Near-Optimal Best-of-Both-Worlds Algorithms for Decoupled Exploration and Exploitation in Multi-armed Bandits

Hibiki Sekiya
The University of Tokyo
Tokyo, Japan
s-hibiki@g.ecc.u-tokyo.ac.jp

Shinji Ito
The University of Tokyo and RIKEN AIP
Tokyo, Japan
shinji@mist.i.u-tokyo.ac.jp

Abstract

We study the decoupled multi-armed bandit problem, where the player selects one arm to exploit (incurring its loss) and one arm to explore (observing its loss) at each round. We propose two algorithms: D-Exp3-Ada, based on Exp3 with an AdaGrad learning rate, and D-TINF-SPM, based on FTRL with Tsallis entropy and the stability-penalty matching learning rate. Here, K denotes the number of arms and T the time horizon. In stochastic regimes, we assume a unique optimal arm $i^\star$, and let Δ_i denote the suboptimality gap of arm i. D-Exp3-Ada achieves $\mathcal{O}(\sqrt{KT\log K})$ regret in adversarial regimes and $\mathcal{O}(\sum_{i\neq i^\star}\frac{\log K}{\Delta_i})$ in stochastic regimes, combining simplicity of algorithm design with strong guarantees. D-TINF-SPM achieves the minimax optimal $\mathcal{O}(\sqrt{KT})$ in adversarial regimes and $\mathcal{O}(\min\{\sqrt{K\sum_{i\neq i^\star}\frac{1}{\Delta_i^2}}, \sum_{i\neq i^\star}\frac{\log K}{\Delta_i}\})$ in stochastic regimes. Finally, we derive instance-dependent lower bounds under natural monotonicity and permutation-invariance assumptions on regret upper bounds, proving that our algorithms are optimal up to $\mathcal{O}((\log K)^2)$ within this class.

1 Introduction

The multi-armed bandit (MAB) problem is a fundamental framework for studying the exploration-exploitation trade-off in sequential decision making (Lai and Robbins, 1985; Auer et al., 2002a,b; Lattimore and Szepesvári, 2020). In the MAB problem, a player selects an action (arm) from a set of K available actions at each round, subsequently observing and incurring the loss associated with the selected action. The losses are generated either adversarially or stochastically, depending on problem settings. The goal of the player is to find a strategy that minimizes the regret, the difference between the cumulative loss of the player and that of the best fixed action in hindsight.

We consider the problem of decoupled exploration and exploitation in multi-armed bandits, a variant of the standard MAB problem introduced by Avner et al. [2012], in which the player is allowed to choose one action to play blindly and one action to observe without suffering its loss at each round. These two actions may differ or coincide. This setting is motivated by full information problems with restricted data access, where the loss of any action can be accessed in principle, but the player can only observe the loss of one action due to the cost of data acquisition. From a practical perspective, such scenarios arise in various domains. For instance, in Ultra-Wide Band (UWB) communications, a learner transmits data on one channel (exploitation) while sensing a different frequency band (exploration) to avoid mutual interference [Avner et al., 2012]. Similarly, in robotics, an agent acts and learns in a simulator (exploration) to preserve the safety and integrity of the physical robot (exploitation) [Zhao et al., 2020]. Additionally, in healthcare, ethical regulations enforce a decoupled structure: clinical trials serve as the exploration arm to validate new treatments on a small cohort,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Table 1: Comparison of regret bounds for the decoupled setting. T denotes the time horizon, K is the number of arms, $i^\star$ denotes the optimal arm, and Δ_i is the suboptimality gap of arm i.

Algorithm	Stochastic Regimes	Adversarial Regimes
Rouyer and Seldin, 2020	$\mathcal{O}\left(\sqrt{\frac{K}{\Delta_{\min}}\sum_{i\neq i^\star}\frac{1}{\Delta_i}}\right)$	$\mathcal{O}(\sqrt{KT})$
Jin et al., 2023 *	$\mathcal{O}\left(\sqrt{K\sum_{i:\Delta_i>0}\frac{1}{\Delta_i^2}}\right)$	$\mathcal{O}\left(\sqrt{KT}\right)$
D-Exp3-Ada	$\mathcal{O}\left(\sum_{i\neq i^\star}\frac{\log K}{\Delta_i}\right)$	$\mathcal{O}\left(\sqrt{KT\log K}\right)$
D-TINF-SPM†	$\mathcal{O}\left(\min\left\{\sqrt{K\sum_{i\neq i^\star}\frac{1}{\Delta_i^2}}, \sum_{i\neq i^\star}\frac{\log K}{\Delta_i}\right\}\right)$	$\mathcal{O}\left(\sqrt{KT}\right)$

† The displayed D-TINF-SPM stochastic bound is a simplified consequence of the sharper bound $\mathcal{O}(C_1(1+\log(\sqrt{KC_2}/C_1)))$ with $C_1=\sum_{i\neq i^\star}1/\Delta_i$ and $C_2=\sum_{i\neq i^\star}1/\Delta_i^2$ (Theorem 5.4).

* This stochastic bound holds without assuming a unique optimal arm. The other stochastic bounds assume uniqueness.

while general medical practice acts as the exploitation arm, providing the best-known care to the public [Villar et al., 2015].

The decoupled setting can be viewed as an intersection of standard bandits and pure exploration (Even-Dar et al., 2006; Gabillon et al., 2011). Consequently, algorithms designed for best-arm identification, which primarily focus on exploration to identify the optimal arm, may appear applicable to this setting. However, applying an algorithm developed for best-arm identification to the exploration selection does not achieve the optimal regret bound. Similarly, explore-then-commit policies [Garivier et al., 2016], which partition exploration and exploitation into distinct phases with each round dedicated exclusively to one task, do not achieve the optimal regret bound.

The decoupled setting also bridges the gap between full information and bandit setups. While the decoupled setting is as hard as the standard MAB in adversarial regimes, the regret bound in stochastic regimes is time-independent, as in full information problems.

Regret bounds for the decoupled setting have been studied in both stochastic and adversarial regimes, and in adversarial regimes, the minimax optimal bound is known. Avner et al. [2012] established a lower bound of $\Omega(\sqrt{KT})$ for the regret in the decoupled setting under adversarial regimes, which coincides with that of the standard MAB problem. Rouyer and Seldin [2020] proposed an algorithm that achieves an $\mathcal{O}(\sqrt{KT})$ regret upper bound in adversarial regimes.

While the adversarial setting is well-understood, several key challenges remain in other aspects. First, a gap exists between the upper and lower bounds in stochastic regimes, implying that the existing algorithms are suboptimal or the lower bound itself is loose. The algorithm proposed by Rouyer and Seldin [2020] achieves $\mathcal{O}\left(\sqrt{\frac{K}{\Delta_{\min}}\sum_{i\neq i^\star}\frac{1}{\Delta_i}}\right)$. Subsequently, Jin et al. [2023] refined the stochastic regret bound of the same algorithm to $\mathcal{O}\left(\sqrt{K\sum_{i:\Delta_i>0}\frac{1}{\Delta_i^2}}\right)$ without requiring a unique optimal arm. Nevertheless, a substantial gap can remain between this upper bound and the conjectured instance-dependent lower bound of $\Omega\left(\sum_{i:\Delta_i>0}\frac{1}{\Delta_i}\right)$, which is yet to be proven. Second, achieving optimal "Best-of-Both-Worlds" (BOBW) performance (Bubeck and Slivkins, 2012; Zimmert and Seldin, 2021; Ito et al., 2024) remains a challenge. While algorithms capable of handling both regimes exist [Rouyer and Seldin, 2020], no single strategy is known to attain the tightest possible regret bounds in both stochastic and adversarial environments simultaneously. These limitations motivate the need for a more refined analysis and a unified algorithmic framework.

This paper makes the following contributions. First, we propose two algorithms for the decoupled setting. We first introduce D-Exp3-Ada, based on the standard Exp3 algorithm [Auer et al., 2002b] with an AdaGrad learning rate [Duchi et al., 2011], which combines simplicity of algorithm design with strong theoretical guarantees. To further close the optimality gap, we introduce D-TINF-SPM, based on the Follow the Regularized Leader (FTRL) framework with Tsallis entropy regularization and the stability-penalty matching (SPM) learning rate [Ito et al., 2024]. The stochastic regret bounds stated below assume a unique optimal arm $i^\star$. We prove that D-Exp3-Ada achieves $\mathcal{O}\left(\sqrt{KT\log K}\right)$ regret in adversarial regimes and $\mathcal{O}\left(\sum_{i\neq i^\star}\frac{\log K}{\Delta_i}\right)$ in stochas-

tic regimes, while D-TINF-SPM achieves the minimax optimal $\mathcal{O}\left(\sqrt{KT}\right)$ regret in adversarial regimes and $\mathcal{O}\left(\min\left\{\sqrt{K\sum_{i\neq i^\star}\frac{1}{\Delta_i^2}}, \sum_{i\neq i^\star}\frac{\log K}{\Delta_i}\right\}\right)$ in stochastic regimes (see Table 1). The improvement over the prior stochastic bound is quantified by a factor of $\widetilde{\Theta}\left(\sqrt{K/m_{\text{eff}}}\right)$, where $m_{\text{eff}} = (\sum_{i\neq i^\star}\Delta_i^{-1})^2/(\sum_{i\neq i^\star}\Delta_i^{-2})$ is the effective number of hard arms (see Remark 5.5).

Second, we derive instance-dependent lower bounds for the decoupled setting. To the best of our knowledge, the only instance-dependent lower bound in the literature is $\Omega(\log K/\Delta)$ for uniform gaps, which follows from the full-information setting and does not exploit the structure of the decoupled problem. We strengthen this to $\Omega(K/\Delta)$ for uniform gaps (i.e., $\Delta_i = \Delta$ for all $i \neq i^\star$), improving the dependence on K from logarithmic to linear. More generally, when m arms have gap Δ and the remaining arms have a large constant gap, we prove a lower bound of $\Omega(m/\Delta)$. Combining these results, we establish an instance-dependent lower bound of $\Omega\left(\frac{1}{\log K}\sum_{i\neq i^\star}\frac{1}{\Delta_i}\right)$ for any algorithm with a monotonically non-increasing and permutation-invariant regret bound. These are natural structural assumptions that serve a purpose analogous to the consistency assumption in standard MAB (e.g., Lattimore and Szepesvári [2020, Chapter 16]), essentially excluding pathological policies. Consequently, this establishes that, among algorithms satisfying these mild conditions, D-Exp3-Ada and D-TINF-SPM are optimal up to a factor of $\mathcal{O}\left((\log K)^2\right)$. Taken together, our upper and lower bounds give a two-sided characterization of the stochastic decoupled setting: decoupling replaces the $\Omega(\log T)$ stochastic regret that is unavoidable in standard MAB by a T-independent cost of order $\sum_{i\neq i^\star} 1/\Delta_i$, i.e., linear in the number of statistically hard arms, and our results determine this cost up to a factor of $\mathcal{O}((\log K)^2)$, whereas the best previous upper bounds exceed it by a factor polynomial in K, namely $\widetilde{\Theta}(\sqrt{K/m_{\text{eff}}})$.

Achieving these results requires overcoming several technical challenges inherent to the decoupled setting. For algorithm construction, the increased flexibility of selecting two actions separately, combined with their interdependence, necessitates re-establishing the validity of existing MAB algorithms and adapting their analysis. In particular, applying the stability-penalty matching (SPM) learning rate [Ito et al., 2024] requires determining appropriate parameter settings and verifying the required conditions; notably, the condition $h_{t+1} = \mathcal{O}(h_t)$, which Ito et al. [2024] enforce through a hybrid regularizer, can be proved for the pure Tsallis regularizer in our decoupled design (see the proof of Theorem 5.4 in Appendix). Both algorithms rest on a design principle that is specific to decoupling: allocate a constant fraction (one half) of the exploration probability to the arm that is currently most likely to be exploited, $i_t^\star$, and spread the remainder over the other arms. This keeps the loss estimator of the exploited arm stable and makes the FTRL stability term proportional to the instantaneous regret, $z_t = \mathcal{O}(r_t)$ with $r_t = \sum_{i\neq i^\star}\Delta_i p_{ti}$, which is what yields T-independent stochastic regret. This is impossible in the standard MAB, where the coupling $q_t = p_t$ forces z_t to be of order K however concentrated p_t is, and it also differs from the exploration $q_t \propto p_t^{2/3}$ of Rouyer and Seldin [2020], under which the total exploration of the arms other than $i_t^\star$ vanishes as p_t concentrates. The same principle should also be useful in other settings in which the information-gathering action is chosen separately from the loss-incurring one, such as purchased feedback or simulator queries. Moreover, for the stochastic analysis of D-TINF-SPM, existing SPM techniques would introduce a $\log T$ factor that is unavoidable in standard bandits but unnecessary in the decoupled setting; we develop a three-phase decomposition analysis that utilizes this exploration design to obtain T-independent bounds (see Section 5.3). For lower bounds, the freedom in selecting decoupled actions prevents direct application of standard MAB techniques, as both exploration and exploitation actions must be evaluated jointly.

2 Related work

The decoupled setting was introduced by Avner et al. [2012], who established a lower bound of $\Omega(\sqrt{KT})$ in adversarial regimes. They also proposed an Exp3-based algorithm that achieves $\mathcal{O}(\sqrt{KT\log K})$ regret in adversarial regimes and $\mathcal{O}\left(\sqrt{T\log K}\right)$ regret in stochastic regimes, while further providing high-probability guarantees and tighter bounds that adapt to specific structures in adversarial environments. However, the optimal learning rate differs between stochastic and adversarial regimes, thus requiring prior knowledge of the environment. Rouyer and Seldin [2020]

proposed an algorithm based on the FTRL framework with Tsallis entropy regularization with the power $\alpha = 2/3$, which achieves $\mathcal{O}(\sqrt{KT})$ regret in adversarial regimes and $\mathcal{O}\left(\sqrt{\frac{K}{\Delta_{\min}}\sum_{i\neq i^\star}\frac{1}{\Delta_i}}\right)$ regret in stochastic regimes. Jin et al. [2023] subsequently refined the stochastic regret bound of the same algorithm to $\mathcal{O}\left(\sqrt{K\sum_{i:\Delta_i>0}\frac{1}{\Delta_i^2}}\right)$ without requiring a unique optimal arm, and also established guarantees for corrupted settings. Recently, Kim et al. [2026] proposed a Follow-the-Perturbed-Leader (FTPL)-based approach that maintains the same order of regret bounds as Rouyer and Seldin [2020]. Their method achieves substantial computational efficiency, with an empirical speedup of approximately 20 times compared to the Tsallis-INF algorithm of Rouyer and Seldin [2020]. While this computational improvement is highly valuable, their approach does not fundamentally improve the theoretical order of the regret bound itself. As a result, a fundamental theoretical challenge remains unresolved: although these existing algorithms handle both regimes without prior knowledge, there is still a clear gap between the upper and lower bounds in stochastic regimes.

A separate line of work, known as Best-of-Both-Worlds (BOBW) algorithms (Bubeck and Slivkins, 2012; Zimmert and Seldin, 2021; Ito et al., 2024), aims to perform optimally in both stochastic and adversarial environments without knowing the regime. Of particular relevance to us is the stability-penalty matching (SPM) learning rate of Ito et al. [2024], which we adapt to the decoupled setting; standard SPM analyses, however, would introduce an unwanted $\log T$ factor in the stochastic regime (Section 5.3). Background on BOBW and SPM, and a more detailed comparison with Avner et al. [2012] and Ito et al. [2024], are deferred to Appendix.

3 Problem setup

We consider a repeated game with K arms ($K \geq 2$). At each round $t = 1, 2, \ldots, T$, the environment selects a loss vector $\ell_t \in [0,1]^K$, and the player selects an action A_t for exploitation and an action B_t for exploration. These two actions may differ or coincide. The player then incurs the loss ℓ_{t,A_t} without observing it, while observing ℓ_{t,B_t} without incurring this loss.

In the oblivious adversarial setting, the loss vectors ℓ_t are fixed arbitrarily before the game begins.

In the stochastic setting, the loss vectors ℓ_t are drawn independently and identically distributed (i.i.d.) from an unknown distribution with mean $\mu \in [0,1]^K$.

We measure the performance of an algorithm by the pseudo-regret defined as follows:

$$R_T \coloneqq \mathbb{E}\left[\sum_{t=1}^{T}\ell_{t,A_t}\right] - \min_{i\in[K]}\mathbb{E}\left[\sum_{t=1}^{T}\ell_{t,i}\right] = \mathbb{E}\left[\sum_{t=1}^{T}\langle p_t - e_{i^\star}, \ell_t\rangle\right], \tag{1}$$

where e_i is the i-th standard basis vector in $\mathbb{R}^K$ and $i^\star \coloneqq \operatorname*{argmin}_{i\in[K]}\mathbb{E}\left[\sum_{t=1}^{T}\ell_{t,i}\right]$. Also, $p_t \in \Delta^{K-1} \coloneqq \left\{p \in [0,1]^K \,\middle|\, \|p\|_1 = 1\right\}$ is the distribution from which the action A_t is chosen, i.e., $\Pr\left[A_t = i \middle| \mathcal{H}_{t-1}\right] = p_{ti}$. Here, $\mathcal{H}_{t-1}$ denotes the history up to round $t-1$, i.e., $\mathcal{H}_{t-1} = \{(A_s, B_s, \ell_{s,B_s})_{s=1}^{t-1}\}$. Note that $\mathcal{H}_{t-1}$ does not include ℓ_{s,A_s}, as this quantity is not observed by the algorithm.

We also define the distribution $q_t \in \Delta^{K-1}$ from which the action B_t is chosen, i.e., $\Pr\left[B_t = i \middle| \mathcal{H}_{t-1}\right] = q_{ti}$.

In the stochastic setting, we denote the suboptimality gap of arm i by $\Delta_i \coloneqq \mu_i - \mu_{i^\star}$. We assume that the optimal arm is unique. This assumption is used only for the stochastic regret upper bounds in Section 5; see Remark F.4 in Appendix for where it enters the analysis and Appendix G.6 for experiments with multiple optimal arms. Using these notations, the pseudo-regret can be rewritten as

$$R_T = \sum_{t=1}^{T}\sum_{i\neq i^\star}\Delta_i\mathbb{E}\left[p_{ti}\right]. \tag{2}$$

Throughout this paper, we use $a \lesssim b$ to denote $a \leq Cb$ for some universal constant $C > 0$, and $a \asymp b$ to denote that both $a \lesssim b$ and $b \lesssim a$ hold.

4 Algorithms

All algorithms considered in this work are based on the Follow the Regularized Leader (FTRL) framework. At each round t, the exploitation distribution p_t from which action A_t is sampled is computed as

$$p_t = \operatorname*{argmin}_{p \in \Delta^{K-1}} \left\{ \langle p, \sum_{s=1}^{t-1} \hat{\ell}_s \rangle + \frac{1}{\eta_t} \psi(p) \right\}, \tag{3}$$

where $\psi : \Delta^{K-1} \to \mathbb{R}$ is a regularization function, $\eta_t > 0$ is the learning rate, and $\hat{\ell}_t$ is an importance-weighted loss estimator. Throughout this work, we employ the following estimator:

$$\hat{\ell}_{ti} = \frac{\ell_{ti} \mathbb{1}\{B_t = i\}}{q_{ti}}, \tag{4}$$

where B_t is the exploration arm sampled from the distribution $q_t \in \Delta^{K-1}$.

We consider three algorithms within this framework: Tsallis-INF-$2/3$ [Rouyer and Seldin, 2020] and our two proposed algorithms, D-Exp3-Ada and D-TINF-SPM. These algorithms share the same importance-weighted estimator construction (4) but differ in three key components: (1) the regularization function ψ, (2) the exploration distribution q_t, and (3) the learning rate update rule for η_t. The differences in learning rate adaptation are particularly significant: Tsallis-INF-$2/3$ employs a decaying learning rate $\eta_t \propto 1/\sqrt{t}$, whereas D-Exp3-Ada utilizes an AdaGrad-style adaptive rate [Duchi et al., 2011], and D-TINF-SPM adopts the stability-penalty matching (SPM) learning rate [Ito et al., 2024].

4.1 Regularization functions

All three algorithms employ entropy-based regularization. Tsallis-INF-$2/3$ and D-TINF-SPM use Tsallis entropy with parameter $\alpha \in (0, 1)$:

$$\psi_\alpha(p) = -\frac{1}{\alpha} \sum_{i=1}^{K} (p_i^\alpha - p_i). \tag{5}$$

Specifically, Tsallis-INF-$2/3$ sets $\alpha = 2/3$, while D-TINF-SPM sets $\alpha = 1/2$. D-Exp3-Ada employs negative Shannon entropy:

$$\phi(p) = \sum_{i=1}^{K} p_i \log p_i. \tag{6}$$

Note that Shannon entropy can be viewed as the limit of Tsallis entropy as $\alpha \to 1$ (after absorbing a $\frac{1}{1-\alpha}$ factor into the learning rate; see Remark F.1 in Appendix); consequently, the analysis of D-Exp3-Ada can be derived as a limiting case of the Tsallis entropy framework.

4.2 Exploration distributions

The exploration distribution q_t determines how the exploration arm B_t is sampled. Let $i_t^\star = \operatorname*{argmax}_i p_{ti}$ denote the arm with the highest exploitation probability at round t (ties are broken by choosing the smallest index). Both D-Exp3-Ada and D-TINF-SPM employ the following parametric form of the exploration distribution:

$$q_{ti_t^\star} = \frac{1}{2}, \quad q_{ti} = \frac{p_{ti}^{1-\alpha/2}}{2 \sum_{j \neq i_t^\star} p_{tj}^{1-\alpha/2}} \quad (i \neq i_t^\star), \tag{7}$$

where $\alpha = 1$ for D-Exp3-Ada and $\alpha = 1/2$ for D-TINF-SPM. In contrast, Tsallis-INF-$2/3$ does not distinguish the arm $i_t^\star$ and sets $q_{ti} \propto p_{ti}^{2/3}$ for all $i \in [K]$.

Table 2: Comparison of algorithm components. Here, $i_t^\star = \operatorname*{argmax}_i p_{ti}$.

Component	Tsallis-INF-$2/3$ [Rouyer and Seldin, 2020]	D-Exp3-Ada (This work)	D-TINF-SPM (This work)
Regularizer ψ	Tsallis (5), $\alpha = 2/3$	Shannon (6)	Tsallis (5), $\alpha = 1/2$
Learning rate η_t	$\propto K^{-1/6}/\sqrt{t}$	AdaGrad-style (9)	Stability-penalty-matching (10)
q_{ti} $(i = i_t^\star)$	$\propto p_{ti}^{2/3}$	$1/2$	$1/2$
q_{ti} $(i \neq i_t^\star)$	$\propto p_{ti}^{2/3}$	$\propto p_{ti}^{1/2}$	$\propto p_{ti}^{3/4}$

4.3 Learning rate update rules

The learning rate schedules differ substantially across algorithms. For all algorithms, the regret analysis relies on the following standard FTRL bound (Lemma 5.1)

$$R_T \lesssim \mathbb{E}\left[\sum_{t=1}^{T} \eta_t z_t + \frac{h_1}{\eta_1} + \sum_{t=2}^{T} \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) h_t\right], \tag{8}$$

where z_t is the stability term (measuring the sensitivity of the FTRL update to the loss estimator) and h_t is the penalty term (an upper bound on $-\psi(p_t)$). The specific forms of z_t and h_t depend on the regularizer and are defined below for each algorithm.

Tsallis-INF-$2/3$ [Rouyer and Seldin, 2020]. This algorithm employs a decaying learning rate $\eta_t \propto K^{-1/6}/\sqrt{t}$.

D-Exp3-Ada. This algorithm adapts the learning rate using an AdaGrad-style rule:

$$\eta_t = \frac{1}{\beta_t}, \quad \beta_t = \sqrt{\frac{2}{\bar{h}}}\sqrt{K + \sum_{s=1}^{t-1} z_s}, \tag{9}$$

where $\bar{h} = \log K$ is a constant penalty bound. The stability term is defined as $z_t = \left(\sum_{i \neq i_t^\star} \sqrt{p_{ti}}\right)^2$.

D-TINF-SPM. This algorithm employs the SPM learning rate Rule 2 of Ito et al. [2024]:

$$\eta_t = \frac{1}{\beta_t}, \quad \beta_t = \beta_{t-1} + \frac{z_{t-1}}{\beta_{t-1} h_{t-1}}, \quad \beta_1 = 16, \tag{10}$$

where $\beta_1 = 16$ is chosen to satisfy the technical conditions required for the SPM analysis (see Section 5 for details). The stability and penalty terms are defined as:

$$z_t = \frac{1}{1-\alpha}\left(\sum_{i \neq i_t^\star} p_{ti}^{1-\alpha/2}\right)^2, \quad h_t = \frac{1}{\alpha}\sum_{i \neq i_t^\star} p_{ti}^{\alpha}. \tag{11}$$

When $h_t \equiv \bar{h}$ is constant, the SPM rule reduces to the AdaGrad-style $\beta_t \asymp \sqrt{\sum_{s<t} z_s/\bar{h}}$ used by D-Exp3-Ada; SPM generalizes this by allowing time-varying h_t (see Remark F.2 in Appendix).

4.4 Summary of algorithm components

Table 2 summarizes the key differences among the three algorithms.

The choices in Table 2 are motivated by the regret analysis in Section 5. In particular, fixing $q_{t,i_t^\star} = 1/2$ is not strictly optimal but is a well-justified approximation of the exact minimizer (Remark A.3 in Appendix); the essential structural requirement is to allocate $\Theta(1)$ probability mass to $i_t^\star$, which prevents the importance-weighted estimator $\hat{\ell}_{t,i} = \ell_{t,i}\mathbf{1}\{B_t = i\}/q_{t,i}$ from blowing up to order K.

Computationally, D-Exp3-Ada's update has the closed form $p_{ti} \propto \exp(-\eta_t \sum_{s<t} \hat{\ell}_{si})$ from Shannon-entropy regularization, while D-TINF-SPM's Tsallis-entropy update can be computed efficiently by Newton's method as in Section 3.3 of Zimmert and Seldin [2021].

5 Regret upper bounds

Throughout this section, all lemmas and theorems whose proofs are not provided in the main text and are not attributed to prior work are proved in the Appendix.

5.1 Basic analysis of FTRL

We first present a fundamental analysis of FTRL with a general convex Legendre regularizer, which covers both the Tsallis entropy (D-TINF-SPM) and the Shannon entropy (D-Exp3-Ada).

Lemma 5.1. *Let $\{p_t\}_{t=1}^T$ be the sequence generated by the FTRL algorithm, as defined in* (3)*. For FTRL with a convex Legendre regularizer ψ and learning rate $\eta_t > 0$, the following holds for any $T \geq 1$:*

$$R_T \leq \mathbb{E}\left[\sum_{t=1}^T S_t + \frac{h_1}{\eta_1} + \frac{\psi(e_{i^\star})}{\eta_T} + \sum_{t=2}^T \left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right) h_t\right], \tag{12}$$

where $S_t \coloneqq \left\langle \hat{\ell}_t, p_t - p_{t+1} \right\rangle - \frac{1}{\eta_t} D\left(p_{t+1}, p_t\right)$ is the stability term *and $h_t \coloneqq -\psi(p_t)$ is the* penalty term*. Here, $D(p,q) \coloneqq \psi(p) - \psi(q) - \langle \nabla\psi(q), p - q\rangle$ is the Bregman divergence induced by ψ.*

We first analyze the stability term S_t by applying Lemma 21 of Ito et al. [2024], restated as Lemma A.1 in Appendix. For D-Exp3-Ada, which uses Shannon entropy, we define $z_t = \left(\sum_{i \neq i_t^\star} \sqrt{p_{ti}}\right)^2$; for D-TINF-SPM, we define $z_t = \frac{1}{1-\alpha}\left(\sum_{i \neq i_t^\star} p_{ti}^{1-\alpha/2}\right)^2$ (see (11)).

Lemma 5.2 (Stability Bounds)**.** *The stability term S_t is bounded as follows:*

(i) ***Shannon entropy (D-Exp3-Ada):*** *Let $\{p_t, \eta_t, \hat{\ell}_t\}_{t=1}^T$ be generated by D-Exp3-Ada with Shannon entropy regularization. Then, for any $t \geq 1$,*

$$\mathbb{E}\left[S_t \middle| \mathcal{H}_{t-1}\right] \lesssim \eta_t \left(\sum_{i \neq i_t^\star} \sqrt{p_{ti}}\right)^2. \tag{13}$$

(ii) ***Tsallis entropy (D-TINF-SPM):*** *Let $\{p_t, \eta_t, \hat{\ell}_t\}_{t=1}^T$ be generated by D-TINF-SPM with Tsallis entropy regularization with parameter $\alpha \in (0,1)$. Assume that $\beta_1 \geq 8/(1-\alpha)$. Then, for any $t \geq 1$,*

$$\mathbb{E}\left[S_t \middle| \mathcal{H}_{t-1}\right] \lesssim \frac{\eta_t}{1-\alpha} \left(\sum_{i \neq i_t^\star} p_{ti}^{1-\alpha/2}\right)^2. \tag{14}$$

The proof of (i) is provided in Appendix (Lemma A.4), and the proof of (ii) applies Lemma A.1 to our setting.

The penalty term is bounded differently for D-Exp3-Ada and D-TINF-SPM, reflecting their distinct regularization functions.

Penalty bounds for D-Exp3-Ada (Shannon entropy). For D-Exp3-Ada with Shannon entropy regularization ϕ, we use a constant penalty bound $\bar{h} = \log K$. This bound is justified as follows:

$$-\phi(p_t) = -\sum_{i=1}^K p_{ti} \log p_{ti} \leq \log K,$$

where the inequality holds with equality when p_t is the uniform distribution, which maximizes the Shannon entropy over the probability simplex.

Penalty bounds for D-TINF-SPM (Tsallis entropy). For D-TINF-SPM with Tsallis entropy regularization ψ_α, we use a time-varying penalty bound $h_t = \frac{1}{\alpha}\sum_{i\neq i_t^\star} p_{ti}^\alpha$, which satisfies

$$-\psi_\alpha(p_t) = \frac{1}{\alpha}\left(\sum_{i=1}^K p_{ti}^\alpha - 1\right) \leq \frac{1}{\alpha}\sum_{i\neq i_t^\star} p_{ti}^\alpha = h_t,$$

where the inequality holds since $p_{ti_t^\star}^\alpha \leq 1$ for $\alpha \in (0,1)$.

5.2 Analysis of D-Exp3-Ada

D-Exp3-Ada employs Shannon entropy regularization with AdaGrad-style learning rate (9). By Lemma 5.2(i) and the penalty bound $\bar{h} = \log K$ established above, the stability term satisfies $\mathbb{E}\left[S_t \middle| \mathcal{H}_{t-1}\right] \lesssim \eta_t z_t$ with $z_t = \left(\sum_{i\neq i_t^\star}\sqrt{p_{ti}}\right)^2$. From the standard FTRL bound (Lemma 5.1), we have (8). The AdaGrad-style learning rate (9) is designed to balance the stability and penalty terms. The detailed analysis is provided in Appendix, which yields the following regret upper bound.

Theorem 5.3 (Regret Upper Bound for D-Exp3-Ada)**.** *D-Exp3-Ada achieves $R_T = \mathcal{O}\left(\sqrt{KT\log K}\right)$ in adversarial regimes and $R_T = \mathcal{O}\left(C_1 \log K\right)$ in stochastic regimes, where $C_1 = \sum_{i\neq i^\star}\frac{1}{\Delta_i}$.*

Proof sketch. Standard AdaGrad-style analysis yields $R_T = \mathcal{O}(\mathbb{E}[\sqrt{\bar{h}\sum_t z_t}] + \sqrt{K\log K})$. The adversarial bound follows from $z_t \leq K$. For the stochastic bound, set $r_t = \sum_{i\neq i^\star}\Delta_i p_{ti}$ so that $\mathbb{E}[\sum_t r_t] = R_T$; Cauchy–Schwarz gives $z_t \leq C_1 r_t$, and the resulting self-bounding inequality $R_T \lesssim \sqrt{\bar{h}\cdot C_1 \cdot R_T}$ yields $R_T = \mathcal{O}(\bar{h}\cdot C_1) = \mathcal{O}(C_1 \log K)$. See Appendix for details. □

The T-independent stochastic regret hinges on $z_t = \mathcal{O}(r_t)$, which the decoupled setting makes possible by letting us choose q_t freely; in classical MAB, the coupling $q_t = p_t$ forces $z_t = \mathcal{O}(K)$, and even $\mathcal{O}(\log T)$ stochastic regret with Shannon entropy then appears difficult (see Remark F.3 in Appendix).

5.3 Analysis of D-TINF-SPM

D-TINF-SPM employs Tsallis entropy regularization with $\alpha = 1/2$ and the SPM learning rate (10). By Lemma 5.2(ii) and the penalty bound $h_t = \frac{1}{\alpha}\sum_{i\neq i_t^\star} p_{ti}^\alpha$ established above, the standard FTRL bound (8) yields $R_T \leq \mathbb{E}\left[\sum_{t=1}^T \eta_t z_t + \frac{h_1}{\eta_1} + \sum_{t=2}^T\left(\frac{1}{\eta_t} - \frac{1}{\eta_{t-1}}\right)h_t\right]$, where z_t and h_t are as defined in (11).

Adapting Lemma 9 of Ito et al. [2024], the SPM learning rate gives, under four explicit conditions (measurability and uniform boundedness of z_t, h_t, the stability bound, the penalty bound, and $h_{t+1} = \mathcal{O}(h_t)$), $R_T = \mathcal{O}(\mathbb{E}[\sum_t z_t/\sqrt{\sum_{s\leq t} z_s/h_s + \bar{z}/\beta_1 + \beta_1 h_1}])$ (see Lemma A.2 and its proof in Appendix; we set $\beta_1 = 16$ as the smallest integer satisfying the conditions). A key technical challenge in the stochastic regime is to bound $\sum_t z_t/\sqrt{\sum_{s\leq t} z_s/h_s}$ without an extra $\log T$ factor: existing SPM analyses partition the horizon into $\mathcal{O}(\log T)$ phases before applying Cauchy–Schwarz, which is incompatible with the T-independent stochastic regret targeted here. We instead introduce a three-phase decomposition tailored to the decoupled setting, sketched below.

Theorem 5.4 (Regret Upper Bound for D-TINF-SPM)**.** *D-TINF-SPM achieves $R_T = \mathcal{O}\left(\sqrt{KT}\right)$ in adversarial regimes. In stochastic regimes, $R_T = \mathcal{O}\left(C_1\left(1+\log\left(\frac{\sqrt{KC_2}}{C_1}\right)\right)\right) \leq \mathcal{O}\left(\min\left\{\sqrt{KC_2}, C_1\log K\right\}\right)$, where $C_1 = \sum_{i\neq i^\star}\frac{1}{\Delta_i}$ and $C_2 = \sum_{i\neq i^\star}\frac{1}{\Delta_i^2}$.*

Proof sketch. For the adversarial regime, applying Cauchy–Schwarz to the first term gives $\sum_t z_t/\sqrt{\sum_{s\leq t} z_s/h_s} \leq \sqrt{\bar{h}\bar{z}T}$ with $\bar{h}, \bar{z} = \mathcal{O}(\sqrt{K})$, which yields $R_T = \mathcal{O}(\sqrt{KT})$.

For the stochastic regime, the key technical challenge is to avoid an extra $\log T$ factor that arises when standard SPM analyses partition the horizon into $\mathcal{O}(\log T)$ phases. The new step is to define

the phases by hitting times of $\sum_{s\le t} z_s/h_s$ and to set the second threshold to a sufficiently large constant multiple of C_2, which removes the $\log T$ phase decomposition and lets the resulting tail term be absorbed by self-bounding. Concretely, let $T_1 \le T_2 \le T$ be the hitting times at which the cumulative quantity $\sum_{s\le t} z_s/h_s$ reaches C_1^2/K and $L \cdot C_2$ (with L a sufficiently large universal constant), with the convention that the corresponding time is set to T if a threshold is not reached. Here $C_1 = \sum_{i\neq i^\star} 1/\Delta_i$ and $C_2 = \sum_{i\neq i^\star} 1/\Delta_i^2$. Combining a path-wise three-phase decomposition with $z_t \lesssim \sqrt{C_2}\, r_t$ and $z_t h_t \lesssim C_1 r_t$ (with $r_t := \sum_{i\neq i^\star} \Delta_i p_{ti}$, so that $\mathbb{E}[\sum_t r_t] = R_T$) and Jensen's inequality then gives the self-bounding inequality $R_T \lesssim C_1 + \sqrt{\log(KC_2/C_1^2)}\sqrt{C_1 R_T} + \frac{1}{2}R_T$. Solving for R_T yields the stated bound. Full details, including the verification of the SPM lemma's technical conditions, are deferred to Appendix. □

Remark 5.5 (Quantifying the Improvement in the Upper Bound). The quantity $m_{\text{eff}} = C_1^2/C_2 \in [1, K-1]$ can be interpreted as the effective number of hard arms. For example, m_{eff} approaches 1 as one suboptimality gap tends to zero while all other suboptimality gaps remain fixed. When all suboptimal arms have the same gap, $m_{\text{eff}} = K-1$. The previously known stochastic bound $\mathcal{O}(\sqrt{KC_2})$ can be written as $\mathcal{O}(C_1\sqrt{K/m_{\text{eff}}})$, whereas Theorem 5.4 yields $R_T = \mathcal{O}\left(C_1\left(1+\log\sqrt{K/m_{\text{eff}}}\right)\right)$. Suppressing logarithmic factors, the ratio of the prior and new rate expressions is $\widetilde{\Theta}(\sqrt{K/m_{\text{eff}}})$. This ratio is $\widetilde{\Theta}(\sqrt{K})$ on natural instances dominated by a few small-gap arms, and it shrinks to $\mathcal{O}(1)$ only when $m_{\text{eff}} = \Theta(K)$, as in the uniform-gap case.

We also provide an alternative proof of the stochastic regret bound $\mathcal{O}\left(\sqrt{K\sum_{i\neq i^\star}\frac{1}{\Delta_i^2}}\right)$ established by Jin et al. [2023] for Tsallis-INF-$2/3$, under the assumption of a unique optimal arm. Furthermore, we show that the regret bound of D-TINF-SPM is no worse, up to constant factors, than the bounds of both D-Exp3-Ada and Tsallis-INF-$2/3$ for any instance. A detailed comparison of the algorithms is provided in Appendix.

6 Regret lower bounds

In this section, we establish asymptotic lower bounds for the decoupled MAB problem in stochastic regimes. These lower bounds characterize the fundamental difficulty of the decoupled setting, where information is gathered only through the exploration arm while regret is incurred by the exploitation arm. Our approach relies on information-theoretic arguments: we construct alternative environments by swapping the roles of the optimal arm and a suboptimal arm, and show that any algorithm must incur sufficient regret to distinguish between these environments.

Our key technical tool is a KL-divergence decomposition (Lemma D.2 in Appendix), which adapts Lemma 15.1 of Lattimore and Szepesvári [2020] to the decoupled setting. The crucial difference from the classical version is that the KL divergence depends only on the number of times each arm is chosen for *exploration* (denoted $T_i^{(B)}$), rather than the total number of pulls. Using this lemma, we establish lower bounds for different gap structures.

When the suboptimality gaps are uniform, i.e., $\Delta_i = \Delta$ for all $i \neq i^\star$, we obtain the following tight lower bound.

Theorem 6.1 (Lower Bound for Uniform Gaps)**.** *For any decoupled Bernoulli bandit environment μ with K arms where $\Delta_i = \Delta$ for all $i \neq i^\star$, there exists an environment μ', which is either μ itself or obtained by swapping the roles of some suboptimal arm i and the optimal arm $i^\star$ in μ, such that the regret on μ' satisfies*

$$R_T = \Omega\left(\frac{K}{\Delta}\right)$$

for $T \ge \frac{4(K-1)^2}{9C^2\Delta^2 K}$, where $C > 0$ is some constant. Here, μ is assumed to lie in $[c, 1-c]^K$ for some constant $c \in (0, \frac{1}{2})$.

The proof uses Lemma D.2 and standard information-theoretic arguments (Appendix). The lower bound scales linearly with K and inversely with Δ, matching our upper bounds up to logarithmic factors.

We next handle heterogeneous gaps where only a subset of arms are hard to distinguish from the optimum.

Theorem 6.2 (Lower Bound for Heterogeneous Gaps)**.** *Let $c \in (0, 1/2)$, $m \in \{1, \dots, K-1\}$, and $\Delta \in (0, 1/2 - c)$. For any decoupled Bernoulli bandit environment μ with K arms where m arms have suboptimality gap Δ and the remaining $K - m - 1$ arms have suboptimality gap at least $\frac{1}{2} - c$, there exists an environment μ', which is either μ itself or obtained by swapping the roles of some suboptimal arm i with gap Δ and the optimal arm $i^\star$ in μ, such that the regret on μ' satisfies*

$$R_T = \Omega\left(\frac{m}{\Delta}\right)$$

for $T \geq \frac{4m^2}{9C^2\Delta^2(m+1)}$, where $C > 0$ is some constant. Here, μ is assumed to lie in $[c, 1-c]^K$ for some constant $c \in (0, \frac{1}{2})$.

The proof follows the same technique as Theorem 6.1 (Appendix). The bound depends only on the m hard arms, since easy arms can be quickly identified and avoided. Note that the hard environment μ' in Theorems 6.1 and 6.2 may be μ itself (see Remark 6.4 below).

Using a harmonic sum inequality (Lemma D.4 in Appendix), we can derive from Theorem 6.2 a general lower bound that demonstrates the near-optimality of our algorithms.

Corollary 6.3 (Near-Optimality of Our Algorithms)**.** *Fix $c \in (0, 1/2)$. For any algorithm and for any $\underline{\Delta} = (\underline{\Delta}_1, \dots, \underline{\Delta}_{K-1}) \in [0, 1/2 - c]^{K-1}$, there exists a Bernoulli bandit environment with optimal arm $i^\star$ such that the suboptimality gaps satisfy $\Delta_i \geq \underline{\Delta}_i$ for all $i \neq i^\star$ (after reordering arms if necessary), and the expected regret of the algorithm for sufficiently large T is lower bounded by*

$$R_T = \Omega\left(\frac{1}{\log K} \sum_{i=1}^{K-1} \frac{1}{\underline{\Delta}_i}\right).$$

Read in terms of regret upper bounds, Corollary 6.3 says that any regret upper bound that is monotone in the gap vector Δ and invariant to relabeling of the suboptimal arms must be $\Omega\big((1/\log K)\sum_{i \neq i^\star} 1/\Delta_i\big)$ uniformly over Δ (formal restatement in Appendix D.1). Hence our $\mathcal{O}(C_1 \log K)$ stochastic upper bounds are optimal up to a factor of $\mathcal{O}((\log K)^2)$ within this class. A detailed asymptotic comparison with prior work is also deferred to Appendix.

Remark 6.4 (Scope of the hard environment). In Theorems 6.1 and 6.2, the hard environment μ' may be the base environment μ itself; the theorems do not guarantee that a swapped environment distinct from μ is hard, which would be a stronger statement. This is inherent to the proofs, which compare μ and its swaps with a common reference environment in which the candidate arms are tied, and lower-bound the total regret over this family; such an averaging argument identifies some hard member of the family but not which one. The weaker formulation suffices for near-optimality: all members of the family have the same multiset of suboptimality gaps, so a permutation-invariant regret upper bound takes the same value on all of them, and a single hard member lower-bounds it. Corollary 6.3 and its reading in terms of regret upper bounds above use only one hard member of the family, together with the monotonicity of the upper bound in the gaps.

7 Conclusion

We proposed two BOBW algorithms for the decoupled MAB problem (D-Exp3-Ada and D-TINF-SPM) and derived the first non-trivial instance-dependent lower bounds, showing our algorithms are optimal up to $\mathcal{O}((\log K)^2)$. Closing this gap, removing the unique-optimal-arm assumption from our stochastic upper bounds (e.g., via an SPM variant that treats the set of optimal arms as a single coordinate; see Remark F.4 in Appendix), and extending the framework to richer settings (e.g., contextual or linear bandits) are interesting future directions.

Acknowledgments and Disclosure of Funding

SI is supported by JSPS KAKENHI Grant Number JP25K03184 and by JST PRESTO, Japan, Grant Number JPMJPR2511.

References

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

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002b.

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

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. In *Proceedings of the 25th Conference on Learning Theory*, pages 42.1–42.23, 2012.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011.

Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, 7(39):1079–1105, 2006.

Victor Gabillon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Sébastien Bubeck. Multi-bandit best arm identification. In *Advances in Neural Information Processing Systems*, pages 2222–2230, 2011.

Aurélien Garivier, Tor Lattimore, and Emilie Kaufmann. On explore-then-commit strategies. In *Advances in Neural Information Processing Systems*, pages 784–792, 2016.

Shinji Ito. On optimal robustness to adversarial corruption in online decision problems. In *Advances in Neural Information Processing Systems*, pages 7409–7420, 2021a.

Shinji Ito. Parameter-free multi-armed bandit algorithms with hybrid data-dependent regret bounds. In *Proceedings of the 34th Conference on Learning Theory*, pages 2552–2583, 2021b.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Nearly optimal best-of-both-worlds algorithms for online learning with feedback graphs. In *Advances in Neural Information Processing Systems*, pages 28631–28643, 2022.

Shinji Ito, Taira Tsuchiya, and Junya Honda. Adaptive learning rate for follow-the-regularized-leader: Competitive analysis and best-of-both-worlds. In *Proceedings of the 37th Conference on Learning Theory*, pages 2522–2563, 2024.

Tiancheng Jin, Junyan Liu, and Haipeng Luo. 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*, pages 30918–30978, 2023.

Chaiwon Kim, Jongyeong Lee, and Min-hwan Oh. Follow-the-perturbed-leader for decoupled bandits: Best-of-both-worlds and practicality. In *Proceedings of the 43rd International Conference on Machine Learning*, pages 58328–58358, 2026.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

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

Chloé Rouyer and Yevgeny Seldin. Tsallis-INF for decoupled exploration and exploitation in multi-armed bandits. In *Proceedings of the 33rd Conference on Learning Theory*, pages 3227–3249, 2020.

Sofía S. Villar, Jack Bowden, and James Wason. Multi-armed bandit models for the optimal design of clinical trials: Benefits and challenges. *Statistical Science*, 30(2):199–215, 2015.

Wenshuai Zhao, Jorge Peña Queralta, and Tomi Westerlund. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In *Proceedings of the IEEE Symposium Series on Computational Intelligence*, pages 737–744, 2020.

Julian Zimmert and Yevgeny Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021.

A Proofs for upper bound analysis

A.1 Restated lemmas

We state here two lemmas referenced from the body. Lemma A.1 is quoted from Ito et al. [2024]; Lemma A.2 is an adaptation of Lemma 9 there, and we give its complete proof in Appendix A.3.

Lemma A.1 (Lemma 21 in Ito et al. [2024])**.** *Fix arbitrary $i^* \in [K]$ and $q \in \Delta^{K-1}$. If $\ell_i \geq -\frac{1-\alpha}{4} q_i^{\alpha-1}$ for all $i \in [K] \setminus \{i^*\}$ and $\ell_{i^*} \leq \frac{1-\alpha}{4}(1-q_{i^*})^{\alpha-1}$, then for any $p \in \Delta^{K-1}$,*

$$\langle \ell, q - p \rangle - D(p, q) \leq \frac{4}{1-\alpha} \left(\sum_{i \neq i^*} q_i^{2-\alpha} \ell_i^2 + (1-q_{i^*})^{2-\alpha} \ell_{i^*}^2 \right). \tag{15}$$

Lemma A.2 (SPM Regret Bound; adapted from Lemma 9 in Ito et al. [2024])**.** *Consider FTRL with regularizer $\frac{1}{\eta_t}\psi$, where ψ is a convex Legendre function satisfying $\psi(e_i) \leq 0$ for all $i \in [K]$ (e.g., the Tsallis entropy ψ_α in* (5)*, for which $\psi_\alpha(e_i) = 0$), and learning rate $\eta_t = 1/\beta_t$, where*

$$\beta_1 > 0, \qquad \beta_{t+1} = \beta_t + \frac{z_t}{\beta_t h_t}.$$

Let $\mathcal{H}_{t-1}$ denote the history before round t, and let S_t denote the stability term in Lemma 5.1. Suppose that the following conditions hold:

1. *$z_t, h_t > 0$ are $\mathcal{H}_{t-1}$-measurable and satisfy $z_t \leq \bar{z}$ and $h_t \leq \bar{h}$ for constants $\bar{z}, \bar{h} > 0$.*
2. *For a constant $c_\alpha > 0$ depending only on α, $\mathbb{E}[S_t \mid \mathcal{H}_{t-1}] \leq c_\alpha \eta_t z_t$.*
3. *$-\psi(p_t) \leq h_t$.*
4. *$h_{t+1} \leq \kappa h_t$ for a universal constant $\kappa \geq 1$.*

Then the regret satisfies

$$R_T = \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^{T} \frac{z_t}{\sqrt{\sum_{s=1}^{t} z_s/h_s}} + \frac{\bar{z}}{\beta_1} + \beta_1 h_1 \right] \right),$$

with constant depending only on c_α and κ.

A.2 Proof of Lemma 5.1

Proof. Let $\frac{\psi(p)}{\eta_t} = \psi_t(p)$, $D_t = \frac{1}{\eta_t} D$. We also define $p_{T+1} = p_T, \eta_{T+1} = \eta_T$ for notational convenience. Considering the regret definition (1), we have

$$\begin{aligned} R_T &= \mathbb{E}\left[\sum_{t=1}^{T} \left\langle p_t - e_{i^\star}, \hat{\ell}_t \right\rangle \right] \\ &\leq \mathbb{E}\left[\sum_{t=1}^{T} \left(\left\langle p_t - p_{t+1}, \hat{\ell}_t \right\rangle + \left\langle p_{t+1} - e_{i^\star}, \hat{\ell}_t \right\rangle \right) \right]. \end{aligned} \tag{16}$$

To prove inequality (12), it suffices to evaluate the second term $\sum_{t=1}^{T} \left\langle p_{t+1} - e_{i^\star}, \hat{\ell}_t \right\rangle$. Let $\Phi_t(p) = \sum_{s=1}^{t-1} \langle p, \hat{\ell}_s \rangle + \psi_t(p)$ be the regularized objective, such that $p_t = \operatorname*{argmin}_{p \in \Delta^{K-1}} \Phi_t(p)$. By the definition of p_t, for any $p \in \Delta^{K-1}$, we have

$$\left\langle \nabla \psi_t(p_t) + \sum_{s=1}^{t-1} \hat{\ell}_s, p - p_t \right\rangle \geq 0.$$

Substituting $p = p_{t+1}$, we obtain

$$\left\langle \sum_{s=1}^{t-1} \hat{\ell}_s, p_t - p_{t+1} \right\rangle \leq \langle \nabla \psi_t(p_t), p_{t+1} - p_t \rangle.$$

Using the definition of the Bregman divergence $D_t(p_{t+1}, p_t) = \psi_t(p_{t+1}) - \psi_t(p_t) - \langle \nabla\psi_t(p_t), p_{t+1} - p_t\rangle$, this implies:

$$\left\langle \sum_{s=1}^{t-1} \hat{\ell}_s, p_t - p_{t+1} \right\rangle \leq \psi_t(p_{t+1}) - \psi_t(p_t) - D_t(p_{t+1}, p_t).$$

Now, we evaluate the difference $\Phi_{t+1}(p_{t+1}) - \Phi_t(p_t)$:

$$\begin{aligned}
\Phi_{t+1}(p_{t+1}) - \Phi_t(p_t) &= \left(\sum_{s=1}^{t} \langle p_{t+1}, \hat{\ell}_s\rangle + \psi_{t+1}(p_{t+1})\right) - \left(\sum_{s=1}^{t-1} \langle p_t, \hat{\ell}_s\rangle + \psi_t(p_t)\right) \\
&= \langle p_{t+1}, \hat{\ell}_t\rangle + \left\langle p_{t+1} - p_t, \sum_{s=1}^{t-1} \hat{\ell}_s \right\rangle + \psi_{t+1}(p_{t+1}) - \psi_t(p_t) \\
&\geq \langle p_{t+1}, \hat{\ell}_t\rangle - (\psi_t(p_{t+1}) - \psi_t(p_t) - D_t(p_{t+1}, p_t)) + \psi_{t+1}(p_{t+1}) - \psi_t(p_t) \\
&= \langle p_{t+1}, \hat{\ell}_t\rangle + \psi_{t+1}(p_{t+1}) - \psi_t(p_{t+1}) + D_t(p_{t+1}, p_t).
\end{aligned}$$

Summing this inequality from $t = 1$ to T, the left-hand side telescopes to $\Phi_{T+1}(p_{T+1}) - \Phi_1(p_1)$. Since p_{T+1} minimizes Φ_{T+1}, we have $\Phi_{T+1}(p_{T+1}) \leq \Phi_{T+1}(e_{i^\star}) = \sum_{t=1}^{T} \langle e_{i^\star}, \hat{\ell}_t\rangle + \psi_{T+1}(e_{i^\star})$. Also, $\Phi_1(p_1) = \psi_1(p_1)$. Thus:

$$\sum_{t=1}^{T} \langle e_{i^\star}, \hat{\ell}_t\rangle + \psi_{T+1}(e_{i^\star}) - \psi_1(p_1) \geq \sum_{t=1}^{T} \langle p_{t+1}, \hat{\ell}_t\rangle + \sum_{t=1}^{T} (\psi_{t+1}(p_{t+1}) - \psi_t(p_{t+1})) + \sum_{t=1}^{T} D_t(p_{t+1}, p_t).$$

Rearranging the terms, we successfully bound the second term of the regret:

$$\sum_{t=1}^{T} \langle p_{t+1} - e_{i^\star}, \hat{\ell}_t\rangle \leq \psi_{T+1}(e_{i^\star}) - \psi_1(p_1) - \sum_{t=1}^{T} (\psi_{t+1}(p_{t+1}) - \psi_t(p_{t+1})) - \sum_{t=1}^{T} D_t(p_{t+1}, p_t).$$

□

A.3 Proof of Lemma A.2

Proof. Write $a_t \coloneqq z_t/h_t$ and $A_t \coloneqq \sum_{s=1}^{t} a_s$ (with $A_0 = 0$), so that the update rule reads $\beta_{t+1} = \beta_t + a_t/\beta_t$. Since $z_t, h_t > 0$, the sequence $(\beta_t)_t$ is non-decreasing, i.e., $1/\eta_t - 1/\eta_{t-1} = \beta_t - \beta_{t-1} \geq 0$. Hence we may replace $-\psi(p_t)$ in Lemma 5.1 by its upper bound h_t from Condition 3, and together with $\psi(e_{i^\star}) \leq 0$ we obtain

$$R_T \leq \mathbb{E}\left[\sum_{t=1}^{T} S_t + \beta_1 h_1 + \sum_{t=2}^{T} (\beta_t - \beta_{t-1}) h_t\right]. \tag{17}$$

Stability terms. The learning rate β_t is a function of $z_{1:t-1}$ and $h_{1:t-1}$, so β_t and z_t are $\mathcal{H}_{t-1}$-measurable by Condition 1. By the tower rule and Condition 2, $\mathbb{E}[\sum_{t=1}^{T} S_t] = \sum_{t=1}^{T} \mathbb{E}[\mathbb{E}[S_t \mid \mathcal{H}_{t-1}]] \leq c_\alpha \mathbb{E}[\sum_{t=1}^{T} z_t/\beta_t]$.

Penalty terms. The penalty sum matches the stability sum through the update rule: for $t \geq 2$, by Condition 4,

$$(\beta_t - \beta_{t-1}) h_t = \frac{z_{t-1}}{\beta_{t-1} h_{t-1}} h_t \leq \kappa \frac{z_{t-1}}{\beta_{t-1}}.$$

This is the only step that uses Condition 4. Summing over t and combining with (17), we obtain

$$R_T \leq (c_\alpha + \kappa)\, \mathbb{E}\left[\sum_{t=1}^{T} \frac{z_t}{\beta_t}\right] + \beta_1 \mathbb{E}[h_1]. \tag{18}$$

From β_t to $\sqrt{A_t}$. It remains to show the deterministic inequality

$$\sum_{t=1}^{T} \frac{z_t}{\beta_t} \le \sqrt{\frac{3}{2}} \sum_{t=1}^{T} \frac{z_t}{\sqrt{A_t}} + \frac{2\bar{z}}{\beta_1}. \tag{19}$$

Squaring the update rule gives $\beta_{t+1}^2 = \beta_t^2 + 2a_t + a_t^2/\beta_t^2 \ge \beta_t^2 + 2a_t$, and hence $\beta_t^2 \ge \beta_1^2 + 2A_{t-1} \ge 2A_{t-1}$ for all $t \ge 1$. We call round t *regular* if $a_t \le \beta_t^2$ and *irregular* otherwise. In a regular round, $A_t = A_{t-1} + a_t \le \beta_t^2/2 + \beta_t^2 = \frac{3}{2}\beta_t^2$, and thus $z_t/\beta_t \le \sqrt{3/2}\, z_t/\sqrt{A_t}$. In an irregular round, $\beta_{t+1} = \beta_t + a_t/\beta_t > 2\beta_t$. Let $t_1 < t_2 < \cdots$ be the irregular rounds. Since $(\beta_t)_t$ is non-decreasing, $\beta_{t_{k+1}} \ge \beta_{t_k+1} > 2\beta_{t_k}$, and hence $\beta_{t_k} \ge 2^{k-1}\beta_1$. Therefore, by Condition 1, the irregular rounds contribute at most $\sum_{k\ge 1} \bar{z}/(2^{k-1}\beta_1) \le 2\bar{z}/\beta_1$. This proves (19), and substituting it into (18) yields

$$R_T \le (c_\alpha + \kappa) \left(\sqrt{\frac{3}{2}}\, \mathbb{E}\left[\sum_{t=1}^{T} \frac{z_t}{\sqrt{\sum_{s=1}^{t} z_s/h_s}} \right] + \frac{2\bar{z}}{\beta_1} \right) + \beta_1 \mathbb{E}[h_1],$$

which is the claimed bound with a constant depending only on c_α and κ. □

A.4 Proof of Lemma 5.2(ii)

Proof. We prove this lemma by applying Lemma A.1 with $i^\star = i_t^\star$, $p = p_{t+1}$, $q = p_t$, and $\ell = \eta_t \hat{\ell}_t$. We first verify that the conditions of Lemma A.1 are satisfied.

$$\begin{aligned}
\eta_t \hat{\ell}_{ti} &\ge 0 \ge -\frac{1-\alpha}{4} p_{ti}^{\alpha-1} \quad (i \ne i_t^\star) \\
\eta_t \hat{\ell}_{ti_t^\star} &\le \eta_1 \frac{1}{q_{ti_t^\star}} \\
&= \eta_1 \cdot 2 \\
&\le \frac{1-\alpha}{8} \cdot 2 \le \frac{1-\alpha}{4} (1 - p_{ti_t^\star})^{\alpha-1}
\end{aligned}$$

Consequently, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\langle \hat{\ell}_t, p_t - p_{t+1} \rangle - \frac{1}{\eta_t} D(p_{t+1}, p_t) \,\middle|\, \mathcal{H}_{t-1} \right] \\
&\le \mathbb{E}\left[\frac{4}{1-\alpha} \frac{1}{\eta_t} \left(\sum_{i \ne i_t^\star} p_{ti}^{2-\alpha} (\eta_t \hat{\ell}_{ti})^2 + (1 - p_{ti_t^\star})^{2-\alpha} (\eta_t \hat{\ell}_{ti_t^\star})^2 \right) \,\middle|\, \mathcal{H}_{t-1} \right] \\
&\le \mathbb{E}\left[\frac{4}{1-\alpha} \frac{1}{\eta_t} \left(\sum_{i \ne i_t^\star} p_{ti}^{2-\alpha} (\eta_t \hat{\ell}_{ti})^2 + \left(\sum_{i \ne i_t^\star} p_{ti} \right)^{2-\alpha} (\eta_t \hat{\ell}_{ti_t^\star})^2 \right) \,\middle|\, \mathcal{H}_{t-1} \right]
\end{aligned}$$

By the definition of $\hat{\ell}_t$, we obtain

$$\mathbb{E}\left[(\hat{\ell}_{ti})^2 \,\middle|\, \mathcal{H}_{t-1} \right] = \mathbb{E}\left[\frac{\ell_{ti}^2 \mathbb{1}\{B_t = i\}}{q_{ti}^2} \,\middle|\, \mathcal{H}_{t-1} \right] = \frac{\ell_{ti}^2}{q_{ti}} \le \frac{1}{q_{ti}}$$

for any $i \in [K]$. Since p_t and η_t are $\mathcal{H}_{t-1}$-measurable, we obtain

$$\leq \mathbb{E}\left[\frac{4}{1-\alpha}\frac{1}{\eta_t}\left(\sum_{i\neq i_t^\star} p_{ti}^{2-\alpha}(\eta_t\hat{\ell}_{ti})^2 + \left(\sum_{i\neq i_t^\star} p_{ti}\right)^{2-\alpha}(\eta_t\hat{\ell}_{ti_t^\star})^2\right)\middle|\mathcal{H}_{t-1}\right]$$

$$\leq \frac{4\eta_t}{1-\alpha}\left(\sum_{i\neq i_t^\star}\frac{p_{ti}^{2-\alpha}}{q_{ti}} + \frac{\left(\sum_{i\neq i_t^\star} p_{ti}\right)^{2-\alpha}}{q_{ti_t^\star}}\right) \tag{20}$$

$$= \frac{4\eta_t}{1-\alpha}\left(2\left(\sum_{i\neq i_t^\star} p_{ti}^{1-\frac{\alpha}{2}}\right)^2 + 2\left(\sum_{i\neq i_t^\star} p_{ti}\right)^{2-\alpha}\right) \tag{21}$$

$$\lesssim \frac{\eta_t}{1-\alpha}\left(\sum_{i\neq i_t^\star} p_{ti}^{1-\alpha/2}\right)^2 \tag{22}$$

Inequality (21) follows from the definition of q_t:

$$q_{ti} = \begin{cases} \dfrac{p_{ti}^{1-\alpha/2}}{2\sum_{j\neq i_t^\star} p_{tj}^{1-\alpha/2}} & (i \neq i_t^\star) \\ \dfrac{1}{2} & (i = i_t^\star). \end{cases}$$

Inequality (22) follows from the inequality $(\sum_i a_i)^c \leq \sum_i a_i^c$ (which holds for any $a_i \geq 0$ and $c \in [0,1]$) by setting $c = 1-\alpha/2$. □

Remark A.3. By minimizing (20), we obtain the optimal q_t:

$$q_{ti} = \frac{p_{ti}^{1-\alpha/2}}{\sum_{j\neq i_t^\star} p_{tj}^{1-\alpha/2} + \left(\sum_{j\neq i_t^\star} p_{tj}^{2-\alpha}\right)^{1/2}},$$

$$q_{ti_t^\star} = \frac{\left(\sum_{j\neq i_t^\star} p_{tj}^{2-\alpha}\right)^{1/2}}{\sum_{j\neq i_t^\star} p_{tj}^{1-\alpha/2} + \left(\sum_{j\neq i_t^\star} p_{tj}^{2-\alpha}\right)^{1/2}}.$$

This exact form makes the subsequent analysis intractable. In the regime where p_t concentrates on a single arm (which naturally occurs after sufficient exploration), $q_{ti_t^\star}$ in the above expression approaches $1/2$, justifying our choice of $q_{ti_t^\star} = 1/2$ as an approximation.

A.5 Proofs for D-Exp3-Ada

Lemma A.4. *For D-Exp3-Ada, the stability term is bounded as follows:*

$$\mathbb{E}\left[\langle\hat{\ell}_t, p_t - p_{t+1}\rangle - \frac{1}{\eta_t}D\left(p_{t+1}, p_t\right)\middle|\mathcal{H}_{t-1}\right] \lesssim \eta_t\left(\sum_{i\neq i_t^\star}\sqrt{p_{ti}}\right)^2.$$

Proof. The proof adapts techniques from Lemma 1 in the supplementary material of Ito [2021a]. For any $x \in \mathbb{R}$, we have

$$\langle\hat{\ell}_t, p_t - p_{t+1}\rangle - \frac{1}{\eta_t}D\left(p_{t+1}, p_t\right) = \langle\hat{\ell}_t - x\mathbf{1}, p_t - p_{t+1}\rangle - \frac{1}{\eta_t}D\left(p_{t+1}, p_t\right) \tag{23}$$

$$\leq \max_{p\in\mathbb{R}^K_{>0}}\left\{\langle\hat{\ell}_t - x\mathbf{1}, p_t - p\rangle - \frac{1}{\eta_t}D\left(p, p_t\right)\right\} \tag{24}$$

$$\leq \frac{\eta_t}{2}\|\hat{\ell}_t - x\mathbf{1}\|^2_{\nabla^2\psi(z)^{-1}},$$

where the last inequality holds for $z = \theta p_t + (1-\theta)p$ with some $\theta \in (0,1)$, as shown by Lattimore and Szepesvári [2020] in Theorem 26.13. Equality (23) holds by $\langle x\mathbf{1}, p_t \rangle = x$. Setting the gradient of the objective function in (24) to zero, we obtain $p_i = p_{ti} \exp\left(-\eta_t \left(\hat{\ell}_{ti} - x\right)\right)$. Here, we define $x = \hat{\ell}_{ti_t^\star}$. By the definition of $\hat{\ell}_t$, we have $\hat{\ell}_{ti} - x \geq -\hat{\ell}_{ti_t^\star} \geq -\frac{\ell_{ti_t^\star}}{q_{ti_t^\star}} \geq -2$. Thus, we have

$$\begin{aligned} p_i = p_{ti} \exp\left(-\eta_t \left(\hat{\ell}_{ti} - x\right)\right) &\leq p_{ti} \exp(\eta_t \cdot 2) \\ &\leq 3 p_{ti}, \end{aligned}$$

where the last inequality holds since $2\eta_1 = 2\sqrt{\frac{\log K}{2K}} \leq 1$. Since $\nabla^2 \psi(z) = \mathrm{diag}(z_1^{-1}, \ldots, z_K^{-1})$, we have

$$\begin{aligned} \|\hat{\ell}_t - x\mathbf{1}\|^2_{\nabla^2 \psi(z)^{-1}} &= \sum_{i=1}^K \left(\hat{\ell}_{ti} - x\right)^2 z_i \\ &\leq 3 \sum_{i=1}^K \left(\hat{\ell}_{ti} - x\right)^2 p_{ti}. \end{aligned}$$

We then obtain

$$\begin{aligned} \mathbb{E}\left[\sum_{i=1}^K \left(\hat{\ell}_{ti} - \hat{\ell}_{ti_t^\star}\right)^2 p_{ti} \middle| \mathcal{H}_{t-1}\right] &= \mathbb{E}\left[\sum_{i \neq i_t^\star} \left(\mathbb{1}[B_t = i] \frac{\ell_{ti}^2}{q_{ti}^2} + \mathbb{1}[B_t = i_t^\star] \frac{\ell_{ti_t^\star}^2}{q_{ti_t^\star}^2}\right) p_{ti} \middle| \mathcal{H}_{t-1}\right] \\ &\leq \sum_{i \neq i_t^\star} \frac{\ell_{ti}^2}{q_{ti}} p_{ti} + \frac{\ell_{ti_t^\star}^2}{q_{ti_t^\star}} \sum_{i \neq i_t^\star} p_{ti} \qquad (25) \\ &\leq 2 \left(\sum_{i \neq i_t^\star} \sqrt{p_{ti}}\right)^2 + 2 \sum_{i \neq i_t^\star} p_{ti} \qquad (26) \\ &\leq 4 \left(\sum_{i \neq i_t^\star} \sqrt{p_{ti}}\right)^2. \end{aligned}$$

Inequality (25) holds since $\mathbb{E}\left[\, \mathbb{1}[B_t = i] \middle| \mathcal{H}_{t-1}\right] = q_{ti}$. Inequality (26) follows from the definition of q_t. □

Lemma A.5. *D-Exp3-Ada achieves the following regret upper bound.*

$$R_T = \mathcal{O}\left(\mathbb{E}\left[\sqrt{\bar{h} \sum_{t=1}^T z_t} + \frac{z_{\max}}{\beta_1} + \beta_1 \bar{h}\right]\right). \qquad (27)$$

Proof. The proof follows a similar approach to Lemma 9 in Ito et al. [2024], but it is an AdaGrad-style argument with the constant penalty bound $\bar{h}$ rather than an application of the SPM bound (Lemma A.2). By Lemma 5.1 applied to ϕ (for which $\phi(e_{i^\star}) = 0$ and $-\phi(p_t) \leq \bar{h}$), Lemma A.4, and the tower rule (η_t and z_t are $\mathcal{H}_{t-1}$-measurable), we have

$$R_T \lesssim \mathbb{E}\left[\sum_{t=1}^T \frac{z_t}{\beta_t} + \beta_1 \bar{h} + \sum_{t=2}^T (\beta_t - \beta_{t-1}) \bar{h}\right].$$

We bound the expression inside the expectation path-wise. The following relationship holds by the definition of β_t.

$$\begin{aligned}
(\beta_t - \beta_{t-1})\,\bar{h} &= \sqrt{2\bar{h}}\left(\sqrt{K + \sum_{s=1}^{t-1} z_s} - \sqrt{K + \sum_{s=1}^{t-2} z_s}\right) \\
&= \sqrt{2\bar{h}}\frac{z_{t-1}}{\sqrt{K + \sum_{s=1}^{t-1} z_s} + \sqrt{K + \sum_{s=1}^{t-2} z_s}} \\
&\leq \sqrt{2\bar{h}}\frac{z_{t-1}}{2\sqrt{K + \sum_{s=1}^{t-2} z_s}} \\
&= \frac{z_{t-1}}{\beta_{t-1}}.
\end{aligned}$$

Consequently, we obtain

$$\begin{aligned}
R_T &\lesssim \frac{z_1}{\beta_1} + \beta_1\bar{h} + \sum_{t=2}^{T}\left(\frac{z_t}{\beta_t} + \frac{z_{t-1}}{\beta_{t-1}}\right) \\
&\leq \frac{z_1}{\beta_1} + \beta_1\bar{h} + 2\sum_{t=1}^{T}\frac{z_t}{\beta_t} \\
&\leq 2\frac{z_1}{\beta_1} + \beta_1\bar{h} + 2\sqrt{2\bar{h}\sum_{t=1}^{T} z_t},
\end{aligned}$$

where the last inequality holds as follows. Let $S_t = \sum_{s=1}^{t} z_s$. We obtain

$$\begin{aligned}
\sum_{t=1}^{T}\frac{z_t}{\beta_t} &\leq \sqrt{\frac{\bar{h}}{2}}\sum_{t=2}^{T}\frac{z_t}{\sqrt{K + \sum_{s=1}^{t-1} z_s}} + \frac{z_1}{\beta_1} \\
&\leq \sqrt{\frac{\bar{h}}{2}}\sum_{t=2}^{T}\frac{z_t}{\sqrt{K + S_t - z_{\max}}} + \frac{z_1}{\beta_1} \\
&\leq \sqrt{\frac{\bar{h}}{2}}\cdot 2\sum_{t=2}^{T}\frac{z_t}{\sqrt{K + S_t - z_{\max}} + \sqrt{K + S_{t-1} - z_{\max}}} + \frac{z_1}{\beta_1} \\
&= \sqrt{2\bar{h}}\sum_{t=2}^{T}\left(\sqrt{K + S_t - z_{\max}} - \sqrt{K + S_{t-1} - z_{\max}}\right) + \frac{z_1}{\beta_1} \\
&= \sqrt{2\bar{h}}\left(\sqrt{K + S_T - z_{\max}} - \sqrt{K + S_1 - z_{\max}}\right) + \frac{z_1}{\beta_1} \\
&\leq \sqrt{2\bar{h}S_T} + \frac{z_1}{\beta_1}.
\end{aligned}$$

□

Lemma A.6. *For D-Exp3-Ada, the following bounds hold:*

1. *In adversarial regimes,*

$$\mathbb{E}\left[\sqrt{\bar{h}\sum_{t=1}^{T} z_t}\right] = \mathcal{O}\left(\sqrt{TK\log K}\right). \tag{28}$$

2. *In stochastic regimes,*

$$\mathbb{E}\left[\sqrt{\bar{h}\sum_{t=1}^{T} z_t}\right] = \mathcal{O}\left(\sqrt{\sum_{i\neq i^\star}\frac{\log K}{\Delta_i}R_T}\right). \tag{29}$$

Proof. For the adversarial regime, we use the uniform bound $z_t \lesssim K$ to obtain

$$\mathbb{E}\left[\sqrt{\bar{h}\sum_{t=1}^{T} z_t}\right] \lesssim \sqrt{\bar{h}TK} = \mathcal{O}\left(\sqrt{TK\log K}\right).$$

For the stochastic regime, we analyze z_t more carefully. Define $r_t = \sum_{i\neq i^\star} \Delta_i p_{ti}$. By the definition of z_t and Cauchy-Schwarz inequality, we have

$$\begin{aligned}
z_t = \left(\sum_{i\neq i_t^\star} \sqrt{p_{ti}}\right)^2 &\leq \left(\sum_{i\neq i^\star} \sqrt{p_{ti}}\right)^2 \\
&= \left(\sum_{i\neq i^\star} \sqrt{\frac{\Delta_i}{\Delta_i}}\sqrt{p_{ti}}\right)^2 \\
&\leq \left(\sum_{i\neq i^\star} \frac{1}{\Delta_i}\right)\left(\sum_{i\neq i^\star} \Delta_i p_{ti}\right) \\
&= \left(\sum_{i\neq i^\star} \frac{1}{\Delta_i}\right) r_t.
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathbb{E}\left[\sqrt{\bar{h}\sum_{t=1}^{T} z_t}\right] &\leq \mathbb{E}\left[\sqrt{\bar{h}\left(\sum_{i\neq i^\star} \frac{1}{\Delta_i}\right)\sum_{t=1}^{T} r_t}\right] \\
&\leq \sqrt{\bar{h}\left(\sum_{i\neq i^\star} \frac{1}{\Delta_i}\right)\mathbb{E}\left[\sum_{t=1}^{T} r_t\right]} \\
&= \sqrt{\sum_{i\neq i^\star} \frac{\log K}{\Delta_i} R_T},
\end{aligned}$$

where the second inequality is Jensen's inequality for the concave function $x \mapsto \sqrt{x}$, and the last equality uses $\mathbb{E}[\sum_{t=1}^{T} r_t] = R_T$, which follows from (2). □

Lemma A.7. *For D-Exp3-Ada, the following bound holds:*

$$\mathbb{E}\left[\frac{\bar{z}}{\beta_1} + \beta_1 \bar{h}\right] = \mathcal{O}\left(\sqrt{K\log K}\right),$$

with $\bar{z} = K - 1$.

Proof. By the definition of β_1 in (9), we have

$$\beta_1 = \sqrt{\frac{2}{\bar{h}}}\sqrt{K} = \sqrt{\frac{2K}{\log K}}.$$

Thus,

$$\begin{aligned}
\frac{z_{\max}}{\beta_1} + \beta_1 \bar{h} &= \frac{K-1}{\sqrt{2K/\log K}} + \sqrt{\frac{2K}{\log K}}\log K \\
&= (K-1)\sqrt{\frac{\log K}{2K}} + \sqrt{2K\log K} \\
&= \mathcal{O}(\sqrt{K\log K}).
\end{aligned}$$

□

We are now ready to prove Theorem 5.3.

Proof of Theorem 5.3. We apply the AdaGrad-style bound of Lemma A.5 (not the SPM bound of Lemma A.2). Its ingredients are the stability bound $\mathbb{E}[S_t \mid \mathcal{H}_{t-1}] = \mathcal{O}(\eta_t z_t)$ of Lemma A.4, the constant penalty bound $-\phi(p_t) \le \bar{h} = \log K$, and the uniform bound $z_t \le z_{\max} = K - 1$ used in Lemma A.7. Combining the lemmas established above, we obtain the desired regret upper bounds as follows:

1. **Adversarial regimes:** By Lemmas A.5 and A.6,

$$\begin{aligned} R_T &= \mathcal{O}\left(\mathbb{E}\left[\sqrt{\bar{h}\sum_{t=1}^{T} z_t}\right] + \sqrt{K\log K}\right) \\ &= \mathcal{O}\left(\sqrt{KT\log K} + \sqrt{K\log K}\right) \\ &= \mathcal{O}\left(\sqrt{KT\log K}\right). \end{aligned}$$

2. **Stochastic regimes:** By Lemmas A.5 and A.6,

$$\begin{aligned} R_T &= \mathcal{O}\left(\mathbb{E}\left[\sqrt{\bar{h}\sum_{t=1}^{T} z_t}\right] + \sqrt{K\log K}\right) \\ &= \mathcal{O}\left(\sqrt{\sum_{i\neq i^\star} \frac{\log K}{\Delta_i} R_T} + \sqrt{K\log K}\right) \\ &= \mathcal{O}\left(\sum_{i\neq i^\star} \frac{\log K}{\Delta_i} + \sqrt{K\log K}\right) \\ &= \mathcal{O}\left(\sum_{i\neq i^\star} \frac{\log K}{\Delta_i}\right), \end{aligned}$$

   where the third equality follows from the fact that $x \le a\sqrt{x} + b$ implies $x \le a^2 + 2b$ for $a, b, x \ge 0$, and the last equality follows from the fact that $\sum_{i\neq i^\star} \frac{1}{\Delta_i} \ge \frac{K-1}{\max_i \Delta_i} \ge \frac{K-1}{1}$, which implies that for sufficiently large K, the first term dominates.

□

A.6 Proofs for D-TINF-SPM

Lemma A.8. *For D-TINF-SPM, the following inequality holds for any* $1 \le T_1 \le T_2 \le T$*:*

$$\begin{aligned} \sum_{t=1}^{T} \frac{z_t}{\sqrt{\sum_{s=1}^{t} z_s/h_s}} \lesssim \sqrt{K\sum_{t=1}^{T_1} \frac{z_t}{h_t}} + \sqrt{\log\left(\frac{\sum_{t=1}^{T_2} z_t/h_t}{\sum_{t=1}^{T_1} z_t/h_t}\right)} \sqrt{\sum_{t=T_1+1}^{T_2} z_t h_t} \\ + \left(\sum_{t=1}^{T_2} \frac{z_t}{h_t}\right)^{-1/2} \sum_{t=T_2+1}^{T} z_t. \end{aligned}$$

Proof. Write $a_t \coloneqq z_t/h_t$ and $A_t \coloneqq \sum_{s=1}^{t} a_s$, and split the sum into the three ranges $t \le T_1$, $T_1 < t \le T_2$, and $t > T_2$.

First range. Since $z_t = h_t a_t \le h_{\max} a_t$ with $h_{\max} \coloneqq \max_t h_t = \max_t \frac{1}{\alpha}\sum_{i\neq i_t^\star} p_{ti}^{\alpha} \le \frac{K^{1-\alpha}}{\alpha} = 2\sqrt{K}$ (with $\alpha = 1/2$), and $a_t/\sqrt{A_t} \le 2(\sqrt{A_t} - \sqrt{A_{t-1}})$ (with $A_0 = 0$),

$$\sum_{t=1}^{T_1} \frac{z_t}{\sqrt{A_t}} \le h_{\max} \sum_{t=1}^{T_1} \frac{a_t}{\sqrt{A_t}} \le 2h_{\max}\sqrt{A_{T_1}} \le 4\sqrt{K \sum_{t=1}^{T_1} \frac{z_t}{h_t}}.$$

Second range. Writing $z_t = \sqrt{a_t}\sqrt{z_t h_t}$ and applying the Cauchy–Schwarz inequality and Lemma A.12,

$$\sum_{t=T_1+1}^{T_2} \frac{z_t}{\sqrt{A_t}} \le \sqrt{\sum_{t=T_1+1}^{T_2} \frac{a_t}{A_t}} \sqrt{\sum_{t=T_1+1}^{T_2} z_t h_t} \le \sqrt{\log\left(\frac{\sum_{t=1}^{T_2} z_t/h_t}{\sum_{t=1}^{T_1} z_t/h_t}\right)} \sqrt{\sum_{t=T_1+1}^{T_2} z_t h_t}. \tag{30}$$

(If $T_1 = T_2$, this range is empty.)

Third range. Since $A_t \ge A_{T_2}$ for $t > T_2$, we have $\sum_{t=T_2+1}^{T} z_t/\sqrt{A_t} \le A_{T_2}^{-1/2} \sum_{t=T_2+1}^{T} z_t$.

Summing the three bounds proves the lemma. □

Lemma A.9 (Bound on z_t/h_t)**.** *For D-TINF-SPM (with $\alpha = 1/2$), the per-step ratio satisfies $z_t/h_t \le 1$ for all $t \ge 1$.*

Proof. Recall $z_t = \frac{1}{1-\alpha}\big(\sum_{i\neq i_t^\star} p_{ti}^{1-\alpha/2}\big)^2$ and $h_t = \frac{1}{\alpha}\sum_{i\neq i_t^\star} p_{ti}^{\alpha}$ from (11). With $\alpha = 1/2$ these become $z_t = 2(\sum_{i\neq i_t^\star} p_{ti}^{3/4})^2$ and $h_t = 2\sum_{i\neq i_t^\star} p_{ti}^{1/2}$. Writing $p_{ti}^{3/4} = p_{ti}^{1/4}\cdot p_{ti}^{1/2}$ and applying the Cauchy-Schwarz inequality,

$$\Big(\sum_{i\neq i_t^\star} p_{ti}^{3/4}\Big)^2 \;=\; \Big(\sum_{i\neq i_t^\star} p_{ti}^{1/4}\cdot p_{ti}^{1/2}\Big)^2 \;\le\; \Big(\sum_{i\neq i_t^\star} p_{ti}^{1/2}\Big)\Big(\sum_{i\neq i_t^\star} p_{ti}\Big).$$

Therefore $z_t/h_t \le \sum_{i\neq i_t^\star} p_{ti} \le 1$. □

Lemma A.10 (Bound on z_t)**.** *For D-TINF-SPM, the following inequality holds for any $t \ge 1$:*

$$z_t \lesssim \sqrt{C_2} r_t,$$

where $C_2 = \sum_{i\neq i^\star} \frac{1}{\Delta_i^2}$ and r_t is defined in the proof of Lemma A.6.

Proof.

$$\begin{aligned} z_t &= \frac{1}{1-\alpha}\left(\sum_{i\neq i_t^\star} p_{ti}^{1-\alpha/2}\right)^2 \\ &\le 2C_2^{1/2} r_t, \end{aligned}$$

where the last inequality follows from Inequality (32) with $\alpha = \frac{1}{2}$. □

Let $\gamma = 1 - \alpha/2$, $C_p = \sum_{i \neq i^\star} \frac{1}{\Delta_i^p}$, $r_t = \sum_{i \neq i^\star} \Delta_i p_{ti}$, and $x \in [0, 1)$. By Hölder's inequality, we have

$$\begin{aligned}
\sum_{i \neq i_t^\star} p_{ti}^\gamma &\leq \sum_{i \neq i^\star} p_{ti}^\gamma \\
&= \sum_{i \neq i^\star} (p_{ti}^{\frac{\gamma - x}{1-x}})^{1-x} p_{ti}^x \\
&\leq (\sum_{i \neq i^\star} p_{ti}^{\frac{\gamma - x}{1-x}})^{1-x} \\
&= \left(\sum_{i \neq i^\star} \left(\frac{1}{\Delta_i^{\frac{\gamma - x}{1-\gamma}}} \right)^{\frac{1-\gamma}{1-x}} (\Delta_i p_{ti})^{\frac{\gamma - x}{1-x}} \right)^{1-x} \\
&\leq \left(C_{\frac{\gamma - x}{1-\gamma}} \right)^{1-\gamma} r_t^{\gamma - x}. \qquad (31)
\end{aligned}$$

Thus, choosing $x = \gamma - \frac{1}{2} \in [0, 1)$, we obtain

$$\begin{aligned}
\left(\sum_{i \neq i_t^\star} p_{ti}^\gamma \right)^2 &\leq \left(C_{\frac{\gamma - x}{1-\gamma}} \right)^{2(1-\gamma)} r_t^{2(\gamma - x)} \\
&= C_{1/\alpha}^\alpha r_t. \qquad (32)
\end{aligned}$$

Lemma A.11 (Bound on $z_t h_t$)**.** *For D-TINF-SPM, the following inequality holds for any $t \geq 1$:*

$$z_t h_t \lesssim C_1 r_t. \qquad (33)$$

Proof.

$$z_t h_t = \frac{1}{\alpha(1-\alpha)} \left(\sum_{i \neq i_t^\star} p_{ti}^{1-\alpha/2} \right)^2 \sum_{i \neq i_t^\star} p_{ti}^\alpha.$$

By Inequality (31), we have

$$\left(\sum_{i \neq i_t^\star} p_{ti}^{1-\alpha/2} \right)^2 \leq \left(C_{\frac{\gamma - x}{1-\gamma}} \right)^{2(1-\gamma)} r_t^{2(\gamma - x)},$$

where $\gamma = 1 - \alpha/2$ and $x \in [0, 1)$. By an argument analogous to Inequality (31), for $y \in [0, 1)$, we have

$$\begin{aligned}
\sum_{i \neq i_t^\star} p_{ti}^\alpha &\leq \sum_{i \neq i^\star} (p_{ti}^{\frac{\alpha - y}{1-y}})^{1-y} p_{ti}^y \\
&\leq \left(\sum_{i \neq i^\star} p_{ti}^{\frac{\alpha - y}{1-y}} \right)^{1-y} \\
&= \left(\sum_{i \neq i^\star} \left(\frac{1}{\Delta_i^{\frac{\alpha - y}{1-\alpha}}} \right)^{\frac{1-\alpha}{1-y}} (\Delta_i p_{ti})^{\frac{\alpha - y}{1-y}} \right)^{1-y} \\
&\leq \left(C_{\frac{\alpha - y}{1-\alpha}} \right)^{1-\alpha} r_t^{\alpha - y}.
\end{aligned}$$

Thus, we obtain

$$\begin{aligned}
\frac{1}{\alpha(1-\alpha)} \left(\sum_{i \neq i_t^\star} p_{ti}^{1-\alpha/2} \right)^2 \sum_{i \neq i_t^\star} p_{ti}^\alpha &\leq \frac{1}{\alpha(1-\alpha)} \left(C_{\frac{\gamma - x}{1-\gamma}} \right)^{2(1-\gamma)} \left(C_{\frac{\alpha - y}{1-\alpha}} \right)^{1-\alpha} r_t^{2(\gamma - x) + \alpha - y} \\
&= 4 C_1 r_t,
\end{aligned}$$

by substituting $\alpha = 1/2$, $x = 1 - \alpha = 1/2$ and $y = 2\alpha - 1 = 0$. These values are chosen to satisfy the following two equations:

$$\begin{cases} 2(\gamma - x) + \alpha - y = 1, \\ \dfrac{\gamma - x}{1 - \gamma} = \dfrac{\alpha - y}{1 - \alpha}. \end{cases}$$

□

Lemma A.12. *For $\{a_t\}_{t=1}^T$ with $a_t \geq 0$ and any $\tau \in \{1, \ldots, T-1\}$, the following inequality holds:*

$$\sum_{t=\tau+1}^{T} \frac{a_t}{\sum_{s=1}^{t} a_s} \leq \log\left(\sum_{t=1}^{T} a_t\right) - \log\left(\sum_{t=1}^{\tau} a_t\right).$$

Proof. Let $S_t = \sum_{s=1}^{t} a_s$. Then,

$$\begin{aligned} \sum_{t=\tau+1}^{T} \frac{a_t}{\sum_{s=1}^{t} a_s} &= \sum_{t=\tau+1}^{T} \frac{a_t}{S_t} \\ &= \sum_{t=\tau+1}^{T} \frac{S_t - S_{t-1}}{S_t} \\ &= \sum_{t=\tau+1}^{T} \int_{S_{t-1}}^{S_t} \frac{1}{S_t}\, dS \\ &\leq \sum_{t=\tau+1}^{T} \int_{S_{t-1}}^{S_t} \frac{1}{S}\, dS \\ &= \int_{S_\tau}^{S_T} \frac{1}{S}\, dS \\ &= \log(S_T) - \log(S_\tau). \end{aligned}$$

□

Lemma A.13. *For D-TINF-SPM, the following bound holds:*

$$\mathbb{E}\left[\frac{\bar{z}}{\beta_1} + \beta_1 h_1\right] = \mathcal{O}\left(K^{1/2}\right),$$

with $\bar{z} = \frac{1}{1-\alpha} K^\alpha$.

Proof.

$$\begin{aligned} \mathbb{E}\left[\frac{z_{\max}}{\beta_1} + \beta_1 h_1\right] &= \frac{K^\alpha}{\beta_1(1-\alpha)} + \beta_1 \mathbb{E}\left[\frac{1}{\alpha} \sum_{i \neq i_t^\star} p_{1,i}^\alpha\right] \\ &\leq \frac{K^\alpha}{\beta_1(1-\alpha)} + \beta_1 \frac{K^{1-\alpha}}{\alpha} \\ &= \mathcal{O}\left(K^{1/2}\right). \end{aligned}$$

□

Lemma A.14 (Adapted from Lemma 24 in Ito et al. [2024])**.** *For D-TINF-SPM, the following inequality holds for any $i \in [K]$ and $t \geq 1$:*

$$p'_{ti} \leq 2p_{ti},$$

where $p'_t = \operatorname{argmin}_{p \in \Delta^{K-1}} \left\{ \left\langle \sum_{s=1}^{t-1} \hat{\ell}_s, p \right\rangle + \beta_{t+1} \psi(p) \right\}$.

Proof. By applying Lemma 24 in Ito et al. [2024] with $q = p_t, r = p'_t, \ell = 0, \beta = \beta_t, \beta' = \beta_{t+1}$, we obtain the desired inequality. To apply this lemma, we must verify that the conditions of Lemma 24 in Ito et al. [2024] are satisfied. Here, $\omega = \sqrt{2}$ and $q_* = \min\{1 - \max_i q_i, \max_i q_i\}$ are defined in Lemma 24 in Ito et al. [2024].

1. $\|\ell\|_\infty \leq \frac{1-\omega^{\alpha-1}}{2}\beta q_*^{\alpha-1}$
2. $0 \leq \beta' - \beta \leq \left(1 - \omega^{\alpha-1}\right)\beta$

Condition 1 is satisfied since $\|\ell\|_\infty = 0$. We verify that Condition 2 holds as follows:

$$\begin{aligned}
\beta' - \beta &= \frac{z_t}{\beta_t h_t} \\
&\leq \frac{1}{\beta_1} \frac{\left(\sum_{i \neq i_t^\star} p_{ti}^{3/4}\right)^2}{\sum_{i \neq i_t^\star} p_{ti}^{1/2}} \\
&\leq \frac{1}{\beta_1} \frac{\sum_{i \neq i_t^\star} p_{ti}^{1/2} \sum_{i \neq i_t^\star} p_{ti}}{\sum_{i \neq i_t^\star} p_{ti}^{1/2}} \qquad (34) \\
&\leq \frac{1}{\beta_1} \\
&\leq (1 - \omega^{\alpha-1})\beta_1 \qquad (35) \\
&\leq (1 - \omega^{\alpha-1})\beta,
\end{aligned}$$

where we use $\alpha = 1/2$. Inequality (34) follows from the Cauchy-Schwarz inequality. The validity of Inequality (35) for the case $\alpha = 1/2$ can be verified by substituting $\beta_1 = 16$ and $\omega = \sqrt{2}$. □

Lemma A.15 (Adapted from Lemma 25 in Ito et al. [2024])**.** *For D-TINF-SPM, the following inequality holds for any $i \in [K]$ and $t \geq 1$:*

$$p_{t+1,i} \leq 2p'_{ti},$$

where $p'_t = \operatorname*{argmin}_{p \in \Delta^{K-1}} \left\{ \left\langle \sum_{s=1}^{t-1} \hat{\ell}_s, p \right\rangle + \beta_{t+1}\psi(p) \right\}$.

Proof. The proof follows the same approach as Lemma 25 in Ito et al. [2024], adapted to our setting. We prove $p_{t+1,i} \leq \omega p'_{ti}$ for some $\omega \geq 2$. By the definition of p_{t+1} and p'_t, we have the following equations.

$$\begin{cases} \sum_{s=1}^{t-1} \hat{\ell}_{si} - \beta_{t+1}(p'_{ti})^{\alpha-1} - \lambda' = 0 \\ \sum_{s=1}^{t} \hat{\ell}_{si} - \beta_{t+1} p_{t+1,i}^{\alpha-1} - \lambda'' = 0 \end{cases},$$

where $\lambda', \lambda'' \in \mathbb{R}$ are some constants. By subtracting these equations, we obtain

$$\beta_{t+1} p_{t+1,i}^{\alpha-1} = \beta_{t+1}(p'_{ti})^{\alpha-1} + \hat{\ell}_{ti} - \lambda. \qquad (36)$$

Since $p_{t+1}, p'_t \in \Delta^{K-1}$, we have $0 \leq \lambda \leq \|\hat{\ell}_t\|_\infty$. Here, we define $g(x) \coloneqq \beta_{t+1} x^{\alpha-1}$. When $i \neq B_t$, $g(p_{t+1,i}) = g(p'_{ti}) - \lambda$ holds, and thus $p_{t+1,i} \geq p'_{ti}$ since $g(x)$ is monotone non-increasing. This implies that $p_{t+1,B_t} \leq p'_{t,B_t} \leq \omega p'_{t,B_t}$, so it suffices to show $p_{t+1,i} \leq \omega p'_{ti}$ for $i \neq B_t$. We show this by a case analysis on the value of q_{t,B_t}.

- $q_{t,B_t} \geq \varepsilon$ (The value of ε ($0 < \varepsilon \leq 1/2$) is determined later):
  By the definition of $\hat{\ell}_t$, we have

$$\lambda \leq \|\hat{\ell}_t\|_\infty \leq \frac{1}{q_{t,B_t}} \leq \frac{1}{\varepsilon}.$$

From equation (36), we obtain

$$p_{t+1,i}^{\alpha-1} = {p'_{ti}}^{\alpha-1} - \lambda/\beta_{t+1} \geq \omega^{\alpha-1}{p'_{ti}}^{\alpha-1},$$

where the last inequality holds when

$$\beta_{t+1} \geq \frac{1}{(1-\omega^{\alpha-1})\varepsilon}. \tag{37}$$

Thus, $p_{t+1,i} \leq \omega p'_{ti}$.

- $q_{t,B_t} < \varepsilon$:
  Let $i'_\star \in \operatorname*{argmax}_i p'_{ti}$. First, we consider the case where $i = i'_\star$. Then, $i'_\star \neq B_t$ holds since

$$q_{ti'_\star} = q_{ti_t^\star} = \frac{1}{2} \geq \varepsilon > q_{t,B_t}.$$

  The first equality holds from the definition of p'_t and p_t.

  Moreover, for $i \neq B_t$, we have following inequality.

$$\begin{aligned} p_{t+1,i} - p'_{ti} &\leq \sum_{j \neq B_t} (p_{t+1,j} - p'_{tj}) \\ &= p'_{t,B_t} - p_{t+1,B_t} \\ &\leq p'_{t,B_t}, \end{aligned} \tag{38}$$

  where the first inequality holds since $p_{t+1,j} - p'_{tj} \geq 0$ for $j \neq B_t$. Based on this, we evaluate the value of $\gamma := \frac{p_{t+1,i'_\star}}{p'_{ti'_\star}}$.

$$\begin{aligned} \gamma &= 1 + \frac{p_{t+1,i'_\star} - p'_{ti'_\star}}{p'_{ti'_\star}} \\ &\leq 1 + \frac{p'_{t,B_t}}{p'_{ti'_\star}} \qquad (39) \\ &\leq 2 \leq \omega, \qquad (40) \end{aligned}$$

  where the second inequality holds by the definition of $i'_\star$. Inequality (39) holds by Inequality (38).

  Next, we consider the case where $i \neq i'_\star$. From equation (36), we have

$$\begin{aligned} g(p'_{ti}) - g(p_{t+1,i}) &= \lambda \\ &= g(p'_{ti'_\star}) - g(p_{t+1,i'_\star}) \\ &= g(p'_{ti'_\star}) - g(\gamma p'_{ti'_\star}) \\ &\leq g(p'_{ti}) - g(\gamma p'_{ti}), \end{aligned}$$

  where the last inequality holds since $g(x) - g(\gamma x) = \beta_{t+1}(1-\gamma^{\alpha-1})x^{\alpha-1}$ is monotone non-increasing (as $\alpha - 1 < 0$ and $\gamma \geq 1$). Thus, we have $g(p_{t+1,i}) \geq g(\gamma p'_{ti})$, which implies

$$p_{t+1,i} \leq \gamma p'_{ti} \leq \omega p'_{ti}.$$

Finally, we verify that β_t as defined by D-TINF-SPM satisfies Condition (37) with $\omega = 2$ and $\varepsilon = \frac{1}{2}$.

$$\begin{aligned} \frac{1}{(1-\omega^{\alpha-1})\varepsilon} &= \frac{1}{(1-1/\sqrt{2})1/2} \\ &= 4 + 2\sqrt{2} \\ &\leq \beta_1. \end{aligned}$$

□

Proof of Theorem 5.4. First, we verify the four conditions of Lemma A.2 for D-TINF-SPM. Condition 1 holds since z_t and h_t are deterministic functions of the $\mathcal{H}_{t-1}$-measurable p_t, with $\bar{z} = \bar{h} = 2\sqrt{K}$ by Hölder's inequality. Condition 2 is Lemma 5.2 (ii), whose hypothesis $\beta_1 \geq 8/(1-\alpha) = 16$ is satisfied by the algorithm's choice $\beta_1 = 16$. Condition 3 is the penalty bound for D-TINF-SPM in Section 5.1, and $\psi_\alpha(e_i) = 0$ for all i. Condition 4 holds with $\kappa = 2$, as follows: By Lemmas A.14 and A.15, we have $p_{t+1,i} \leq 4p_{ti}$ for every i. Since $i^\star_{t+1}$ maximizes $p_{t+1,i}$ over i,

$$\begin{aligned} h_{t+1} &= \frac{1}{\alpha}\left(\sum_i p_{t+1,i}^\alpha - p_{t+1,i^\star_{t+1}}^\alpha\right) \\ &\leq \frac{1}{\alpha}\left(\sum_i p_{t+1,i}^\alpha - p_{t+1,i^\star_t}^\alpha\right) \\ &\leq \frac{4^\alpha}{\alpha}\sum_{i\neq i^\star_t} p_{t,i}^\alpha = 4^\alpha h_t = 2h_t. \end{aligned}$$

We note that Ito et al. [2024] enforce this condition by using a hybrid regularizer, whereas here it holds for the pure Tsallis regularizer; the proof exploits the decoupled exploration design $q_{t,i^\star_t} = 1/2$ (see the proof of Lemma A.15).

Thus, the conditions of Lemma A.2 hold for D-TINF-SPM.

1. Adversarial regimes: Since $h_s \leq \bar{h}$, we have $\sum_{s=1}^t z_s/h_s \geq \bar{h}^{-1}\sum_{s=1}^t z_s$, and hence, path-wise,

$$\sum_{t=1}^T \frac{z_t}{\sqrt{\sum_{s=1}^t z_s/h_s}} \leq \sqrt{\bar{h}}\sum_{t=1}^T \frac{z_t}{\sqrt{\sum_{s=1}^t z_s}} \leq 2\sqrt{\bar{h}\sum_{t=1}^T z_t} \leq 2\sqrt{\bar{h}\bar{z}T} = 4\sqrt{KT},$$

   where the second inequality is the standard bound $\sum_t x_t/\sqrt{\sum_{s\leq t} x_s} \leq 2\sqrt{\sum_t x_t}$ for $x_t > 0$. By Lemmas A.2 and A.13, we obtain

$$R_T = \mathcal{O}\left(\sqrt{KT} + \sqrt{K}\right) = \mathcal{O}\left(\sqrt{KT}\right).$$

2. Stochastic regimes: The proof uses three deterministic (path-wise) inequalities, namely the first-term three-phase decomposition (Lemma A.8), $z_t \leq 2\sqrt{C_2}\, r_t$ (Lemma A.10), and $z_t h_t \lesssim C_1 r_t$ (Lemma A.11), together with a single probabilistic step (Jensen's inequality and $\mathbb{E}[\sum_t r_t] = R_T$). Self-bounding then gives the stated bound.

   By Lemmas A.2 and A.13, for any $1 \leq T_1 \leq T_2 \leq T$,

$$R_T = \mathcal{O}\left(\mathbb{E}\left[\sum_{t=1}^T \frac{z_t}{\sqrt{\sum_{s=1}^t z_s/h_s}}\right] + K^{1/2}\right).$$

   To bound the inner sum we apply Lemma A.8 *path-wise* with stopping times defined as follows. Let $A_t \coloneqq \sum_{s=1}^t z_s/h_s$, fix the thresholds

$$V_1 \coloneqq \frac{C_1^2}{K}, \qquad V_2 \coloneqq L \cdot C_2,$$

   where $L > 0$ is a universal constant fixed below, and define the hitting times

$$T_1 \coloneqq \min\{t \geq 1 : A_t \geq V_1\}, \qquad T_2 \coloneqq \min\{t \geq T_1 : A_t \geq V_2\},$$

   with the convention that the minimum equals T when the corresponding threshold is never reached on $[1, T]$. Since $C_1^2 \leq KC_2$ by Cauchy–Schwarz, $V_1 \leq V_2$, so $T_1 \leq T_2 \leq T$ holds path-wise. By Lemma A.9, each per-step increment $z_t/h_t \leq 1$, so when a threshold is reached the corresponding hitting time overshoots it by at most 1; thus, path-wise,

$$A_{T_1} \leq V_1 + 1, \qquad A_{T_2} \leq V_2 + 1. \tag{41}$$

(When the threshold for T_j is not reached, $T_j = T$ and $A_{T_j} = A_T < V_j$, so (41) holds trivially in that case as well.)

Path-wise, Lemma A.8 together with (41) and the bound $A_{T_2}^{-1/2} \le V_2^{-1/2}$ (which holds when $T_2 < T$, and is trivial when $T_2 = T$ since $\sum_{t>T_2} z_t = 0$ in that case) give

$$\sum_{t=1}^{T} \frac{z_t}{\sqrt{\sum_{s=1}^{t} z_s/h_s}} \lesssim \sqrt{K(V_1+1)} + \sqrt{1+\log(V_2/V_1)} \sqrt{\sum_{t=T_1+1}^{T_2} z_t h_t} + V_2^{-1/2} \sum_{t=T_2+1}^{T} z_t. \tag{42}$$

The first term is bounded by $\sqrt{KV_1} + \sqrt{K} = C_1 + \sqrt{K}$. Taking expectations of (42), applying Jensen's inequality $\mathbb{E}[\sqrt{X}] \le \sqrt{\mathbb{E}[X]}$ to the middle term, and using the path-wise inequalities $z_t h_t \lesssim C_1 r_t$ (Lemma A.11) and $z_t \le 2\sqrt{C_2}\, r_t$ (Lemma A.10) together with $\mathbb{E}[\sum_t r_t] = R_T$, we obtain

$$\mathbb{E}\left[\sum_{t=1}^{T} \frac{z_t}{\sqrt{\sum_{s=1}^{t} z_s/h_s}}\right] \lesssim C_1 + \sqrt{K} + \sqrt{1+\log(V_2/V_1)}\sqrt{C_1 R_T} + 2V_2^{-1/2}\sqrt{C_2}\, R_T.$$

Substituting $V_2 = LC_2$, the last term equals $(2/\sqrt{L})\, R_T$. Choose $L = 16$, so that $2/\sqrt{L} = 1/2$ and the last term is at most $R_T/2$. Combining with the earlier display gives

$$R_T \le \mathcal{O}\Big(C_1 + \sqrt{K} + \sqrt{1+\log(V_2/V_1)}\sqrt{C_1 R_T}\Big) + \tfrac{1}{2} R_T,$$

and using $C_1 \ge K - 1$ (so $\sqrt{K} = \mathcal{O}(C_1)$), the $\sqrt{K}$ term is absorbed into C_1. For the logarithmic factor, since $L = 16$ is a universal constant and $C_1^2 \le KC_2$,

$$1 + \log(V_2/V_1) = 1 + \log(LKC_2/C_1^2) = \mathcal{O}\Big(1 + \log(\sqrt{KC_2}/C_1)\Big).$$

Self-bounding (solving the quadratic-in-$\sqrt{R_T}$ inequality after absorbing $\frac{1}{2} R_T$) yields

$$R_T = \mathcal{O}\left(C_1\left(1 + \log\left(\frac{\sqrt{KC_2}}{C_1}\right)\right)\right). \tag{43}$$

Since $C_1 \ge \frac{1}{\Delta_{\min}}$ and $C_2 \le \frac{K}{\Delta_{\min}^2}$, we have $R_T = \mathcal{O}\left(C_1 \log K\right)$. When $\Delta_i = \mathcal{O}\left(\Delta_{\min}\right)$ for all $i \ne i^\star$, we have $C_1 \ge C\frac{K}{\Delta_{\min}}$ for some constant $C > 0$, which denotes $R_T = \mathcal{O}\left(C_1\right)$.

□

B Alternative proof of the stochastic regret bound for Tsallis-INF-$2/3$

In this section, we provide an alternative proof of the stochastic regret bound for Tsallis-INF-$2/3$ established by Jin et al. [2023]. Our proof assumes a unique optimal arm and refines the original analysis of Rouyer and Seldin [2020] using Hölder's inequality. In Rouyer and Seldin [2020], they used Tsallis-INF with $\alpha = \frac{2}{3}$ and $\eta_t = \frac{2K^{1/2-\alpha}}{\sqrt{t}}$, which we denote as Tsallis-INF-$2/3$.

Lemma B.1. *For Tsallis-INF-$2/3$, the following regret bound holds for arbitrary $T_0 \ge 1$.*

$$R_T = \mathcal{O}\left(\sqrt{KT_0} + \sum_{t=T_0+1}^{T} \frac{K^{\frac{1}{6}}}{\sqrt{t}} \sum_{i \ne i^\star} \mathbb{E}\left[p_{ti}\right]^{\frac{2}{3}}\right).$$

Proof. By the Proof of Theorem 4 in Rouyer and Seldin [2020], we have

$$\text{stability} \leq 2\sqrt{KT_0} + \sum_{i \neq i^\star} \sum_{t=T_0+1}^{T} \frac{CK^{\frac{1}{6}}}{\sqrt{t}} \mathbb{E}\left[p_{ti}\right]^{\frac{2}{3}},$$
$$\text{penalty} \leq \frac{9K^{\frac{1}{6}}}{8} \sum_{i \neq i^\star} \sum_{t=T_0+1}^{T} \frac{\mathbb{E}\left[p_{ti}\right]^{\frac{2}{3}}}{\sqrt{t}} + \frac{9\sqrt{KT_0}}{4} + \frac{9\sqrt{K}}{4},$$

where $C > 0$ is some constant. Thus, we have

$$\begin{aligned} R_T &\leq \text{stability} + \text{penalty} \\ &= \mathcal{O}\left(\sqrt{KT_0} + \sum_{t=T_0+1}^{T} \frac{K^{\frac{1}{6}}}{\sqrt{t}} \sum_{i \neq i^\star} \mathbb{E}\left[p_{ti}\right]^{\frac{2}{3}}\right). \end{aligned}$$

□

Theorem B.2. *For Tsallis-INF-$2/3$, the following regret bound holds in stochastic regimes:*

$$R_T = \mathcal{O}\left(\sqrt{K \sum_{i \neq i^\star} \frac{1}{\Delta_i^2}}\right).$$

Proof. By Hölder's inequality, we have

$$\begin{aligned} \sum_{t=T_0+1}^{T} \frac{K^{\frac{1}{6}}}{\sqrt{t}} \sum_{i \neq i^\star} \mathbb{E}\left[p_{ti}\right]^{\frac{2}{3}} &\leq \sum_{t=T_0+1}^{T} \frac{K^{\frac{1}{6}}}{\sqrt{t}} \left(\sum_{i \neq i^\star} \frac{1}{\Delta_i^2}\right)^{\frac{1}{3}} \left(\sum_{i \neq i^\star} \Delta_i \mathbb{E}\left[p_{ti}\right]\right)^{\frac{2}{3}} \\ &= \sum_{t=T_0+1}^{T} \frac{K^{\frac{1}{6}}}{\sqrt{t}} C_2^{\frac{1}{3}} \mathbb{E}[r_t]^{\frac{2}{3}} \\ &\leq K^{\frac{1}{6}} C_2^{\frac{1}{3}} \left(\sum_{t=T_0+1}^{T} \frac{1}{t^{3/2}}\right)^{\frac{1}{3}} \left(\sum_{t=T_0+1}^{T} \mathbb{E}[r_t]\right)^{\frac{2}{3}} \\ &\leq \left(2\sqrt{K} C_2 \frac{1}{\sqrt{T_0}}\right)^{\frac{1}{3}} R_T^{\frac{2}{3}}, \end{aligned}$$

where $C_2 = \sum_{i \neq i^\star} \frac{1}{\Delta_i^2}$ and $r_t = \sum_{i \neq i^\star} \Delta_i p_{ti}$. The last inequality holds because $\sum_{t=T_0+1}^{T} \frac{1}{t^{3/2}} \leq \int_{T_0}^{\infty} \frac{1}{t^{3/2}} dt = \frac{2}{\sqrt{T_0}}$. Thus, we have

$$\begin{aligned} R_T &= \mathcal{O}\left(\sqrt{KT_0} + \left(\sqrt{K} C_2 \frac{1}{\sqrt{T_0}}\right)^{\frac{1}{3}} R_T^{\frac{2}{3}}\right) \\ &= \mathcal{O}\left(\sqrt{KT_0} + \sqrt{K} C_2 \frac{1}{\sqrt{T_0}}\right) \end{aligned}$$

By optimizing $T_0 = C_2$, we obtain

$$R_T = \mathcal{O}\left(\sqrt{KC_2}\right) = \mathcal{O}\left(\sqrt{K \sum_{i \neq i^\star} \frac{1}{\Delta_i^2}}\right).$$

□

B.1 Comparison of regret upper bounds

First, we compare the regret upper bounds of D-Exp3-Ada, D-TINF-SPM, and Tsallis-INF-$2/3$ in stochastic regimes.

1. D-Exp3-Ada: $R_T = \mathcal{O}\left(C_1 \log K\right)$
2. D-TINF-SPM: $R_T = \mathcal{O}\left(C_1 \left(1 + \log \frac{\sqrt{KC_2}}{C_1}\right)\right)$
3. Tsallis-INF-$2/3$: $R_T = \mathcal{O}\left(\sqrt{KC_2}\right)$

Here, $C_1 = \sum_{i \neq i^\star} \frac{1}{\Delta_i}$ and $C_2 = \sum_{i \neq i^\star} \frac{1}{\Delta_i^2}$.

First, we demonstrate that the regret bound of D-TINF-SPM is at least as tight as that of D-Exp3-Ada and Tsallis-INF-$2/3$ in the order sense (up to constants):

$$\begin{aligned} C_1 \left(1 + \log \frac{\sqrt{KC_2}}{C_1}\right) &\leq C_1 \left(1 + \log \frac{K/\Delta_{\min}}{1/\Delta_{\min}}\right) \qquad (44) \\ &= C_1 (1 + \log K) \\ &\leq \left(\frac{1}{\log 2} + 1\right) C_1 \log K, \end{aligned}$$

where Inequality (44) holds since $C_1 \geq \frac{1}{\Delta_{\min}}$ and $C_2 \leq \frac{K}{\Delta_{\min}^2}$. This indicates the bound for D-TINF-SPM is no larger than that for D-Exp3-Ada up to constant factors. Moreover, using the inequality $1 + \log x \leq x$ for $x > 0$, we obtain

$$\begin{aligned} C_1 \left(1 + \log \frac{\sqrt{KC_2}}{C_1}\right) &\leq C_1 \left(\frac{\sqrt{KC_2}}{C_1}\right) \\ &= \sqrt{KC_2}, \end{aligned}$$

which demonstrates that the bound for D-TINF-SPM is no larger than that for Tsallis-INF-$2/3$ up to constant factors.

Next, we turn to the comparison between D-Exp3-Ada and Tsallis-INF-$2/3$. Whether the regret bound for D-Exp3-Ada is tighter than that for Tsallis-INF-$2/3$ depends on the values of Δ_i, K.

- $\Delta_i = \Delta$ for all $i \neq i^\star$:
  In this case, the regret bound for Tsallis-INF-$2/3$ is tighter than that for D-Exp3-Ada.

  $$\begin{aligned} \sqrt{K \sum_{i \neq i^\star} \frac{1}{\Delta_i^2}} &\leq \frac{K}{\Delta} \\ \sum_{i \neq i^\star} \frac{\log K}{\Delta_i} &= \frac{(K-1)\log K}{\Delta} \end{aligned}$$

  Thus, we have

  $$\mathcal{O}\left(\sqrt{K \sum_{i \neq i^\star} \frac{1}{\Delta_i^2}}\right) \subsetneq \mathcal{O}\left(\sum_{i \neq i^\star} \frac{\log K}{\Delta_i}\right).$$

- $\Delta_{\min}$ is sufficiently small compared to the other Δ_i:
  In this case, the regret bound for D-Exp3-Ada is tighter than that for Tsallis-INF-$2/3$. Here, we assume that $\sum_{i \neq i^\star} \frac{1}{\Delta_i} \leq C \frac{1}{\Delta_{\min}}$ for some constant $C > 0$.

  $$\begin{aligned} \sum_{i \neq i^\star} \frac{\log K}{\Delta_i} &\leq C \frac{\log K}{\Delta_{\min}} \\ \sqrt{K \sum_{i \neq i^\star} \frac{1}{\Delta_i^2}} &\geq \frac{\sqrt{K}}{\Delta_{\min}} \end{aligned}$$

Thus, we obtain

$$\mathcal{O}\left(\sum_{i\neq i^\star}\frac{\log K}{\Delta_i}\right)\subsetneq\mathcal{O}\left(\sqrt{K\sum_{i\neq i^\star}\frac{1}{\Delta_i^2}}\right).$$

C Detailed comparison of algorithms

We provide a detailed comparison of the regret bounds for D-Exp3-Ada, D-TINF-SPM, and the algorithm of Rouyer and Seldin [2020] (denoted as Tsallis-INF-$2/3$).

C.1 Stochastic regimes

In stochastic regimes, the three algorithms achieve the following regret bounds:

1. D-Exp3-Ada: $R_T=\mathcal{O}\left(\sum_{i\neq i^\star}\frac{\log K}{\Delta_i}\right)=\mathcal{O}(C_1\log K)$
2. D-TINF-SPM: $R_T=\mathcal{O}\left(C_1\log K\right)$, and when $\Delta_i=\mathcal{O}(\Delta_{\min})$ for all i, $R_T=\mathcal{O}(C_1)$
3. Tsallis-INF-$2/3$ (analysis of Jin et al. [2023]; see also Theorem B.2): $R_T=\mathcal{O}\left(\sqrt{K\sum_{i\neq i^\star}\frac{1}{\Delta_i^2}}\right)=\mathcal{O}(\sqrt{KC_2})$

We now compare these bounds in different scenarios:

Comparison between D-Exp3-Ada and Tsallis-INF-$2/3$:

- When all gaps are approximately equal (i.e., $\Delta_i\asymp\Delta$ for all $i\neq i^\star$), we have

$$C_1\log K=\mathcal{O}\left(\frac{K\log K}{\Delta}\right),$$
$$\sqrt{KC_2}=\mathcal{O}\left(\frac{\sqrt{K^2}}{\Delta}\right)=\mathcal{O}\left(\frac{K}{\Delta}\right).$$

  In this regime, the bound for Tsallis-INF-$2/3$ is tighter than that for D-Exp3-Ada by a factor of $\log K$.
- When gaps are non-uniform, the comparison depends on the specific gap structure. If $\sum_{i\neq i^\star}\frac{1}{\Delta_i^2}\ll\left(\sum_{i\neq i^\star}\frac{1}{\Delta_i}\right)^2$, then Tsallis-INF-$2/3$ achieves a tighter bound.

Comparison between D-TINF-SPM and Tsallis-INF-$2/3$:

- For uniform gaps, the coarser bound for D-TINF-SPM evaluates to $\mathcal{O}(K\log K/\Delta)$, while Tsallis-INF-$2/3$ achieves $R_T=\mathcal{O}(K/\Delta)$. However, uniform gaps satisfy $\Delta_i=\mathcal{O}(\Delta_{\min})$, so the refined bound of D-TINF-SPM in Theorem 5.4 applies and gives $R_T=\mathcal{O}(C_1)=\mathcal{O}(K/\Delta)$, matching Tsallis-INF-$2/3$ up to constants.
- More generally, when $\Delta_i=\mathcal{O}(\Delta_{\min})$ for all $i\neq i^\star$, D-TINF-SPM achieves $R_T=\mathcal{O}(C_1)=\mathcal{O}(K/\Delta_{\min})$, which matches the bound of Tsallis-INF-$2/3$ up to constant factors.
- In the heterogeneous gap setting (Theorem 6.2), D-TINF-SPM achieves $R_T=\mathcal{O}(m\log K/\Delta)$, while Tsallis-INF-$2/3$ achieves $R_T=\mathcal{O}(\sqrt{mK}/\Delta)$. Thus D-TINF-SPM is strictly tighter when $m\log K<\sqrt{mK}$, i.e., when $m<K/(\log K)^2$.

Comparison between D-Exp3-Ada and D-TINF-SPM:

- Both algorithms achieve $R_T=\mathcal{O}(C_1\log K)$ in general stochastic settings.
- When $\Delta_i=\mathcal{O}(\Delta_{\min})$ for all $i\neq i^\star$, D-TINF-SPM improves to $R_T=\mathcal{O}(C_1)$, which is strictly tighter than the bound of D-Exp3-Ada by a factor of $\log K$.

C.2 Adversarial regimes

In adversarial regimes, the three algorithms achieve:

1. D-Exp3-Ada: $R_T = \mathcal{O}\left(\sqrt{KT \log K}\right)$
2. D-TINF-SPM: $R_T = \mathcal{O}\left(\sqrt{KT}\right)$ (minimax optimal)
3. Tsallis-INF-$2/3$: $R_T = \mathcal{O}\left(\sqrt{KT}\right)$ (minimax optimal)

Both D-TINF-SPM and Tsallis-INF-$2/3$ achieve the minimax optimal regret bound, while D-Exp3-Ada is suboptimal by a factor of $\sqrt{\log K}$.

C.3 Summary

D-TINF-SPM achieves the best overall performance:

- Minimax optimal in adversarial regimes
- Near-optimal in stochastic regimes, with the tightest bounds when gaps are approximately uniform
- Reduces the upper-lower bound gap from $\mathcal{O}(\sqrt{K})$ (achieved by Tsallis-INF-$2/3$) to $\mathcal{O}(\log K)$ in the heterogeneous gap setting

D Proofs for lower bound analysis

D.1 Equivalent reading of Corollary 6.3 and asymptotic comparison

Corollary 6.3 can be equivalently stated in terms of monotone, permutation-invariant regret upper bounds.

Remark D.1. If an algorithm achieves a regret upper bound $\overline{R}(\Delta)$ that is monotonically non-increasing in Δ (i.e., $\overline{R}(\Delta) \geq \overline{R}(\Delta')$ whenever $\Delta_i \leq \Delta'_i$ for all i) and permutation invariant, then $\overline{R}(\Delta) = \Omega\big(\frac{1}{\log K} \sum_{i \neq i^\star} 1/\Delta_i\big)$.

To compare our upper bounds with the lower bound of Theorem 6.2, we focus on the asymptotic regime where K and m are fixed and $\Delta \to 0$. In this setting, D-Exp3-Ada achieves $R_T = \mathcal{O}(m \log K/\Delta)$, D-TINF-SPM achieves $R_T = \mathcal{O}(m(1 + \log(K/m))/\Delta)$, and the algorithm of Rouyer and Seldin [2020] achieves $R_T = \mathcal{O}(\sqrt{mK}/\Delta)$. Combined with the lower bound $\Omega(m/\Delta)$, this shows that our algorithms reduce the upper-lower bound gap from $\sqrt{K/m}$ (achieved by prior work) to $\log K$ (D-Exp3-Ada) and $1 + \log(K/m)$ (D-TINF-SPM).

D.2 KL-divergence decomposition lemma

We restate the lemma adapted from Lemma 15.1 of Lattimore and Szepesvári [2020] and used in Section 6.

Lemma D.2. *Let $\nu = (P_1, \ldots, P_K)$ be the loss distributions associated with one K-armed bandit, and let $\nu' = (P'_1, \ldots, P'_K)$ be those of another K-armed bandit. Fix a policy π and let $\mathbb{P}_\nu = \mathbb{P}_{\nu\pi}$ and $\mathbb{P}_{\nu'} = \mathbb{P}_{\nu'\pi}$ be the probability measures on the canonical bandit model induced by the T-round interconnection of π with ν and ν', respectively. Then*

$$D(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) = \sum_{i=1}^{K} \mathbb{E}_\nu[T_i^{(B)}] D(P_i, P'_i), \tag{45}$$

where $T_i^{(B)}$ is the number of times arm i is chosen as the exploration arm up to round T.

D.3 Proof of Lemma D.2

Proof. The proof follows the same approach as Lemma 15.1 in Lattimore and Szepesvári [2020]. Let x_t be the observed loss at round t.

$$\begin{aligned}&\mathbb{P}_\nu(a_1, b_1, x_1, \ldots, a_T, b_T, x_T)\\ =&\prod_{t=1}^{T} \pi(a_t, b_t | a_1, b_1, x_1, \ldots, a_{t-1}, b_{t-1}, x_{t-1}) P_{b_t}(x_t).\end{aligned}$$

Consequently, we have

$$\log \frac{d\mathbb{P}_\nu}{d\mathbb{P}_{\nu'}}(a_1, b_1, x_1, \ldots, a_T, b_T, x_T) = \sum_{t=1}^{T} \log \frac{P_{b_t}(x_t)}{P'_{b_t}(x_t)}.$$

Following the same reasoning as Lemma 15.1,

$$\begin{aligned}D(\mathbb{P}_\nu, \mathbb{P}_{\nu'}) &= \mathbb{E}_\nu \left[\log \frac{d\mathbb{P}_\nu}{d\mathbb{P}_{\nu'}}(A_1, B_1, X_1, \ldots, A_T, B_T, X_T) \right] \\ &= \mathbb{E}_\nu \left[\sum_{t=1}^{T} \log \frac{P_{B_t}(X_t)}{P'_{B_t}(X_t)} \right] \\ &= \mathbb{E}_\nu \left[\sum_{t=1}^{T} D(P_{B_t}, P'_{B_t}) \right] \\ &= \sum_{i=1}^{K} \mathbb{E}_\nu[T_i^{(B)}] D(P_i, P'_i).\end{aligned}$$

□

D.4 General lower bound

We first present a general lower bound that applies to arbitrary gap structures. This result is weaker than the specific lower bounds presented in the main text (Theorems 6.1 and 6.2), but provides a general characterization.

Theorem D.3 (General Lower Bound)**.** *For any stochastic decoupled bandit environment μ with K arms where losses are Bernoulli distributed, there exists an environment μ', which is constructed by swapping the roles of some suboptimal arm i and the optimal arm $i^\star$ in μ, such that the regret satisfies*

$$R_T = \Omega\left(\frac{K\Delta_{\min}}{\sum_{i \neq i^\star} \Delta_i^2} \right),$$

for $T \geq \frac{4(K-2)^2 \Delta_{\min}^2}{9(C\Delta_{\min})^2 (K-1) \sum_{i \neq i^\star} \Delta_i^2}$, where $C > 0$ is some constant. Here, μ is assumed to lie in $[c, 1-c]^K$ for some constant $c \in (0, \frac{1}{2})$.

Proof. We consider the following K kinds of environments.

- $\mu^{(i^\star)}$: Arbitrary environment where the suboptimality gap of arm i is Δ_i and the optimal arm is $i^\star$.
- $\mu^{(i)}$ ($i \neq i^\star$): We construct $\mu^{(i)}$ by swapping the loss distributions of arm i and arm $i^\star$ in $\mu^{(i^\star)}$.

$$\begin{cases} \mu_j^{(i)} = \mu_j^{(i^\star)} & (j \neq i, i^\star) \\ \mu_i^{(i)} = \mu_{i^\star}^{(i^\star)} \\ \mu_{i^\star}^{(i)} = \mu_i^{(i^\star)} \end{cases}$$

Let $\mathbb{E}_i[\cdot]$ be the expectation when the environment is $\mu^{(i)}$, and let $T_i^{(A)}$ be the number of times arm i is chosen as the exploitation arm up to round T. We define $\mathbb{P}^{(i)} = \mathbb{P}_{\mu^{(i)}}$.

$$\begin{aligned}
\mathbb{E}_i[T_i^{(A)}] - \mathbb{E}_{i^\star}[T_i^{(A)}] &= \int T_i^{(A)}(\mathbb{P}^{(i)}(\omega) - \mathbb{P}^{(i^\star)}(\omega))d\omega \\
&\leq T \sup_{A\in\mathcal{F}}(\mathbb{P}^{(i)}(A) - \mathbb{P}^{(i^\star)}(A)) \\
&= T\delta(\mathbb{P}^{(i)}, \mathbb{P}^{(i^\star)}) \\
&\leq T\sqrt{\frac{1}{2}D(\mathbb{P}^{(i^\star)}, \mathbb{P}^{(i)})} \qquad (46) \\
&= T\sqrt{\frac{1}{2}\sum_{j=1}^{K}\mathbb{E}_{i^\star}[T_j^{(B)}]D(P_j^{(i^\star)}, P_j^{(i)})} \qquad (47) \\
&= T\sqrt{\frac{1}{2}\left(\mathbb{E}_{i^\star}[T_i^{(B)}]D(P_i^{(i^\star)}, P_i^{(i)}) + \mathbb{E}_{i^\star}[T_{i^\star}^{(B)}]D(P_{i^\star}^{(i^\star)}, P_{i^\star}^{(i)})\right)} \\
&= CT\Delta_i\sqrt{\mathbb{E}_{i^\star}[T_i^{(B)}] + \mathbb{E}_{i^\star}[T_{i^\star}^{(B)}]}, \qquad (48)
\end{aligned}$$

where $C > 0$ is a constant and $\mathcal{F}$ is the σ-algebra generated by the history H_T. Inequality (46) holds by Pinsker's inequality. Equality (47) holds by Lemma D.2. The last equality holds since the losses associated with $\mu^{(i)}, \mu^{(i^\star)}$ are Bernoulli distributed as follows:

$$\begin{aligned}
D(P_i^{(i^\star)}, P_i^{(i)}) &= \mu^{(i)}\log\frac{\mu^{(i)}}{\mu^{(i^\star)}} + (1-\mu^{(i)})\log\frac{1-\mu^{(i)}}{1-\mu^{(i^\star)}} \\
&= \left(-\Delta_i + \frac{\Delta_i^2}{2\mu^{(i^\star)}}\right) + \left(\Delta_i + \frac{\Delta_i^2}{2(1-\mu^{(i^\star)})}\right) + \mathcal{O}(\Delta_i^3) \qquad (49) \\
&\leq \frac{\Delta_i^2}{2c(1-c)} + \mathcal{O}(\Delta_i^3) \qquad (50) \\
&= \mathcal{O}(\Delta_i^2),
\end{aligned}$$

where Equality (49) holds by Taylor expansion. Inequality (50) holds since $\mu^{(i^\star)} \in [c, 1-c]$. By the same argument, an analogous bound holds for $D(P_{i^\star}^{(i^\star)}, P_{i^\star}^{(i)})$, yielding $\mathcal{O}(\Delta_i^2)$ as well.

We define $R_T^{(i)}$ as the expected regret when the environment is $\mu^{(i)}$.

$$\begin{aligned}
\sum_{i\neq i^\star} R_T^{(i)} &= \sum_{i\neq i^\star}\left(\sum_{j\neq i, i^\star}\Delta_j\mathbb{E}_i[T_j^{(A)}] + \Delta_i\mathbb{E}_i[T_{i^\star}^{(A)}]\right) \\
&\geq \sum_{i\neq i^\star}\Delta_{\min}\sum_{j\neq i}\mathbb{E}_i[T_j^{(A)}] \\
&= \sum_{i\neq i^\star}\Delta_{\min}(T - \mathbb{E}_i[T_i^{(A)}]) \\
&\geq \sum_{i\neq i^\star}\Delta_{\min}\left(T - \mathbb{E}_{i^\star}[T_i^{(A)}] - CT\Delta_i\sqrt{\mathbb{E}_{i^\star}[T_i^{(B)}] + \mathbb{E}_{i^\star}[T_{i^\star}^{(B)}]}\right) \qquad (51) \\
&\geq (K-1)\Delta_{\min}T - \Delta_{\min}T - CT\Delta_{\min}\sqrt{\sum_{i\neq i^\star}\Delta_i^2\sum_{i\neq i^\star}(\mathbb{E}_{i^\star}[T_i^{(B)}] + \mathbb{E}_{i^\star}[T_{i^\star}^{(B)}])} \qquad (52) \\
&= (K-2)\Delta_{\min}T - CT\Delta_{\min}\sqrt{\sum_{i\neq i^\star}\Delta_i^2(T + (K-2)\mathbb{E}_{i^\star}[T_{i^\star}^{(B)}])} \\
&\geq (K-2)\Delta_{\min}T - CT^{3/2}\Delta_{\min}\sqrt{(K-1)\sum_{i\neq i^\star}\Delta_i^2}.
\end{aligned}$$

Inequality (51) holds by Inequality (48). Inequality (52) holds by Cauchy-Schwarz inequality. Let $f(T) = AT - BT^{3/2}$ denote the derived lower bound function, where $A = (K-2)\Delta_{\min}, B = C\Delta_{\min}\sqrt{(K-1)\sum_{i\neq i^\star}\Delta_i^2}$. By elementary optimization, $f(T)$ attains its maximum at $T_{\max} = \frac{4A^2}{9B^2}$. Since $f(T)$ strictly decreases for $T > T_{\max}$ and the cumulative expected pseudo-regret R_T is monotonically non-decreasing over time, for any $T \geq T_{\max}$, we have

$$\sum_{i\neq i^\star} R_T^{(i)} \geq \sum_{i\neq i^\star} R_{\lfloor T_{\max}\rfloor}^{(i)} \geq f(\lfloor T_{\max}\rfloor).$$

Thus, for $T \geq T_{\max} = \frac{4(K-2)^2\Delta_{\min}^2}{9(C\Delta_{\min})^2(K-1)\sum_{i\neq i^\star}\Delta_i^2}$, the following holds:

$$\begin{aligned}
\sum_{i\neq i^\star} R_T^{(i)} &\geq f(\lfloor T_{\max}\rfloor) \\
&= \Omega\left(f(T_{\max})\right) \\
&= \Omega\left(\frac{4}{27}\frac{(K-2)^3\Delta_{\min}^3}{(C\Delta_{\min})^2(K-1)\sum_{i\neq i^\star}\Delta_i^2}\right) \\
&= \Omega\left(\frac{K^2\Delta_{\min}}{\sum_{i\neq i^\star}\Delta_i^2}\right).
\end{aligned}$$

Then, we obtain

$$\exists i \in \{1,\ldots,K\}, i \neq i^\star, R_T^{(i)} = \Omega\left(\frac{K\Delta_{\min}}{\sum_{j\neq i^\star}\Delta_j^2}\right),$$

for $T \geq \frac{4(K-2)^2\Delta_{\min}^2}{9(C\Delta_{\min})^2(K-1)\sum_{i\neq i^\star}\Delta_i^2}$. □

D.5 Proof of Theorem 6.1

Proof. Fix an arbitrary environment $\mu \in [c, 1-c]^K$ satisfying the assumption of the theorem, with optimal arm $i^\star$, and let $\ell := \mu_{i^\star} + \Delta$, so that $\mu_{i^\star} = \ell - \Delta$ and $\mu_j = \ell$ for all $j \neq i^\star$. We consider the $K+1$ kinds of environments $\mu^{(i)}$ ($i = 0, 1, \ldots, K$). Let $\mu_i^{(0)} = \ell$ (for all i) and let $\mu_j^{(i)} = \ell$ ($j \neq i$), $\mu_i^{(i)} = \ell - \Delta$, so that arm i is the unique optimal arm in $\mu^{(i)}$ with suboptimality gap Δ on every other arm. By construction, $\mu^{(i^\star)} = \mu$, and for every $i \neq i^\star$, $\mu^{(i)}$ is obtained from μ by swapping the roles of arm i and arm $i^\star$. Hence it suffices to show that $R_T^{(i)} = \Omega(K/\Delta)$ for some $i \in [K]$, where $R_T^{(i)}$ denotes the regret in $\mu^{(i)}$. All means involved, ℓ and $\ell - \Delta$, are means of μ and thus lie in $[c, 1-c]$.

Each $\mu^{(i)}$ differs from $\mu^{(0)}$ only in the i-th coordinate. Hence, by Lemma D.2 and the bound $D(\mathrm{Ber}(p), \mathrm{Ber}(q)) \leq (p-q)^2/(q(1-q))$ on the Kullback–Leibler divergence between Bernoulli distributions,

$$D(\mathbb{P}^{(0)}, \mathbb{P}^{(i)}) = \mathbb{E}_0[T_i^{(B)}]\, D(\mathrm{Ber}(\ell), \mathrm{Ber}(\ell - \Delta)) \leq \mathbb{E}_0[T_i^{(B)}]\frac{\Delta^2}{c(1-c)},$$

where we used $\ell - \Delta \in [c, 1-c]$. As in (46), Pinsker's inequality then gives

$$\mathbb{E}_i[T_i^{(A)}] - \mathbb{E}_0[T_i^{(A)}] \leq T\sqrt{\frac{1}{2}D(\mathbb{P}^{(0)}, \mathbb{P}^{(i)})} \leq CT\Delta\sqrt{\mathbb{E}_0[T_i^{(B)}]}, \tag{53}$$

where $C = 1/\sqrt{2c(1-c)}$.

Thus, we have

$$\begin{aligned}
\sum_{i=1}^{K} R_T^{(i)} &= \sum_{i=1}^{K} \sum_{j \neq i} \Delta \mathbb{E}_i[T_j^{(A)}] \\
&= \sum_{i=1}^{K} \Delta \left(T - \mathbb{E}_i[T_i^{(A)}] \right) \\
&\geq \sum_{i=1}^{K} \Delta (T - \mathbb{E}_0[T_i^{(A)}] - CT\Delta \sqrt{\mathbb{E}_0[T_i^{(B)}]}) \qquad (54) \\
&\geq \Delta \left(KT - T - C\Delta \sqrt{K} T^{\frac{3}{2}} \right) \qquad (55) \\
&\geq \Delta \frac{4(K-1)^3}{27C^2\Delta^2 K} \qquad (56) \\
&= \Omega \left(\frac{K^2}{\Delta} \right), \qquad (57)
\end{aligned}$$

where inequality (54) holds by Inequality (53). Inequality (55) holds by Cauchy-Schwarz inequality. Inequality (56) holds when $T \geq \frac{4(K-1)^2}{9C^2\Delta^2 K}$, which follows from the same reasoning as the proof of Theorem D.3.

Then, we obtain

$$\exists i \in \{1, \ldots, K\}, R_T^{(i)} = \Omega \left(\frac{K}{\Delta} \right),$$

when $T \geq \frac{4(K-1)^2}{9C^2\Delta^2 K}$. The environment $\mu' \coloneqq \mu^{(i)}$ is μ itself if $i = i^\star$, and is obtained from μ by swapping arms i and $i^\star$ otherwise, which proves the theorem. □

D.6 Proof of Theorem 6.2

Proof. Fix an arbitrary environment $\mu \in [c, 1-c]^K$ satisfying the assumption of the theorem. Let $i^\star$ be its optimal arm, $a \coloneqq \mu_{i^\star}$, H the set of the m arms with suboptimality gap Δ (so that $\mu_j = a + \Delta$ for $j \in H$), and $S \coloneqq H \cup \{i^\star\}$; every arm $j \notin S$ satisfies $\mu_j - a \geq \frac{1}{2} - c > \Delta$. We consider the $m+2$ kinds of environments $\mu^{(0)}$ and $\mu^{(i)}$ ($i \in S$). Let $\mu_j^{(0)} = a + \Delta$ ($j \in S$) and $\mu_j^{(0)} = \mu_j$ ($j \notin S$), i.e., $\mu^{(0)}$ is obtained from μ by raising the mean of arm $i^\star$ to $a + \Delta$. For $i \in S$, we define $\mu^{(i)}$ by setting $\mu_j^{(i)} = \mu_j^{(0)}$ ($j \neq i$) and $\mu_i^{(i)} = a$. By construction, $\mu^{(i^\star)} = \mu$, and for every $i \in H$, $\mu^{(i)}$ is obtained from μ by swapping the roles of arm i and arm $i^\star$. In $\mu^{(i)}$, arm i is the unique optimal arm, the arms in $S \setminus \{i\}$ have suboptimality gap Δ, and the arms outside S have suboptimality gap at least $\frac{1}{2} - c > \Delta$. All means involved are means of μ and thus lie in $[c, 1-c]$.

Each $\mu^{(i)}$ differs from $\mu^{(0)}$ only in the i-th coordinate, where the means are $a + \Delta$ and $a \in [c, 1-c]$. Hence, exactly as in the proof of Theorem 6.1, we have

$$\mathbb{E}_i[T_i^{(A)}] - \mathbb{E}_0[T_i^{(A)}] \leq CT\Delta \sqrt{\mathbb{E}_0[T_i^{(B)}]}, \qquad (58)$$

where $C = 1/\sqrt{2c(1-c)}$.

Since every arm other than i has suboptimality gap at least Δ in $\mu^{(i)}$, and since $\sum_{i\in S}\mathbb{E}_0[T_i^{(A)}] \le T$ and $\sum_{i\in S}\mathbb{E}_0[T_i^{(B)}] \le T$, we have

$$\begin{aligned}
\sum_{i\in S} R_T^{(i)} &\ge \sum_{i\in S} \Delta\left(T - \mathbb{E}_i[T_i^{(A)}]\right) \\
&\ge \Delta \sum_{i\in S}\left(T - \mathbb{E}_0[T_i^{(A)}] - CT\Delta\sqrt{\mathbb{E}_0[T_i^{(B)}]}\right) \\
&\ge \Delta\left(mT - C\Delta\sqrt{m+1}T^{3/2}\right) \qquad (59)\\
&\ge \Delta\frac{4m^3}{27C^2\Delta^2(m+1)} \qquad (60)\\
&= \Omega\left(\frac{m^2}{\Delta}\right),
\end{aligned}$$

where Inequality (59) holds by Cauchy-Schwarz inequality. Inequality (60) holds when $T \ge \frac{4m^2}{9C^2\Delta^2(m+1)}$, which follows from the same reasoning as the proof of Theorem D.3.

Then, since $|S| = m+1 \le 2m$, we obtain

$$\exists i \in S, R_T^{(i)} = \Omega\left(\frac{m}{\Delta}\right), \qquad (61)$$

when $T \ge \frac{4m^2}{9C^2\Delta^2(m+1)}$. The environment $\mu' \coloneqq \mu^{(i)}$ is μ itself if $i = i^\star$, and is obtained from μ by swapping $i^\star$ and the gap-Δ arm i otherwise, which proves the theorem. □

D.7 Proof of Corollary 6.3

Lemma D.4. *For any nonnegative sequence $\{a_i\}_{i=1}^n$, the following inequality holds:*

$$\left(\sum_{j=1}^n \frac{1}{j}\right)\cdot \max_{i\le n}\{ia_i\} \ge \sum_{i=1}^n a_i. \qquad (62)$$

Consequently, $\max_{i\le n}\{ia_i\} = \Omega\left(\frac{1}{\log(n+1)}\sum_{i=1}^n a_i\right)$.

Proof.

$$\sum_{i=1}^n a_i = \sum_{i=1}^n \frac{1}{i}\cdot ia_i \le \sum_{i=1}^n \frac{1}{i}\cdot \max_{j\le n}\{ja_j\}. \qquad (63)$$

□

Combining this lemma with Theorem 6.2 yields the proof of Corollary 6.3.

Proof of Corollary 6.3. Without loss of generality, we suppose that $(\underline{\Delta}_i)_{i\le K-1}$ is monotonically non-decreasing: $\underline{\Delta}_1 \le \underline{\Delta}_2 \le \cdots \le \underline{\Delta}_{K-1}$. Applying Lemma D.4 with $a_i = \frac{1}{\underline{\Delta}_i}$, there exists m such that

$$\frac{m}{\underline{\Delta}_m} = \Omega\left(\frac{1}{\log K}\sum_{i=1}^{K-1}\frac{1}{\underline{\Delta}_i}\right). \qquad (64)$$

From Theorem 6.2 with this m and $\Delta = \underline{\Delta}_m$, applied to the base environment given in (i) below with $i^\star = 1$ (the hard environment it provides is this base environment or one of its swaps, i.e., again of the form (i) with some $1 \le i^\star \le m+1$), for any algorithm, there exists a Bernoulli bandit environment such that:

(i) The mean loss vector μ is set as follows:

$$\mu_i = \begin{cases} \frac{1}{2} & i = i^\star \\ \frac{1}{2} + \underline{\Delta}_m & i \in \{1, \dots, m+1\} \setminus \{i^\star\} \\ 1 - c & m+2 \le i \le K, \end{cases}$$

where $1 \le i^\star \le m+1$.

(ii) The expected regret of the algorithm for sufficiently large T is lower bounded by $\Omega\left(\frac{m}{\underline{\Delta}_m}\right)$.

By applying (64), we have $R_T = \Omega\left(\frac{1}{\log K} \sum_{i=1}^{K-1} \frac{1}{\underline{\Delta}_i}\right)$ in this environment. Moreover, we set

$$\sigma(i) = \begin{cases} i & 1 \le i \le i^\star - 1 \\ i+1 & i^\star \le i \le K-1 \\ i^\star & i = K, \end{cases}$$

which satisfies $\Delta_{\sigma(i)} = \underline{\Delta}_m \ge \underline{\Delta}_i$ for $1 \le i \le m$ and $\Delta_{\sigma(i)} = \frac{1}{2} - c \ge \underline{\Delta}_i$ for $m+1 \le i \le K-1$ (the latter using $\underline{\Delta}_i \in [0, 1/2 - c]$). Thus, this environment satisfies the required conditions. □

Remark D.5. Our regret upper bounds of $\sum_{i \neq i^\star} \frac{\log K}{\Delta_i}$ immediately imply the upper bounds of $\sum_{i=1}^{K-1} \frac{\log K}{\underline{\Delta}_i}$ for any environments satisfying the condition (i) in Corollary 6.3. Thus, Corollary 6.3 shows that our regret upper bounds are optimal up to a factor of $\mathcal{O}((\log K)^2)$. We also note that Corollary 6.3 implies the following statement as well: If an algorithm achieves a regret upper bound of $\omega(\Delta)$ for all $\Delta = (\Delta_i)_{i \neq i^\star}$, where the function ω is assumed to be monotonically non-increasing (i.e., $\omega(\Delta) \ge \omega(\Delta')$ if $\Delta_i \le \Delta'_i$ for all i) and permutation invariant (i.e., if Δ' is given by $\Delta'_i = \Delta_{\sigma(i)}$ for some permutation σ, then $\omega(\Delta) = \omega(\Delta')$), then $\omega(\Delta) = \Omega\left(\frac{1}{\log K} \sum_{i \neq i^\star} \frac{1}{\Delta_i}\right)$. In fact, Corollary 6.3 implies that, for any Δ, there exists $\Delta' \ge \Delta$ such that $\omega(\Delta') = \Omega\left(\frac{1}{\log K} \sum_{i \neq i^\star} \frac{1}{\Delta_i}\right)$. From the assumption of monotone non-increasingness of ω, this implies $\omega(\Delta) = \Omega\left(\frac{1}{\log K} \sum_{i \neq i^\star} \frac{1}{\Delta_i}\right)$.

E Additional related work

Here we provide additional background on Best-of-Both-Worlds (BOBW) algorithms and the SPM framework, complementing the brief discussion in Section 2.

There has been substantial interest in developing algorithms that perform optimally in both stochastic and adversarial environments without prior knowledge of the regime, known as "Best-of-Both-Worlds" (BOBW) algorithms (Bubeck and Slivkins, 2012; Zimmert and Seldin, 2021; Ito et al., 2024). A prominent approach is the FTRL framework with Tsallis entropy regularization, as exemplified by the Tsallis-INF algorithm of Zimmert and Seldin [2021], which achieves optimal BOBW performance for the standard MAB problem. Recently, Ito et al. [2024] proposed an adaptive learning rate scheme called stability-penalty matching (SPM), which formulates a general method for tuning FTRL's learning rate. This scheme contributes to constructing BOBW algorithms in various settings such as multi-armed bandits, graph bandits, linear bandits, and contextual bandits.

Although the SPM framework provides a powerful methodology for constructing BOBW algorithms, its application to the decoupled setting has not been previously studied. Unlike standard MAB problems, the decoupled setting requires managing two distinct actions (exploration and exploitation), which complicates the direct application of standard SPM analysis. Furthermore, while Avner et al. [2012] established a lower bound in adversarial regimes, lower bounds in stochastic regimes have not been thoroughly investigated. In this paper, we first propose D-Exp3-Ada, which is based on Exp3 with an AdaGrad learning rate, corresponding to FTRL with Shannon entropy. We demonstrate that this approach achieves regret bounds competitive with the results of Rouyer and Seldin [2020]. To achieve tighter bounds, we further propose D-TINF-SPM, which employs Tsallis entropy regularization with the SPM learning rate. Crucially, we also establish the first non-trivial instance-dependent lower bound for the decoupled setting.

F Remarks on the algorithm definitions

These remarks accompany Section 4.

Remark F.1 (On the exact limit as $\alpha \to 1$). Strictly speaking, the limit as $\alpha \to 1$ of the Tsallis regularizer ψ_α in (5) yields 0 rather than the negative Shannon entropy. If the coefficient were $\frac{1}{1-\alpha}$ instead of $\frac{1}{\alpha}$, the limit would converge to the negative Shannon entropy. In our framework, the learning rate naturally absorbs this coefficient. Therefore, the FTRL update remains well-defined and formally reduces to the Shannon entropy case when $\alpha = 1$, justifying our unified analysis.

Remark F.2 (SPM as a generalization of the AdaGrad-style learning rate). The SPM learning rate (10) can be viewed as a generalization of the AdaGrad-style rule (9). When $h_t = \bar{h}$ for all t (i.e., the penalty term is constant), the SPM update rule yields $\beta_t^2 \asymp \sum_{s=1}^{t-1} z_s/\bar{h}$, recovering the AdaGrad-style learning rate $\beta_t \asymp \sqrt{\sum_{s=1}^{t-1} z_s/\bar{h}}$. The SPM framework extends this by allowing time-varying penalty terms h_t, which enables tighter regret bounds when h_t varies significantly across rounds.

Remark F.3 (Achieving T-independent stochastic regret). The T-independent stochastic regret relies fundamentally on the bound $z_t = \mathcal{O}(r_t)$. In the decoupled setting, we can freely design the exploration distribution q_t to satisfy this relation. In contrast, classical MAB rigidly couples exploration and exploitation ($q_t = p_t$), resulting in $z_t = \mathcal{O}(K)$. Due to this difference in z_t, related Shannon-entropy-based BoBW analyses in classical bandits still incur extra logarithmic factors; for instance, Ito et al. [2022] and Jin et al. [2023] achieve only $\mathcal{O}((\log T)^2)$ in the corresponding settings, suggesting that even the milder $\mathcal{O}(\log T)$ stochastic regret with Shannon entropy in classical MAB is hard to obtain.

Remark F.4 (On the unique-optimal-arm assumption). Neither algorithm uses the assumption that the optimal arm is unique; it enters only the self-bounding step of the stochastic analysis. There, the stability and penalty terms are controlled by the instantaneous regret $r_t = \sum_{i \neq i^\star} \Delta_i p_{ti}$, e.g., $z_t \leq 2\sqrt{C_2}\, r_t$ and $z_t h_t \lesssim C_1 r_t$ for D-TINF-SPM (Lemmas A.10 and A.11), and $z_t \leq C_1 r_t$ for D-Exp3-Ada (proof of Lemma A.6). With multiple optimal arms, probability mass on optimal arms other than the current maximizer $i_t^\star$ contributes to z_t and h_t, which exclude only $i_t^\star$, but not to r_t, so that r_t no longer dominates them. The assumption is not needed for our lower bounds (Section 6) or for the adversarial guarantees of Theorems 5.3 and 5.4. For the $\mathcal{O}(\sqrt{KC_2})$ rate, Jin et al. [2023] showed that the assumption can be removed. Their technique, like that of Ito [2021b], decomposes the regret into contributions of suboptimal arms, optimal arms, and a residual term, and bounds the optimal-arm part independently of the gaps using arm-dependent learning rates. This does not apply directly to SPM, which couples all arms through a single scalar learning rate; the SPM framework of Ito et al. [2024] also assumes a unique optimal arm. A natural route to our C_1-type rates without uniqueness is a variant of the SPM update that treats the set of optimal arms as a single coordinate, which we leave for future work. Empirically, the unmodified algorithms remain stable with multiple optimal arms (Appendix G.6).

G Numerical experiments

This section reports numerical experiments that validate the theoretical regret bounds derived in Section 5. We compare our proposed algorithms (D-Exp3-Ada and D-TINF-SPM) with two existing decoupled-bandit algorithms in both stochastic and adversarial environments. The experiments confirm that our gap-dependent bounds translate into observable regret saturation as T grows, while existing algorithms keep accumulating regret over the evaluated horizons. We also confirm that our algorithms remain competitive in adversarial environments, supporting the Best-of-Both-Worlds nature of the proposed methods.

G.1 Experimental setup

Algorithms. We compare the following four algorithms:

- D-Exp3-Ada (ours): Shannon entropy with AdaGrad-style learning rate (9).
- D-TINF-SPM (ours): Tsallis entropy with $\alpha = 1/2$ and the SPM learning rate (10) with $\beta_1 = 16$.

Table 3: Cumulative regret in the stochastic environment with $K = 10$, one hard arm with $\Delta = 0.02$, and eight easy arms with gap 0.5. Mean $\pm$ standard deviation over 100 runs. Smallest mean per row in bold.

T	D-Exp3-Ada (ours)	D-TINF-SPM (ours)	Tsallis-INF-$2/3$	Kim et al.
10^3	87 ± 29	111 ± 26	$\mathbf{62 \pm 19}$	71 ± 23
$5 \cdot 10^3$	113 ± 39	156 ± 42	$\mathbf{96 \pm 32}$	106 ± 35
10^4	131 ± 57	179 ± 60	$\mathbf{123 \pm 51}$	130 ± 50
$5 \cdot 10^4$	$\mathbf{150 \pm 92}$	212 ± 100	200 ± 129	193 ± 102
10^5	$\mathbf{150 \pm 92}$	217 ± 104	233 ± 202	217 ± 115
$2 \cdot 10^5$	$\mathbf{150 \pm 92}$	219 ± 105	253 ± 218	237 ± 119
$5 \cdot 10^5$	$\mathbf{150 \pm 92}$	220 ± 106	269 ± 222	261 ± 122

- Tsallis-INF-$2/3$ [Rouyer and Seldin, 2020]: Tsallis entropy with $\alpha = 2/3$ and decaying learning rate $\eta_t \propto K^{-1/6}/\sqrt{t}$.
- Kim et al. [Kim et al., 2026]: FTPL-based algorithm with Pareto perturbations (Pareto shape parameter 3.0).

All algorithms are implemented in Python following the descriptions in the corresponding papers.

Performance metric. For each run, we report the cumulative pseudo-regret $\sum_{t=1}^{T} \Delta_{A_t}$, where A_t is the exploitation arm chosen at round t and Δ_i is the suboptimality gap of arm i (with respect to the best arm at round t in the non-stationary setting of Section G.7). We report mean and standard deviation across multiple independent runs (each with a different random seed for both algorithm and environment). For the stochastic experiments in Sections G.2–G.5, each random seed produces a single trajectory that is run up to the largest T in the table and the cumulative regret is recorded at every reported horizon along the way; this means that for a given algorithm and seed, the value at a smaller T is the prefix sum of the value at a larger T. Consequently, once the cumulative regret has saturated (no further suboptimal exploitation choices are made), the recorded values at all subsequent horizons are exactly equal for that seed, and the table entries can repeat down a column to the displayed precision. The switching adversarial experiment in Section G.7 instead runs an independent simulation for each T, since the loss process depends on T through the switch time $T/2$.

G.2 Stochastic environment with a single hard arm

We first study a heterogeneous-gap stochastic environment with a single "hard" arm and several "easy" arms, the regime in which our gap-dependent bounds are tightest and the asymptotic advantage over $\sqrt{KT}$-type bounds is most pronounced.

Setting. Bernoulli bandit with $K = 10$ arms: the optimal arm has mean loss 0.5, one "hard" arm has mean loss $0.5 + 0.02$ (gap $\Delta = 0.02$), and the remaining eight "easy" arms have mean loss 1.0 (gap 0.5). We run each algorithm for $T \in \{10^3, 5 \times 10^3, 10^4, 5 \times 10^4, 10^5, 2 \times 10^5, 5 \times 10^5\}$ and report results averaged over 100 independent runs.

Results. Table 3 and Figure 1 summarize cumulative regret. The proposed D-Exp3-Ada exhibits clear regret saturation: its cumulative regret stabilizes around 150 for $T \geq 5 \times 10^4$, consistent with the gap-dependent bound $\mathcal{O}(C_1 \log K)$ in Theorem 5.3. The proposed D-TINF-SPM also saturates (around 220), exhibiting the same qualitative trend at a higher level. In contrast, the regrets of Tsallis-INF-$2/3$ and Kim et al. continue to grow with T, exceeding the regret of D-Exp3-Ada at $T \approx 5 \times 10^4$ and that of D-TINF-SPM from $T = 2 \times 10^5$ on, and reaching 269 and 261 at $T = 5 \times 10^5$, respectively.

Discussion. The crossover behavior in Table 3 matches our theoretical prediction: D-Exp3-Ada and D-TINF-SPM attain T-independent regret in stochastic environments (Theorems 5.3 and 5.4); the stochastic bound for Tsallis-INF-$2/3$ established by Jin et al. [2023] is also T-independent but grows as $\mathcal{O}(\sqrt{KC_2})$, which is larger than $\mathcal{O}(C_1 \log K)$ in this instance. For small T, the constants in the gap-dependent bound dominate and our algorithms incur slightly higher regret than Tsallis-INF-$2/3$.

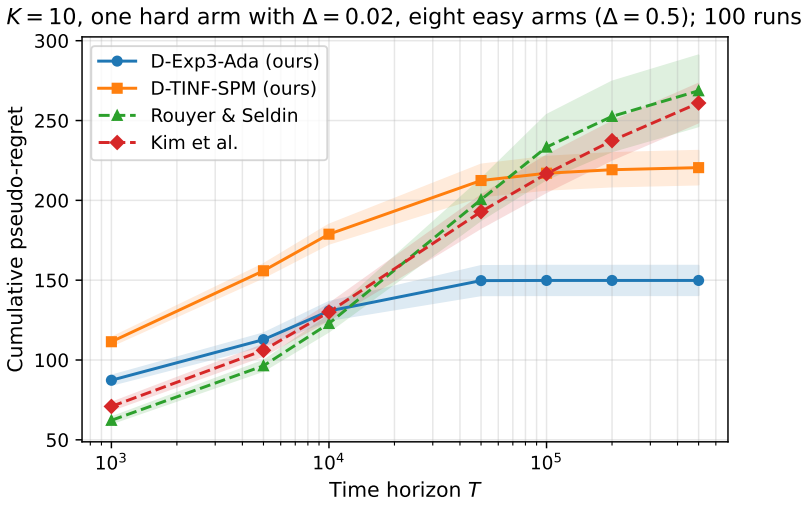


Figure 1: Cumulative regret vs. T in the stochastic environment with $K = 10$, one hard arm with $\Delta = 0.02$, and eight easy arms with gap 0.5. Shaded bands indicate one standard error over 100 runs. D-Exp3-Ada's regret clearly saturates while the regrets of Tsallis-INF-$2/3$ and Kim et al. continue to grow.

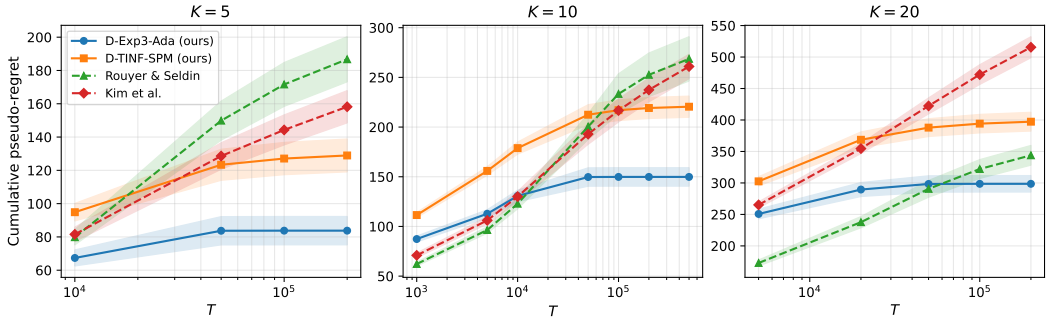


Figure 2: Cumulative regret vs. T for $K \in \{5, 10, 20\}$, with one hard arm ($\Delta = 0.02$) and the remaining arms with gap 0.5. Shaded bands indicate one standard error. The crossover point at which D-Exp3-Ada overtakes Tsallis-INF-$2/3$ shifts to larger T as K grows.

The empirical regret of Tsallis-INF-$2/3$ keeps growing over the evaluated horizons, while the regrets of our algorithms saturate; once the saturation level becomes smaller than the observed regret of the baselines, our algorithms become better (here, from $T \approx 5 \times 10^4$ for D-Exp3-Ada and from $T = 2 \times 10^5$ for D-TINF-SPM, whose saturation level is higher).

G.3 Effect of the number of arms K

To examine how regret scales with the number of arms, we vary $K \in \{5, 10, 20\}$ while keeping the gap structure (one hard arm with $\Delta = 0.02$, the rest with gap 0.5) fixed. Table 4 reports cumulative regret at the longest horizon evaluated for each K, and Figure 2 shows the regret curves. The crossover point shifts to a larger T as K grows, but D-Exp3-Ada eventually overtakes Tsallis-INF-$2/3$ in all three cases. D-TINF-SPM saturates for all three values of K, but at a higher level than D-Exp3-Ada; at $T = 2 \cdot 10^5$ it is below both baselines for $K \in \{5, 10\}$, whereas for $K = 20$ its saturation level (397) exceeds the regret of Tsallis-INF-$2/3$ (344) at this horizon. This is consistent with the upper bound $\mathcal{O}(C_1 \log K)$ for D-Exp3-Ada and the $\mathcal{O}(\sqrt{KC_2})$ bound for Tsallis-INF-$2/3$ [Jin et al., 2023]: the latter scales with K as $\sqrt{K}$, so a larger K enlarges the saturation level of both algorithms but in different ways, and the crossover shifts to larger T.

Table 4: Effect of K on cumulative regret at $T = 2 \cdot 10^5$ in the stochastic environment with one hard arm ($\Delta = 0.02$) and the remaining arms with gap 0.5. Mean $\pm$ standard deviation over 100 runs.

K	D-Exp3-Ada (ours)	D-TINF-SPM (ours)	Tsallis-INF-$2/3$	Kim et al.
5	$\mathbf{84 \pm 85}$	129 ± 98	187 ± 135	158 ± 96
10	$\mathbf{150 \pm 92}$	219 ± 105	253 ± 218	237 ± 119
20	$\mathbf{299 \pm 130}$	397 ± 148	344 ± 156	516 ± 166

Table 5: Effect of the hard arm's gap Δ on cumulative regret with $K = 10$ and eight easy arms with gap 0.5. Mean $\pm$ standard deviation over 100 runs.

Δ	T	D-Exp3-Ada (ours)	D-TINF-SPM (ours)	Tsallis-INF-$2/3$	Kim et al.
0.01	$5 \cdot 10^4$	$\mathbf{174 \pm 104}$	237 ± 101	202 ± 102	192 ± 98
0.01	$2 \cdot 10^5$	$\mathbf{209 \pm 165}$	296 ± 168	337 ± 245	290 ± 165
0.02	$5 \cdot 10^4$	$\mathbf{150 \pm 92}$	212 ± 100	200 ± 129	193 ± 102
0.02	$2 \cdot 10^5$	$\mathbf{150 \pm 92}$	219 ± 105	253 ± 218	237 ± 119
0.05	$5 \cdot 10^4$	$\mathbf{129 \pm 69}$	178 ± 65	151 ± 78	154 ± 62
0.05	$2 \cdot 10^5$	$\mathbf{129 \pm 69}$	179 ± 65	160 ± 79	168 ± 63
0.1	$5 \cdot 10^4$	$\mathbf{120 \pm 43}$	165 ± 53	125 ± 118	134 ± 55
0.1	$2 \cdot 10^5$	$\mathbf{120 \pm 43}$	166 ± 53	129 ± 118	139 ± 55

G.4 Effect of the hard arm's gap Δ

We next vary the hard-arm gap $\Delta \in \{0.01, 0.02, 0.05, 0.1\}$ while keeping $K = 10$ and the easy-arm structure fixed. Table 5 reports cumulative regret at $T = 5 \cdot 10^4$ and $T = 2 \cdot 10^5$, and Figure 3 visualizes the regret at $T = 2 \cdot 10^5$. As Δ decreases, the saturation level of D-Exp3-Ada increases (consistent with the $1/\Delta$ dependence in the bound). At $T = 2 \cdot 10^5$, D-Exp3-Ada outperforms Tsallis-INF-$2/3$ for all four values of Δ, with the relative advantage becoming larger as Δ decreases. For $\Delta = 0.1$, the gap between D-Exp3-Ada and Tsallis-INF-$2/3$ is small (120 vs 129), since the saturation levels of the two bounds are close in magnitude when Δ is large. D-TINF-SPM saturates for all four values of Δ at a higher level than D-Exp3-Ada; at $T = 2 \cdot 10^5$ it is below Tsallis-INF-$2/3$ for $\Delta \leq 0.02$ but not for $\Delta \in \{0.05, 0.1\}$, where all algorithms are already close to saturation. This is consistent with the asymptotic nature of the gap-dependent advantage: the proposed algorithms are most beneficial when $T \gg \mathrm{const} \cdot K/\Delta^2$.

G.5 Multiple hard arms

We also examine the case where there are multiple hard arms: $K = 10$, three hard arms with gaps $\Delta \in \{0.01, 0.02, 0.05\}$, six easy arms with gap 0.5, and one optimal arm. In this regime, $C_1 = \sum_{i \neq i^\star} 1/\Delta_i = 1/0.01 + 1/0.02 + 1/0.05 + 6/0.5 = 182$ is much larger than in Section G.2, so the saturation level of D-Exp3-Ada and D-TINF-SPM is correspondingly higher. Nonetheless, D-Exp3-Ada still overtakes Tsallis-INF-$2/3$ at $T \approx 5 \cdot 10^4$ and remains the best at $T = 2 \cdot 10^5$ (Table 6), whereas the regret of D-TINF-SPM stays above that of Tsallis-INF-$2/3$ over the evaluated horizons. The advantage is smaller than in the single-hard-arm setting (Section G.2), consistent with the theory: gap-dependent bounds are most advantageous when the number of hard arms is small.

G.6 Effect of multiple optimal arms

To examine the behavior of our algorithms when the optimal arm is not unique (an assumption of our stochastic upper bounds; see Remark F.4), we vary the number of optimal arms $m^\star \in \{1, 2, 3\}$ while fixing $K = 10$. All losses are Bernoulli, and the optimal arms have mean loss 0.3. There are three hard suboptimal arms with gap 0.1, while the remaining $K - m^\star - 3$ arms have gap 0.4. We use deterministic tie-breaking when selecting $i_t^\star$ (the smallest index). Each instance is evaluated over 20 runs, each run being a single trajectory of length 10^5 recorded at the four horizons. In this experiment, we report the expected pseudo-regret $\sum_{t=1}^T \langle p_t, \Delta \rangle$ (actions are still sampled, and only the sampled exploration arm is used for the estimator), which has lower variance than $\sum_{t=1}^T \Delta_{A_t}$. Table 7 reports this cumulative regret at four horizons, together with the number of

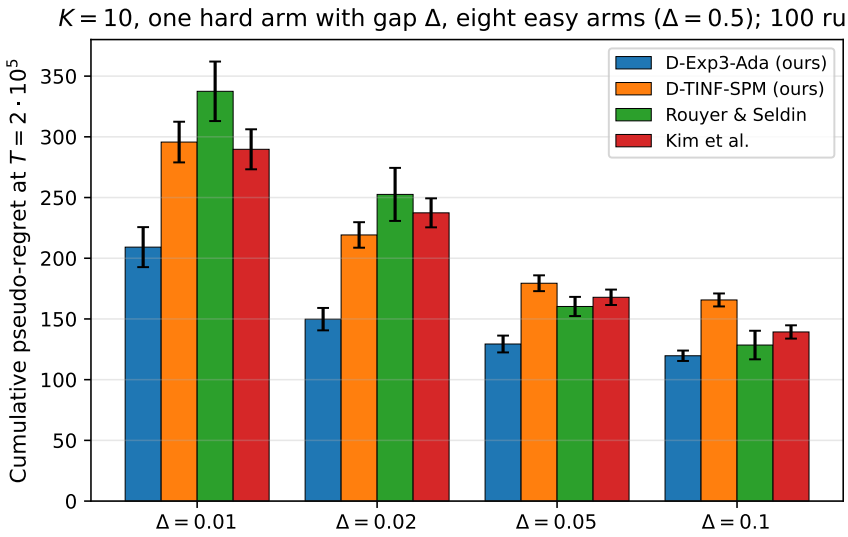


Figure 3: Cumulative regret at $T = 2 \cdot 10^5$ for varying hard-arm gap Δ, with $K = 10$ and eight easy arms ($\Delta = 0.5$). Error bars show one standard error over 100 runs. D-Exp3-Ada wins for all four Δ values, with the largest advantage at small Δ.

Table 6: Cumulative regret in a stochastic environment with three hard arms ($\Delta \in \{0.01, 0.02, 0.05\}$), six easy arms (gap 0.5), and $K = 10$ arms total. Mean $\pm$ standard deviation over 100 runs.

T	D-Exp3-Ada (ours)	D-TINF-SPM (ours)	Tsallis-INF-2/3	Kim et al.
$5 \cdot 10^3$	116 ± 32	157 ± 39	$\mathbf{100 \pm 32}$	109 ± 30
$2 \cdot 10^4$	211 ± 81	269 ± 92	$\mathbf{202 \pm 79}$	221 ± 78
$5 \cdot 10^4$	$\mathbf{304 \pm 155}$	381 ± 169	317 ± 148	343 ± 150
10^5	$\mathbf{381 \pm 259}$	476 ± 277	421 ± 240	454 ± 234
$2 \cdot 10^5$	$\mathbf{436 \pm 375}$	552 ± 387	512 ± 319	556 ± 290

rounds $t \in (T/2, T]$ at $T = 10^5$ in which $i_t^\star$ differs from $i_{t-1}^\star$, averaged over runs. Although with ties the maximizer $i_t^\star$ keeps changing (last column), D-Exp3-Ada's regret is exactly as flat as in the unique case, and in fact lower. D-TINF-SPM remains stable and far below the minimax adversarial rate, with a slow residual drift whose increments decelerate and stay roughly constant per doubling of T, consistent with at most logarithmic growth in T.

G.7 Robustness in adversarial environments

To probe the Best-of-Both-Worlds property of our algorithms beyond the stationary stochastic regime, we evaluate them on a non-stationary environment where the identity of the optimal arm changes during the horizon. Strictly speaking, the theoretical adversarial bound of order $\sqrt{KT}$ in this paper is against the best fixed arm in hindsight, while the experiment below reports a dynamic-regret-style quantity that uses a stronger comparator than the fixed-arm one; we use it as a more challenging stress test, and we also derive from it the regret against the best fixed arm.

Setting. Bernoulli bandit with $K = 10$ arms. The mean losses are $\mu_1 = 0.4$ and $\mu_i = 0.5$ for $i \neq 1$ during $t \in [1, T/2]$ (so arm 1 is optimal), then $\mu_2 = 0.4$ and $\mu_i = 0.5$ for $i \neq 2$ during $t \in (T/2, T]$ (so arm 2 is optimal). The reported regret is computed against the best arm at each round (dynamic regret). Each $T \in \{5 \cdot 10^3, 10^4, 5 \cdot 10^4, 10^5\}$ is averaged over 100 independent runs, each of which is a separate simulation of length T with the switch at $T/2$.

Results. Table 8 reports the cumulative dynamic regret in this switching environment. At every horizon, all four algorithms incur regret close to $\Delta \cdot T/2 = 0.05\,T$ (with $\Delta = 0.1$), and our D-Exp3-Ada and D-TINF-SPM are within a factor of 1.25 of the best baseline (Tsallis-INF-2/3). This

Table 7: Effect of the number of optimal arms $m^\star$ on cumulative expected pseudo-regret with $K = 10$. The optimal arms have mean loss 0.3, three suboptimal arms have gap 0.1, and the remaining arms have gap 0.4. Mean $\pm$ standard deviation over 20 runs; the last column is the mean number of changes of $i_t^\star$ over $t \in (5 \cdot 10^4, 10^5]$.

Algorithm	$m^\star$	$T = 10^4$	$T = 3 \cdot 10^4$	$T = 6 \cdot 10^4$	$T = 10^5$	changes of $i_t^\star$
D-Exp3-Ada	1	142 ± 41	142 ± 41	142 ± 41	142 ± 41	0
D-Exp3-Ada	2	119 ± 49	120 ± 50	121 ± 49	122 ± 49	36.4
D-Exp3-Ada	3	111 ± 36	114 ± 36	115 ± 36	116 ± 36	69.5
D-TINF-SPM	1	173 ± 45	179 ± 47	181 ± 48	181 ± 48	0
D-TINF-SPM	2	169 ± 44	196 ± 53	212 ± 56	223 ± 58	37.0
D-TINF-SPM	3	150 ± 27	186 ± 33	209 ± 35	223 ± 37	75.0

Table 8: Cumulative dynamic regret in a switching adversarial environment with $K = 10$ arms (loss gap 0.1), where the optimal arm switches at $t = T/2$; each T is a separate simulation of length T. Mean $\pm$ standard deviation over 100 runs. The regret against the best fixed arm in hindsight is obtained by subtracting $0.05\,T$.

T	D-Exp3-Ada (ours)	D-TINF-SPM (ours)	Tsallis-INF-$2/3$	Kim et al.
$5 \cdot 10^3$	390 ± 43	376 ± 48	$\mathbf{342 \pm 61}$	347 ± 56
10^4	719 ± 96	674 ± 95	$\mathbf{586 \pm 122}$	$\mathbf{586 \pm 113}$
$5 \cdot 10^4$	2559 ± 292	2629 ± 293	$\mathbf{2242 \pm 497}$	2450 ± 376
10^5	4818 ± 490	5007 ± 429	$\mathbf{4325 \pm 961}$	4760 ± 580

reflects a structural property of the experiment rather than a failure mode of particular algorithms. At the end of Phase 1, the expected cumulative loss of the previously optimal arm is smaller than that of the new optimal arm by $\Delta \cdot T/2$; in Phase 2 this difference shrinks at rate Δ per round and vanishes only at $t = T$. Algorithms whose exploitation distribution is determined by cumulative (estimated) losses, such as all four algorithms here, therefore keep favouring the previously optimal arm during most of Phase 2, and their dynamic regret is of order $\Delta \cdot T/2$, i.e., linear in T, regardless of their adversarial guarantees. Against the best fixed arm in hindsight, which is the comparator in our $\sqrt{KT}$ bounds, the picture is different. Since arms 1 and 2 have the same expected cumulative loss $0.45\,T$, the regret against the best fixed arm equals the reported dynamic regret minus exactly $\Delta \cdot T/2$. For $T = 5 \cdot 10^3, 10^4, 5 \cdot 10^4, 10^5$, it is at most 140, 219, 129, and 7, respectively, for our algorithms (and negative in some cases, since an algorithm that partially switches to the new optimal arm can outperform the best fixed arm), well below $\sqrt{KT} \approx 224$, 316, 707, and 1000. Thus, on this non-stationary stress test, our algorithms are within a small constant factor of the existing decoupled bandit baselines in dynamic regret and incur small regret against the best fixed arm, while in stochastic regimes (Section G.2) they exhibit gap-dependent saturation. This combination is the empirical counterpart of the Best-of-Both-Worlds guarantee proven in Theorems 5.3 and 5.4.

G.8 Summary

The empirical results support the theoretical findings of Section 5 and the lower bound (Corollary 6.3). In the stochastic settings we examined, our D-Exp3-Ada and D-TINF-SPM exhibit regret saturation as T grows (in the multiple-hard-arm setting, whose smallest gap is 0.01, saturation is not yet complete at $T = 2 \cdot 10^5$), while the regrets of the existing algorithms keep growing over the evaluated horizons when the hard-arm gap is small. D-Exp3-Ada outperforms the existing algorithms in all stochastic settings once T is moderately large. D-TINF-SPM saturates at a higher level than D-Exp3-Ada; it outperforms the existing algorithms at long horizons in the single-hard-arm settings with $\Delta = 0.02$ and $K \leq 10$, but not uniformly across the settings. The advantage is more pronounced when the hard-arm gap is small and when the time horizon is long, while the crossover point shifts to larger T as K grows. Even in the most challenging stochastic setting we studied (multiple hard arms with a wide range of gaps), D-Exp3-Ada eventually overtakes Tsallis-INF-$2/3$ although the crossover requires larger T. In adversarial / non-stationary environments, all algorithms incur dynamic regret of the same order, and our algorithms incur small regret against the best fixed arm, supporting the Best-of-Both-Worlds nature of the proposed methods.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and the contribution paragraphs of the introduction (Section 1, with Table 1) explicitly state our contributions (two BoBW algorithms with their regret bounds, instance-dependent lower bounds, and the resulting near-optimality up to $\mathcal{O}((\log K)^2)$), and they correspond to Theorems 5.3, 5.4, 6.1, 6.2 and Corollary 6.3. The unique-optimal-arm assumption underlying the stochastic upper bounds is stated in the abstract and introduction, and the scope of the lower bound is qualified by the natural monotonicity and permutation-invariance assumptions on regret upper bounds.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the remaining $\mathcal{O}((\log K)^2)$ gap between our upper and lower bounds (Section 7), the structural assumptions (monotonicity, permutation invariance) under which the lower bound applies (Corollary 6.3 and the discussion that follows), the scope of the hard environment in the lower bounds (Remark 6.4), and the unique-optimal-arm assumption of the stochastic upper bounds, including where it enters the analysis and experiments without it (Remark F.4 and Appendix G.6). Empirically, the saturation level of D-TINF-SPM is higher than that of D-Exp3-Ada and is not uniformly below the baselines at the evaluated horizons (Appendix G). Regarding computational cost, D-Exp3-Ada has a closed-form update, whereas D-TINF-SPM requires solving the Tsallis-entropy FTRL problem by Newton's method in each round (Section 4).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.

- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: All theorems and lemmas are numbered and cross-referenced; the assumptions are stated explicitly (e.g., the unique optimal arm for the stochastic upper bounds in Section 3, $\mu \in [c, 1-c]$ in Theorems 6.1 and 6.2, monotonicity and permutation invariance in Corollary 6.3, and the four conditions of the SPM lemma, Lemma A.2, each verified for D-TINF-SPM). The main body provides proof sketches for the key results (Theorems 5.3 and 5.4); full proofs and supporting lemmas are deferred to Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experiment section (Appendix G) specifies the full setup: Bernoulli losses, the gap structures, K, the horizons T, the number of independent runs, how runs are recorded (a single trajectory per seed in the stochastic experiments, and a separate simulation for each T with the switch at $T/2$ in the switching experiment), the performance metric, and all four algorithms with their hyperparameters ($\beta_1 = 16$ for D-TINF-SPM, $\eta_t \propto K^{-1/6}/\sqrt{t}$ for Tsallis-INF-$2/3$, and Pareto shape 3 for the FTPL baseline). Code with fixed random seeds and per-run raw regret data are provided as supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often

one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.

- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: A code-and-data archive is included as supplementary material. It contains the algorithm implementations, the experiment driver, the raw per-run regret data underlying every reported number and figure, and a README describing the directory layout and the exact commands to reproduce all results. No external datasets are needed.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

Justification: For each experiment we report the gap structure, K, the set of horizons T, the number of independent runs, and the algorithm-specific hyperparameters; the choices of these hyperparameters follow directly from the theoretical specifications in Section 4 for our algorithms and from the original papers for the baselines.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: All tables report mean $\pm$ standard deviation across independent runs (100 runs in the five main experiments and 20 runs per configuration in the multiple-optimal-arms experiment). The shaded bands and error bars in the figures denote one standard error of the mean. The variability captured is the algorithm/environment randomness (the seeds for the algorithm's internal randomization and the Bernoulli loss draws are independent across runs).

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: All experiments run on a single CPU machine; the five main experiments (100 runs each, time horizons up to 5×10^5) complete in approximately 3 hours when parallelized over 8 CPU cores, and the multiple-optimal-arms experiment (20 runs per configuration, horizon 10^5) takes about 15 minutes on a single core. The supplementary code uses Python's standard `concurrent.futures` for parallelization and requires no GPU.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This is foundational theoretical work on multi-armed bandits with synthetic Bernoulli simulations only; it involves no human subjects, personal data, or sensitive content, and conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: The paper studies a foundational online-learning problem; the contributions are theoretical regret bounds and synthetic numerical experiments, with no direct path to deployments that would have specific positive or negative societal impacts beyond the general benefits of improved decision-making algorithms.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper releases neither pretrained models nor scraped data; it releases only algorithm implementations and synthetic experiment outputs, which pose no misuse risk.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [N/A]

Justification: We do not use any external datasets or pretrained models. The baseline algorithms we re-implement (Tsallis-INF of Rouyer and Seldin [2020] and the FTPL-based algorithm of Kim et al. [2026]) are cited at every appearance.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The released code and synthetic experiment outputs are documented in a README that maps each table and figure in the paper to the source data file and the code that produced it, and includes the exact commands to reproduce all experiments.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.

- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or human-subjects research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects, so no IRB review is applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not a component of the methods studied in this paper. They were used as assistants for drafting and editing text and for implementing and running the numerical experiments with the proposed and baseline algorithms; the authors reviewed all such outputs and take full responsibility for the content of the paper.

    Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.