner observes X_t , chooses an action $A_t \in [K]$,
 154 and an observation set $O_t \subseteq [K]$.
- 155 4. The learner incurs loss $\ell_t(X_t, A_t) + c|O_t|$, where ℓ_t is
 156 a loss function that depends on the environment param-
 157 eters $(\theta_{t,a})_{a \in [K]}$, $c > 0$ is the known unit cost of ob-
 158 servation, and $|O_t|$ is the cardinality of the observation
 159 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 $X \sim \mathcal{D}$, $\|X\|_2 \leq X_{\max}$ almost surely.
2. For any $t \in [T], a \in [K]$, $\|\theta_{t,a}\|_2 \leq \Theta_{\max}$.
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 $\Sigma = \mathbb{E}_{X \sim \mathcal{D}}[XX^\top] \succ 0$ 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 $\sum_{t=1}^T \max_{a \in [K]} \|\theta_{t,a} - \theta_a\|_2 \leq C$. 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 π_T^* by

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

where potential randomness of the loss distribution. The
 learners’ objective is to minimize the expected cumulative

165 regret against π_T^* ,

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

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

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

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

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

184 and the minimum sub-optimality gap

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

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

196 We denote $\mathcal{H}_t = \sigma(X_s, A_s, O_s, \{l_s(X_s, o)\}_{o \in O_s}, s \leq t)$
197 the filtration generated by all past contexts, actions, and
198 observed losses. Finally, we use equivalently the notation
199 $a = \mathcal{O}(b)$ or $a \lesssim b$ when there exists a constant $\omega > 0$
200 such that $a \leq \omega b$, where ω is independent of the following
201 problem-dependent quantities: $T, d, K, \Sigma, \mathcal{D}, C, \Delta_{\min}$.

202 3. ALGORITHM

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

211 To obtain loss estimates adapted to the linear contextual
212 setting, we follow the approach of Kuroki et al. (2024), con-
213 structing importance-weighted regression estimates of the
214 losses. For computational efficiency, we employ the *Matrix
215 Geometric Resampling* (MGR) method (Neu & Bartók,
216 2013; Bartók et al., 2014; Kuroki et al., 2024), which guar-
217antees tractability while controlling both the bias and variance
218 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?]

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

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 $s \leq t - 1$,

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

where $\tilde{\theta}_{s,a}$ 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

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

where c, K , and λ_{\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 E.8), in particular at the beginning of the interaction (since this term does not scale up with t). This value is adjusted by the parameter λ_{\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

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

where $\Sigma_{t,a} := \mathbb{E}[\mathbb{1}_{a \in O_t} X_t X_t^\top | \mathcal{H}_{t-1}]$ [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 2 (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

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

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

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

[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

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

#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 λ_{\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

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

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 F.1. 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.

275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329

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

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

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

```

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

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

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

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

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

This result shows that Algorithm 1 achieves the minimax-optimal $\mathcal{O}(T^{2/3})$ 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 $c \rightarrow 0$, 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 $\frac{1}{\Delta_{\min}}$ factor compared to the full-information bounds in Kuroki et al. (2024). But on the contrary, our algorithm has a better $\log T$ dependence, thus our bound is better if T is significantly larger than $\frac{1}{\Delta_{\min}}$. 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 λ_{\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

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

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 λ_{\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 D. 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

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

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

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

Then, in Lemma F.2 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) $R_T \leq \sum_{t=1}^T \mathbb{E} \left[\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) h_t + \frac{z_t \eta_t}{\gamma_t} + \gamma_t \right] + \bar{\beta} \bar{h}$,
- (ii) $\mathbb{E}[h_{t+1} | \mathcal{H}_t] \leq 2\mathbb{E}[h_t | \mathcal{H}_{t-1}]$ 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

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

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) $\sqrt{z_t h_t} \leq \sqrt{\rho}(1 - q_t(\pi_T^*(X_t) | X_t))$, and
- (iv) $u_t h_t \leq \rho(1 - q_t(\pi_T^*(X_t) | X_t))$,

then, for $T \geq \tau := \frac{1}{\Delta_{\min}^3} + \frac{C}{\Delta_{\min}^2}$ it holds that

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

with

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

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 C.

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 E.5 (in Appendix), which is similarly to the proof of Tsuchiya & Ito (2024, Theorem 8).

We prove condition (ii) in Lemma E.8. The proof consists in applying Lemma 15 from Tsuchiya & Ito (2024) (restated as Lemma E.7) 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 $\mathbb{E}[\langle X_t, \hat{\theta}_{t,a} \rangle^2] \leq \frac{1}{\lambda_{\min}^2 p_t}$, which contrasts with the original bound $\mathbb{E}[\ell_t^2] \leq \frac{1}{p_t}$ in the non-contextual case. Since Lemma E.7 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 D, where we carefully establish that each condition of the lemma holds in our setting. \square

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 $p_t = \frac{1}{cK}(\sqrt{z_t \eta_t} + u_t \eta_t)$, 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. (11)) 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

$$\begin{aligned} \sum_{t=1}^T \mathbb{E}\left[\frac{z_t \eta_t}{\gamma'_t} + \gamma_t \right] &\leq \sum_{t=1}^T \mathbb{E}\left[\frac{1}{cK} \sqrt{\frac{z_t}{\beta_t}} + cK \left(\sqrt{\frac{z_t}{\beta_t}} + \frac{u_t}{\beta_t} \right) \right] \\ &\leq \left(\frac{1}{cK} + cK \right) \sum_{t=1}^T \mathbb{E}\left[2 \sqrt{\frac{z_t}{\beta_t}} + \frac{u_t}{\beta_t} \right]. \end{aligned}$$

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 Δ_{\min} is often zero, which implies that the regret bound remains at $\Theta(T^{2/3})$, 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 Δ_{\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.

- 495 Kirschner, J., Lattimore, T., and Krause, A. Information
496 directed sampling for linear partial monitoring. In *Proceedings of the Conference on Learning Theory*, volume
497 125, pp. 2328–2369. PMLR, 2020.
- 498
- 499 Kuroki, Y., Rumi, A., Tsuchiya, T., Vitale, F., and Cesa-
500 Bianchi, N. Best-of-both-worlds algorithms for linear
501 contextual bandits. In Dasgupta, S., Mandt, S., and Li, Y.
502 (eds.), *Proceedings of the 27th International Conference
503 on Artificial Intelligence and Statistics*, volume 238 of
504 *Proceedings of Machine Learning Research*, pp. 1216–
505 1224. PMLR, 2024.
- 506
- 507 Langford, J. and Zhang, T. The epoch-greedy algorithm for
508 multi-armed bandits with side information. In *Advances
509 in Neural Information Processing Systems 20*, pp. 817–
510 824, 2007.
- 511
- 512 Lattimore, T. and Szepesvári, C. The end of optimism?
513 an asymptotic analysis of finite-armed linear bandits. In
514 *Proceedings of the 20th International Conference on Arti-
515 ficial Intelligence and Statistics*, volume 54, pp. 728–737.
516 PMLR, 2017.
- 517 Lattimore, T. and Szepesvári, C. *Bandit Algorithms*. Cam-
518 bridge University Press, 2020.
- 519
- 520 Li, L., Chu, W., Langford, J., and Schapire, R. E. A
521 contextual-bandit approach to personalized news arti-
522 cle recommendation. In *Proceedings of the 19th Interna-
523 tional Conference on World Wide Web*, pp. 661–670.
524 ACM, 2010. doi: 10.1145/1772690.1772758.
- 525
- 526 Mannor, S. and Shamir, O. From bandits to experts: On
527 the value of side-observations. In *Advances in Neural
528 Information Processing Systems 24*, pp. 684–692, 2011.
- 529
- 530 Neu, G. Explore no more: Improved high-probability
531 regret bounds for non-stochastic bandits. *arXiv preprint
532 arXiv:1506.03271*, 2015.
- 533
- 534 Neu, G. and Bartók, G. An efficient algorithm for
535 learning with semi-bandit feedback. *arXiv preprint
536 arXiv:1305.2732*, 2013.
- 537
- 538 Neu, G. and Olkhovskaya, J. Efficient and robust algorithms
539 for adversarial linear contextual bandits. In *Proceedings
540 of the Conference on Learning Theory*, volume 125, pp.
3049–3068. PMLR, 2020.
- 541
- 542 Olkhovskaya, J., Mayo, J. J., van Erven, T., Neu, G., and
543 Wei, C.-Y. First- and second-order bounds for adver-
544 sarial linear contextual bandits. In *Advances in Neural
545 Information Processing Systems 36*, 2023.
- 546
- 547 Rouyer, C., van der Hoeven, D., Cesa-Bianchi, N., and
548 Seldin, Y. A near-optimal best-of-both-worlds algorithm
549 for online learning with feedback graphs. In *Advances in
550 Neural Information Processing Systems 35*, 2022.
- 551
- 552 Saha, A. and Gaillard, P. Versatile dueling bandits: Best-of-
553 both world analyses for learning from relative preferences.
554 In *Proceedings of the International Conference on Ma-
555 chine Learning*, pp. 19011–19026. PMLR, 2022.
- 556
- 557 Seldin, Y. and Slivkins, A. One practical algorithm for both
558 stochastic and adversarial bandits. In *Proceedings of the
559 31st International Conference on Machine Learning*, pp.
1287–1295. JMLR.org, 2014.
- 560
- 561 Seldin, Y., Bartlett, P., Crammer, K., and Abbasi-Yadkori,
562 Y. Prediction with limited advice and multiarmed bandits
563 with paid observations. In Xing, E. and Jebara, T. (eds.),
564 *Proceedings of the 31st International Conference on Ma-
565 chine Learning*, volume 32 of *Proceedings of Machine
566 Learning Research*, pp. 280–287, Beijing, China, 2014.
567 PMLR.
- 568
- 569 Shalev-Shwartz, S. Online learning and online convex opti-
570 mization. *Foundations and Trends in Machine Learning*,
4(2):107–194, 2012.
- 571
- 572 Tsuchiya, T. and Ito, S. A simple and adaptive learning rate
573 for ftrl in online learning with minimax regret of
 $\theta(t^{2/3})$ and its application to best-of-both-worlds.
574 *NeurIPS*, 2024.
- 575
- 576 Tsuchiya, T., Ito, S., and Honda, J. Further adaptive best-of-
577 both-worlds algorithm for combinatorial semi-bandits. In
578 *Proceedings of the International Conference on Artificial
579 Intelligence and Statistics*, pp. 8117–8144. PMLR, 2023.
- 580
- 581 Yun, D., Proutière, A., Ahn, S., Shin, J., and Yi, Y. Multi-
582 armed bandit with additional observations. *Proceedings
583 of ACM Measurement and Analysis of Computing Sys-
584 tems*, 2(1):13:1–13:22, 2018. doi: 10.1145/3179416.
- 585
- 586 Zimmert, J. and Marinov, T. V. Productive bandits: Im-
587 portance weighting no more. In *Advances in Neural
588 Information Processing Systems 37*, pp. 85360–85388,
589 2024.
- 590
- 591 Zimmert, J. and Seldin, Y. Tsallis-inf: An optimal algorithm
592 for stochastic and adversarial bandits, 2022.

A. To-Do Dorian

- We have to cleanly re-write the proof with the ghost sample technique, which is probably going to simplify some things (e.g., Lemma 6, maybe the property is only needed for the ghost context). To do that, work directly from Lemma 3 of Neu & Olkhovskaya (2020). We will probably be able to get the dimension from the standard trace trick.
- In Eq. (12), if we want to use Lemma 2 we have to upper bounds the probabilities by 1 and get K instead of $\sum q_a^{2-\alpha}$. Otherwise, we need a bound on $\mathbb{E}[q_{t,a}^{2-\alpha} \hat{\ell}_{t,a}^2]$ directly.
- In the MGR concentration and the proof of Lemma 2 we wrongly do as if the observation probability was uniform across all arms, the only way to make it right seems to use $\min_{x \in \mathcal{X}} p_t(x)$. Several errors of this type can be corrected by introducing a forced exploration.

In addition, below are some things that I don't understand and some remarks

- In those papers they also use what is our Lemma 2 on the MGR-induced losses directly, not on the “unbiased estimates losses”.
- I really want to understand what “should” be the right dependency in d , if not from a formal lower bound at least from an intuitive point of view.
- Kuroki et al. have a worst stochastic bound in $(\log(T))^3 + d(\log(T))^2$. I'm wondering why and if we should have the same or if Taira's framework prevents us from that.

B. Temp Antoine: Regret analysis

Recall the definition of the regret

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

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 $\mathbb{E}[|O_t| | \mathcal{H}_{t-1}, X_t] = K p_t(X_t)$ 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 $\hat{\theta}_{t,a} = \Sigma_{t,a}^{-1} X_t \ell_t(X_t, a) \mathbf{1}_{\{a \in O_t\}}$ and note that $\mathbb{E}[\hat{\theta}_{t,a} | \mathcal{H}_{t-1}] =$

605 $\theta_{t,a}$. By Lemma E.1, we can further rewrite the regret as

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

622 where $\tilde{\theta}_{t,a}$ are MGR estimate defined as $\tilde{\theta}_{t,a} = \Sigma_t^+ X_t \ell_t(X_t, a) \mathbb{1}_{\{a \in O_t\}}$ (see Eq. 5). For any context $x \in \mathcal{X}$, we define
 623 the auxiliary regret
 624

$$\tilde{R}_T(x) := \sum_{t=1}^T \sum_{a=1}^K (q_t(a | x) - \pi_T^*(a | x)) \langle x, \tilde{\theta}_{t,a} \rangle + cK \sum_{t=1}^T p_t(x)$$

625 and the bias induced by MGR
 626

$$\text{bias}_{\text{MGR}} := \sum_{t=1}^T \max_{a \in [K]} \left| \mathbb{E} \left[\langle X_0, \hat{\theta}_{t,a} - \tilde{\theta}_{t,a} \rangle \right] \right|.$$

633 With this notation, we can write the inequality above as
 634

$$R_T \leq \mathbb{E}_{X_0 \sim \mathcal{D}} [\tilde{R}_T(X_0)] + 2 \text{bias}_{\text{MGR}}.$$

635 We bound the two terms separately. In Lemma F.2 we prove that the bias induced by MGR is upper bounded as
 636

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

637 Thus, it remains to upper bound $\mathbb{E}_{X_0 \sim \mathcal{D}} [\tilde{R}_T(X_0)]$. Let us fix a context $x \in \mathcal{X}$ and recall that at time t , the distribution
 638 $q_t(\cdot | x)$ is
 639

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

640 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,
 641 Exercise 28.12), we have
 642

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

C. PROOF OF LEMMA 4.2

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

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

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

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

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

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

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

$$(iv) \quad u_t h_t \leq \rho(1 - q_t(\pi_T^*(X_t) | X_t)),$$

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

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

with

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

Proof of Lemma 4.2. The argument follows the same general structure of the proof of Theorem 7 in (Tsuchiya & Ito, 2024). We first define

$$\gamma'_t := \gamma_t - \frac{u_t}{\beta_t} = \sqrt{\frac{z_t}{\beta_t}}.$$

Starting from the regret decomposition given by Assumption (i) of the lemma, we have:

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

(by Assumption (i) of the lemma)

$$= \mathbb{E} \left[\sum_{t=1}^T \left((\beta_t - \beta_{t-1}) h_t + \frac{z_t \eta_t}{\gamma_t} + \gamma_t \right) \right] + \bar{\beta} \bar{h}$$

where we used $\beta_t = 1/\eta_t$ and $\bar{h} = \max_{p \in \Delta_K} H_{\bar{\alpha}}(p)$. We now replace γ_t by $\gamma'_t \leq \gamma_t$ to simplify the analysis (this may loosen the bound slightly but keeps the algebra tractable):

$$\begin{aligned} R_T &\leq \mathbb{E} \left[\sum_{t=1}^T \left((\beta_t - \beta_{t-1}) h_t + \frac{z_t \eta_t}{\gamma'_t} + \gamma_t \right) \right] + \bar{\beta} \bar{h} \\ &= \sum_{t=1}^T \mathbb{E} \left[(\beta_t - \beta_{t-1}) h_t + \frac{z_t \eta_t}{\gamma'_t} + \gamma_t \right] + \bar{\beta} \bar{h} \end{aligned}$$

Bounding the first term. We first upper bound $\sum_{t=1}^T \mathbb{E}[(\beta_t - \beta_{t-1}) h_t]$. By the tower rule and the fact that $(\beta_t - \beta_{t-1})$ is \mathcal{H}_{t-1} -measurable, we have

$$\sum_{t=1}^T \mathbb{E}[(\beta_t - \beta_{t-1}) h_t] = \sum_{t=1}^T \mathbb{E}[(\beta_t - \beta_{t-1}) \mathbb{E}[h_t | \mathcal{H}_{t-1}]] .$$

Then, by Assumption (ii), which ensures $\mathbb{E}[h_t | \mathcal{H}_{t-1}] \leq 2 \mathbb{E}[h_{t-1} | \mathcal{H}_{t-2}]$, we get:

$$\sum_{t=1}^T \mathbb{E}[(\beta_t - \beta_{t-1}) h_t] \leq 2 \sum_{t=1}^T \mathbb{E}[(\beta_t - \beta_{t-1}) h_{t-1}] .$$

Bounding the remaining terms. Using the definitions of γ_t and γ'_t , namely $\gamma'_t = \sqrt{z_t / \beta_t}$ and $\gamma_t = \gamma'_t + u_t / \beta_t$, we have

$$\begin{aligned} \sum_{t=1}^T \mathbb{E} \left[\frac{z_t \eta_t}{\gamma'_t} + \gamma_t \right] &= \sum_{t=1}^T \mathbb{E} \left[\sqrt{\frac{z_t}{\beta_t}} + \left(\sqrt{\frac{z_t}{\beta_t}} + \frac{u_t}{\beta_t} \right) \right] \\ &\leq \sum_{t=1}^T \mathbb{E} \left[2 \sqrt{\frac{z_t}{\beta_t}} + \frac{u_t}{\beta_t} \right] . \end{aligned} \tag{11}$$

Combining the two results above, we obtain:

$$R_T \leq \mathbb{E}[F(\beta_{1:T}, z_{1:T}, u_{1:T}, h_{0:T-1})] + \mathbb{E}[\bar{\beta} \bar{h}],$$

where we define

$$F(\beta_{1:T}, z_{1:T}, u_{1:T}, h_{1:T}) := \sum_{t=1}^T \left((\beta_t - \beta_{t-1}) h_t + 2 \sqrt{\frac{z_t}{\beta_t}} + \frac{u_t}{\beta_t} \right) .$$

Note that the regret upper bound we obtained at this step involves the sequence $h_{0:T-1}$, and not $h_{1:T}$ as in the above definition.

Adversarial regime Using Lemma E.9, we obtain that, for any $\varepsilon \geq 1/T$,

$$\begin{aligned} F(\beta_{1:T}, z_{1:T}, u_{1:T}, h_{0:T-1}) &\leq \left(\left[\sum_{t=1}^T \sqrt{z_t h_t} \right] \log(\varepsilon T) \right)^{2/3} \\ &\quad + \sqrt{\left[\sum_{t=1}^T u_t h_t \right] \log(\varepsilon T)} \\ &\quad + \left(\frac{\sqrt{z_{\max} h_1}}{\varepsilon} \right)^{2/3} + \sqrt{\frac{u_{\max} h_1}{\varepsilon}} + \kappa. \end{aligned}$$

Substituting this into the previous inequality gives the claimed regret bound for the adversarial case:

$$R_T \leq \left(\mathbb{E} \left[\sum_{t=1}^T \sqrt{z_t h_t} \right] \log(\varepsilon T) \right)^{2/3} + \sqrt{\mathbb{E} \left[\sum_{t=1}^T u_t h_t \right] \log(\varepsilon T)} \\ + \left(\frac{\sqrt{z_{\max} h_1}}{\varepsilon} \right)^{2/3} + \sqrt{\frac{u_{\max} h_1}{\varepsilon}} + \kappa.$$

Setting $\varepsilon = 1/T$ yields the desired bound in the adversarial regime.

Stochastic regime. We now turn to the stochastic case, under Assumptions (iii)–(iv). Define

$$\varrho_0(\pi_T^*) := \sum_{t=1}^T (1 - q_t(\pi_T^*(X_t) | X_t)).$$

By Assumptions (iii)–(iv),

$$\mathbb{E} \left[\sum_{t=1}^T \sqrt{z_t h_t} \right] \leq \sqrt{\rho} \cdot \varrho_0(\pi_T^*), \\ \mathbb{E} \left[\sum_{t=1}^T u_t h_t \right] \leq \rho \cdot \varrho_0(\pi_T^*).$$

Furthermore, Lemma 21 of (Kuroki et al., 2024) gives the lower bound

$$R_T \geq \frac{\Delta_{\min}}{2} \mathbb{E}[\varrho_0(\pi^*)] - 2C.$$

Balancing both bounds using any $\lambda \in (0, 1]$, and applying the inequalities $ax^2 - bx^3 \leq \frac{4a^3}{27b^2}$ and $ax - bx^2 \leq \frac{a^2}{4b}$ (for $a \geq 0$, $b > 0$), we obtain after simplification:

$$R_T \lesssim \frac{(1+\lambda)^3}{\lambda^2} \cdot \frac{\rho \log(\varepsilon T)}{\Delta_{\min}^2} + \frac{(1+\lambda)^2}{\lambda} \cdot \frac{\rho \log(\varepsilon T)}{\Delta_{\min}} \\ + \left(\frac{\sqrt{z_{\max} h_1}}{\varepsilon} \right)^{2/3} + \sqrt{\frac{u_{\max} h_1}{\varepsilon}} + \kappa + 2\lambda C.$$

Choosing $\lambda = \Theta \left(\left(\frac{\rho \log(\varepsilon T)}{C} \right)^{1/3} \right)$ and setting $\varepsilon = 1/(\rho^2/\Delta_{\min}^3 + C\rho/\Delta_{\min}) \leq 1/T$ gives, for $T \geq \tau := \frac{1}{\Delta_{\min}^3} + \frac{C}{\Delta_{\min}}$,

$$R_T \lesssim \frac{\rho}{\Delta_{\min}^2} \log_+ (T \Delta_{\min}^3) + \left(\frac{C^2 \rho}{\Delta_{\min}^2} \log_+ \left(\frac{T \Delta_{\min}}{C} \right) \right)^{1/3} \\ + \left((z_{\max} h_1)^{1/3} + \sqrt{u_{\max} h_1} \right) \left(\frac{1}{\Delta_{\min}^3} + \frac{C}{\Delta_{\min}} \right)^{2/3} + \kappa,$$

which concludes the proof. \square

D. PROOF OF THEOREM 4.1

We build on Lemma 4.2, presented and proved in Appendix C, to prove Theorem 4.1 by verifying that Algorithm 1 satisfies conditions (i)–(iv) of the lemma. We recall the theorem below, before presenting its proof.

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

$$826 \quad R_T \lesssim \left(\frac{cKd^2 \log K}{\lambda_{\min}^2} \right)^{1/3} T^{2/3} \\ 827 \quad + \sqrt{\frac{\max(c, 1)d \log K \cdot T}{\lambda_{\min}}} + \kappa \\ 828 \\ 829 \\ 830 \\ 831$$

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

$$833 \quad \kappa = \sqrt{\frac{cKd^2 \log K}{\lambda_{\min}^2}} + \frac{\max(c, 1)d \log K}{\lambda_{\min}} \\ 834 \quad + \frac{\max(c, 1)K \log K}{\lambda_{\min}^2} + \frac{32Kd\sqrt{c}}{(1-\alpha)^2 \sqrt{\beta_1} \min(1, \lambda_{\min})}. \\ 835 \\ 836 \\ 837 \\ 838 \\ 839$$

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

$$841 \quad R_T \lesssim \frac{d\sqrt{\max(c, 1)K \log K}}{\lambda_{\min} \Delta_{\min}^2} \cdot \log(T \Delta_{\min}^3) \\ 842 \quad + \left(\frac{C^2 d \sqrt{\max(c, 1)K \log K}}{\lambda_{\min} \Delta_{\min}^2} \cdot \log\left(\frac{T \Delta_{\min}}{C}\right) \right)^{1/3} \\ 843 \quad + \kappa + \kappa', \text{ where we further define} \\ 844 \quad \kappa' = \left(\left(\frac{cKd^2 \log K}{\lambda_{\min}^2} \right)^{1/3} + \sqrt{\frac{\max(c, 1)d \log K}{\lambda_{\min}}} \right) \\ 845 \quad \times \left(\frac{1}{\Delta_{\min}^3} + \frac{C}{\Delta_{\min}} \right)^{2/3}. \\ 846 \\ 847 \\ 848 \\ 849 \\ 850 \\ 851 \\ 852 \\ 853 \\ 854$$

855 *Proof.* #DB: Tldr: all this part is shit because it mixes using fixed context and current context, with ambiguous notation that
856 doesn't allow to understand in which case we are.# We verify that Algorithm 1 satisfies the four conditions of Lemma 4.2.
857

858 Throughout this proof, we work with the *exact* loss estimates $\hat{\theta}_{t,a}$ defined in Eq. (3), rather than their MGR approximations
859 $\tilde{\theta}_{t,a}$ used in the algorithmic description. This distinction is only technical and does not affect the regret order, since
860 Lemma F.2 guarantees that the cumulative bias introduced by the MGR approximation remains uniformly bounded.

861 **Condition (i).** By definition of the importance-sampled loss (Eq. (3)), for any $a \in [K]$ we have

$$862 \quad |\hat{\ell}_{t,a} \eta_t| \leq \frac{\ell_{t,a} \eta_t}{p_t \lambda_{\min}} \leq \frac{1}{u_t \lambda_{\min}} \leq \frac{1-\alpha}{8} \cdot \frac{1}{\min(q_{t,a_t^*}, 1-q_{t,a_t^*})^{1-\alpha}}.$$

863 Hence, the scaled losses $\hat{\ell}_t \eta_t$ satisfy the condition of Lemma E.5, presented in Appendix C, which provides an upper bound
864 on the penalty term $\langle q_t - q_{t+1}, \hat{\ell}_t(x) \rangle - D_t(q_{t+1}, q_t)$ appearing in the standard FTRL regret decomposition.
865

866 **#DB:** Imo, has to be re-written from the start with the ghost sample x and with the estimates $\hat{\theta}_{t,a}$ actually used by FTRL.#
867 Since the regret is defined by

$$868 \quad R_T = \mathbb{E} \left[\sum_{t=1}^T (\langle X_t, \theta_{t,A_t} \rangle - \langle X_t, \theta_{t,\pi^*(X_t)} \rangle) + cK \sum_{t=1}^T p_t \right],$$

869 and $\hat{\theta}_{t,a}$ is an unbiased estimator of $\theta_{t,a}$, we can equivalently write
870

$$871 \quad R_T = \mathbb{E} \left[\sum_{t=1}^T (\langle X_t, \hat{\theta}_{t,A_t} \rangle - \langle X_t, \hat{\theta}_{t,\pi^*(X_t)} \rangle) + cK \sum_{t=1}^T p_t \right].$$

#DB: This is wrong without using the ghost sample technique, cause the estimates depend on X_t (with some X_0 , it's OK).#

Fix any context $x \in \mathbb{R}^d$. Applying Lemma E.6, we obtain:

$$\begin{aligned} \sum_{t=1}^T (\langle x, \hat{\theta}_{t,A_t} \rangle - \langle x, \hat{\theta}_{t,\pi^*(x)} \rangle) &\leq \underbrace{\sum_{t=1}^T (\psi_t(q_{t+1}(\cdot|x)) - \psi_{t+1}(q_{t+1}(\cdot|x)))}_{\text{stability}} \\ &+ \underbrace{\sum_{t=1}^T (\langle q_t(\cdot|x) - q_{t+1}(\cdot|x), \hat{\ell}_t(x) \rangle - D_t(q_{t+1}(\cdot|x), q_t(\cdot|x))) + A + \bar{\beta}h}_{\text{penalty}}, \end{aligned}$$

where $A = \psi_{T+1}(\pi^*(\cdot|x)) - \psi_1(q_1(\cdot|x)) \leq \beta_1 \log K$ is independent of T and will be ignored in the sequel (together with $\bar{\beta}h$).

Bounding the stability term. By the definition of ψ_t , we have

$$\sum_{t=1}^T (\psi_t(q_{t+1}(\cdot|x)) - \psi_{t+1}(q_{t+1}(\cdot|x))) \leq \sum_{t=1}^T \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) h_{t+1}.$$

#DB: Now here this is were messing up with the notation is confusing: above check which components should depend on x .# Reindexing $t \mapsto t - 1$ yields the equivalent form

$$\sum_{t=1}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) h_t.$$

Bounding the penalty term. Using Lemma E.5 together with Lemma E.4, we have

$$\begin{aligned} \sum_{t=1}^T (\langle q_t - q_{t+1}, \hat{\ell}_t(x) \rangle - D_t(q_{t+1}, q_t)) &= \sum_{t=1}^T \frac{1}{\eta_t} \left(\langle q_t - q_{t+1}, \hat{\ell}_t(x) \eta_t \rangle - D_t(q_{t+1}, q_t) \right) \\ &\leq \sum_{t=1}^T \frac{4\eta_t}{1-\alpha} \left(q_{t,a_t^*}^{2-\alpha} \hat{\ell}_{t,a_t^*}^2 + \sum_{a \neq a_t^*} q_{t,a}^{2-\alpha} \hat{\ell}_{t,a}^2 \right) \\ &\leq \sum_{t=1}^T \frac{4d^2\eta_t}{p_t(1-\alpha)\lambda_{\min}^2} \left(q_{t,a_t^*}^{2-\alpha} + \sum_{a \neq a_t^*} q_{t,a}^{2-\alpha} \right). \end{aligned} \tag{12}$$

#DB: Isn't there some $\min(q, 1-q)$ at some point in lemma 3? Otherwise why split between a_t^* and $a \neq a_t^*$. Not consistent#

#DB: More important: lemma 2 bounds $\mathbb{E}[\hat{\ell}_{t,a}^2]$ while here you should have a bound on $\mathbb{E}[q_{t,a}^{2-\alpha} \cdot \hat{\ell}_{t,a}^2]$. Transposing previous works we would work on $\mathbb{E}[\sum_{a=1}^K q_{t,a}^{2-\alpha} \cdot \hat{\ell}_{t,a}^2]$ directly.#

Taking expectations over the random context X_t , we obtain

$$\begin{aligned} \mathbb{E}[R_T] &\leq \mathbb{E} \left[\sum_{t=1}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) h_t \right. \\ &\quad \left. + \sum_{t=1}^T \frac{4d^2\eta_t}{p_t(1-\alpha)\lambda_{\min}^2} \left(q_{t,a_t^*}^{2-\alpha} + \sum_{a \neq a_t^*} q_{t,a}^{2-\alpha} \right) + cK \sum_{t=1}^T p_t \right], \end{aligned}$$

which matches the required structure of condition (i).

#DB: It seems we should not be doing that but instead use a ghost sample to define the regret as in (Tsuchiya & Ito, 2024; Kuroki et al., 2024).#

935 Condition (ii). Condition (ii) follows directly from Lemma E.8, presented and proved in Appendix C, which guarantees
 936 that

$$\mathbb{E}[h_{t+1} \mid \mathcal{H}_t] \leq 2\mathbb{E}[h_t \mid \mathcal{H}_{t-1}], \quad \forall t \geq 1.$$

937 938 939 **Conditions (iii) and (iv).** Lemma 13 of (Tsuchiya & Ito, 2024) provides an upper bound on the entropy term,
 940

$$h_t \leq \frac{1}{\alpha}(K-1)^{1-\alpha}(1-q_{t,a_t^*})^\alpha,$$

941 942 943 where $a_t^* := \arg \max_{a \in [K]} \langle X_t, \theta_{t,a} \rangle$ denotes the optimal arm for context X_t . Moreover, using the definitions of z_t and u_t ,
 944 we obtain:

$$\begin{aligned} z_t &= \frac{4cKd^2}{(1-\alpha)\lambda_{\min}^2} \left(\sum_{a \neq a_t^*} q_{t,a}^{2-\alpha} + (\min(q_{t,a_t^*}, 1-q_{t,a_t^*}))^{2-\alpha} \right) \\ &\leq \frac{8cKd^2}{(1-\alpha)\lambda_{\min}^2} (1-q_{t,a_t^*})^{2-\alpha}. \end{aligned}$$

952 Combining the bounds on h_t and z_t yields

$$\begin{aligned} z_t h_t &\leq \frac{8cKd^2(K-1)^{1-\alpha}}{\alpha(1-\alpha)\lambda_{\min}^2} (1-q_{t,a_t^*})^2, \\ u_t h_t &\leq \frac{8d \max(c, 1)}{(1-\alpha)\alpha} (K-1)^{1-\alpha} (1-q_{t,a_t^*}). \end{aligned}$$

953 Hence, both conditions are satisfied with

$$\begin{aligned} \sqrt{z_t h_t} &\leq \sqrt{\rho} (1-q_{t,a_t^*}), \\ u_t h_t &\leq \rho (1-q_{t,a_t^*}), \end{aligned}$$

954 955 956 957 958 where

$$\rho := \frac{d}{\lambda_{\min}} \max \left(\sqrt{\frac{8cK(K-1)^{1-\alpha}}{\alpha(1-\alpha)}}, \frac{8 \max(c, 1)}{(1-\alpha)\alpha} (K-1)^{1-\alpha} \right).$$

959 **Conclusion.** Having verified conditions (i)–(iv), we can invoke Lemma 4.2 to conclude that Algorithm 1 enjoys a
 960 Best-of-Both-Worlds (BoBW) regret guarantee. To make the constants explicit, note that

$$h_{\max} \leq \frac{K^{1-\alpha}}{\alpha}, \quad z_{\max} = \mathcal{O}\left(\frac{cKd^2}{(1-\alpha)\lambda_{\min}^2}\right), \quad u_{\max} = \mathcal{O}\left(\frac{d \max(c, 1)}{1-\alpha}\right).$$

961 Plugging these into Lemma 4.2 gives

$$\text{Adversarial regime: } R_T = \mathcal{O}\left(\left(\frac{cKd^2}{\lambda_{\min}^2}\right)^{1/3} T^{2/3} + \sqrt{\frac{d \max(c, 1) T}{\lambda_{\min}}}\right),$$

$$\text{Corrupted stochastic regime: } R_T = \mathcal{O}\left(\frac{d\sqrt{\max(c, 1)K}}{\lambda_{\min}\Delta_{\min}^2} \log(T\Delta_{\min}^3) + \left(\frac{C^2 d \sqrt{\max(c, 1)K}}{\lambda_{\min}\Delta_{\min}^2} \log \frac{T\Delta_{\min}}{C}\right)^{1/3}\right).$$

962 963 964 965 966 967 968 969 970 971 972 973 974 975 976 977 978 979 980 981 982 983 984 985 986 987 988 989 This completes the proof of Theorem 4.1. \square

E. Technical tools

Lemma E.1 (Neu & Olkhovskaya, 2020, Lemma 3). Let π^* 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 $\mathbb{E}[\widehat{\theta}_{t,a} | \mathcal{H}_{t-1}] = \theta_{t,a}$. Then, it holds that

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

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

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

[Antoine: Some of the lemmas below should stay in this section, but the most important ones should probably move to the appendix “Omitted Proofs”.]

#DB: Below we have a problem, $\lambda_{\min}(\Sigma_{t,a}) \geq p_t \lambda_{\min}$ has no reason to hold and this should be fixed.#

#DB: Additionally, we haven’t sorted the dimension issue yet. What should be the right dependence in d ? It’s not so clear, but it should appear in the squared loss if we follow previous works: there, d come from using a trace bound instead of our inequalities below. Some of our steps are probably wrong.#

#DB: The following result should bound $\mathbb{E}[\sum_a q_{t,a}^{2-\alpha} \widehat{\ell}_{t,a}^2]$ or whatever is needed in the part of the proof where it’s used.#

Lemma E.4. Let $X_t \in \mathbb{R}^d$ be a random context and fix any arm $a \in [K]$.

Under the assumptions of Section 2, we have $\|X_t\|_2 \leq 1$ almost surely, and the loss function satisfies $-1 \leq \ell_t(X_t, a) \leq 1$.

We also recall that $\Sigma_{t,a}$ is a positive definite matrix such that

$$\lambda_{\min}(\Sigma_{t,a}) \geq p_t \lambda_{\min},$$

#DB: We should probably adopt a notation $p_t(x)$, a lot of confusion comes from using p_t for many different things. The above line is clearly wrong (if p_t is $p_t(X_t)$). It is true with $p_{t,\min} := \min_{x \in \mathcal{X}} p_t(x)$ though, and could be replaced by $\frac{\gamma}{K}$ if we add a forced exploration γ .# and that the importance-weighted estimator is given by

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

where $\mathbb{P}(a \in O_t) = p_t$. Then,

$$\mathbb{E}[\langle X_t, \widehat{\theta}_{t,a} \rangle^2] \leq \frac{1}{\lambda_{\min}^2(\Sigma_{t,a})}.$$

Proof. We know that the smallest eigenvalue of $\Sigma_{t,a}$ is $\geq p_t \lambda_{\min}$.

$$\|\Sigma_{t,a}^{-1}\|_2 \leq \frac{1}{\lambda_{\min}(\Sigma_{t,a})} \leq \frac{1}{p_t \lambda_{\min}}.$$

Therefore,

$$\|\widehat{\theta}_{t,a}\|_2 = \|\Sigma_{t,a}^{-1} X_t \ell_t(X_t, a) \mathbf{1}\{a \in O_t\}\|_2 \leq \frac{\|X_t\|_2}{p_t \lambda_{\min}} \mathbf{1}\{a \in O_t\}.$$

By the Cauchy–Schwarz inequality,

$$\langle X_t, \widehat{\theta}_{t,a} \rangle^2 \leq \|X_t\|_2^2 \|\widehat{\theta}_{t,a}\|_2^2 \leq \frac{\|X_t\|_2^4}{p_t^2 \lambda_{\min}^2} \mathbf{1}\{a \in O_t\}.$$

Taking expectations and using $\mathbb{E}[\mathbf{1}\{a \in O_t\}] = p_t$, we obtain

$$\mathbb{E}[\langle X_t, \widehat{\theta}_{t,a} \rangle^2] \leq \frac{\mathbb{E}[\|X_t\|_2^4]}{p_t \lambda_{\min}^2}.$$

1045 Since $\|X_t\|_2 \leq \sqrt{d}$ almost surely, it follows that $\mathbb{E}[\|X_t\|_2^4] \leq 1$, and hence
 1046
 1047
 1048

$$\mathbb{E}[\langle X_t, \hat{\theta}_{t,a} \rangle^2] \leq \frac{1}{\lambda_{\min}^2 p_t}.$$

1049
 1050
 1051
 1052
 1053

#DB: (minor for now) Write a few sentences between lemmas. Also, here we just do an unstructured list, but actually the above lemma is crucial, while the following is just a restatement of an existing result. Might be nice to structure this. #

1054 **Lemma E.5** (Lemma 14 in (Tsuchiya & Ito, 2024)). Let $q \in \mathcal{P}_K$ and let $\bar{I} \in \arg \max_{i \in [K]} q_i$. Let $l \in \mathbb{R}^K$ be such that, for
 1055 all $i \in [K]$,

$$|l_i| \leq \frac{1-\alpha}{4} \cdot \frac{1}{\min(q_{\bar{I}}, 1-q_{\bar{I}})^{1-\alpha}}.$$

1056 Then, the following bound holds:
 1057
 1058

$$\max_{p \in \mathcal{P}_K} \{ \langle l, q - p \rangle - D_{-H_\alpha}(p, q) \} \leq \frac{4}{1-\alpha} \left(\sum_{i \neq \bar{I}} q_i^{2-\alpha} l_i^2 + \min(q_{\bar{I}}, 1-q_{\bar{I}})^{2-\alpha} l_{\bar{I}}^2 \right)$$

1060
 1061
 1062
 1063
 1064 **Lemma E.6.** Let $x \in \mathbb{R}^d$ be any fixed context. For each $t \geq 1$, let $q_t(\cdot|x) \in \Delta_K$ be the distribution used to sample A_t given
 1065 x , and let $\pi^*(\cdot|x) \in \Delta_K$ be any comparator policy. Let $(\psi_t)_{t \geq 1}$ be a sequence of σ -strongly convex regularizers on Δ_K ,
 1066 and let $D_t(\cdot, \cdot)$ denote the Bregman divergence induced by ψ_t . Denoting $\hat{\ell}_t(x) \in \mathbb{R}^K$ for the vector of estimated losses at
 1067 context x , with $[\hat{\ell}_t(x)]_a := \langle x, \hat{\theta}_{t,a} \rangle$. Then
 1068

$$\begin{aligned} \mathbb{E} \left[\sum_{t=1}^T \left(\langle x, \hat{\theta}_{t,A_t} \rangle - \langle x, \hat{\theta}_{t,\pi^*(x)} \rangle \right) \right] &\leq \mathbb{E} \left[\sum_{t=1}^T (\psi_t(q_{t+1}(\cdot|x)) - \psi_{t+1}(q_{t+1}(\cdot|x))) \right] \\ &\quad + \mathbb{E} \left[\sum_{t=1}^T \left(\langle q_t(\cdot|x) - q_{t+1}(\cdot|x), \hat{\ell}_t(x) \rangle - D_t(q_{t+1}(\cdot|x), q_t(\cdot|x)) \right) \right] \\ &\quad + \mathbb{E} [\psi_{T+1}(\pi^*(\cdot|x)) - \psi_1(q_1(\cdot|x))] + \bar{\beta} \bar{h}, \end{aligned}$$

1069 where $\bar{h} := \max_{p \in \Delta_K} H_{\bar{\alpha}}(p)$ and $\bar{\beta} \geq 0$ is the coefficient that upper-bounds the change of regularizer in our setting.
 1070
 1071

#DB: This lemma above uses a fixed context, so is suited for a proof that uses the ghost sample technique. It cannot work
 1072 for the proof as it is right now.#

#DB: Has to be modified if forced exploration is introduced. See (44) in (Kuroki et al., 2024) for their result.#

1073 *Proof.* Conditionally on x , $A_t \sim q_t(\cdot|x)$, hence
 1074

$$\mathbb{E} \left[\langle x, \hat{\theta}_{t,A_t} \rangle \mid x \right] = \sum_{a \in [K]} q_t(a|x) \langle x, \hat{\theta}_{t,a} \rangle.$$

1075 Therefore,
 1076

$$\begin{aligned} \mathbb{E} \left[\sum_{t=1}^T \left(\langle x, \hat{\theta}_{t,A_t} \rangle - \langle x, \hat{\theta}_{t,\pi^*(x)} \rangle \right) \right] &= \mathbb{E} \left[\sum_{t=1}^T \sum_{a \in [K]} (q_t(a|x) - \pi^*(a|x)) \langle x, \hat{\theta}_{t,a} \rangle \right] \\ &= \mathbb{E} \left[\sum_{t=1}^T \langle q_t(\cdot|x) - \pi^*(\cdot|x), \hat{\ell}_t(x) \rangle \right]. \end{aligned}$$

1077 We now invoke the standard FTRL regret decomposition with time-varying regularizers (see, e.g. Exercise 28.12 in (Lattimore
 1078 & Szepesvári, 2020)): #DB: I think this works only with the estimate that is actually used in FTRL. So, for us, $\tilde{\theta}_{t,a}$ # for any
 1079

1100 $q \in \Delta_K$,

$$\begin{aligned} 1102 \quad & \sum_{t=1}^T \langle q_t(\cdot|x) - q, \hat{\ell}_t(x) \rangle \leq \psi_{T+1}(q) - \psi_1(q_1) + \sum_{t=1}^T (\psi_t(q_{t+1}(\cdot|x))) - \psi_{t+1}(q_{t+1}(\cdot|x)) \\ 1103 \quad & + \sum_{t=1}^T (\langle q_t(\cdot|x) - q_{t+1}(\cdot|x), \hat{\ell}_t(x) \rangle - D_t(q_{t+1}(\cdot|x), q_t(\cdot|x))). \end{aligned}$$

1108 Choosing $q = \pi^*(\cdot|x)$ and taking expectations yields the claim, with the additional additive term $\bar{\beta} \bar{h}$ accounting for the
 1109 regularizer variation bound used in our setup. \square

1112 **Lemma E.7** (Lemma 15 of (Tsuchiya & Ito, 2024)). Let $l, L \in \mathbb{R}^K$, let $q, r \in \mathcal{P}_k$ be:

$$1114 \quad q \in \arg \min_{p \in \mathcal{P}_k} \{ \langle L, p \rangle + \beta(-H_\alpha(p)) + \bar{\beta}(-H_{\bar{\alpha}}(p)) \}$$

$$1118 \quad r \in \arg \min_{p \in \mathcal{P}_k} \{ \langle L + l, p \rangle + \beta'(-H_\alpha(p)) + \bar{\beta}(-H_{\bar{\alpha}}(p)) \}$$

1120 for the Tsallis entropy H_α and $0 < \beta < \beta'$. Suppose also that

$$1123 \quad \|l\|_\infty \leq \max\left(\frac{1 - (\sqrt{2})^{\alpha-1}}{2} q_*^{\alpha-1} \beta, \frac{1 - (\sqrt{2})^{\bar{\alpha}-1}}{2} q_*^{\bar{\alpha}-1} \bar{\beta}\right)$$

$$1127 \quad 0 \leq \beta' - \beta \leq \max((1 - (\sqrt{2})^{\alpha-1})\beta, \frac{1 - (\sqrt{2})^{\bar{\alpha}-1}}{\sqrt{2}} q_*^{\bar{\alpha}-\alpha} \bar{\beta})$$

1130 Then it holds that $H_\alpha(r) \leq 2H_\alpha(q)$.

1132 **#DB:** Check carefully this thing below, cause there might be again the issue of treating some probabilities as fixed while
 1133 they're not. Note that the ghost sample technique might require that we only have to control the entropy of a fixed context
 1134 X_0 , which might simplify things.#

1136 **Lemma E.8.** Algorithm 1 satisfies, for all $t \geq 1$,

$$1137 \quad \mathbb{E}[h_{t+1} | \mathcal{H}_t] \leq 2\mathbb{E}[h_t | \mathcal{H}_{t-1}].$$

1140 *Proof.* We first control the key quantities appearing in Lemma E.7. Recall that $\beta_t = 1/\eta_t$, $\gamma_t = \sqrt{z_t/\beta_t} + u_t/\beta_t$, and
 1141 $h_t = \frac{1}{\alpha} \sum_{i=1}^K (q_{t,i}^\alpha - q_{t,i})$.

1144 **Step 1: Bounding $\sqrt{z_t}$ and h_t .** By definition of z_t we have

$$1146 \quad \sqrt{z_t} = \sqrt{\frac{4cKd^2}{1-\alpha} \left(\sum_{i \neq I_t} q_{t,i}^{2-\alpha} + q_{t,a_t^*}^{2-\alpha} \right)} \leq \frac{2d\sqrt{Kc}}{\sqrt{1-\alpha}} q_{t,a_t^*}^{1-\frac{\alpha}{2}}.$$

1150 In addition, from the properties of the Tsallis entropy (see, e.g., Lemma 13 of (Tsuchiya & Ito, 2024)),

$$1152 \quad h_t = \frac{1}{\alpha} \sum_{i=1}^K (q_{t,i}^\alpha - q_{t,i}) \geq \frac{1 - (1/2)^{1-\alpha}}{\alpha} q_{t,a_t^*}^\alpha \geq \frac{1 - \alpha}{4\alpha} q_{t,a_t^*}^\alpha.$$

1155 **Step 2: Bounding the variation of β_t .** From Equation 9,

$$1157 \quad \beta_{t+1} - \beta_t = \frac{2}{h_t} \sqrt{\frac{z_t}{\beta_t}} + \frac{u_t}{h_t \beta_t}.$$

1160 Plugging in the bounds on $\sqrt{z_t}$ and h_t gives

$$\begin{aligned} 1162 \quad \beta_{t+1} - \beta_t &\leq \frac{16\alpha d \sqrt{Kc}}{\sqrt{\beta_t}(1-\alpha)^{3/2}} q_{t,a_t^*}^{1-\frac{3\alpha}{2}} + \frac{32\alpha d \max(c,1)}{\sqrt{\beta_t}(1-\alpha)^2 \lambda_{\min}} q_{t,a_t^*}^{1-2\alpha} \\ 1164 \quad &\leq \alpha \bar{\beta} q_{t,a_t^*}^{1-\frac{3\alpha}{2}} + \frac{\alpha \bar{\beta}}{\lambda_{\min}} q_{t,a_t^*}^{1-2\alpha} \\ 1167 \quad &\leq 2 \frac{(1-\bar{\alpha})}{\min(1, \lambda_{\min})} \bar{\beta} q_{t,a_t^*}^{\bar{\alpha}-\alpha} \leq 2 \frac{1-(\sqrt{2})^{\bar{\alpha}-1}}{\sqrt{2}} \bar{\beta} q_{t,a_t^*}^{\bar{\alpha}-\alpha}. \end{aligned}$$

1169 Hence, $\beta_{t+1} - \beta_t$ satisfies the second condition of Lemma E.7.

1171 **Step 3: Bounding the loss magnitude.** For any fixed context x and arm $i \in [K]$,

$$\begin{aligned} 1174 \quad |\hat{\ell}_{t+1,i}(x)| &= |\langle x, \hat{\theta}_{t+1,i} \rangle| \leq \frac{d}{\lambda_{\min} p_t} \leq \frac{d}{\lambda_{\min}} \cdot \frac{\beta_t}{u_t} \\ 1175 \quad &= \frac{1-\alpha}{8} \cdot \frac{\beta_t}{q_{t,a_t^*}^{1-\alpha}} \leq \frac{1-(\sqrt{2})^{\alpha-1}}{2} \cdot \frac{\beta_t}{q_{t,a_t^*}^{1-\alpha}}. \end{aligned}$$

1178 This matches the first smoothness condition of Lemma E.7.

1181 **Step 4: Applying Lemma E.7.** Since both smoothness conditions are satisfied, the lemma implies

$$1183 \quad H_\alpha(q_{t+1}) \leq 2 H_\alpha(q_t),$$

1185 and therefore $h_{t+1} \leq 2h_t$ whenever the context remains fixed.

1187 Taking conditional expectations and using the stationarity of the context distribution then yields

$$1188 \quad \mathbb{E}[h_{t+1} | \mathcal{H}_t] \leq 2 \mathbb{E}[h_t | \mathcal{H}_{t-1}],$$

1191 which completes the proof. \square

1192 **Lemma E.9** (Slight adaptation of Theorem 6 of (Tsuchiya & Ito, 2024)). *For all $\varepsilon \geq 1/T$, it holds that*

$$\begin{aligned} 1195 \quad &F(\beta_{1:T}, z_{1:T}, u_{1:T}, h_{0:T-1}) \\ 1197 \quad &\lesssim \min \left\{ \left(\sum_{t=1}^T \sqrt{z_t h_t \log(\varepsilon T)} \right)^{2/3} \right. \\ 1199 \quad &\quad \left. , \left(\frac{\sqrt{z_{\max} h_{\max}}}{\varepsilon} \right)^{2/3}, \left(\sum_{t=1}^T \sqrt{z_t h_{\max}} \right)^{2/3} \right\} \\ 1201 \quad &+ \min \left\{ \sqrt{\sum_{t=1}^T u_t h_t \log(\varepsilon T)}, \frac{\sqrt{u_{\max} h_{\max}}}{\varepsilon}, \sum_{t=1}^T u_t h_{\max} \right\} \\ 1203 \quad &+ \sqrt{\frac{z_{\max}}{\beta_1}} + \frac{u_{\max}}{\beta_1} + \beta_1 h_1 \end{aligned}$$

1210 *Proof.* This slight adaptation originates from a minor modification of Lemma 4 in (Tsuchiya & Ito, 2024), where in the first
 1211 line of equation (24) we instead bound:

$$1213 \quad F(\beta_{1:T}, z_{1:T}, u_{1:T}, h_{0:T-1}) \leq 2\sqrt{\frac{z_1}{\beta_1}} + \frac{u_1}{\beta_1} + \beta_1 h_1 \\ 1214 \\ 1215 \quad + \sum_{t=2}^T \left(2\sqrt{\frac{z_t}{\beta_t}} + \frac{u_t}{\beta_t} + (\beta_t - \beta_{t-1})h_{t-1} \right).$$

1218 After this adjustment, the remainder of the proof proceeds identically.

1220 □

1221
 1222
 1223
 1224
 1225
 1226
 1227
 1228
 1229
 1230
 1231
 1232
 1233
 1234
 1235
 1236
 1237
 1238
 1239
 1240
 1241
 1242
 1243
 1244
 1245
 1246
 1247
 1248
 1249
 1250
 1251
 1252
 1253
 1254
 1255
 1256
 1257
 1258
 1259
 1260
 1261
 1262
 1263
 1264

F. MATRIX GEOMETRIC RESAMPLING

Before detailing Algorithm 2, we elaborate on why using the parameter estimates from Eq. (3) [Antoine: probably meant Equation (4)?] would be untractable in practice. To prove this point, we detail the computation of the exact covariance matrix $\Sigma_{t,a}$, which involves evaluating the following conditional expectation

$$\begin{aligned}\Sigma_{t,a} &= \mathbb{E}_t [\mathbb{1}_{\{a \in O_t\}} X_t X_t^\top] \\ &= \sum_{x \in \mathcal{X}} \mathbb{P}[X_t = x, a \in O_t | \mathcal{H}_{t-1}] x x^\top \\ &= \sum_{x \in \mathcal{X}} \mathbb{P}[X_t = x | \mathcal{H}_{t-1}] \underbrace{\mathbb{P}[a \in O_t | X_t = x, \mathcal{H}_{t-1}]}_{=p_t(x)} x x^\top.\end{aligned}$$

The challenge lies in evaluating the conditional observation probability $p_t(x)$. Note that in Algorithm 1, p_t was defined unambiguously since it was the observation probability corresponding to the (unique) fixed context X_t , computed after it is revealed. Here, $p_t(x)$ is derived following the same steps, but computed as if context x was observed instead of X_t . Doing so requires performing all computations leading to Eq. (8) separately for each possible context $x \in \mathcal{X}$. This results in a computational complexity proportional to the size of the context space, $|\mathcal{X}|$, which becomes quickly prohibitive when \mathcal{X} is large. In addition, we can note that each individual computation requires solving an optimization problem to obtain the FTRL sampling probability (Eq. (2)).

To circumvent this limitation, we follow Neu & Bartók (2013); Neu (2015); Kuroki et al. (2024) and use Matrix Geometric Resampling (MGR) to efficiently approximate the *inverse* of the matrix $\Sigma_{t,a}$ directly. It does not need to compute the FTRL sampling allocation over all possible contexts but only on a carefully chosen number of sampled contexts, and only use matrix products (costing $\mathcal{O}(d^2)$) but no matrix inversion (costing $\mathcal{O}(d^3)$). We recall this procedure in Algorithm 2 below. In the pseudo-code, we denote by $\mathcal{B}(p)$ the Bernoulli distribution with parameter p .

Algorithm 2 Matrix Geometric Resampling (MGR).

```

1: Input: Sampler of the context distribution  $\mathcal{D}$ , number of iterations  $M_t$ .
2: Initialize  $\Sigma_t^+ \leftarrow \frac{1}{2}I$ ,  $A_0 \leftarrow I$ .
3: for  $i = 1$  to  $M_t$  do
4:   Sample  $X \sim \mathcal{D}$ .
5:   Compute probability of observation  $p$  as in Step 5 of Algorithm 1 if  $X_t$  was equal to  $X$ .
6:   Sample  $b \sim \mathcal{B}(p)$ .
7:   Compute  $B_i \leftarrow b X X^\top$ .
8:   Compute  $A_i \leftarrow A_{i-1} \left( I - \frac{1}{2} B_i \right)$ .
9:   Update  $\Sigma_t^+ \leftarrow \Sigma_t^+ + \frac{1}{2} A_i$ .
10: end for
11: Return  $\Sigma_t^+$ .
```

We now introduce the technical results related to the cost and approximation guarantees of the MGR procedure, which will be used in the regret analysis (see the proof sketch in Section 4).

Lemma F.1 (Adapted from Lemma 9 of Kuroki et al., 2024). Denote $\widehat{\theta}_{t,a} = \Sigma_{t,a}^{-1} X_t \ell_t(X_t, a) \mathbb{1}_{\{a \in O_t\}}$ and $\widetilde{\theta}_{t,a} = \Sigma_{t,a}^+ X_t \ell_t(X_t, a) \mathbb{1}_{\{a \in O_t\}}$, where $\Sigma_{t,a}^+$ is obtained via Algorithm 2 with the number of iterations M_t tuned as in Eq. (6). Then, for any arm $a \in [K]$ and round $t \geq 1$, it holds that

$$\left| \mathbb{E} \left[\langle X_t, \widetilde{\theta}_{t,a} - \widehat{\theta}_{t,a} \rangle \mid \mathcal{H}_{t-1} \right] \right| \leq \exp \left(- \frac{p_{t,\min} \lambda_{\min}}{2K} M_t \right).$$

#DB: $p_{t,\min}$ can be replaced using forced exploration.#

Proof. [Antoine: Expectations should be conditional, and the estimators are incorrect (should be $\mathbb{1}_{\{a \in O_t\}}$ instead)] Let $\|\cdot\|_{\text{op}}$ denote the operator norm. Denote by $\widehat{\Sigma}_{t,a}^+$ the random matrix output by the MGR procedure in Algorithm 2. Under

independence assumptions of the geometric resampling steps, we have

$$\mathbb{E} \left[\prod_{j=1}^i \left(I - \frac{1}{2} B_j \right) \right] = \left(I - \frac{1}{2} \Sigma_{t,a} \right)^i,$$

and consequently,

$$\mathbb{E} \left[\widehat{\Sigma}_{t,a}^+ \right] = \frac{1}{2} \sum_{i=0}^{M_t} \left(I - \frac{1}{2} \Sigma_{t,a} \right)^i = \Sigma_{t,a}^{-1} - \left(I - \frac{1}{2} \Sigma_{t,a} \right)^{M_t} \Sigma_{t,a}^{-1}.$$

Using this, we compute the expectation of the biased estimator

$$\begin{aligned} \mathbb{E} \left[\widetilde{\theta}_{t,a} \right] &= \mathbb{E} \left[\widehat{\Sigma}_{t,a}^+ X_t \ell_t(X_t, a) \mathbb{1}_{\{A_t=a\}} \right] \\ &= \mathbb{E} \left[\widehat{\Sigma}_{t,a}^+ \right] \cdot \mathbb{E} \left[X_t \langle X_t, \theta_{t,a} \rangle \mathbb{1}_{\{A_t=a\}} \right] \\ &= \mathbb{E} \left[\widehat{\Sigma}_{t,a}^+ \right] \cdot \mathbb{E} \left[X_t X_t^\top \mathbb{1}_{\{A_t=a\}} \right] \cdot \theta_{t,a} \\ &= \left(\Sigma_{t,a}^{-1} - \left(I - \frac{1}{2} \Sigma_{t,a} \right)^{M_t} \Sigma_{t,a}^{-1} \right) \cdot \Sigma_{t,a} \cdot \theta_{t,a} \\ &= \theta_{t,a} - \left(I - \frac{1}{2} \Sigma_{t,a} \right)^{M_t} \theta_{t,a}. \end{aligned}$$

Hence, the bias is given by

$$\mathbb{E} \left[\widetilde{\theta}_{t,a} - \widehat{\theta}_{t,a} \right] = - \left(I - \frac{1}{2} \Sigma_{t,a} \right)^{M_t} \theta_{t,a}.$$

We then bound the inner product as

$$\begin{aligned} \left| \mathbb{E} \left[\langle X_t, \widetilde{\theta}_{t,a} - \widehat{\theta}_{t,a} \rangle \mid \mathcal{H}_{t-1} \right] \right| &\leq \|X_t\|_2 \|\theta_{t,a}\|_2 \left\| \left(I - \frac{1}{2} \Sigma_{t,a} \right)^{M_t} \right\|_{\text{op}} \\ &\leq X_{\max} \Theta_{\max} \left(1 - \frac{p_{t,\min} \lambda_{\min}}{2K} \right)^{M_t} \\ &\leq \exp \left(- \frac{p_{t,\min} \lambda_{\min}}{2K} M_t \right), \end{aligned}$$

where we used $\|X_t\|_2 \leq X_{\max}$, $\|\theta_{t,a}\|_2 \leq \Theta_{\max}$, and the bound $\Sigma_{t,a} \succeq \frac{p_{t,\min} \lambda_{\min}}{K} I$ in the third inequality (since each arm is observed with probability p_t). **#DB: Why do we have a K^{-1} factor here?** [Antoine: should not be here] \square

Lemma F.2. *The cumulative bias introduced by the MGR approximation is uniformly bounded as*

$$\sum_{t=1}^T \max_{a \in [K]} \left| \mathbb{E} \left[\langle X_t, \widetilde{\theta}_{t,a} - \widehat{\theta}_{t,a} \rangle \right] \right| \leq \frac{\pi^2}{6}.$$

Proof. From Lemma F.1 and the definition $M_t = \lceil \frac{4K}{p_{t,\min} \lambda_{\min}} \log t \rceil$, we obtain, conditionally on \mathcal{H}_{t-1} ,

$$\left| \mathbb{E} \left[\langle X_t, \widetilde{\theta}_{t,a} - \widehat{\theta}_{t,a} \rangle \mid \mathcal{H}_{t-1} \right] \right| \leq \exp \left(- \frac{p_{t,\min} \lambda_{\min}}{2K} M_t \right) \leq \frac{1}{t^2}.$$

Taking total expectation and maximizing over $a \in [K]$ yields

$$\max_{a \in [K]} \left| \mathbb{E} \left[\langle X_t, \widetilde{\theta}_{t,a} - \widehat{\theta}_{t,a} \rangle \right] \right| \leq \frac{1}{t^2}.$$

We finally obtain the result by summing over t . \square

G. TIME AND SPACE COMPLEXITY OF ALGORITHM 1

#DB: Propagate the computations below in the main text. #

In this section, we detail the computation of the memory requirement and computation time of Algorithm 1, presented at the end of Section 3 of the paper.

At each round t , the algorithm stores the tuple (X_t, A_t, p_t, q_t) , which is of negligible size $\mathcal{O}(d + K)$, together with the parameter estimates $\hat{\theta}_{t,a}$ for all $a \in [K]$ and $t \leq T$, which must be kept across rounds to enable information reuse. This requires a total of $\mathcal{O}(dKT)$ memory. In addition, at each round t , computing the MGR approximation requires storing $\Sigma_t^+ \in \mathbb{R}^{d \times d}$, which leads to an additional temporary cost of $\mathcal{O}(d^2)$ during the computation of that round. Therefore, with Equation (13), the total space complexity is

$$\mathcal{O}(dKT + d^2).$$

Per-round computational cost. Each round involves two main computational steps: (i) solving the FTRL update via convex optimization, and (ii) performing Matrix Geometric Resampling (MGR).

FTRL update. In practice, we solve the FTRL objective up to precision $\varepsilon_t = \mathcal{O}(1/t^2)$ so that the cumulative optimization error remains finite. Because of that, we ignored this term in the regret bound of Theorem 4.1, which assumes that the computation of the sampling distribution is exact.

Using projected gradient descent, the number of iterations required at round t is $\mathcal{O}(\log t)$, and each iteration costs $\mathcal{O}(d \log d)$. Hence, the total cost over T rounds, that we denote by $\text{Comp}_T^{\text{FTRL}}$, satisfies

$$\mathcal{O}\left(\sum_{t=1}^T d \log t \log d\right) = \mathcal{O}(Td \log T \log d).$$

Matrix Geometric Resampling. We recall that M_t denotes the number of resampling steps performed at round t . Let us assume that $1/p_t = \mathcal{O}(t)$, which essentially corresponds to assuming that the logarithmic regret bound of Theorem 4.1 is also a lower bound, which is reasonable from an information-theoretic perspective. Then, since by Eq. (6) we defined M_t such that

$$M_t = \mathcal{O}\left(\frac{Kt \log t}{\lambda_{\min}}\right), \quad (13)$$

we can define the total computational cost of the MGR procedure at round t as

$$\Gamma_t^{\text{MGR}} \lesssim d^2 M_t \lesssim \frac{Kd^2 t \log t}{\lambda_{\min}}.$$

Summing over all rounds up to T yields a total computation time $\text{Comp}_T^{\text{MGR}}$ satisfying

$$\text{Comp}_T^{\text{MGR}} = \sum_{t=1}^T \Gamma_t^{\text{MGR}} \lesssim \frac{Kd^2}{\lambda_{\min}} \sum_{t=1}^T t \log t \lesssim \frac{Kd^2 T^2 \log T}{\lambda_{\min}}.$$

Thus, since the MGR procedure has to be fully rerun at each iteration, its total computational cost scales quadratically in T .

Overall complexity. Combining the two components of the algorithm, we obtain a total computational cost of order $\text{Comp}_T^{\text{total}} = \text{Comp}_T^{\text{FTRL}} + \text{Comp}_T^{\text{MGR}}$, satisfying

$$\text{Comp}_T^{\text{total}} \lesssim Td \log T \log d + \frac{Kd^2 T^2 \log T}{\lambda_{\min}}.$$

Treating λ_{\min}^{-1} as a numerical constant, and remarking that the second term (MGR steps) dominates, we obtain that the overall running time of the algorithm scales as

$$\text{Comp}_T^{\text{total}} \lesssim Kd^2 T^2 \log T.$$