LC-Tsallis-INF:
Generalized Best-of-Both-Worlds Linear Contextual Bandits

Masahiro Kato[1,2]
[1] Mizuho-DL Financial Technology   [2] The University of Tokyo

Shinji Ito[2,3]
[3] RIKEN AIP

masahiro-kato@fintec.co.jp   shinji@mist.i.u-tokyo.ac.jp

Abstract

We investigate the *linear contextual bandit problem* with independent and identically distributed (i.i.d.) contexts. In this problem, we aim to develop a *Best-of-Both-Worlds* (BoBW) algorithm with regret upper bounds in both stochastic and adversarial regimes. We develop an algorithm based on *Follow-The-Regularized-Leader* (FTRL) with Tsallis entropy, referred to as the α-*Linear-Contextual (LC)-Tsallis-INF*. We show that its regret is at most $O(\log(T))$ in the stochastic regime under the assumption that the suboptimality gap is uniformly bounded from below, and at most $O(\sqrt{T})$ in the adversarial regime. Furthermore, our regret analysis is extended to more general regimes characterized by the *margin condition* with a parameter $\beta \in (1, \infty]$, which imposes a milder assumption on the suboptimality gap. We show that the proposed algorithm achieves $O\left(\log(T)^{\frac{1+\beta}{2+\beta}} T^{\frac{1}{2+\beta}}\right)$ regret under the margin condition.

1 Introduction

This study investigates the linear contextual bandit problem (Abe & Long, 1999), which has been extensively studied across various domains, including sequential treatment allocation (Tewari & Murphy, 2017) and online advertising (Li et al., 2010). Depending on the domain, the behavior of loss and reward varies, and environment models have been developed to represent these differing behaviors. The primary models include the stochastic regime and the adversarial regime.

In the stochastic regime, losses are generated from a fixed distribution (Abbasi-yadkori et al., 2011), whereas in the adversarial regime, losses are chosen to maximize the regret. For the stochastic regime, Abbasi-yadkori et al. (2011) proposes OFUL. For the adversarial regime, Neu & Olkhovskaya (2020) proposes RealLinExp3. It is crucial to determine beforehand whether the environment is closer to a stochastic or adversarial setting to select an appropriate algorithm. However, it is often challenging to make this determination.

To address this difficulty, algorithms that perform well in both regimes, known as best-of-both-worlds (BoBW) algorithms (Bubeck & Slivkins, 2012), have garnered attention. BoBW algorithms play a crucial role in real-world applications. For example, in online advertising, an advertisement algorithm is typically designed for a stochastic environment and deployed accordingly. Modeling the environment as stochastic corresponds to assuming that customer responses follow a fixed stationary probability distribution. However, customer responses can be influenced by the advertisements themselves. Some customers may respond less favorably after repeatedly seeing the same advertisements, thereby violating the stochastic assumption. In such cases, a purely stochastic algorithm may no longer perform well. Deploying an algorithm designed for an adversarial environment could mitigate this issue, as it is robust to worst-case scenarios. However, adversarial algorithms tend to be overly conservative and pessimistic, leading to suboptimal performance in environments that are neither fully stochastic nor fully adversarial. In such intermediate cases, a best-of-both-worlds algorithm proves effective, offering robustness while maintaining competitive performance.

For linear contextual bandits, Kuroki et al. (2024) develops BoBW algorithms based on a black-box approach by Dann et al. (2023a), achieving $O(\log(T))$ regret in

stochastic regimes. However, these algorithms may not be practical due to implementation difficulties, as mentioned by Kuroki et al. (2024). To bypass this issue, Kuroki et al. (2024) and Kato & Ito (2023) propose algorithms based on follow-the-regularized-leader (FTRL) with Shannon entropy regularization. These FTRL-based approaches are simple and computationally tractable, but their regrets remain suboptimal, at least $\log^2(T)$ for the number of rounds T.

Thus, for linear contextual bandits, the only BoBW algorithm with $O(\log(T))$ regret bounds relies on a complex black-box approach, and whether a simpler FTRL-based algorithm can achieve the same goal has remained an open question. This study resolves this question by providing an FTRL-based BoBW algorithm with regret bounds tightly dependent on T. To construct such algorithms, we employ FTRL with Tsallis entropy regularization. Furthermore, we consider a generalized setting for linear contextual bandits by introducing the margin condition (Li et al., 2021). For our algorithm, we derive regret upper bounds and show that in the stochastic regime, it is of order $O(\log(T))$.

1.1 Problem setting

We suppose that there are T rounds and K arms, denoted by $[T] := \{1, 2, \ldots, T\}$ and $[K] := \{1, 2, \ldots, K\}$, respectively. In each round $t \in [T]$, each arm $a \in [K]$ is associated with a loss $\ell_t(a, X_t)$ given a context $X_t \in \mathcal{X}$, where $\mathcal{X}$ is an arbitrary context space. Here, X_t is generated from a distribution $\mathcal{G}$ and induces a d-dimensional arm-dependent *feature* vector $\phi(a, X_t) \in \mathcal{Z} \subset \mathbb{R}^d$ for each arm $a \in [K]$ (see Assumption 2.2), where $\phi(a, \cdot)$ is a mapping from an arm-independent context to an arm-dependent feature. The loss $\ell(a, x)$ follows linear models with $\phi(a, x)$.

Assumption 1.1 (Linear models). For all $a \in [K]$ and any $x \in \mathcal{X}$, the following holds:

$$\ell_t(a, x) = \langle \phi(a, x), \theta_t \rangle + \varepsilon_t(a),$$

where $\theta_t \in \Theta$ is a d-dimensional parameter with a space $\Theta \subset \mathbb{R}^d$, $\varepsilon_t(a)$ satisfies $\mathbb{E}\left[\varepsilon_t(a) \mid X_t, \mathcal{F}_{t-1}\right] = 0$, and $\mathcal{F}_{t-1} = \left\{\left(X_s, A_s, \ell_s(A_s, X_s)\right)\right\}_{s=1}^{t-1}$.

The decision-maker follows a *policy* in arm selection. Let Π be the set of all possible policies $\pi : \mathcal{X} \to \mathcal{P}_K := \left\{u = (u_1 \ u_2 \ \ldots \ u_K)^\top \in [0,1]^K \mid \sum_{k=1}^K u_k = 1\right\}$ with its a-th element $\pi(a \mid x)$. Then, we consider sequential decision-making with the following steps in each $t \in [T]$:

1. Nature decides θ_t based on $\mathcal{F}_{t-1}$.
2. The decision-maker observes a context $X_t \in \mathcal{X}$, generated from a fixed distribution $\mathcal{G}$.
3. Based on the observed context X_t, the decision-maker selects a policy $\pi_t(X_t) \in \mathcal{P}_K$.

4. The decision-maker chooses an action $A_t \in [K]$ with probability $\pi_t(A_t \mid X_t)$.
5. The decision-maker incurs a loss $\ell_t(A_t, X_t)$.

The goal of the decision-maker is to select actions to minimize the total loss $\sum_{t=1}^T \ell_t(A_t, X_t)$.

The performance is measured via the regret, defined as

$$R_T := \max_{\rho : \mathcal{X} \to [K]} \mathbb{E}\left[\sum_{t=1}^T \left\{\ell_t(A_t, X_t) - \ell_t(\rho(X_t), X_t)\right\}\right],$$

where the expectation is taken over the randomness of policies and the contexts, $\{X_t\}_{t \in [T]}$, and losses, $\{\ell_t(\cdot, X_t)\}_{t \in [T]}$. The optimal policy ρ_T^* is defined as

$$\rho_T^* = \arg\min_{\rho : \mathcal{X} \to [K]} \mathbb{E}\left[\sum_{t=1}^T \left\langle \phi(\rho(X_t), X_t), \theta_t \right\rangle\right].$$

Thus, the regret becomes

$$R_T = \mathbb{E}\left[\sum_{t=1}^T \left\langle \phi(A_t, X_t) - \phi(\rho_T^*(X_t), X_t), \theta_t \right\rangle\right].$$

1.2 Contributions

This study aims to propose a practical BoBW algorithm with $O(\log(T))$ regret in the stochastic regime without using the black-box framework. We focus on the use of Tsallis entropy instead of Shannon entropy. Section 3 presents our algorithm, the α-Linear-Contextual-Tsallis-INF (α-LC-Tsallis-INF), which is the FTRL with α-Tsallis entropy regularization. Our algorithm's regret satisfies $O(\log(T))$. Additionally, compared to the algorithm using the black-box approach by Kuroki et al. (2024), our algorithm is easier to implement and has a tighter dependence regarding K.

A key component of our algorithm is the inverse of the covariance matrix of an arm-dependent feature, multiplied by a policy π_t in each round $t \in [T]$, denoted as Σ_t (Section 2.1). Throughout this study, we assume that Σ_t can be computed exactly. When it is approximated using finite samples, the Matrix Geometric Resampling (MGR) method proposed by Neu & Olkhovskaya (2020) may be employed. Although we do not derive the regret bound when using the MGR, we provide a brief overview in Section 3.

Our study also addresses general stochastic regimes depending on a margin condition with various parameters β. Kato & Ito (2023) and Kuroki et al. (2024) only discuss a case where there exists a positive lower bound for the suboptimality gap, which is considered restrictive in linear contextual bandits. In contrast, we consider a milder assumption on the suboptimality gap, called a margin condition, and characterize the problem difficulty using a parameter $\beta \in (0, +\infty)$.

We then derive regret upper bounds of the 1/2-LC-Tsallis-INF (α-LC-Tsallis-INF with $\alpha = 1/2$) for each regime defined in Section 4. In an adversarial regime, the regret is given as $O\left(\sqrt{d\sqrt{K}T} + \frac{1}{\lambda}\left(\sqrt{K} + \frac{\sqrt{\sqrt{K}T}}{\sqrt{d}}\right)\right)$, where λ is a parameter that characterizes Σ^{-1} (Assumption 2.4) and a generalization of $\lambda_{\min}$ in the existing studies. In a stochastic regime with a margin condition, the regret is given as $O\left(\left(\frac{1+\beta}{\beta\Delta_*}\right)^{\frac{\beta}{2+\beta}} \left(Ld\sqrt{K}\log(T)\right)^{\frac{1+\beta}{2+\beta}} T^{\frac{1}{2+\beta}} + \kappa\right)$, where $\kappa = O\left(\sqrt{\frac{KL}{\lambda}} + \frac{\sqrt{K}L}{\lambda d}\log(T)\right)$, and L is the lowest probability of contexts (see Assumption 2.3). In an adversarial regime with a self-bounding constraint, the regret is given as $O\left(\frac{Ld\sqrt{K}}{\Delta_*}\log(T) + \sqrt{\frac{CL\sqrt{K}d}{\Delta_*}\log_+\left(\frac{\Delta_*T}{CL}\right)} + \kappa\right)$, where $\log_+(x) \coloneqq \max\{1, \log(x)\}$ and recall that β is a parameter of a margin condition.

The advantage of using Tsallis entropy lies in its improved regret bounds, particularly in the dependency on T in the stochastic regime. When using Shannon entropy, the regret does not achieve the optimal $\log(T)$ dependency. For example, Kuroki et al. (2024) develops a best-of-both-worlds algorithm with a regret bound of $\log(T)^2$ using Shannon entropy.

We also discuss the difference between settings with arm-dependent and arm-independent features in Section 4.2. While the former is the setting employed in this study and most existing studies in linear contextual bandits, the latter is a setting used in Neu & Olkhovskaya (2020), Kuroki et al. (2024), and Kato & Ito (2023). We point out that if we can derive a regret in either of the settings, we can obtain a regret in the other. We highlight that a tighter regret can be achieved if we first derive a regret under a setting with arm-dependent features and then transform it to a setting with arm-independent features. In Table 1, we report the upper bound for RealLinExp3 from Neu & Olkhovskaya (2020) after applying the regret transformation described in Section 4.2. By Theorem 2 of Neu & Olkhovskaya (2020), their original regret bound is $O\left(\sqrt{TK\max\{d, \frac{\log T}{\lambda}\}\log K}\right)$. Because their setting is based on arm-independent features (differing from our setup with arm-dependent features), we applied the transformation. By applying Theorem 4.5, the parameter d is replaced with dK. Consequently, we obtain the regret upper bound shown in Table 1.

In summary, our contributions lie in the proposition of an FTRL-based algorithm whose upper bound tightly depends on T, and the analysis under a margin condi-

tion and the arm-dependent feature setting. In Table 1, we compare our algorithm's regrets with existing ones.

1.3 Related work

For linear contextual bandits, Kuroki et al. (2024) and Kato & Ito (2023) propose FTRL-based BoBW algorithms. Kuroki et al. (2024) develop several BoBW algorithms using the black-box framework approach by Dann et al. (2023a). Among them, they propose a BoBW reduction of RealEXP3 by Neu & Olkhovskaya (2020), showing a regret of order $O(\log(T))$ in the stochastic regime. They also present another algorithm by combining the black-box framework with the continuous exponential weights algorithm investigated by Olkhovskaya et al. (2023), which has a regret of $O(\log^5(T))$ in the stochastic regime and $O(\log(T)dK\sqrt{\Lambda^*})$ in the adversarial regime, where Λ^* denotes the cumulative second moment of the losses incurred by the algorithm.

While the black-box framework provides a tight $O(\log(T))$ regret regarding T, limitations have been reported. For example, Kuroki et al. (2024) mention that "it may not be practical to implement." Compared to the black-box framework, "FTRL with Shannon entropy regularization is a much more practical algorithm" (Kuroki et al., 2024). Kuroki et al. (2024) and Kato & Ito (2023) show that the algorithm has a regret of $O\left(\frac{K}{\Delta_*}\left(dK + \frac{\log(T)}{\lambda_{\min}}\right)\log\left(KT\right)\log(T)\right)$ in the linear contextual adversarial regime with a self-bounding constraint, where $\lambda_{\min}$ is the smallest eigenvalue of a feature covariance matrix induced by an exploratory policy (see Assumption 2.4). Note that we transformed their original regrets derived for the arm-independent feature setting to the ones for the arm-dependent feature setting by using our Theorem 4.5 in Section 4.2. They also show a regret of $O\left(\sqrt{TK\left(dK + \frac{\log(T)}{\lambda_{\min}}\right)\log(T)\log(K)}\right)$ in the adversarial regime. Note that their FTRL with Shannon entropy considers a case where Σ_t^{-1}, the inverse of the feature covariance matrix, is approximated by finite samples, and they assume that features are arm-independent (see Section 4.2). Here, the approximation error of Σ_t^{-1} affects the regrets. If Σ_t^{-1} can be exactly computed, we can remove $\log(T)$ in $\log(T)/\lambda_{\min}$, and the regret in the stochastic regime becomes of order $O(\log^2(T))$ regarding T instead of $\log^3(T)$.

BoBW algorithms based on FTRL with Shannon entropy usually incur regret of at least $O(\log^2(T))$ (Ito et al., 2022; Tsuchiya et al., 2023a; Kong et al., 2023). As a potential solution to this issue, using Tsallis entropy instead of Shannon entropy has been shown to be effective, achieving success in multi-armed ban-

Table 1: Comparison of regrets. We categorize regrets based on regimes. We place a ✓ in the "$\sqrt{C}$" column if the regret depends on the corruption level $C \geq 0$ in the presence of adversarial corruption. We transformed the regret from the arm-independent setting to the arm-dependent setting using Theorem 4.5. For the BoBW with Shannon entropy regularization, we also present a regret derived for the arm-dependent setting by ourselves.

	Regret		$\sqrt{C}$
	Stochastic	Adversarial	
1/2-LC-Tsallis-INF Section 4.1	$O\left(\frac{1}{\Delta_*} L d\sqrt{K}\log(T) + \kappa\right)$	$O\left(\sqrt{d\sqrt{K}T} + \kappa\right)$	✓
	$O\left(\left(\frac{1+\beta}{\beta\Delta_*}\right)^{\frac{\beta}{2+\beta}}\left(Ld\sqrt{K}\log(T)\right)^{\frac{1+\beta}{2+\beta}}T^{\frac{1}{2+\beta}} + \kappa\right)$ under a margin condition with $\beta \in (0,\infty]$. $\kappa = O\left(\sqrt{\frac{KL}{\lambda}} + \frac{\sqrt{K}L}{\lambda d}\log(T)\right)$		-
FTRL with Shannon entropy (Kuroki et al., 2024) (Kato & Ito, 2023) Section 4.2 and Appendix D	$O\left(\frac{1}{\Delta_*}K\left(dK + \frac{\log(T)}{\lambda_{\min}}\right)\log(KT)\log(T)\right)$	$O\left(\sqrt{\log(T)\log(K)TK\left(dK + \frac{\log(T)}{\lambda_{\min}}\right)}\right)$	✓
	under transformation using our Thm 4.5.		
	$O\left(\frac{1}{\Delta_*}\left(d + \frac{1}{\lambda}\right)\log(KT)\log(T)\right)$	$O\left(\sqrt{\log(T)T\left(d + \frac{1}{\lambda}\right)}\log(KT)\right)$	
	from Appendix D.		
BoBW reduction of the RealLinExp3 (Prop. 8, Kuroki et al., 2024)	$O\left(\frac{1}{\Delta_*}K^2\left(dK + \frac{1}{\lambda_{\min}}\right)^2\log(K)\log(T)\right)$	$O\left(\sqrt{TK^2\left(dK + \frac{1}{\lambda_{\min}}\right)^2\log(K)}\right)$	✓
MWU-LC (Thm. 1, Kuroki et al., 2024)	$O\left(\frac{1}{\Delta_*}d^2K^4\log^2(dK^2T)\log^3(T)\right)$	$O\left(dK^2\sqrt{\Lambda^*}\log(T)\right)$	✓
Logdet-FTRL (poly(K,d,T) in computation)	-	$O\left(d^2\sqrt{T}\log(T)\right)$	-
Linear EXP4 (T^d in computation) (Liu et al., 2023)	-	$O\left(d\sqrt{T\log(T)}\right)$	-
OFUL (Abbasi-yadkori et al., 2011)	$O\left(d\log(1/\delta)/\Delta_*\right)$ with probability $1-\delta$ $O\left(d\log(T)/\Delta_*\right)$ with probability $1-1/T$	- -	- -
RealLinExp3 (Neu & Olkhovskaya, 2020)	-	$O\left(\sqrt{TK\max\left\{dK, \frac{\log(T)}{\lambda_{\min}}\right\}\log(K)}\right)$	-

dits (Zimmert & Seldin, 2021; Masoudian & Seldin, 2021; Jin et al., 2023), combinatorial semi-bandits (Zimmert et al., 2019), dueling bandits (Saha & Gaillard, 2022), and graph bandits (Rouyer et al., 2022). However, for linear contextual bandits, even when restricted to the adversarial setting, algorithms using Tsallis entropy are not known, and the application and analysis of this approach are not straightforward.

Notation Let $\langle \cdot, \cdot \rangle$ denote inner products in Euclidean space and let $\|\cdot\|_2$ denote the ℓ_2 norm.

2 Preliminaries

This section provides assumptions on our problem. This study considers a setting where the arm-dependent feature is given, and the parameters are arm-independent.

2.1 Boundedness of variables

We first assume the boundedness of the variables.

Assumption 2.1 (Bounded loss)**.** We assume that $|\langle z, \theta \rangle| \leq 1$ for all $z \in \mathcal{Z}$ and $\theta \in \Theta$. In addition, we assume that $|\ell_t(a,x)| = |\langle \phi(a,x), \theta_t \rangle + \varepsilon_t(a)| \leq 1$ holds for all $a \in [K]$, $x \in \mathcal{X}$ and $t \in [T]$.

The parameter θ_t is generated in different ways according to the data-generating process (DGP). We define regimes of the DGP in Section 2.3.

2.2 Assumptions on contexts

Contexts and feature map We first assume that an i.i.d. random variable X_t is generated from a fixed distribution $\mathcal{G}$ over the support $\mathcal{X}$. We refer to X_t as contexts. Features are obtained from X_t through a feature map that transforms a context $x \in \mathcal{X}$ into a feature $\phi(a,x) \in \mathcal{Z} \subseteq \mathbb{R}^d$, where $\mathcal{Z}$ is a d-dimensional feature space. We assume that ϕ is known.

For any $p = \{p_x\}_{x \in \mathcal{X}}$, a set of conditional distributions $p_x \in \mathcal{P}_K$ given $x \in \mathcal{X}$, we define a matrix $\Sigma(p) \in \mathbb{R}^{d \times d}$ by $\Sigma(p) := \mathbb{E}_{X_0 \sim \mathcal{G}, a \sim p_x}\left[\phi(a, X_0)\left(\phi(a, X_0)\right)^\top\right]$, where X_0 is a sample from $\mathcal{G}$. This matrix plays an important role in constructing unbiased estimators of loss and in the analysis of regret. We assume that $\Sigma^{-1}(p)$ can be computed exactly. Below, we summarize the assumptions about the contexts.

Assumption 2.2 (Contextual distribution)**.** (i) Context $X_t \in \mathcal{X}$ is an i.i.d. random variable from a contextual distribution $\mathcal{G}$ with support $\mathcal{X}$. (ii) There is

a *known* feature map $\phi : [K] \times \mathcal{X} \to \mathcal{Z}$, which maps $x \in \mathcal{X}$ to feature $\phi(a, x) \in \mathcal{Z} \subset \mathbb{R}^d$. (iii) For any $p \in \Pi$, $\Sigma^{-1}(p)$ is exactly computable.

Conditions (i) and (ii) are standard assumptions in adversarial linear contextual bandits (Neu & Olkhovskaya, 2020) and are essential for the algorithm design.

Under condition (i), the regret analysis can be reduced to evaluating the pointwise regret for each x. Similar analytical techniques have also been used in previous studies, such as Neu & Olkhovskaya (2020).

Although condition (iii) may appear restrictive, it is not necessarily stronger than assumptions in existing studies, such as Neu & Olkhovskaya (2020), Kuroki et al. (2024), and Kato & Ito (2023), which approximate $\Sigma^{-1}(p)$ using finite samples from a known $\mathcal{G}$ and the MGR algorithm. For instance, if the support is finite and its probability mass function ($g(x)$ in Assumption 2.3) is known, we can compute the exact $\Sigma^{-1}(p)$. If there are infinite samples from $\mathcal{G}$, then we can approximate $g(x)$ using these samples.

Moreover, the MGR method is computationally inefficient in the BoBW algorithms proposed by Kuroki et al. (2024) and Kato & Ito (2023). While the MGR is computationally efficient in adversarial linear contextual bandits (Neu & Olkhovskaya, 2020), the computational costs are not bounded in those BoBW algorithms. We discuss this point further in Section 3.3.

Contexts in a stochastic regime In a stochastic regime with adversarial corruption, we additionally make the following assumption on contexts.

Assumption 2.3 (Finite support)**.** (i) The context space $\mathcal{X}$ is a finite set of size $S = |\mathcal{X}| \in \mathbb{N}$. (ii) There exists a constant $L \geq S$ such that $g(x) := \mathbb{P}_{X_0 \sim \mathcal{G}}(X_0 = x) \geq 1/L$ holds for all $x \in \mathcal{X}$.

This assumption is reasonable in some applications. For instance, in advertising, arm selections are often based on contexts provided by personal information, such as age or gender. When considering regret minimization for a specific group (e.g., people in the US), we can often obtain its distribution from publicly available data, such as demographics. Moreover, when contexts are continuous, practitioners often discretize them and select arms based on groups for ease of implementation.

Exploration policy We use an *exploration policy* that satisfies Assumption 2.4 in the algorithm design.

Assumption 2.4 (Exploration policy)**.** There exists a constant $\lambda > 0$ and a set $e^* = \{e^*_x\}_{x \in \mathcal{X}}$ of distributions over $[K]$ such that $z^\top \Sigma(e^*)^{-1} z \leq 1/\lambda$ for all $z \in \mathcal{Z}$.

The value of λ in this assumption can be interpreted as a generalized or a relaxed version of the smallest eigenvalue $\lambda_{\min}$ of $\Sigma(e^*)$, which is used in assumptions of existing studies (Neu & Olkhovskaya, 2020; Kuroki et al., 2024). In fact, if $\|z\|_2 \leq 1$ holds for all $z \in \mathcal{Z}$, $\max_{z \in \mathcal{Z}} z^\top (\Sigma(e^*))^{-1} z$ is at most $1/\lambda_{\min}$. In addition, if we choose e^*_x to be a G-optimal design (see, e.g., Chapter 21 by Lattimore & Szepesvári (2020)) for $\mathcal{Z}_x = \{\phi(a, x) \in \mathcal{Z} \mid a \in [K]\}$, we may set $\lambda = 1/(dL)$. In fact, it holds that $\phi(a, x)^\top (\Sigma(e^*))^{-1} \phi(a, x) = \phi(a, x)^\top \left(\sum_{x' \in \mathcal{X}} g(x') \Sigma(e^*_{x'}) \right)^{-1} \phi(a, x) \leq \phi(a, x)^\top (g(x) \Sigma(e^*_x))^{-1} \phi(a, x) \leq d/g(x) \leq dL$ for all $a \in [K]$ and $x \in \mathcal{X}$. As discussed in Note 6 of Section 21.2 in Lattimore & Szepesvári (2020), we can define the G-optimal design even when $\mathcal{Z}$ does not span the entire space. For example, we can use Moore-Penrose pseudoinverses in place of inverse matrices.

2.3 DGP: stochastic and adversarial regimes

We define three regimes for the DGP of $\{\theta_t\}_{t \in [T]}$: an adversarial regime, a stochastic regime with a margin condition, and an adversarial regime with a self-bounding constraint.

(1) Adversarial regime First, we introduce the adversarial regime, where we do not make any assumptions about the behavior of the nature. In this case, it is known that the lower bound is $O(\sqrt{T})$ when there is no context (Auer et al., 2002).

Note that adversarial linear contextual bandits can be defined in various ways. For example, some studies consider adversarial contexts and fixed losses (Chu et al., 2011; Abbasi-yadkori et al., 2011). Meanwhile, other studies address adversarial contexts and adversarial losses (Kanade & Steinke, 2014; Hazan et al., 2016). This study focuses exclusively on contextual bandits with i.i.d. contexts and adversarial losses, as studied by Rakhlin & Sridharan (2016) and Syrgkanis et al. (2016). This study follows the setting of Neu & Olkhovskaya (2020), Kato & Ito (2023), and Kuroki et al. (2024).

(2) Stochastic regime with a margin condition Next, we define a *margin condition*, which is often assumed in linear contextual bandits to characterize the difficulty of the problem instance (Li et al., 2021). Since this section focuses on a stochastic regime, regression coefficients are fixed, and we denote them by θ_0; that is, $\theta_1 = \cdots = \theta_t = \theta_0$. Note that under a stochastic regime, it holds that

$$R_T = \mathbb{E}\left[\sum_{t=1}^{T} \sum_{a \in [K]} \Delta(a \mid X_t) \pi_t(a \mid X_t)\right] \tag{1}$$

$$\geq \mathbb{E}\left[\sum_{t=1}^{T} \Delta(X_t)\big(1 - \pi_t(\rho^*_T(X_t) \mid X_t)\big)\right],$$

where

$$\Delta(a \mid x) = \left\langle \phi(a,x) - \phi(\rho_T^*(x),x), \theta_0 \right\rangle \quad \text{and}$$

$$\Delta(x) = \min_{a \neq \rho_T^*(x)} \left\{ \left\langle \phi(a,x), \theta_0 \right\rangle - \left\langle \phi(\rho_T^*(x),x), \theta_0 \right\rangle \right\}.$$

Based on this suboptimality gap $\Delta(x)$, we define a stochastic regime with a margin condition.

Definition 2.5 (Stochastic regime with a margin condition)**.** Consider the stochastic regime with fixed regression coefficient $\theta_0 \in \Theta$, where for all $t \in [T]$, all $a \in [K]$ and any $x \in \mathcal{X}$, the loss is generated as $\ell_t(a,x) = \left\langle \phi(a,x), \theta_0 \right\rangle + \varepsilon_t(a)$. Furthermore, there exists a universal constant $\Delta_* > 0$ and $\beta \in (0,+\infty]$, such that for any $h \in [0, \Delta_*]$, it holds that

$$\mathbb{P}\left(\Delta(X_t) \leq h\right) \leq \frac{1}{2}\left(\frac{h}{\Delta_*}\right)^\beta.$$

Note that when $\beta = \infty$, $\Delta(x) \geq \Delta_*$ holds for any $x \in \mathcal{X}$ and $t \in [T]$. A margin condition in Li et al. (2021) restricts the range of h as $[D\sqrt{\log(d)/T}, \Delta_*]$ for some universal constant $D > 0$, and they derive matching lower and upper bounds.

The margin condition is one of the minimal assumptions for achieving $o(\sqrt{T})$-regret in linear contextual bandits. For instance, the lower bound in Theorem 1 by Li et al. (2021) suggests that the achievable regret bounds can be characterized by the margin parameter. If the condition does not hold, the regret becomes $\Omega(\sqrt{T})$.

(3) Adversarial regime with a self-bounding constraint This section defines a regime with an adversarial corruption. Let $\Delta_* > 0$ be a universal constant, as used in Definition 2.5. If $\Delta(x) \geq \Delta_*$ holds for any $x \in \mathcal{X}$ in a stochastic regime, then the regret can be lower bounded as $R_T \geq \Delta_* \mathbb{E}\left[\sum_{t=1}^T \left(1 - \pi_t\left(\rho_T^*(X_t)\right)\right)\right]$. Based on this intuition, we define an *adversarial regime with a self-bounding constraint* below, as well as introduced by Kuroki et al. (2024) and Kato & Ito (2023).

Definition 2.6 (Adversarial regime with a self-bounding constraint)**.** We say that the DGP is in a (Δ_*, C, T)-adversarial regime with a self-bounding constraint for some $\Delta_*, C > 0$ if the regret R_T is lower bounded as

$$R_T \geq \mathbb{E}\left[\sum_{t=1}^T \sum_{a \neq \rho_T^*(X_t)} \Delta_t(a \mid X_t)\pi_t(a \mid X_t)\right] - C,$$

where $\Delta_t(a \mid x) = \left\langle \phi(a,x), \theta_t \right\rangle - \left\langle \phi(\rho_T^*(x),x), \theta_t \right\rangle$, and for any $x \in \mathcal{X}$ and $a \in [K] \setminus \{\rho_T^*(x)\}$, $\Delta_t(a \mid x) \geq \Delta_*$ holds.

An adversarial regime with a self-bounding constraint encompasses several important settings. See examples in Kato & Ito (2023). Note that in an adversarial regime, there may exist $a \in [K]$ and $x \in \mathcal{X}$ such that $\Delta_t(x) \coloneqq \left\langle \phi(a,x), \theta_t \right\rangle - \left\langle \phi(\rho_T^*(x),x), \theta_t \right\rangle < 0$. This is because $\Delta(a \mid x)$ and $\Delta_s(a \mid x)$ can take a different value for some $a \in [K]$ and $x \in \mathcal{X}$ if $t \neq s$.

3 Algorithm: α-LC-Tsallis-INF

This section provides an algorithm for linear contextual bandits with adversarial corruption. We refer to our algorithm as the α-Linear-Contextual (LC)-Tsallis-INF because it modifies the Tsallis-INF (Zimmert & Seldin, 2021), an FTRL-based algorithm with Tsallis entropy regularization without contexts. Here, $\alpha \in (0,1)$ is a parameter of Tsallis-entropy. While our algorithm is defined for general α, we show the regret bounds only for $\alpha = 1/2$. The pseudo-code is shown in Algorithm 1.

When selecting arm a with probability $\pi_t(a \mid x)$ given $x \in \mathcal{X}$, we denote $\Sigma((\pi_t(x))_{x \in \mathcal{X}})$ by Σ_t, equal to

$$\Sigma_t = \mathbb{E}_{X_t \sim \mathcal{G}}\left[\sum_{a \in [K]} \pi_t(a \mid X_t)\phi(a, X_t)\left(\phi(a, X_t)\right)^\top\right].$$

3.1 Regression coefficient estimator

We define an estimator of the parameter θ_t as

$$\widehat{\theta}_t \coloneqq \widehat{\theta}_t\left(\Sigma_t^{-1}\right) \coloneqq \Sigma_t^{-1}\phi(A_t, X_t)\ell_t(A_t, X_t). \quad (2)$$

This estimator is unbiased for θ_t since it holds that

$$\mathbb{E}\left[\widehat{\theta}_t(\Sigma_t^{-1}) \mid \mathcal{F}_{t-1}\right] = \Sigma_t^{-1}\Sigma_t\theta_t = \theta_t.$$

Here, recall that Σ_t is defined as an expectation taken w.r.t. X_0 and a given π_t (as defined in Section 2.2).

In each round t, using $\widehat{\theta}_t$, we construct an estimator of the loss $\ell_t(x)$ as $\widehat{\ell}_t(x) \coloneqq \left(\langle \widehat{\theta}_t, \phi(a, X_t)\rangle\right)_{a \in [K]}$. This estimator is unbiased for the true loss $\ell_t(x)$.

3.2 FTRL with Tsallis entropy regularization

By using the estimator $\widehat{\theta}_t\left(\Sigma_t^{-1}\right)$, we define the α-LC-Tsallis-INF. In each round $t \in [T]$, the α-LC-Tsallis-INF selects an arm with the following policy:

$$\pi_t(X_t) \coloneqq (1 - \gamma_t)q_t(X_t) + \gamma_t e^*(X_t), \quad (3)$$

where recall that e^* is an exploration policy defined in Assumption 2.4, and

$$q_t(x) = \underset{q \in \mathcal{P}_K}{\arg\min}\left\{\sum_{s=1}^{t-1} \left\langle \widehat{\ell}_s(x), q \right\rangle + \frac{1}{\eta_{t-1}}\psi(q)\right\}, \quad (4)$$

$$\psi(q(x)) \coloneqq \frac{1}{\alpha}\left(1 - \sum_{a \in [K]} q(a \mid x)^\alpha\right),$$

$$\widetilde{\eta}_t \coloneqq \frac{K^{1/4}}{\sqrt{dt}}, \quad \gamma_t \coloneqq \frac{128L\eta_t^2}{\lambda}\left(\leq \frac{1}{2}\right),$$

$$\eta_t \coloneqq \begin{cases} \min\left\{\widetilde{\eta}_t, \ \frac{1}{16}\sqrt{\frac{\lambda}{L}}\right\} & \text{if Assumption 2.3 holds} \\ \min\left\{\widetilde{\eta}_t, \ \frac{\lambda}{16}\right\} & \text{otherwise} \end{cases}.$$

Here, the regularizer $\psi(q(x))$ is referred to as the α-Tsallis entropy (Tsallis, 1988).

3.3 Approximation of Σ_t^{-1}

If Σ_t^{-1} is not exactly computable, we can use the MGR to approximate it using finite samples from $\mathcal{G}$ (Section 2.1). The MGR is a computationally efficient algorithm for approximating $\Sigma^{-1}(p)$ in the adversarial regime, as demonstrated by Neu & Olkhovskaya (2020).

However, in the BoBW setting, the MGR incurs a significantly higher computational cost (Kuroki et al., 2024; Kato & Ito, 2023). Approximating $\Sigma^{-1}(p)$ using the MGR from finite samples drawn from $\mathcal{G}$ results in a computational cost of order $O\left(d^2 K M_t^2\right)$, where M_t^2 is a term that increases at least linearly with t. While the computational burden of the MGR is manageable in the adversarial regime (Neu & Olkhovskaya, 2020), it becomes substantial when aiming for a BoBW regret guarantee. For example, if we use the MGR in the FTRL with Shannon entropy regularization proposed by Kuroki et al. (2024) and Kato & Ito (2023), M_t can be $\Omega(\sqrt{t})$, which results in a computational cost of at least order t with respect to t.

In contrast, if contexts have finite support with bounded probabilities (Assumption 2.3), it is possible to compute $\Sigma^{-1}(p)$ exactly with an acceptable computational cost, as assumed in this study. Specifically, when the probability is lower bounded by L, $\Sigma^{-1}(p)$ can be computed with a cost of order $O(d^2 K|\mathcal{S}|)$.

Thus, if our primary interest lies in achieving BoBW guarantees, the MGR may not be the most computationally efficient choice. Compared to the MGR, our approach, which relies on Assumption 2.3, offers a more computationally feasible alternative.

4 Regret analysis for the $1/2$-LC-Tsallis-INF

This section provides regrets of the α-LC-Tsallis-INF with $\alpha = 1/2$ (1/2-LC-Tsallis-INF).

First, we show the following upper bound, from which

Algorithm 1 α-LC-Tsallis-INF.

Parameter: Learning rate $\eta_1, \eta_2, \ldots, \eta_T > 0$.
for $t = 1, \ldots, T$ **do**
  Observe X_t.
  Draw $A_t \in [K]$ following the policy $\pi_t(X_t) \coloneqq (1 - \gamma_t)q_t(X_t) + \gamma_t e^*(X_t)$ defined in (3).
  Observe the loss $\ell_t(A_t, X_t)$.
  Compute $\widehat{\theta}_t$.
end for

---

we derive upper bounds in adversarial and stochastic regimes. We prove Theorem 4.1 in Appendix B.

Theorem 4.1 (General regret bounds). *Consider the $1/2$-LC-Tsallis-INF. Assumptions 2.2 and 2.4–2.1 hold. Then, the regret satisfies*

$$R_T = O\left(\mathbb{E}\left[\sum_{t=1}^T \frac{\sqrt{d\sqrt{K}\omega_t}}{\sqrt{t}}\right] + \widetilde{\kappa}\right), \quad \text{where } \widetilde{\kappa} \coloneqq$$

$$\begin{cases} O\left(\sqrt{\frac{KL}{\lambda}} + \frac{\sqrt{K}L\log(T)}{\lambda d}\right) & \text{if Assumption 2.3 holds} \\ O\left(\frac{1}{\lambda}\left(\sqrt{K} + \frac{\sqrt{\sqrt{K}T}}{\sqrt{d}}\right)\right) & \text{otherwise} \end{cases},$$

and $\omega_t \in [0, 1]$ is given as

$$\omega_t \coloneqq \max_{x \in \mathcal{X}} \min_{a \in [K]} \{1 - q_t(a \mid x)\}.$$

Using this general regret bound, we derive a corresponding bound under an adversarial regime in Theorem 4.2. In Theorem 4.3, we present a regret bound under a stochastic regime with a margin condition. Lastly, in Theorem 4.4, we develop a regret bound under a stochastic regime that achieves an order of $\log(T)$.

In these derivations, the term ω_t plays a pivotal role, corresponding to an upper bound of the probabilities of choosing suboptimal arms. For the adversarial regime, the regret is derived without placing any restrictions on ω_t, whereas for the stochastic regime, we leverage specific properties of ω_t. In particular, we employ a self-bounding technique, which tightens the regret bound by exploiting the coherence between the regret and ω_t in the stochastic regime.

4.1 Regret upper bound in each regime

Because $\omega_t \leq 1$, by replacing ω_t with 1 in Theorem 4.1, we can directly obtain a regret upper bound in the adversarial regime in the following theorem.

We show the proof in Appendix A and B.4. As well as Kato & Ito (2023) and Kuroki et al. (2024), we employ the self-bounding technique (Zimmert & Seldin, 2021; Wei & Luo, 2018; Masoudian & Seldin, 2021) and an entropy-adaptive update rule for learning rates,

which have been proven effective in providing BoBW guarantees for online learning in feedback graph contexts (Ito et al., 2022), multi-armed bandits (Jin et al., 2023), partial monitoring (Tsuchiya et al., 2023a), linear bandits (Kong et al., 2023), episodic Markov Decision Processes (MDPs) (Dann et al., 2023b), and sparse bandits (Tsuchiya et al., 2023b).

Theorem 4.2 (Regret upper bound in an adversarial regime)**.** *Consider the 1/2-LC-Tsallis-INF. Assume that the loss is generated under an adversarial regime. Suppose that Assumptions 2.2 and 2.4–2.1 hold. Then, the regret satisfies*

$$R_T = O\left(\sqrt{d\sqrt{K}T} + \frac{1}{\lambda}\left(\sqrt{K} + \frac{\sqrt{\sqrt{K}T}}{\sqrt{d}}\right)\right).$$

Note that this result does not require Assumptions 2.3, which restricts the context support to be finite.

Next, we show a regret upper bound in a stochastic regime with a margin condition, which depends on the parameter $\beta \in (0, \infty]$. Recall that β decides the behavior of $\Delta(x)$. The proof is in Appendix B.6.

Theorem 4.3 (Regret upper bound in a stochastic regime with a margin condition)**.** *Consider the 1/2-LC-Tsallis-INF. Assume that the loss is generated under a stochastic regime with a margin condition (Definition 2.5). Suppose that Assumption 2.2–2.1 hold. Then, the regret satisfies*

$$R_T = O\left(\left(\frac{1+\beta}{\beta\Delta_*}\right)^{\frac{\beta}{2+\beta}}\left(Ld\sqrt{K}\log(T)\right)^{\frac{1+\beta}{2+\beta}}T^{\frac{1}{2+\beta}} + \kappa\right),$$

where $\kappa = O\left(\sqrt{\frac{KL}{\lambda}} + \frac{\sqrt{K}L}{\lambda d}\log(T)\right)$.

For example, when $\beta = \infty$, then the regret is $R_T = O\left(\frac{1}{\Delta_*}Ld\sqrt{K}\log(T)\right)$. When $\beta = 1$, then the regret is $R_T = O\left(\left(\frac{1}{\Delta_*}\right)^{\frac{1}{3}}\left(Ld\sqrt{K}\log(T)\right)^{\frac{2}{3}}T^{\frac{1}{3}}\right)$.

Lastly, we derive a regret upper bound in a linear contextual adversarial regime with a self-bounding constraint, which is a generalization of the stochastic and adversarial regimes under $\beta = \infty$ in a margin condition. We provide the proof in Appendix B.7.

Theorem 4.4 (Regret upper bound in an adversarial regime with a self-bounding constraint)**.** *Consider the 1/2-LC-Tsallis-INF. Assume that the loss is generated under a linear contextual adversarial regime with a self-bounding constraint (Definition 2.6). Suppose that Assumption 2.2–2.1 hold. Then,*

$$R_T = O\left(\frac{Ld\sqrt{K}}{\Delta_*}\log(T) + \kappa + C\right)$$

holds. Moreover, for $\frac{dL\sqrt{K}}{\Delta_*}\left(\log\left(\frac{T\Delta_*^2}{L^2d\sqrt{K}}\right) + 1\right) \leq C \leq \frac{\Delta_*T}{L}$, *the regret satisfies*

$$R_T = O\left(\sqrt{\frac{CL\sqrt{K}d}{\Delta_*}\log_+\left(\frac{\Delta_*T}{CL}\right)} + \kappa\right).$$

Theorem 4.4 implies that our algorithm achieves a tight regret bound of order $\log_+(T)$ with respect to T. Recall that $\log_+(x) = \max\{1, \log(x)\}$.

4.2 Regret Transformation

This study considers a setting where arm-dependent features are observable, while there are also studies investigating arm-independent features (Neu & Olkhovskaya, 2020; Kuroki et al., 2024; Kato & Ito, 2023). We show that (i) a regret derived under either of the problem settings can be transformed into a regret under the other setting, and (ii) transforming a regret under arm-dependent features to one under arm-independent features results in tighter bounds.

Here, we illustrate how to transform the setting with arm-independent features into one with arm-dependent features. We consider the former problems with $\widetilde{d}$-dimensional arm-independent features, $\widetilde{K}$ arms, and $\widetilde{T}$ rounds, denoted by P.Indep $\left(\widetilde{d}, \widetilde{K}, \widetilde{T}\right)$. We also consider the latter problem with d-dimensional features, K arms, and T rounds, denoted by P.Dep (d, K, T).

In the P.Indep $\left(\widetilde{d}, \widetilde{K}, \widetilde{T}\right)$, we observe $\widetilde{d}$-dimensional features $\widetilde{\phi}(X_t)$ at each $t \in [T]$, and the loss follows a linear model $\widetilde{\ell}_t(a, X_t) = \left\langle\widetilde{\theta}_{a,t}, \widetilde{\phi}(X_t)\right\rangle + \widetilde{\varepsilon}_t$, where $\widetilde{\varepsilon}_t$ is an error term, $\widetilde{\phi} : \mathcal{X} \to \mathbb{R}^d$, and $\widetilde{\theta}_{a,t} \in \mathbb{R}^{\widetilde{d}}$. This problem can be transformed into P.Dep $\left(\widetilde{d}\widetilde{K}, \widetilde{K}, \widetilde{T}\right)$ by considering arm-dependent features

$$\phi(a, X_t) = \left(\mathbf{0}_d^\top \cdots \widetilde{\phi}^\top(X_t) \cdots \mathbf{0}_d^\top\right)^\top$$

and the parameters $\theta_t = \left(\widetilde{\theta}_{1,t}^\top \cdots \widetilde{\theta}_{a,t}^\top \cdots \widetilde{\theta}_{K,t}^\top\right)^\top$, where $\mathbf{0}_d$ is the d-dimensional zero vector. Here, the loss follows as $\ell_t(a, X_t) = \langle\theta_t, \phi(a, X_t)\rangle + \varepsilon_t$. Similarly, we can transform the P.Dep (d, K, T) into the P.Indep (dK, K, T).

Based on this fact, we have the following theorem.

Theorem 4.5. *If there exists an algorithm whose regret is $R_T = f(d, K, T)$ in the P.Dep (d, K, T), where $f : \mathbb{N} \times \mathbb{N} \times \mathbb{N} \to \mathbb{R}$, then there exists an algorithm whose regret is given as $\widetilde{R}_T = f(\widetilde{d}\widetilde{K}, \widetilde{K}, \widetilde{T})$ in the P.Indep $\left(\widetilde{d}, \widetilde{K}, \widetilde{T}\right)$. Conversely, if there exists an algorithm whose regret is given as $\widetilde{R}_T = \widetilde{f}\left(\widetilde{d}, \widetilde{K}, \widetilde{T}\right)$ in*

the P.Indep $\left(\widetilde{d}, \widetilde{K}, \widetilde{T}\right)$, where $\widetilde{f} : \mathbb{N} \times \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ is some function of $\widetilde{d}$, $\widetilde{K}$, and $\widetilde{T}$, then there exists an algorithm whose regret is given as $R_T = \widetilde{f}(dK, K, T)$ in the P.Dep (d, K, T).

By using Theorem 4.5, we can transform regrets derived under each problem setting. This implies that if we derive a regret under one setting, we can obtain a regret under the other setting.

However, we note that we can derive tighter regrets by transforming a regret under the arm-dependent feature setting to a regret under the arm-independent feature setting. We confirm this by examining the regrets of the FTRL with Shannon entropy in the stochastic regime.

In the P.Indep $\left(\widetilde{d}, \widetilde{K}, \widetilde{T}\right)$, Kuroki et al. (2024) and Kato & Ito (2023) show that the FTRL with Shannon entropy incurs $O\left(\frac{\widetilde{K}}{\Delta_*}\left(\frac{1}{\lambda} + \widetilde{d}\right)\log(\widetilde{T})\log\left(\widetilde{K}\widetilde{T}\right)\right)$ regret, where we changed $\frac{\log(\widetilde{T})}{\lambda}$ to $\frac{1}{\lambda}$ since Σ_t^{-1} is assumed to be exactly computable. As discussed above, this algorithm incurs $O\left(\frac{K}{\Delta_*}\left(\frac{1}{\lambda} + dK\right)\log(T)\log(KT)\right)$ regret in the P.Dep (d, K, T).

In contrast, in the P.Dep (d, K, T), our Theorem D.4 shows that the FTRL with Shannon entropy incurs $O\left(\frac{1}{\Delta_*}\left(\frac{1}{\lambda} + d\right)\log(T)\log(KT)\right)$ regret, which implies that it incurs $O\left(\frac{1}{\Delta_*}\left(\frac{1}{\lambda} + \widetilde{d}\widetilde{K}\right)\log(\widetilde{T})\log(\widetilde{K}\widetilde{T})\right)$ regret in the P.Indep $\left(\widetilde{d}, \widetilde{K}, \widetilde{T}\right)$.

By comparing the results, we find that we can obtain a tighter upper bound when transforming a regret under the arm-dependent setting to a regret under the arm-independent setting. In fact, if we transform a regret under the arm-independent setting to a regret under the arm-dependent setting, the order of the regret becomes $O(K^2)$ regarding K, which is $O(1)$ in our Theorem D.4.

We present the proof of Theorem 4.5 and more detailed arguments in Appendix G.

4.3 Discussion

Here, we discuss related topics.

Time complexity of solving the optimization problem (4) The solution of (4) can be efficiently computed using the method described in Section 3.3 of Zimmert & Seldin (2021). As mentioned there, it reduces the problem of solving a univariate equation, to which Newton's method can be applied. This method has quadratic convergence, typically requiring fewer than ten iterations to converge with double precision.

Comparison with a simple baseline method For comparison, we consider a baseline method where we run $|\mathcal{X}|$ independent 1/2-Tsallis-INF algorithms, originally proposed for multi-armed bandits without contexts by Zimmert & Seldin (2021), in each context. A similar analysis can be found in Section 18.1 of Lattimore & Szepesvári (2020). In round T, for each context $x \in \mathcal{X}$, we define T_x as the number of times that context x is observed. From the results in Zimmert & Seldin (2021), the regret of the 1/2-Tsallis-INF for each context is given as $O\left(\frac{1}{\Delta_*}K\log(T_x)\right)$ in the stochastic regime and $O\left(\sqrt{KT_x}\right)$ in the adversarial regime. Therefore, if we marginalize the context-wise regret, the total regret is $\sum_{x\in\mathcal{X}} O\left(\frac{1}{\Delta_*}K\log(T_x)\right) = O\left(\frac{1}{\Delta_*}SK\log(T/S)\right)$ in the stochastic regime and $\sum_{x\in\mathcal{X}} O\left(\sqrt{KT_x}\right) = O\left(\sqrt{SKT}\right)$ in the adversarial regime, where $S = |\mathcal{X}|$ and $L \geq S$. In contrast, our 1/2-LC-Tsallis-INF incurs $O\left(\frac{1}{\Delta_*}L\sqrt{K}\max\left\{d, \frac{1}{\lambda d}\right\}\log(T)\right)$ regret in the stochastic regime and $O\left(\sqrt{d\sqrt{K}T}\right)$ regret in the adversarial regime. Comparing these results, we find that the dependency on K is improved in our 1/2-LC-Tsallis-INF in both regimes, while the dependency on L and S $(L \geq S)$ is worse than that of the baseline method. Additionally, our 1/2-LC-Tsallis-INF depends on the feature dimension d, which does not appear in the baseline method.

Our setting can be reduced to a tabular Markov decision process (MDP). However, employing this reduction overlooks the linear structure in the loss function, ultimately compromising efficiency, particularly when the number of actions K is large. More specifically, in the stochastic setting, the upper bound derived under the tabular MDP interpretation is roughly $O\left(\frac{KI\log T}{\Delta_{\min}}\right)$, corresponding to replacing $d\sqrt{K}$ in our upper bound with K, where I is the number of interactions within each episode (Jin & Luo, 2020; Dann et al., 2023b). In the adversarial setting, the upper bound becomes $O(\sqrt{KIT})$, corresponding to replacing $d\sqrt{K}$ in our bound with KI.

5 Conclusion

We presented a BoBW algorithm for linear contextual bandits with regret upper bounds that tightly depend on T. Our proposed algorithm, the α-LC-Tsallis-INF, employs FTRL with Tsallis entropy regularization and achieves $O(\log(T))$ regret in the stochastic regime. Additionally, we derived regret upper bounds under the margin condition and the arm-dependent setting. An important remaining challenge is to improve the dependence on T when $\beta \in (0, \infty)$ in the stochastic regime.

References

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

Abe, N. and Long, P. M. Associative reinforcement learning using linear probabilistic concepts. In *International Conference on Machine Learning (ICML)*, 1999.

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

Bubeck, S. and Slivkins, A. The best of both worlds: Stochastic and adversarial bandits. In *Conference on Learning Theory (COLT)*, 2012.

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

Dann, C., Wei, C.-Y., and Zimmert, J. A blackbox approach to best of both worlds in bandits and beyond. In Neu, G. and Rosasco, L. (eds.), *Conference on Learning Theory (COLT)*, volume 195, pp. 5503–5570. PMLR, 2023a.

Dann, C., Wei, C.-Y., and Zimmert, J. Best of both worlds policy optimization. In *International Conference on Machine Learning (ICML)*, 2023b.

Hazan, E., Koren, T., Livni, R., and Mansour, Y. Online learning with low rank experts. In *Conference on Learning Theory (COLT)*, 2016.

Ito, S., Tsuchiya, T., and Honda, J. Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jin, T. and Luo, H. Simultaneously learning stochastic and adversarial episodic mdps with known transition. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 16557–16566, 2020.

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. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Kanade, V. and Steinke, T. Learning hurdles for sleeping experts. *ACM Transactions on Computation Theory*, 6(3), 2014.

Kato, M. and Ito, S. Best-of-both-worlds linear contextual bandits, 2023. arXiv:2312.16489.

Kong, F., Zhao, C., and Li, S. Best-of-three-worlds analysis for linear bandits with follow-the-regularized-leader algorithm. In *Conference on Learning Theory (COLT)*, 2023.

Kuroki, Y., Rumi, A., Tsuchiya, T., Vitale, F., and Cesa-Bianchi, N. Best-of-both-worlds algorithms for linear contextual bandits. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 238, pp. 1216–1224. PMLR, 2024.

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

Li, K., Yang, Y., and Narisetty, N. N. Regret lower bound and optimal algorithm for high-dimensional contextual linear bandit. *Electronic Journal of Statistics*, 15(2):5652 – 5695, 2021.

Li, L., Chu, W., Langford, J., and Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. In *International Conference on World Wide Web (WWW)*, 2010.

Liu, H., Wei, C.-Y., and Zimmert, J. Bypassing the simulator: Near-optimal adversarial linear contextual bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Masoudian, S. and Seldin, Y. Improved analysis of the tsallis-INF algorithm in stochastically constrained adversarial bandits and stochastic bandits with adversarial corruptions. In *Conference on Learning Theory (COLT)*, 2021.

Neu, G. and Olkhovskaya, J. Efficient and robust algorithms for adversarial linear contextual bandits. In *Conference on Learning Theory (COLT)*, 2020.

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

Rakhlin, A. and Sridharan, K. Bistro: An efficient relaxation-based method for contextual bandits. In *International Conference on Machine Learning (ICML)*, 2016.

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

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

Syrgkanis, V., Luo, H., Krishnamurthy, A., and Schapire, R. E. Improved regret bounds for oracle-based adversarial contextual bandits. In *Advances in*

Neural Information Processing Systems (NeurIPS), 2016.

Tewari, A. and Murphy, S. A. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health: Sensors, Analytic Methods, and Applications*, pp. 495–517, 2017.

Tsallis, C. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52:479–487, 1988.

Tsuchiya, T., Ito, S., and Honda, J. Best-of-both-worlds algorithms for partial monitoring. In *International Conference on Algorithmic Learning Theory (ALT)*, 2023a.

Tsuchiya, T., Ito, S., and Honda, J. Stability-penalty-adaptive follow-the-regularized-leader: Sparsity, game-dependency, and best-of-both-worlds. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023b.

Wei, C.-Y. and Luo, H. More adaptive algorithms for adversarial bandits. In *Conference on Learning Theory (COLT)*, 2018.

Zimmert, J. and Seldin, Y. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(1), 2021.

Zimmert, J., Luo, H., and Wei, C.-Y. Beating stochastic and adversarial semi-bandits optimally and simultaneously. In *International Conference on Machine Learning*, pp. 7683–7692. PMLR, 2019.

Checklist

1. For all models and algorithms presented, check if you include:

    (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]

    (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]

    (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Not Applicable]

2. For any theoretical claim, check if you include:

    (a) Statements of the full set of assumptions of all theoretical results. [Yes]

    (b) Complete proofs of all theoretical results. [Yes]

    (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:

    (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Not Applicable]

    (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Not Applicable]

    (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Not Applicable]

    (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Not Applicable]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:

    (a) Citations of the creator If your work uses existing assets. Not Applicable]

    (b) The license information of the assets, if applicable. [Not Applicable]

    (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]

    (d) Information about consent from data providers/curators. [Not Applicable]

    (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

    (a) The full text of instructions given to participants and screenshots. [Not Applicable]

    (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

    (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Contents

A   Preliminaries for the proof of Theorems 4.1–4.4

This section provides preliminary results for the proof of Theorems 4.1–4.4. We prove Theorems 4.1–4.4 in Appendix B.

A.1   Upper bound by Bregman divergence

Let X_0 be a sample from the context distribution $\mathcal{G}$ independent of $\mathcal{F}_T$. Let $D(p, q)$ denote a Bregman divergence between $p.q \in \Pi$ with respect to ψ_t, defined as

$$D(p, q) := \psi(p) - \psi(q) - \left\langle \nabla \psi(q), p - q \right\rangle.$$

In our proof, the following proposition plays an essential role, which is inspired by Lemma 4.4 in Kato & Ito (2023).

Proposition A.1. *Consider the 1/2-LC-Tsallis-INF with our defined parameters. Then, the regret satisfies*

$$R_T \le \mathbb{E}\left[\sum_{t=1}^{T} \left(\gamma_t + \left\langle \widehat{\ell}_t(X_0), q_t(X_0) - q_{t+1}(X_0) \right\rangle - \frac{1}{\eta_t} D\big(q_{t+1}(X_0), q_t(X_0)\big) - \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) H\big(q_t(X_0)\big) \right) \right],$$

where we define

$$H\big(q(x)\big) := 2 \left(-\sum_{a \in [K]} \sqrt{q(a \mid x)} + 1 \right)$$

and define $1/\eta_0 = 0$ as an exception.

We show the proof below.

Proof of Proposition A.1. Let us define an optimal policy $\pi^* \in \Pi$ as $\pi^*(\rho_T^*(x) \mid x) = 1$ and $\pi^*(a \mid x) = 0$ for all $a \in [K]\backslash\{\rho_T^*(x)\}$ and for all $x \in \mathcal{X}$.

Recall that in (4), we defined q_t as

$$q_t(x) := \underset{q \in \mathcal{P}_K}{\arg\min} \left\{ \sum_{s=1}^{t-1} \left\langle \widehat{\ell}_s(x), q \right\rangle + \frac{1}{\eta_t} \psi(q) \right\}$$

for all $t \ge 2$.

From the definition of our algorithm, we have

$$
\begin{aligned}
R_T &= \mathbb{E}\left[\sum_{t=1}^{T}\left(\ell_t(A_t, X_t) - \ell_t\left(\rho_T^*(X_t), X_t\right)\right)\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\langle \ell_t(X_t), \pi_t(X_t) - \pi^*(X_t)\rangle\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\langle \ell_t(X_t), q_t(X_t) - \pi^*(X_t)\rangle + \sum_{t=1}^{T}\gamma_t\langle \ell_t(X_t), e^*(X_t) - q_t(X_t)\rangle\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\left\langle \ell_t(X_t), q_t(X_t) - \pi^*(X_t)\right\rangle + \sum_{t=1}^{T}\gamma_t\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\left\langle \ell_t(X_0), q_t(X_0) - \pi^*(X_0)\right\rangle + \sum_{t=1}^{T}\gamma_t\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\left\langle \widehat{\ell}_t(X_0), q_t(X_0) - \pi^*(X_0)\right\rangle + \sum_{t=1}^{T}\gamma_t\right].
\end{aligned}
\tag{5}
$$

Next, we show that for all $x \in \mathcal{X}$ and any $p^* \in \mathcal{P}_K$, it holds that

$$
\begin{aligned}
&\sum_{t=1}^{T}\left\langle \widehat{\ell}_t(x), q_t(x) - p^*(x)\right\rangle \\
&\leq \sum_{t=1}^{T}\left(\left\langle \widehat{\ell}_t(x), q_t(x) - q_{t+1}(x)\right\rangle - \frac{1}{\eta_t}D\big(q_{t+1}(x), q_t(x)\big) - \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)H\big(q_t(x)\big)\right).
\end{aligned}
\tag{6}
$$

This result follows from the definition of q_t; that is,

$$
\begin{aligned}
&\left\langle \sum_{t=1}^{T}\widehat{\ell}_t(x), p^*(x)\right\rangle + \frac{1}{\eta_T}\psi\left(p^*(x)\right) \\
&\geq \left\langle \sum_{t=1}^{T}\widehat{\ell}_t(x), q_{T+1}(x)\right\rangle + \frac{1}{\eta_T}\psi\left(q_{T+1}(x)\right) \\
&\geq \left\langle \sum_{t=1}^{T-1}\widehat{\ell}_t(x), q_{T+1}(x)\right\rangle + \left\langle \widehat{\ell}_T(x), q_{T+1}(x)\right\rangle + \frac{1}{\eta_T}\psi\left(q_{T+1}(x)\right) \\
&\geq \left\langle \sum_{t=1}^{T-1}\widehat{\ell}_t(x), q_T(x)\right\rangle + \left\langle \widehat{\ell}_T(x), q_{T+1}(x)\right\rangle + \frac{1}{\eta_T}\psi\left(q_T(x)\right) + \frac{1}{\eta_T}D\left(q_{T+1}, q_T\right) \\
&\geq \sum_{t=1}^{T}\left(\left\langle \widehat{\ell}_t(x), q_{t+1}(x)\right\rangle + \left(\frac{1}{\eta_{t-1}} - \frac{1}{\eta_t}\right)\psi\left(q_t(x)\right) + \frac{1}{\eta_t}D\left(q_{t+1}, q_t\right)\right).
\end{aligned}
$$

Here, we defined $H\left(q_t(x)\right)$ as $\psi\left(q_t(x)\right) = H\left(q_t(x)\right)$ under the 1/2-LC-Tsallis-INF. Combining (6) with (5) yields the statement. $\qquad\square$

A.2 Bregman divergence associated with the Tsallis entropy

The Bregman divergence associated with the α-Tsallis entropy

$$
\psi(q(x)) = -\frac{1}{\alpha}\left(\sum_{a \in [K]}q_t(a \mid x)^{\alpha} + 1\right)
$$

is given as

$$D\big(p(x), q(x)\big) = \frac{1}{\alpha} \sum_{a \in [K]} \big(q(a \mid x)^\alpha + \alpha\big(p(a \mid x) - q(a \mid x)\big)q(a \mid x)^{\alpha-1} - p(a \mid x)^\alpha\big)$$

$$= \sum_{a \in [K]} d\big(p(a \mid x), q(a \mid x)\big),$$

where $d(p, q)$ is defined as

$$d(p, q) := \frac{1}{\alpha}q^\alpha + (p - q)q^{\alpha-1} - \frac{1}{\alpha}p^\alpha \le \frac{1-\alpha}{2}\big(\min\{p, q\}\big)^{\alpha-2}(p - q)^2$$

for all $p, q \in (0, 1)$.

A.3 Upper bound of $\left|\widehat{\ell}_t(a, x)\right|$

For any $d \times d$ real symmetric matrices A and B, denote $A \succeq B$ if and only if $A - B$ is positive-semidefinite and $A \succ B$ if and only if $A - B$ is positive-definite.

Lemma A.2. *For any positive-definite matrix $A \in \mathbb{R}^{d \times d}$ and a real vector $x \in \mathbb{R}^d$, if $A \succeq xx^\top$, it holds that $xA^{-1}x \le 1$.*

Proof. From the assumption of $A \succeq xx^\top$, we have

$$x^\top A^{-1}x = (A^{-1}x)^\top A(A^{-1}x) \ge (A^{-1}x)^\top xx^\top(A^{-1}x) = (x^\top A^{-1}x)^2,$$

which implies $x^\top A^{-1}x \le 1$. □

Lemma A.3. *Suppose that Assumptions 2.2–2.1 hold. Then, for all $x \in \mathcal{X}$ and $a \in [K]$, it holds that*

$$\left|\widehat{\ell}_t(a, x)\right| \le \min\left\{\frac{1}{\lambda\gamma_t}, \sqrt{\frac{1}{\lambda\gamma_t(1 - \gamma_t)q_t(a \mid x)g(x)}}\right\}. \tag{7}$$

Proof. From Assumption 2.1 and the definition of $\widehat{\ell}_t$ in (2), we have

$$\left|\widehat{\ell}_t(a, x)\right| = \left|\left\langle\widehat{\theta}_t, \phi(a, x)\right\rangle\right| \le \left|(\phi(a, x))^\top \Sigma_t^{-1}\phi(A_t, X_t)\right|$$

$$\le \sqrt{(\phi(a, x))^\top \Sigma_t^{-1}\phi(a, x) \cdot (\phi(A_t, X_t))^\top \Sigma_t^{-1}\phi(A_t, X_t)}. \tag{8}$$

From Assumption 2.4, for any $z \in \mathcal{Z}$, we have

$$z^\top \Sigma_t^{-1}z \le z^\top \big(\gamma_t\Sigma(e^*)\big)^{-1} z \le \frac{1}{\gamma_t\lambda}. \tag{9}$$

In addition, as we have

$$\Sigma_t \succeq (1 - \gamma_t)\Sigma(q_t) = (1 - \gamma_t) \sum_{x' \in \mathcal{X}} g(x') \sum_{a' \in [K]} q_t(a' \mid x')\phi(a', x')(\phi(a', x'))^\top$$

$$\succeq (1 - \gamma_t)g(x)q_t(a \mid x)\phi(a, x)(\phi(a, x))^\top,$$

Lemma A.2 yields

$$(\phi(a, x))^\top \Sigma_t^{-1}\phi(a, x) \le \frac{1}{(1 - \gamma_t)q_t(a \mid x)g(x)}. \tag{10}$$

Combining (8), (9), and (10), we obtain (7). □

A.4   Upper bound of $\mathbb{E}\left[\pi_t(a \mid X_0)\widehat{\ell}_t(a \mid X_0)^2 \mid \mathcal{F}_{t-1}\right]$

We upper bound $\mathbb{E}\left[\pi_t(a \mid X_0)\widehat{\ell}_t(a \mid X_0)^2 \mid \mathcal{F}_{t-1}\right]$. This result is inspired by Lemma 6 in Neu & Olkhovskaya (2020).

Lemma A.4. *For all $t \in [T]$ and all $a \in [K]$, our policy satisfies*

$$\mathbb{E}\left[\sum_{a\in[K]} \pi_t(a \mid X_0)\widehat{\ell}_t(a, X_0)^2 \mid \mathcal{F}_{t-1}\right] \leq d.$$

Proof. We can show the statement as follows:

$$
\begin{aligned}
&\mathbb{E}\left[\sum_{a\in[K]} \pi_t(a \mid X_0)\hat{\ell}_t(a, X_0)^2 \mid \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\sum_{a\in[K]} \pi_t(a \mid X_0)\hat{\theta}_t^\top \phi(a, X_0)\phi^\top(a, X_0)\hat{\theta}_t \mid \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\hat{\theta}_t^\top \Sigma_t \hat{\theta}_t \mid \mathcal{F}_{t-1}\right] \\
&\leq \mathbb{E}\left[\ell_t(A_t, X_t)\phi^\top(A_t, X_t)\Sigma_t^{-1}\Sigma_t\Sigma_t^{-1}\phi(A_t, X_t)\ell_t(A_t, X_t) \mid \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\phi^\top(A_t, X_t)\Sigma_t^{-1}\Sigma_t\Sigma_t^{-1}\phi(A_t, X_t) \mid \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\mathrm{tr}(\phi^\top(A_t, X_t)\Sigma_t^{-1}\Sigma_t\Sigma_t^{-1}\phi(A_t, X_t)) \mid \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\mathrm{tr}(\Sigma_t^{-1}\phi(A_t, X_t)\phi^\top(A_t, X_t)) \mid \mathcal{F}_{t-1}\right] \\
&= \mathbb{E}\left[\mathrm{tr}(\Sigma_t^{-1}\Sigma_t) \mid \mathcal{F}_{t-1}\right] \\
&= d.
\end{aligned}
$$

$\square$

A.5   Margin condition

To utilize a margin condition for deriving a regret upper bound, we show the following lemma.

Lemma A.5. *Let $U \in [0, 1]$ and $V \in \mathbb{R}$ be some random variables. Let $\Delta_* > 0$ be some universal constant. For $\beta \in (0, \infty]$, if the random variable U has a mean $\mathbb{E}[U] = \mu$ and the random variable V satisfies*

$$F(h) := \mathbb{P}\big(V \leq h\big) \leq \frac{1}{2}\left(\frac{h}{\Delta_*}\right)^\beta$$

for all $h \in [0, \Delta_*]$, and $\mu \in \left[0, \frac{1}{2}\right]$ holds, then

$$\mathbb{E}\big[UV\big] \geq \frac{\Delta_*\beta}{2(1+\beta)}(2\mu)^{\frac{1+\beta}{\beta}}$$

holds.

Proof. For simplicity, let us assume that V has a density function. This assumption implies that the cumulative density function of V is a continuous and monotonically increasing function. Then, $Z := F(V)$ follows the uniform distribution $\mathrm{Unif}[0, 1]$ over the support $[0, 1]$.

Let us define

$$G(z) := \mathbb{E}[U \mid Z = z].$$

Then, we have $G(z) \in [0, 1]$, and the expected values of U and UV are given as follows:

$$\mathbb{E}\big[U\big] = \mathop{\mathbb{E}}_{W \sim \mathrm{Unif}[0,1]}[G(W)] = \int_0^1 g(W)\mathrm{d}w = \mu, \tag{11}$$

$$\mathbb{E}\big[UV\big] = \mathop{\mathbb{E}}_{W \sim \mathrm{Unif}[0,1]}[G(W)F^{-1}(W)] = \int_0^1 g(w)F^{-1}(w)\mathrm{d}w, \tag{12}$$

where $\mathbb{E}_{W \sim \mathcal{J}}$ denotes an expectation operator for a random variable W under a probability distribution $\mathcal{J}$.

Here, note that F^{-1} is a monotonically decreasing function. Therefore, for $G(w) \in [0,1]$, under (11), a function G^* minimizes (12) if $G^*(w) \coloneqq \mathbb{1}[w \leq \mu]$.

In conclusion, we obtain

$$\mathbb{E}[UV] \geq \int_0^1 G^*(w)F^{-1}(w)\mathrm{d}z = \int_0^\mu F^{-1}(w)\mathrm{d}w. \tag{13}$$

Furthermore, because we assumed $F(h) \leq \frac{1}{2}\left(\frac{h}{\Delta_*}\right)^\beta$ for all $h \in [D, \Delta_*]$, we have $F^{-1}(w) \geq \Delta_*(2w)^{\frac{1}{\beta}}$ for $w \in [F(D), F(\Delta_*)] = \left[\frac{1}{2}\frac{D}{\Delta_*}, \frac{1}{2}\right]$, which implies

$$\int_0^\mu F^{-1}(w)\mathrm{d}w \geq \int_0^\mu \Delta_*(2w)^{\frac{1}{\beta}}\,\mathrm{d}w \geq \int_{\frac{1}{2}\frac{D}{\Delta_*}}^\mu \Delta_*(2w)^{\frac{1}{\beta}}\,\mathrm{d}w = \frac{\Delta_*}{2(1+\beta)}(2\mu)^{\frac{1+\beta}{\beta}}, \tag{14}$$

where we used $0 \leq \mu \leq \frac{1}{2}$. From (13) and (14), we obtain the statement. $\qquad\square$

B Proof of Theorems 4.1–4.4

This section provides the proofs for Theorem 4.1. In Appendix A, we provide preliminary results for the proof. Then, in Appendix B.1, we decompose the regret into the stability and penalty terms. Lastly, in Appendix B.4, we prove Theorem 4.1.

B.1 Stability and penalty decomposition

Following the standard analysis of FTRL methods, we decompose the regret into stability and penalty terms.

Based on the result in Proposition A.1, let us define $r_T(x)$ as

$$r_T(x) \coloneqq \sum_{t=1}^T \left(\gamma_t + \left\langle \widehat{\ell}_t(x), q_t(x) - q_{t+1}(x) \right\rangle - \frac{1}{\eta_t}D(q_{t+1}(x), q_t(x)) - \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)H\big(q_t(x)\big) \right).$$

Using this function $r_T(x)$, we can bound R_T as $R_T \leq \mathbb{E}\big[r_T(X_0)\big]$.

The standard FTRL analysis breaks the pseudo-regret into penalty and stability terms. Following this approach, we decompose the pointwise regret upper bound $r_T(x)$ as follows:

$$r_T(x) = \underbrace{\sum_{t=1}^T \left(\gamma_t + \left\langle \widehat{\ell}_t(x), q_t(x) - q_{t+1}(x) \right\rangle - \frac{1}{\eta_t}D(q_{t+1}(x), q_t(x)) \right)}_{= \text{ stability term}}$$

$$+ \underbrace{\left(-\sum_{t=1}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)H\big(q_t(x)\big) \right)}_{= \text{ penalty term}}. \tag{15}$$

Our remaining task is to derive upper bounds for the following terms and for all $x \in \mathcal{X}$:

$$\text{stability}_t(x) \coloneqq \gamma_t + \left\langle \widehat{\ell}_t(x), q_t(x) - q_{t+1}(x) \right\rangle - \frac{1}{\eta_t}D(q_{t+1}(x), q_t(x)), \tag{16}$$

$$\text{penalty}_t(x) \coloneqq -\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)H\big(q_t(x)\big), \tag{17}$$

In Appendices B.2– B.3, we bound the stability and penalty terms, respectively.

B.2 Bounding the stability term

To bound the stability term, we obtain the following lemma. Recall that for all $x \in \mathcal{X}$ and $q_t(x) \in \mathcal{P}_K$, we defined $a_t^\dagger(x) \in \arg\max_{a \in [K]} q_t(a \mid x)$.

Lemma B.1. *Consider the α-LC-Tsallis-INF. Suppose that Assumptions [2.1]–[2.2], and [2.4] hold. If*

$$\eta_t |\widehat{\ell}_t(a, x)| \leq \frac{1 - \alpha}{4} \left(q_t(a \mid x) \right)^{\alpha - 1}$$

holds for all $a \in [K]$ and for all $x \in \mathcal{X}$, we then have

$$\left\langle \widehat{\ell}_t(x), q_t(x) - q_{t+1}(x) \right\rangle - \frac{1}{\eta_t} D(q_{t+1}(x), q_t(x))$$

$$\leq \frac{4\eta_t}{1 - \alpha} \left(\sum_{a \neq a_t^\dagger(x)} (q_t(a \mid x))^{2-\alpha} \ell_t^2(a \mid x) + \left(\min\left\{ q_t(a_t^\dagger(x) \mid x), 1 - q_t(a_t^\dagger(x) \mid x) \right\} \right)^{2-\alpha} \ell_t^2(a_t^\dagger(x), x) \right).$$

The proof is shown in Appendix [C.1]. By using this lemma, we show an upper bound for the stability term. To provide it, we introduce the following notations: for each $t \in [T]$,

$$a_t^\dagger(x) := \arg\max_{a \in [K]} q_t(a \mid x),$$

$$\omega_t := \sup_{x \in \mathcal{X}} \min_{a \in [K]} \left\{ 1 - q_t(a \mid x) \right\} = \sup_{x \in \mathcal{X}} \left\{ 1 - q_t\left(a_t^\dagger(x) \mid x \right) \right\}.$$

Lemma B.2 (Upper bound for the stability term). *Consider the $1/2$-LC-Tsallis-INF. Assumptions [2.1]–[2.2], and [2.4] hold. It holds for all $x \in \mathcal{X}$ and $t \in [T]$ that*

$$\mathrm{stability}_t(x) \leq 2\eta_t \sqrt{\omega_t} \sum_{a \in [K]} \widehat{\ell}_t(a \mid x)^2 \pi_t(a \mid x).$$

Proof. We first check the condition of Lemma [B.1]. From Lemma [A.3] and the definition of γ_t in [(4)], we have

$$\eta_t \left| \widehat{\ell}_t(a, x) \right| \leq \eta_t \sqrt{\frac{1}{\lambda \gamma_t (1 - \gamma_t) q_t(a \mid x) g(x)}} \leq \eta_t \sqrt{\frac{2L}{\lambda \gamma_t q_t(a \mid x)}} \leq \frac{1}{8} \left(q_t(a \mid x) \right)^{-1/2}.$$

Hence, for any $x \in \mathcal{X}$ we can apply Lemma [B.1] with $\alpha = 1/2$ to obtain the following:

$$\left\langle \widehat{\ell}_t(x), q_t(x) - q_{t+1}(x) \right\rangle - \frac{1}{\eta_t} D(q_{t+1}(x), q_t(x))$$

$$= \frac{1}{\eta_t} \left(\left\langle \eta_t \widehat{\ell}_t(x), q_t(x) - q_{t+1}(x) \right\rangle - D(q_{t+1}(x), q_t(x)) \right)$$

$$\leq \eta_t \left(\sum_{a \neq a_t^\dagger(x)} \left(\widehat{\ell}_t(a \mid x) \right)^2 (q_t(a \mid x))^{3/2} + \left(\widehat{\ell}_t(a_t^\dagger(x) \mid x) \right)^2 \min\left\{ q_t(a_t^\dagger(x) \mid x), 1 - q_t(a_t^\dagger(x) \mid x) \right\}^{3/2} \right)$$

$$\leq \eta_t \left(\sum_{a \neq a_t^\dagger(x)} \left(\widehat{\ell}_t(a \mid x) \right)^2 q_t(a \mid x) \sqrt{\omega_t} + \left(\widehat{\ell}_t(a_t^\dagger(x) \mid x) \right)^2 q_t\left(a_t^\dagger(x) \right) \left(1 - q_t(a_t^\dagger(x) \mid x) \right)^{1/2} \right)$$

$$\leq \sqrt{\omega_t} \eta_t \sum_{a \in [K]} \widehat{\ell}_t^2(a \mid x) q_t(a \mid x) \leq 2\sqrt{\omega_t} \eta_t \sum_{a \in [K]} \widehat{\ell}_t^2(a \mid x) \pi_t(a \mid x),$$

where the second inequality follows from $q_t(a \mid x) \leq 1 - q_t(a_t^\dagger(x) \mid x) \leq \omega_t$ that holds for all $x \in \mathcal{X}$ and $a \in [K] \setminus \{a_t^\dagger(x)\}$, and the last inequality follows from $\pi_t \geq (1 - \gamma_t) q_t \geq q_t/2$.

This completes the proof. $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

B.3 Bounding the penalty term

Next, we bound $\mathrm{penalty}_t(x)$.

Lemma B.3 (Upper bound of the penalty term). *For all $\alpha \in [0, 1]$, the penalty term of α-LC-Tsallis-INF satisfies*

$$\mathrm{penalty}_t(x) = -\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) H\big(q_t(x)\big) \leq 2\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\sqrt{K\omega_t},$$

where recall that the penalty term $\mathrm{penalty}_t(x)$ is defined in (17).

Proof. This statement directly follows from the following inequality: for all $x \in \mathcal{X}$,

$$
\begin{aligned}
-H(q_t(x)) &= 2\left(\sum_{a \in [K]} \sqrt{q_t(a \mid x)} - 1\right) \\
&\leq 2\sum_{a \neq a^\dagger(x)} \sqrt{q_t(a \mid x)} \\
&\leq 2\sqrt{(K-1)\sum_{a \neq a^\dagger(x)} q_t(a \mid x)} \\
&= 2\sqrt{(K-1)(1 - q_t(a_t^\dagger(x) \mid x))} \leq 2\sqrt{K\omega_t}.
\end{aligned}
$$

$\square$

B.4 Proof of Theorem 4.1

Then, we prove Theorem 4.1 as follows.

Proof. From Lemma B.2, we have

$$\mathrm{stability}_t(x) \leq 2\eta_t\sqrt{\omega_t}\sum_{a \in [K]} \widehat{\ell}_t^2(a \mid x)\pi_t(a \mid x).$$

From Lemma B.3, we have

$$\mathrm{penalty}_t(x) \leq 2\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\sqrt{K\omega_t}.$$

Therefore, we can bound the pointwise regret as

$$r_T(x) \leq \sum_{t=1}^{T}\left(2\sqrt{\omega_t}\eta_t\sum_{a \in [K]}\widehat{\ell}_t^2(a \mid x)\pi_t(a \mid x) + 2\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\sqrt{K\omega_t} + \gamma_t\right).$$

We hence have

$$
\begin{aligned}
R_T &\leq \sum_{t=1}^{T}\mathbb{E}\left[2\sqrt{\omega_t}\eta_t\sum_{a \in [K]}\widehat{\ell}_t(a \mid X_0)^2\pi_t(a \mid X_0) + 2\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\sqrt{K\omega_t} + \gamma_t\right] \\
&\leq \sum_{t=1}^{T}\mathbb{E}\left[2d\sqrt{\omega_t}\eta_t + 2\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\sqrt{K\omega_t} + \gamma_t\right], \\
&\leq O\left(\mathbb{E}\left[\sum_{t=1}^{T}d\eta_t\sqrt{\omega_t} + \sum_{t=1}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\sqrt{K}\sqrt{\omega_t}\right] + \sum_{t=1}^{T}\gamma_t\right),
\end{aligned}
\tag{18}
$$

where the second inequality follows from Lemma A.4.

Recall here that when Assumption 2.3 holds, parameters η_t and γ_t are defined as

$$\widetilde{\eta}_t := \frac{K^{1/4}}{\sqrt{dt}}, \quad \eta_t := \min\left\{\widetilde{\eta}_t, \frac{1}{16}\sqrt{\frac{\lambda}{L}}\right\}, \quad \gamma_t := \frac{128L\eta_t^2}{\lambda} \leq \frac{128\sqrt{K}L}{\lambda dt}$$

for $t = 1, 2, \ldots$ and $1/\eta_0 = 0$. In this case, we have

$$\sum_{t=1}^{T} d\eta_t\sqrt{\omega_t} \leq \sum_{t=1}^{T} d\widetilde{\eta}_t\sqrt{\omega_t} = \sqrt{d}K^{1/4}\sum_{t=1}^{T}\frac{\sqrt{\omega_t}}{\sqrt{t}}, \tag{19}$$

and

$$\begin{aligned}
\sum_{t=1}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\sqrt{K}\sqrt{\omega_t} &\leq \sum_{t=2}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\sqrt{K}\sqrt{\omega_t} + \frac{1}{\eta_1}\sqrt{K}\omega_1 \\
&\leq \sum_{t=2}^{T}\left(\frac{1}{\widetilde{\eta}_t} - \frac{1}{\widetilde{\eta}_{t-1}}\right)\sqrt{K}\sqrt{\omega_t} + \max\left\{\frac{1}{\widetilde{\eta}_1}, 16\sqrt{\frac{L}{\lambda}}\right\}\sqrt{K}\omega_1 \\
&= O\left(\sqrt{d}K^{1/4}\sum_{t=1}^{T}\frac{\sqrt{\omega_t}}{\sqrt{t}} + \sqrt{\frac{KL}{\lambda}}\right).
\end{aligned} \tag{20}$$

It also holds that

$$\sum_{t=1}^{T}\gamma_t \leq \frac{128\sqrt{K}L}{\lambda d}\sum_{t=1}^{T}\frac{1}{t} = O\left(\frac{\sqrt{K}L}{\lambda d}\log T\right). \tag{21}$$

When Assumption 2.3 holds, by combining (18), (19), (20), and (21), we obtain

$$R_T = O\left(\mathbb{E}\left[\sum_{t=1}^{T}\frac{\sqrt{\sqrt{K}d\omega_t}}{\sqrt{t}}\right] + \sqrt{\frac{KL}{\lambda}} + \frac{\sqrt{K}L}{\lambda d}\log T\right).$$

When Assumption 2.3 does not hold, recall that η_t is defined as

$$\eta_t := \min\left\{\widetilde{\eta}_t, \frac{\lambda}{16}\right\}.$$

Using an argument similar to the case when Assumption 2.3 holds, we obtain

$$R_T = O\left(\mathbb{E}\left[\sum_{t=1}^{T}\frac{\sqrt{\sqrt{K}d\omega_t}}{\sqrt{t}}\right] + \frac{1}{\lambda}\left(\sqrt{K} + \frac{\sqrt{\sqrt{K}T}}{\sqrt{d}}\right)\right).$$

$\square$

B.5 Proof of Theorem 4.2

Proof. Recall that $\omega_t \leq 1$. By replacing ω_t with 1 in Theorem 4.1, we can directly obtain a regret upper bound in Theorem 4.2. $\square$

B.6 Proof of Theorem 4.3

Lemma B.4. *Under the margin condition given in Definition 2.5, we have*

$$\mathbb{E}\left[\sum_{t=1}^{T}\omega_t\right] \leq LT^{\frac{1}{1+\beta}}\left(\frac{2(1+\beta)R_T}{\beta\Delta^*}\right)^{\frac{\beta}{1+\beta}}. \tag{22}$$

Proof. Under the margin condition, we have

$$R_T \geq \mathbb{E}\left[\sum_{t=1}^{T} \Delta(X_t) \sum_{a \neq \rho_T^*(X_t)} \pi_t(a \mid X_t)\right].$$

From this, by using $a_t^\dagger$, the regret is lower bounded as

$$R_T \geq \mathbb{E}\left[\sum_{t=1}^{T} \Delta(X_t) \sum_{a \neq \rho_T^*(X_t)} \pi_t(a \mid X_t)\right] \geq \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T} \Delta(X_t) \sum_{a \neq \rho_T^*(X_t)} q_t(a \mid X_t)\right]$$

$$\geq \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T} \Delta(X_t)\big(1 - q_t(\rho_T^*(X_t) \mid X_t)\big)\right] \geq \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T} \Delta(X_t)\big(1 - q_t(a_t^\dagger(X_t) \mid X_t)\big)\right],$$

where we used $\pi_t(a \mid x) \geq (1 - \gamma_t)q_t(a \mid x) \geq q_t(a \mid x)/2$ for all $a \in [K]$ and $x \in \mathcal{X}$. Define u_t by

$$u_t = \mathbb{E}\left[1 - q_t\left(a_t^\dagger(X_t) \mid X_t\right)\right]. \tag{23}$$

Then, from Lemma A.5, under the margin condition given in Definition 2.5, we have

$$\mathbb{E}\left[\Delta(X_t)\big(1 - q_t(a_T^\dagger(X_t) \mid X_t)\big)\right] \geq \Delta_* \frac{\beta}{1+\beta}\left(\mathbb{E}\left[1 - q_t\left(a_t^\dagger(X_t) \mid X_t\right)\right]\right)^{\frac{1+\beta}{\beta}} = \Delta_* \frac{\beta}{1+\beta} u_t^{\frac{1+\beta}{\beta}}.$$

Hence, under the margin condition, we have

$$R_T \geq \frac{\beta\Delta_*}{2(1+\beta)} \sum_{t=1}^{T} u_t^{\frac{1+\beta}{\beta}} = \frac{\beta\Delta_* T}{2(1+\beta)} \frac{1}{T} \sum_{t=1}^{T} u_t^{\frac{1+\beta}{\beta}} \tag{24}$$

$$\geq \frac{\beta\Delta_* T}{2(1+\beta)}\left(\frac{1}{T}\sum_{t=1}^{T} u_t\right)^{\frac{1+\beta}{\beta}} = \frac{\beta\Delta_* T^{-\frac{1}{\beta}}}{2(1+\beta)}\left(\sum_{t=1}^{T} u_t\right)^{\frac{1+\beta}{\beta}}, \tag{25}$$

where the second inequality follows from Jensen's inequality. In addition, as we have

$$u_t = \mathbb{E}\left[\sum_{x \in \mathcal{X}} g(x)\left(1 - q_t\left(a_t^\dagger(x) \mid x\right)\right)\right] \geq \frac{1}{L}\mathbb{E}\left[\sum_{x \in \mathcal{X}}\left(1 - q_t\left(a_t^\dagger(x) \mid x\right)\right)\right]$$

$$\geq \frac{1}{L}\mathbb{E}\left[\sup_{x \in \mathcal{X}}\left(1 - q_t\left(a_t^\dagger(x) \mid x\right)\right)\right] = \frac{1}{L}\mathbb{E}\left[\omega_t\right], \tag{26}$$

we have

$$R_T \geq \frac{\beta\Delta_* T^{-\frac{1}{\beta}}}{2(1+\beta)}\left(\frac{1}{L}\mathbb{E}\left[\sum_{t=1}^{T} \omega_t\right]\right)^{\frac{1+\beta}{\beta}}, \tag{27}$$

which implies

$$\mathbb{E}\left[\sum_{t=1}^{T} \omega_t\right] \leq LT^{\frac{1}{1+\beta}}\left(\frac{2(1+\beta)R_T}{\beta\Delta^*}\right)^{\frac{\beta}{1+\beta}}. \tag{28}$$

$\square$

Proof of Theorem 4.3. We start with the upper bound given in Theorem 4.1. From the Cauchy-Schwarz inequality and the Jensen inequality, we have

$$R_T = O\left(\mathbb{E}\left[\sum_{t=1}^{T} \frac{\sqrt{\sqrt{K}d\omega_t}}{\sqrt{t}}\right] + \kappa\right)$$

$$= O\left(\sqrt{\sqrt{K}d}\sqrt{\sum_{t=1}^{T}\frac{1}{t}}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\omega_t\right]+\kappa}\right)$$

$$= O\left(\sqrt{\sqrt{K}d\log(T)}\sqrt{\sum_{t=1}^{T}\mathbb{E}\left[\omega_t\right]+\kappa}\right).$$

From this and Lemma B.4, we have

$$R_T = O\left(\sqrt{\sqrt{K}d\log(T)\cdot LT^{\frac{1}{1+\beta}}\left(\frac{(1+\beta)}{\beta\Delta^*}\right)^{\frac{\beta}{1+\beta}}R_T^{\frac{\beta}{2(1+\beta)}}+\kappa}\right),\tag{29}$$

which implies

$$R_T = O\left(\left(\frac{(1+\beta)}{\beta\Delta^*}\right)^{\frac{\beta}{2+\beta}}\left(L\sqrt{K}d\log(T)\right)^{\frac{1+\beta}{2+\beta}}T^{\frac{1}{2+\beta}}+\kappa\right).\tag{30}$$

We here used the fact that $x \leq ax^v + b$ implies $x = O\left(a^{\frac{1}{1-v}}+b\right)$, which holds for any for $a > 0$, $b \geq 0$, $x \geq 0$ and $v \in (0,1)$,

$\square$

B.7 Proof of Theorem 4.4

Proof of Theorem 4.4. From Definition 2.6 and the fact that $\pi_t(a \mid X_t) \geq \gamma_t q_t(a \mid X_t) \geq \frac{1}{2}q_t(a \mid X_t)$, we have

$$R_T \geq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a\neq\rho_T^*(X_t)}\Delta_t(a \mid X_t)\pi_t(a \mid X_t)\right] - C$$

$$\geq \frac{1}{2}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a\neq\rho_T^*(X_t)}\Delta_t(a \mid X_t)q_t(a \mid X_t)\right] - C$$

$$\geq \frac{1}{2}\Delta_*\mathbb{E}\left[\sum_{t=1}^{T}\sum_{a\neq\rho_T^*(X_t)}q_t(a \mid X_t)\right] - C$$

$$\geq \frac{1}{2}\Delta_*\mathbb{E}\left[\sum_{t=1}^{T}\left(1 - q_t(a_t^\dagger(X_t) \mid X_t)\right)\right] - C$$

$$= \frac{1}{2}\Delta_*\sum_{t=1}^{T}u_t - C \geq \frac{\Delta_*}{2L}\mathbb{E}\left[\sum_{t=1}^{T}\omega_t\right] - C,$$

where the third inequality follows from the assumption that $\Delta(a \mid X_t) \geq \Delta_*$ holds for all $a \neq \rho_T^*(x)$ in posed in Definition 2.6. The value of u_t is defined in (23), and the last inequality follows from (26). From this and Theorem 4.1, by applying Theorem 4 in Masoudian & Seldin (2021) with $K = 2$, $\Delta_i = \frac{\Delta_*}{L}$ and $B = O(\sqrt{d\sqrt{K}})$, we obtain

$$R_T = O\left(\frac{dL\sqrt{K}}{\Delta_*}\log\left(\frac{\Delta_*^2 T}{L^2}\right) + \kappa + C\right).$$

Moreover, for $\frac{dL\sqrt{K}}{\Delta_*}\left(\log\left(\frac{T\Delta_*^2}{L^2d\sqrt{K}}\right)+1\right) \leq C \leq \frac{\Delta_* T}{L}$, we have

$$R_T = O\left(\sqrt{\frac{CL\sqrt{K}d}{\Delta_*}}\left(\sqrt{\log\left(\frac{\Delta_* T}{CL}\right)}+2\right)+W+\kappa\right),$$

where W is a subdominant term given as

$$W := \frac{L\sqrt{K}d}{\Delta_*} \left(\log \left(\frac{\Delta_* T}{CL} \right) + \sqrt{\log \left(\frac{\Delta_* T}{CL} \right)} \right).$$

□

C Proof of Lemma B.1

To show Lemma B.1, we show the following lemmas. The proofs are shown in Appendices C and C.2, respectively.

Lemma C.1. *For $p, q \in [0, 1]$ and $\ell \geq -\frac{1-\alpha}{2} q^{\alpha-1}$, we have*

$$\ell \cdot (q - p) - d(p, q) \leq \frac{2q^{2-\alpha} \ell^2}{1 - \alpha}. \tag{31}$$

Lemma C.2. *Fix $a^\dagger \in [K]$ and $q \in \mathcal{P}_K$. Then, the following holds for $\ell \in \mathbb{R}^K$:*

- *If $\ell(a) \geq -\frac{1-\alpha}{4} q(a)^{\alpha-1}$ for all $a \in [K]$, we then have*

$$\left\langle \ell, q - p \right\rangle - D(p, q) \leq \frac{4}{1 - \alpha} \left(\sum_{a \in [K]} q(a)^{2-\alpha} \ell_t(a)^2 \right) \tag{32}$$

  for all $p \in \mathcal{P}_K$.

- *If $\ell(a) \geq -\frac{1-\alpha}{4} q(a)^{\alpha-1}$ for all $a \in [K] \setminus \{a^\dagger\}$ and $\ell(a^\dagger) \leq \frac{1-\alpha}{4}(1 - q(a^\dagger))^{\alpha-1}$, we then have*

$$\left\langle \ell, q - p \right\rangle - D(p, q) \leq \frac{4}{1 - \alpha} \left(\sum_{a \neq a^\dagger} q(a)^{2-\alpha} \ell(a)^2 + \left(1 - q(a^\dagger) \right)^{2-\alpha} \ell(a^\dagger)^2 \right) \tag{33}$$

  for all $p \in \mathcal{P}_K$.

Lemma C.3. *Fix $a^\dagger \in [K]$ and $q \in \mathcal{P}_K$. If $|\ell(a)| \leq \frac{1-\alpha}{4} q(a)^{\alpha-1}$ for all $a \in [K]$, we have*

$$\left\langle \ell, q - p \right\rangle - D(p, q) \leq \frac{4}{1 - \alpha} \left(\sum_{a \neq a^\dagger} q(a)^{2-\alpha} \ell(a)^2 + \left(\min \{q(a^\dagger), 1 - q(a^\dagger)\} \right)^{2-\alpha} \ell(a^\dagger)^2 \right) \tag{34}$$

for all $p \in \mathcal{P}_K$.

Then, by using Lemma C.3, we prove Lemma B.1 as follows:

Proof of Lemma B.1. From Lemma C.2 with $\ell = \eta_t \widehat{\ell}_t(x)$ and $p = p(x)$, we have we have

$$\left\langle \eta_t \widehat{\ell}_t(x), q(x) - p(x) \right\rangle - D(p(x), q(x))$$

$$\leq \frac{4\eta_t^2}{1 - \alpha} \left(\sum_{a \neq a^\dagger(x)} q(a \mid x)^{2-\alpha} \ell_t(a \mid x)^2 + \left(1 - q(a^\dagger(x) \mid x) \right)^{2-\alpha} \ell_t(a^\dagger(x), x)^2 \right)$$

$$\leq \frac{4\eta_t^2}{1 - \alpha} \left(\sum_{a \neq a^\dagger(x)} q(a \mid x)^{2-\alpha} \ell_t(a \mid x)^2 + \min \{q(a^\dagger(x) \mid x), 1 - q(a^\dagger(x) \mid x)\}^{2-\alpha} \ell_t \left(a^\dagger(x), x \right)^2 \right).$$

By dividing both sides by η_t, we obtain the desired bound. □

C.1   Proof of Lemma C.1

Proof. For all given q and ℓ, the LHS of (31) is concave in p. Hence, this is maximized when

$$\frac{\mathrm{d}}{\mathrm{d}p}\Big\{\ell\cdot\big(q-p\big)-d\,(p,q)\Big\}=-\ell-q^{\alpha-1}+p^{\alpha-1}=0. \tag{35}$$

We then have

$$p=\big(q^{\alpha-1}+\ell\big)^{\frac{1}{\alpha-1}}\le\left(q^{\alpha-1}-\frac{1-\alpha}{2}q^{\alpha-1}\right)^{\frac{1}{\alpha-1}}=q\left(1-\frac{1-\alpha}{2}\right)^{\frac{1}{\alpha-1}}\le 2q, \tag{36}$$

where the first equality follows from (35) and the first inequality follows from the assumption of $\ell\ge-\frac{1-\alpha}{2}q^{\alpha-1}$. Furthermore, from the intermediate value theorem and the fact that $p^{\alpha-2}$ is monotone decreasing in p, we have

$$
\begin{aligned}
|\ell|&=\big|p^{\alpha-1}-q^{\alpha-1}\big|\\
&\ge\min\Big\{\big|(\alpha-1)p^{\alpha-2}\big|,\big|(\alpha-1)q^{\alpha-2}\big|\Big\}|p-q|\\
&=(1-\alpha)\max\{p,q\}^{\alpha-2}|p-q|,
\end{aligned}
$$

where the first inequality follows from (35) and the second inequality follows from the intermediate value theorem. This implies

$$|p-q|\le\frac{1}{1-\alpha}\cdot\max\{p,q\}^{2-\alpha}|\ell|. \tag{37}$$

We then have

$$\ell\cdot(p-q)-d(p,q)\le|\ell||q-p|\le\frac{\ell^2}{1-\alpha}\max\{p,q\}^{2-\alpha}\le\frac{4\ell^2}{1-\alpha}q^{2-\alpha},$$

where the second inequality follows from (37) and the last inequality follows from (36). □

C.2   Proof of Lemma C.2

Proof. We have

$$
\begin{aligned}
&\Big\langle\ell,q-p\Big\rangle-D(p,q)\\
&=\frac{1}{2}\sum_{a\ne a^\dagger(x)}\big(2\ell(a)\cdot\big(q(a)-p(a)\big)-d\big(p(a),q(a)\big)\big)\\
&\quad\frac{1}{2}\left(2\ell_t(a^\dagger)\cdot\big(q(a^\dagger)-p(a^\dagger)\big)-d\big(p(a^\dagger),q(a^\dagger)\big)-\sum_{a\in[K]}d\big(p(a),q(a)\big)\right)\\
&\le\frac{1}{2}\sum_{a\ne a^\dagger}\Big\{2\ell(a)\cdot\big(q(a)-p(a)\big)-d\big(p(a),q(a)\big)\Big\}\\
&\quad+\frac{1}{2}\min\Big\{2\ell(a^\dagger)\cdot\big(q(a^\dagger)-p(a^\dagger)\big)-d\big(p(a^\dagger),q(a^\dagger)\big),\\
&\qquad\qquad 2\ell(a^\dagger)\cdot\big(q(a^\dagger)-p(a^\dagger)\big)-\sum_{a\ne a^\dagger}d\big(p(a),q(a)\big)\Big\}.
\end{aligned}
\tag{38}
$$

From Lemma C.1, if $\ell(a)\ge-\frac{1-\alpha}{2}q(a)^{\alpha-1}$, we have

$$2\ell(a)\cdot\big(q(a)-p(a)\big)-d\,(p(a),q(a))\le\frac{8q(a)^{2-\alpha}\ell(a)^2}{1-\alpha}. \tag{39}$$

Furthermore, we have

$$q(a^\dagger) - p(a^\dagger) = \big(1 - p(a^\dagger)\big) - \big(1 - q(a^\dagger)\big) = \sum_{a \neq a^\dagger} \big(p(a) - q(a)\big).$$

As we have $\big(1 - q(a^\dagger)\big)^{\alpha-1} \leq q(a)^{\alpha-1}$ for all $a \in [K]\backslash\{a^\dagger\}$, if $\ell(a^\dagger) \leq \frac{1-\alpha}{4}\big(1 - q(a^\dagger)\big)^{\alpha-1}$, we then have

$$-\ell(a^\dagger) \geq -\frac{1-\alpha}{4}q(a)^{\alpha-1}$$

for all $a \in [K] \backslash \{a^\dagger(x)\}$. Hence, Lemma C.1 yields

$$2\ell(a^\dagger) \cdot \big(q(a^\dagger) - p(a^\dagger)\big) - \sum_{a \neq a^\dagger} d\big(p(a), q(a)\big)$$

$$= \sum_{a \neq a^\dagger} \big(-2\ell(a^\dagger) \cdot \big(q(a) - p(a)\big) - d\big(p(a), q(a)\big)\big)$$

$$\leq \frac{2}{1-\alpha} \sum_{a \neq a^\dagger} \big(2\ell(a^\dagger)\big)^2 q(a)^{2-\alpha}$$

$$\leq \frac{8}{1-\alpha}\ell(a^\dagger)^2 \left(\sum_{a \neq a^\dagger} q(a)\right)^{2-\alpha}$$

$$= \frac{8}{1-\alpha}\big(1 - q(a^\dagger)\big)^{2-\alpha}(\ell(a^\dagger))^2$$

if $\ell(a^\dagger) \leq \frac{1-\alpha}{4}\big(1 - q(a^\dagger)\big)^{\alpha-1}$. Combining this with (38) and (39), we obtain the desired bounds. $\square$

C.3 Proof of Lemma C.3

Proof. Let us consider two cases: when $q(a^\dagger) \leq 1 - q(a^\dagger)$ and when $q(a^\dagger) < 1 - q(a^\dagger)$. Suppose $q(a^\dagger) \leq 1 - q(a^\dagger)$. Then, from (32) in Lemma C.2, we have

$$\big\langle \ell, q - p \big\rangle - D(p, q) \leq \frac{4}{1-\alpha}\left(\sum_{a \neq a^\dagger} q(a)^{2-\alpha}\ell(a)^2 + \big(q(a^\dagger)\big)^{2-\alpha}\ell(a^\dagger)^2\right)$$

$$\leq \frac{4}{1-\alpha}\left(\sum_{a \neq a^\dagger} q(a)^{2-\alpha}\ell(a)^2 + \big(\min\{q(a^\dagger), 1 - q(a^\dagger)\}\big)^{2-\alpha}\ell(a^\dagger)^2\right).$$

Suppose $q(a^\dagger) > 1 - q(a^\dagger)$. We then have $|\ell(a^\dagger)| \leq \frac{1-\alpha}{4}(q(a^\dagger))^{\alpha-1} < \frac{1-\alpha}{4}(1 - q(a^\dagger))^{\alpha-1}$. Hence, we can use (33) in Lemma C.2 to obtain:

$$\big\langle \ell, q - p \big\rangle - D(p, q) \leq \frac{4}{1-\alpha}\left(\sum_{a \neq a^\dagger} q(a)^{2-\alpha}\ell(a)^2 + \big(1 - q(a^\dagger)\big)^{2-\alpha}\ell(a^\dagger)^2\right)$$

$$\leq \frac{4}{1-\alpha}\left(\sum_{a \neq a^\dagger} q(a)^{2-\alpha}\ell(a)^2 + \big(\min\{q(a^\dagger), 1 - q(a^\dagger)\}\big)^{2-\alpha}\ell(a^\dagger)^2\right),$$

which completes the proof. $\square$

D The BoBW-RealFTRL with arm-dependent features

In this section, we reformulate the FTRL with Shannon entropy under the setting with arm-dependent features and derive its regret upper bound. Note that Kuroki et al. (2024) and Kato & Ito (2023) investigate the FTRL with Shannon entropy under the setting with arm-independent features.

We consider the same formulation in our main text. However, we do not assume the finiteness of contexts in a stochastic regime.

Algorithm 2 BoBW-RealFTRL.

Parameter: Learning rate $\eta_1, \eta_2, \ldots, \eta_T > 0$.
Initialization: Set $\theta_0 = 0$.
for $t = 1, \ldots, T$ **do**
    Observe X_t.
    Draw $A_t \in [K]$ following the policy $\pi_t(X_t) := (1 - \gamma_t)q_t(X_t) + \gamma_t e^*(X_t)$ defined in (40).
    Observe the loss $\ell_t(A_t, X_t)$.
    Compute $\widehat{\theta}_t$.
end for

D.1 The BoBW-RealFTRL

Following Kuroki et al. (2024) and Kato & Ito (2023), we define the BoBW-RealFTRL with the arm-dependent feature as

$$\pi_t(X_t) := (1 - \gamma_t)q_t(X_t) + \gamma_t e^*(X_t), \tag{40}$$

where

$$q_t(x) \in \arg\min_{q \in \mathcal{P}_K} \left\{ \sum_{s=1}^{t-1} \left\langle \widehat{\ell}_s(x), q \right\rangle + \frac{1}{\eta_t} \psi(q) \right\} \quad \text{for } t \geq 2,$$

$$q_1(x) := (1/K \ 1/K \ \cdots \ 1/K)^\top,$$

$$\psi(q(x)) := - \sum_{a \in [K]} q(a \mid x) \log\left(\frac{1}{q(a \mid x)}\right),$$

$$\frac{1}{\eta_{t+1}} := \frac{1}{\eta_t} + \frac{1}{\eta_1} \frac{1}{\sqrt{1 + \left(\log(K)\right)^{-1} \sum_{s=1}^{t} H\left(q_s(X_s)\right)}},$$

$$\frac{1}{\eta_1} := \sqrt{\frac{\left(\frac{1}{\lambda} + d\right) \log(T)}{\log(K)}}, \quad \text{and} \quad \gamma_t := \frac{\eta_t}{\lambda}.$$

D.2 Regret analysis

This section provides upper bounds for the regret of our proposed BoBW-RealFTRL algorithm.

To derive upper bounds, we define the following quantities:

$$Q(\rho_T^* \mid x) = \sum_{t=1}^{T} \left\{ 1 - q_t\left(\rho_T^*(x) \mid x\right) \right\},$$

$$\overline{Q}(\rho_T^*) = \mathbb{E}\left[Q(\rho_T^* \mid X_0) \right].$$

Then, we show the following upper bound, which holds for general cases such as adversarial and stochastic regimes. We show the proof in Appendix E.

Theorem D.1 (General regret bounds). *Consider the BoBW-RealFTRL. Assumptions 2.2–2.1 hold. Then, the decision-maker incurs the following regret:*

$$R_T = O\left(\left(\eta_1 \left(\frac{1}{\lambda} + d\right) \frac{\log(T)}{\sqrt{\log(K)}} + \frac{1}{\eta_1} \sqrt{\log(K)} \right) \sqrt{\log(KT)} \max\left\{ \overline{Q}^{1/2}(\rho_T^*), 1 \right\} \right).$$

For each regime, we derive a specific upper bound. The proof is shown in Appendix F. First, from $\overline{Q}(\rho_T^*) \leq T$, the following regret bound holds without any assumptions on the loss; that is, it holds in an adversarial regime.

Theorem D.2 (Regret upper bound in an adversarial regime)**.** *Consider the BoBW-RealFTRL. Assume that the loss is generated under an adversarial regime. Suppose that Assumption 2.2 and 2.4 hold. Then, the regret satisfies*

$$R_T = O\left(\log(KT)\sqrt{\left(\frac{1}{\lambda}+d\right)\log(T)T}\right).$$

Next, we derive a regret upper bound in a stochastic regime with a margin condition.

Theorem D.3 (Regret upper bound in a stochastic regime with a margin condition)**.** *Consider the BoBW-RealFTRL. Assume that the loss is generated under a stochastic regime with a margin condition (Definition 2.5). Suppose that Assumption 2.2, 2.4–2.1 hold. Then, the regret satisfies*

$$R_T = O\left(\left(\frac{1+\beta}{\beta\Delta_*}\right)^{\frac{\beta}{2+\beta}}\left(\log(KT)\sqrt{\left(\frac{1}{\lambda}+d\right)\log(T)}\right)^{\frac{1+\beta}{2+\beta}}T^{\frac{1}{2+\beta}}\right).$$

We omit the proof because it is almost the same as that for Theorem 4.3 for the 1/2-LC-Tsallis-INF

Furthermore, we derive a regret bound under the linear contextual adversarial regime with a self-bounding constraint. The proof is provided in Appendix F

Theorem D.4 (Regret bounds under the linear contextual adversarial regime with a self-bounding constraint)**.** *Consider the BoBW-RealFTRL. Assume that the loss is generated under a linear contextual adversarial regime with a self-bounding constraint (Definition 2.6). Suppose that Assumption 2.2, 2.4–2.1 hold. Then, the regret satisfies*

$$R_T = O\left(\frac{\left(\frac{1}{\lambda}+d\right)\log(T)\log(KT)}{\Delta_*} + \sqrt{\frac{C\left(\frac{1}{\lambda}+d\right)\log(T)\log(KT)}{\Delta_*}}\right).$$

Note that the BoBW-RealFTRL does not require Assumption 2.3 in stochastic regimes.

E Proof of Theorem D.1

This section provides the proof of Theorems D.1.

As well as Section B.1, based on the result in Proposition A.1, we define

$$r_T(x) :=$$
$$\sum_{t=1}^{T}\left(\gamma_t + \left\langle \widehat{\ell}_t(x), q_t(x) - q_{t+1}(x)\right\rangle - \frac{1}{\eta_t}D(q_{t+1}(x), q_t(x)) - \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\psi\big(q_t(x)\big)\right),$$

where we replace $H(q(x))$ with $\psi(q(x))$. Note that we define $1/\eta_0 = 0$ as an exception. Using this function $r_T(x)$, we can bound R_T as $R_T \leq \mathbb{E}\big[r_T(X_0)\big]$.

We decompose the pointwise regret upper bound $r_T(x)$ as follows:

$$r_T(x) = \underbrace{\sum_{t=1}^{T}\left(\gamma_t + \left\langle \widehat{\ell}_t(x), q_t(x) - q_{t+1}(x)\right\rangle - \frac{1}{\eta_t}D(q_{t+1}(x), q_t(x))\right)}_{= \text{ stability term}}$$

$$+ \underbrace{\left(-\sum_{t=1}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\psi\big(q_t(x)\big)\right)}_{= \text{ penalty term}}.$$

To bound the stability term $\left\langle \widehat{\ell}_t(x), q_t(x) - q_{t+1}(x)\right\rangle - \frac{1}{\eta_t}D(q_{t+1}(x), q_t(x))$, we use the following proposition from Ito et al. (2022).

Proposition E.1 (From Lemma 8 in Ito et al. (2022))**.** *Consider the BoBW-RealFTRL. For all $x \in \mathcal{X}$, all $\ell : \mathcal{X} \to \mathbb{R}^K$ and $p, q \in \Pi$, it holds that*

$$\left\langle \ell_t(x), p(x) - q(x) \right\rangle - \frac{1}{\eta_t} D(q(x), p(x)) \leq \frac{1}{\eta_t} \sum_{a \in [K]} p(a \mid x) \xi \left(\eta_t \ell_t(a, x) \right),$$

where $\xi(x) := \exp(-x) + x - 1$.

By using Proposition E.1, we obtain the following lemma.

Lemma E.2. *The regret for the BoBW-RealFTRL satisfies*

$$R_T \leq \mathbb{E} \left[\sum_{t=1}^{T} \left(\gamma_t + 3\eta_t d + \left(-\sum_{t=1}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \psi\big(q_t(X_0)\big) \right) \right) \right].$$

Proof of Lemma E.2. From Proposition A.1, we have

$$R_T \leq \mathbb{E}\big[r_T(X_0)\big],$$

where

$$r_T(x) = \underbrace{\sum_{t=1}^{T} \left(\gamma_t + \left\langle \widehat{\ell}_t(x), q_t(x) - q_{t+1}(x) \right\rangle - \frac{1}{\eta_t} D(q_{t+1}(x), q_t(x)) \right)}_{= \text{ stability term}}$$

$$+ \underbrace{\left(-\sum_{t=1}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}} \right) \psi\big(q_t(x)\big) \right)}_{= \text{ penalty term}}.$$

To prove the statement, we bound the stability term as

$$\mathbb{E} \left[\left\langle \widehat{\ell}_t(X_0), \pi_t(X_0) - q_{t+1}(X_0) \right\rangle - D_t(q_{t+1}(X_0), \pi_t(X_0)) \right] \leq 3\eta_t d. \tag{41}$$

To show this, from Proposition E.1, we have

$$\left\langle \widehat{\ell}_t(x), \pi_t(x) - q_{t+1}(x) \right\rangle - D_t(q_{t+1}(x), \pi_t(x)) \leq \frac{1}{\eta_t} \sum_{a \in [K]} \pi_t(a \mid x) \xi \left(\eta_t \widehat{\ell}_t(a, x) \right).$$

Here, from Lemma A.3, we have $\eta_t \widehat{\ell}_t(a, x) \geq -\eta_t / (\lambda \gamma_t)$. Additionally, since $\gamma_t = \frac{\eta_t}{\lambda}$, $\eta_t \widehat{\ell}_t(a, x) = -1$ holds. Then, we have

$$\left\langle \widehat{\ell}_t(x), \pi_t(x) - q_{t+1}(x) \right\rangle - D_t(q_{t+1}(x), \pi_t(x))$$

$$\leq \frac{1}{\eta_t} \sum_{a \in [K]} \pi_t(a \mid x) \xi \left(\eta_t \widehat{\ell}_t(a, x) \right)$$

$$\leq \eta_t \sum_{a \in [K]} \pi_t(a \mid x) \widehat{\ell}_t^2(a, x).$$

Lastly, from Lemma A.4, which states that $\mathbb{E} \left[\sum_{a \in [K]} \pi_t(a \mid X_0) \widehat{\ell}_t(a, X_0)^2 \mid \mathcal{F}_{t-1} \right] \leq d$, we have (41). $\qquad \square$

From this result, we obtain the following lemma.

Lemma E.3. *Assume the conditions in Theorem E.2. Consider the BoBW-RealFTRL. Then, we have*

$$R_T \leq \bar{c} \sqrt{\mathbb{E}\left[\sum_{t=1}^{T} \psi(q_t(X_0))\right]},$$

where $\bar{c} = O\left(\eta_1 \left(\frac{1}{\lambda} + d\right) \frac{\log(T)}{\sqrt{\log(K)}} + \frac{1}{\eta_1}\sqrt{\log(K)}\right).$

Proof. We prove the following inequalities:

$$\sum_{t=1}^{T}(\gamma_t + 3\eta_t d) = O\left(\eta_1\left(\frac{1}{\lambda}+d\right)\frac{\log(T)}{\sqrt{\log(K)}}\sqrt{\sum_{t=1}^{T}\psi(q_t(X_t))}\right), \tag{42}$$

$$-\sum_{t=1}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\psi(q_t(x)) = O\left(\beta_1\sqrt{\log(K)}\sqrt{\sum_{t=1}^{T}H(q_t(X_t))}\right). \tag{43}$$

Proof of (42) From $\gamma_t = \frac{\eta_t}{\lambda}$, it holds that

$$\sum_{t=1}^{T}(\gamma_t + 3\eta_t d) = \sum_{t=1}^{T}\left(\frac{\eta_t}{\lambda} + 3\eta_t d\right) = \left(\frac{1}{\lambda}+3d\right)\sum_{t=1}^{T}\eta_t.$$

From $\frac{1}{\eta_{t+1}} := \frac{1}{\eta_t} + \frac{1}{\eta_1}\frac{1}{\sqrt{1+(\log(K))^{-1}\sum_{s=1}^{t}H(q_s(X_s))}}$, we have

$$\frac{1}{\eta_t} = \frac{1}{\eta_1} + \sum_{u=1}^{t-1}\frac{1}{\eta_1}\frac{1}{\sqrt{1+(\log(K))^{-1}\sum_{s=1}^{u}\psi(q_s(X_s))}}$$

$$\geq \frac{1}{\eta_1}\frac{t}{\sqrt{1+(\log(K))^{-1}\sum_{s=1}^{t}\psi(q_s(X_s))}}.$$

Therefore, we have

$$\sum_{t=1}^{T}\eta_t \leq \sum_{t=1}^{T}\eta_1\frac{\sqrt{1+(\log(K))^{-1}\sum_{s=1}^{t}H(q_s(X_s))}}{t}$$

$$\leq \eta_1\left(1+\log(T)\right)\sqrt{1+(\log(K))^{-1}\sum_{t=1}^{T}\psi(q_t(X_t))}.$$

Therefore, we obtain

$$\sum_{t=1}^{T}(\gamma_t + 3\eta_t d) = O\left(\eta_1\left(\frac{1}{\lambda}+d\right)\frac{\log(T)}{\sqrt{\log(K)}}\sqrt{\sum_{t=1}^{T}\psi(q_t(X_t))}\right).$$

Proof of (43) From the definitions of β_t and γ_t, we have

$$-\sum_{t=1}^{T}\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)\psi(q_t(x))$$

$$= \sum_{t=1}^{T}\frac{1}{\eta_1\sqrt{1+(\log(K))^{-1}\sum_{s=1}^{t-1}\psi(q_s(X_s))}}\psi(q_t(X_t))$$

$$= 2\frac{1}{\eta_1}\sqrt{\log(K)}\sum_{t=1}^T \frac{\psi(q_t(X_t))}{\sqrt{\log(K)+\sum_{s=1}^{t-1}\psi(q_s(X_s))}+\sqrt{\log(K)+\sum_{s=1}^{t-1}\psi(q_s(X_s))}}$$

$$\leq 2\frac{1}{\eta_1}\sqrt{\log(K)}\sum_{t=1}^T \frac{\psi(q_t(X_t))}{\sqrt{\log(K)+\sum_{s=1}^{t}\psi(q_s(X_s))}+\sqrt{\log(K)+\sum_{s=1}^{t-1}\psi(q_s(X_s))}}$$

$$\leq 2\frac{1}{\eta_1}\sqrt{\log(K)}\sum_{t=1}^T \frac{\psi(q_t(X_t))}{\sqrt{\sum_{s=1}^{t}\psi(q_s(X_s))}+\sqrt{\sum_{s=1}^{t-1}\psi(q_s(X_s))}}$$

$$= 2\frac{1}{\eta_1}\sqrt{\log(K)}\sum_{t=1}^T \frac{\psi(q_t(X_t))}{\psi(q_t(X_t))}\left\{\sqrt{\sum_{s=1}^{t}\psi(q_s(X_s))}-\sqrt{\sum_{s=1}^{t-1}\psi(q_s(X_s))}\right\}$$

$$= 2\frac{1}{\eta_1}\sqrt{\log(K)}\sum_{t=1}^T \left\{\sqrt{\sum_{s=1}^{t}\psi(q_s(X_s))}-\sqrt{\sum_{s=1}^{t-1}\psi(q_s(X_s))}\right\}$$

$$= 2\frac{1}{\eta_1}\sqrt{\log(K)}\left\{\sqrt{\sum_{s=1}^{T}\psi(q_s(X_s))}-\sqrt{\psi(q_1(X_1))}\right\}$$

$$\leq 2\frac{1}{\eta_1}\sqrt{\log(K)}\sqrt{\sum_{s=1}^{T}\psi(q_s(X_s))}.$$

Inequalities (42) and (43) combined with the inequality in Lemma E.2 yield

$$R_T \leq \mathbb{E}\left[\sum_{t=1}^T\left(\gamma_t+3\eta_t d+\left(-\sum_{t=1}^T\left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}\right)\psi(q_t(X_0))\right)\right)\right]$$

$$= \mathbb{E}\left[\sum_{t=1}^T\left(\gamma_t+3\eta_t d+\left(-\sum_{t=1}^T\left(\frac{1}{\eta_t}-\frac{1}{\eta_{t-1}}\right)\psi(q_t(X_t))\right)\right)\right]$$

$$= \mathbb{E}\left[O\left(\eta_1\left(\frac{1}{\lambda}+d\right)\frac{\log(T)}{\sqrt{\log(K)}}\sqrt{\sum_{t=1}^T\psi(q_t(X_t))}\right)+O\left(\frac{1}{\eta_1}\sqrt{\log(K)}\sqrt{\sum_{t=1}^T\psi(q_t(X_t))}\right)\right]$$

$$= O\left(\eta_1\left(\frac{1}{\lambda}+d\right)\frac{\log(T)}{\sqrt{\log(K)}}\sqrt{\sum_{t=1}^T\mathbb{E}\left[\psi(q_t(X_t))\right]}\right)+O\left(\frac{1}{\eta_1}\sqrt{\log(K)}\sqrt{\sum_{t=1}^T\mathbb{E}\left[\psi(q_t(X_t))\right]}\right).$$

Thus, we obtain the regret bound in Lemma E.3. □

Next, we consider bounding $\sum_{t=1}^T\psi(q_t(x))$ by $Q(\rho_T^* \mid x)$ as shown in the following proposition.

Proposition E.4 (From Lemma 4 in Ito et al. (2022))**.** *For all $\rho^*:\mathcal{X}\to[K]$, the following holds:*

$$\sum_{t=1}^T\psi(q_t(x))\leq Q(\rho^*\mid x)\log\left(\frac{eKT}{Q(\rho^*\mid x)}\right),$$

where e is Napier's constant.

By using the above lemmas and propositions, we prove Theorem D.1.

Proof of Theorem D.1. From Lemma E.4, if $Q(\rho_T^*\mid x)\leq e$, we have $\sum_{t=1}^T\psi(q_t(x))\leq e\log(KT)$ and otherwise, we have $\sum_{t=1}^T\psi(q_t(x))\leq Q(\rho_T^*\mid x)\log(KT)$. Hence, we have $\sum_{t=1}^T\psi(q_t(x))\leq\log(KT)\max\{e,Q(\rho_T^*\mid x)\}$. From

Lemma E.3, we have

$$R_T \leq \bar{c} \sqrt{\sum_{t=1}^{T} \mathbb{E}\left[\psi(q_t(X_0))\right]}$$

$$= O\left(\left(\eta_1 \left(\frac{1}{\lambda} + d\right) \frac{\log(T)}{\sqrt{\log(K)}} + \frac{1}{\eta_1}\sqrt{\log(K)}\right)\sqrt{\log(KT)} \max\left\{\overline{Q}^{1/2}, 1\right\}\right).$$

□

F Proof of Theorem D.4

Proof of Theorem D.4. From the definition of the contextual adversarial regime with a self-bounding constraint, we have

$$R_T \geq \Delta_* \cdot \mathbb{E}\left[\sum_{t=1}^{T} \left(1 - \pi_t(a^*(X_0) \mid X_0)\right)\right] - C$$

$$= \Delta_* \cdot \overline{Q}(\rho_T^*) - C.$$

Therefore, from Lemma E.3, for all $\lambda > 0$, we have

$$R_T = (1+\lambda)R_T - \lambda R_T$$

$$= (1+\lambda)O\left(\bar{c}\sqrt{\log(KT)}\sqrt{\sum_{t=1}^{T} \mathbb{E}\left[\psi(q_t(X_0))\right]}\right) - \lambda R_T$$

$$\leq (1+\lambda)O\left(\bar{c}\sqrt{\log(KT)}\sqrt{\sum_{t=1}^{T} \mathbb{E}\left[\psi(q_t(X_0))\right]}\right) - \lambda \Delta_* \cdot \overline{Q}(\rho_T^*) + \lambda C,$$

where

$$\bar{c} = \eta_1 \left(\frac{1}{\lambda} + d\right) \frac{\log(T)}{\sqrt{\log(K)}} + \frac{1}{\eta_1}\sqrt{\log(K)}.$$

Here, as well as the proof of Theorem 4.1, from Lemma E.4, if

$$Q(\rho_T^*(x) \mid x) \leq e,$$

we have

$$\sum_{t=1}^{T} \psi(q_t(x)) \leq e \log(KT)$$

and otherwise, we have

$$\sum_{t=1}^{T} \psi(q_t(x)) \leq Q(\rho_T^* \mid x) \log(KT).$$

Hence, we have $\sum_{t=1}^{T} \psi(q_t(x)) \leq \log(KT) \max\{e, Q(\rho_T^* \mid x)\}$. Here, to upper bound R_T, it is enough to only consider a case with $Q(\rho_T^* \mid x) \geq e$, and we obtain

$$R_T \leq (1+\lambda)O\left(\bar{c}\sqrt{\log(KT)}\sqrt{\overline{Q}(\rho_T^*)\log(KT)}\right) - \lambda \Delta_* \cdot \overline{Q}(\rho_T^*) + \lambda C$$

$$\leq \frac{O\left(\left\{(1+\lambda)\bar{c}\right\}^2 \sqrt{\log(KT)}\right)}{2\lambda\Delta_*} + \lambda\Delta_*.$$

where the second inequality follows from $a\sqrt{b} - \frac{c}{2}b \le \frac{a^2}{c^2}$ holds for all $a, b, c > 0$. By choosing

$$\lambda = \sqrt{\frac{\bar{c}^2 \log(KT)}{\Delta_*} \Big/ \left(\frac{c^2 \log(KT)}{\Delta_*} + 2C\right)}.$$

Then, we obtain $R_T = O\left(\bar{c}^2 \log(KT)/\Delta_* + \sqrt{C\bar{c}^2 \log(KT)/\Delta_*}\right).$ $\qquad\square$

G Details of regret transformation

This section provides the details of regret transformation in Section 4.2.

We consider two problems of linear contextual bandits: P.Indep $(\widetilde{d}, \widetilde{K}, \widetilde{T})$ and P.Dep (d, K, T), defined below.

P.Indep $(\widetilde{d}, \widetilde{K}, \widetilde{T})$ In this stetting, there are $\widetilde{K}$ arms and $\widetilde{T}$ rounds. We can observe $\widetilde{d}$-dimensional contexts. Let $\widetilde{\phi} : \mathcal{X} \to \mathbb{R}^d$ be a context map. We consider a linear model $\widetilde{\ell}_t(a, X_t) = \left\langle \widetilde{\theta}_{a,t}, \widetilde{\phi}(X_t) \right\rangle + \widetilde{\varepsilon}_t(a)$, where $\widetilde{\theta}_{a,t} \in \mathbb{R}^{\widetilde{d}}$ is a regression coefficient, and $\widetilde{\varepsilon}_t(a)$ is the error term. Then, in each round t, given $\left\{\widetilde{\theta}_{a,t}\right\}_{a \in [K]}$

- We observe $\widetilde{d}$-dimensional contexts $\widetilde{\phi}(X_t)$.

- The loss is generated from a linear model $\widetilde{\ell}_t(a, X_t) = \left\langle \theta_{a,t}, \widetilde{\phi}(X_t) \right\rangle + \widetilde{\varepsilon}_t(a)$.

This problem can be transformed into P.Dep $(\widetilde{d}\widetilde{K}, \widetilde{K}, \widetilde{T})$ by defining the corresponding feature and regression parameters well.

Let us define

$$\phi(a, X_t) = \begin{pmatrix} \mathbf{0}_d \\ \vdots \\ \widetilde{\phi}(X_t) \\ \vdots \\ \mathbf{0}_d \end{pmatrix}, \quad \theta_t = \begin{pmatrix} \widetilde{\theta}_{1,t} \\ \vdots \\ \widetilde{\theta}_{a,t} \\ \vdots \\ \widetilde{\theta}_{K,t} \end{pmatrix}$$

where $\mathbf{0}_d$ is the d-dimensional zero vector.

Here, the loss is generated as $\ell_t(a, X_t) = \langle \theta_t, \phi(a, X_t) \rangle + \varepsilon_t(a)$. Note that θ is a $(\widetilde{d} \times \widetilde{K})$ vector.

P.Dep (d, K, T) This setting is identical to our setting. Therefore, we omit the details of this setting.

As well as the case with P.Indep $(\widetilde{d}, \widetilde{K}, \widetilde{T})$, we can transform the P.Dep (d, J, T) into the P.Indep (dK, K, T).

Let us define

$$\widetilde{\phi}(X_t) = \begin{pmatrix} \phi(1, X_t) \\ \vdots \\ \phi(a, X_t) \\ \vdots \\ \phi(K, X_t) \end{pmatrix}, \quad \widetilde{\theta}_{a,t} = \begin{pmatrix} \mathbf{0}_d \\ \vdots \\ \theta_t \\ \vdots \\ \mathbf{0}_d \end{pmatrix}$$

where $\mathbf{0}_d$ is the d-dimensional zero vector.

Here, the loss is generated as $\widetilde{\ell}_t(a, X_t) = \left\langle \widetilde{\theta}_{a,t}, \widetilde{\phi}(X_t) \right\rangle + \widetilde{\varepsilon}_t(a)$.

Regret transformation theorem   The above construction directly yields Theorem 4.5 as follows.

Proof of Theorem 4.5. Suppose that there exists an algorithm whose regret is $R_T = f(d, K, T)$ in the P.Dep (d, K, T), where $f : \mathbb{N} \times \mathbb{N} \times \mathbb{N} \to \mathbb{R}$.

Consider the P.Indep $(\widetilde{d}, \widetilde{K}, \widetilde{T})$. For this problem, we can apply the algorithm by transforming the arm-independent features and parameters to the ones in the arm-dependent feature setting as

$$
\phi(a, X_t) = \begin{pmatrix} \mathbf{0}_d \\ \vdots \\ \widetilde{\phi}(X_t) \\ \vdots \\ \mathbf{0}_d \end{pmatrix}, \quad \theta_t = \begin{pmatrix} \widetilde{\theta}_{1,t} \\ \vdots \\ \widetilde{\theta}_{a,t} \\ \vdots \\ \widetilde{\theta}_{K,t} \end{pmatrix}
$$

This problem has $\widetilde{d}\widetilde{K}$-dimensional parameter θ_t. Therefore, the regret under this setting becomes $f(\widetilde{d}\widetilde{K}, \widetilde{K}, \widetilde{T})$.

Similarly, we can prove the reverse case. Thus, the proof is complete. $\qquad \square$