# Looking Through the Mirror: Minimax-Optimal Regularized Regrets in Online Learning and Bandits

**Junghyun Lee, Yujun Kim, Chulhee Yun, Se-Young Yun**
Kim Jaechul Graduate School of AI, KAIST
Seoul 02455, Republic of Korea
{jh_lee00, kyujun02, chulhee.yun, yunseyoung}@kaist.ac.kr

## Abstract

We revisit regularized regret minimization under full-information and bandit feedback, where a learner optimizes an objective of the form $\langle r, \pi \rangle - \eta^{-1}\psi(\pi)$ for adversarially chosen rewards $r$, policies $\pi \in \Delta_K$, and a known strongly convex regularizer $\psi(\cdot)$. We establish nearly tight lower and upper bounds on the minimax regularized regret for a broad class of regularizers, specifically those that are relatively strongly-convex and smooth with exponent $p \geq 1$, Legendre, and coordinate-wise separable. For sufficiently large $\eta$, the problem approaches unregularized online learning, recovering the standard $\Theta(\sqrt{T})$ regret; for small $\eta$, the strong convexity of the regularizer dominates, and we identify the regimes that attain $\Theta(\eta \log T)$ regret. The dependence on the number of arms $K$ reflects both the feedback model and the regularizer's geometry, which we nearly match as well. Notably, while prior logarithmic regret lower bounds are largely restricted to specific curved loss functions (e.g., squared loss), our lower bounds hold for ***any*** such regularizer against purely linear adversarial rewards. A key technical contribution is our proof strategy, which leverages a Bregman geometric perspective of regularized regret over the quotient space $\mathbb{R}^K/\mathrm{span}(\{\mathbf{1}\})$ to resolve the translation invariance of the inverse mirror map. By combining this dual geometric formulation with the ***van Trees inequality***, we reduce online learning to Bayesian estimation: a unified framework for establishing logarithmic regret lower bounds.

## 1 Introduction

In sequential decision making with finite actions (Cesa-Bianchi and Lugosi, 2006; Lattimore and Szepesvári, 2020), a learner chooses a policy $\pi \in \Pi = \Delta_K$ over $K$ actions at each step. The expected performance of the learner is measured through $\mathbb{E}_{a\sim\pi}[r_a] = \langle r, \pi \rangle$, where $r = (r_a)_{a\in[K]}$ is a reward vector. In many cases, however, it is often highly desirable to optimize a *regularized* objective of the form $J_\eta(\pi; r) \coloneqq \langle r, \pi \rangle - \eta^{-1}\psi(\pi)$, for some fixed strongly convex regularizer $\psi(\cdot)$ and the regularization coefficient $\eta^{-1}$. For example, regularization in RLHF and LLM alignment (Bai et al., 2022; Casper et al., 2023; Christiano et al., 2017; Gao et al., 2023; Ouyang et al., 2022; Rafailov et al., 2023) is used to control deviation from a reference policy, encourage diversity, and in some cases, prevent overoptimization.

Given this context, the learner sequentially interacts with the environment through the following protocol at each timestep $t = 1, 2, \cdots, T$:

1. The learner chooses a $\pi_t \in \Pi$ and samples $i_t \sim \pi_t$.
2. The environment uses $r_t = (r_{t,i})_{i\in[K]}$ from a reward sequence fixed before the interaction (an oblivious adversary).
3. The learner incurs $J_\eta(\pi_t; r_t)$, and observes either *(i)* $r_t$ or *(ii)* $r_{t,i_t}$.

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

We consider two types of feedback: *(i)* full-information feedback or *(ii)* bandit feedback. In either case, we measure the learner's performance via ***regularized regret*** (which we will refer to as "regret" hereafter) with respect to the best policy in hindsight:

$$\mathrm{Reg}_\eta(T;\{r_t\}) \coloneqq \sum_{t=1}^{T}\left\{J_\eta(\pi^\star(\{r_t\});r_t) - J_\eta(\pi_t;r_t)\right\}, \quad \pi^\star(\{r_t\}) \coloneqq \arg\max_{\pi\in\Pi}\sum_{t=1}^{T} J_\eta(\pi;r_t).$$

When $\eta$ is small, the strongly convex regularizer dominates, and one may hope for $\log T$ regret (Duchi et al., 2010; Hazan et al., 2007); when $\eta$ is large, the problem approaches unregularized online learning, where $\sqrt{T}$ regret is unavoidable (Auer et al., 2002; Cesa-Bianchi and Lugosi, 2006).

**Examples of $\psi$.** We present several representative regularizers that motivate the setting studied in this paper. The "reverse" KL-divergence, $\psi(\cdot) = D_{\mathrm{KL}}(\cdot,\rho)$ for a fixed $\rho \in \mathrm{relint}(\Pi)$, has established itself as the standard regularizer in recent theoretical analyses of online preference learning and RLHF (Cen et al., 2025; Jacob et al., 2022; Ji et al., 2026; Tiapkin et al., 2023, 2024; Wu et al., 2025a,b; Xie et al., 2025; Xiong et al., 2024; Zhao et al., 2025a,b). Recently, a wide variety of strongly convex regularizers beyond reverse KL have also been studied with different motivations: negative Shannon entropy is central to maximum-entropy exploration (Haarnoja et al., 2018; Neu et al., 2017; Ziebart et al., 2008), mixtures of reverse KL divergences lead to more diverse and less biased alignment (Aminian et al., 2025; Le et al., 2025), chi-squared divergence provably avoids overoptimization (Huang et al., 2025), Tsallis entropy is a key component for obtaining best-of-both-world guarantees for stochastic and adversarial bandits (Kato and Ito, 2025; Zimmert and Seldin, 2019, 2021), and Burg entropy is foundational to barrier-based interior-point methods (Abernethy et al., 2012). Broader families, such as $\alpha$-Rényi entropies (Zhang et al., 2026), Csiszár $f$-divergences (Go et al., 2023; Han et al., 2025; Wang et al., 2024a; Zhao et al., 2026), expected convex (Yang et al., 2019), and strongly convex regularizers (Geist et al., 2019), further broaden the scope of applicable information geometries.

**Known results and gaps.** Under the full-information feedback, optimizing the objective consisting of an adversarial component $\langle r_t, \pi_t\rangle$ and a fixed, known regularizer component $\psi(\pi_t)$ has been extensively studied under the name *composite online optimization* in the online learning and optimization communities (Duchi and Singer, 2009; Duchi et al., 2011, 2010; Joulani et al., 2020; McMahan, 2010, 2017; Wang et al., 2024b; Xiao, 2010; Yuan et al., 2021). While $\mathcal{O}(\log T)$ upper bounds are well-established for a wide class of strongly convex regularizers $\psi$ (e.g., via `COMID` of Duchi et al. (2010)), a corresponding $\Omega(\log T)$ lower bound for ***each fixed*** choice of $\psi(\cdot)$ remains open.

Classical logarithmic lower bounds address losses whose curvature is supplied by the adversary. For instance, Abernethy et al. (2008) consider *Strongly Convex Games* where the adversary plays strongly convex loss functions; Abernethy et al. (2009); Gaillard et al. (2019); Takimoto and Warmuth (2000); Vovk (2001) analyze the squared loss in online linear regression; and Ordentlich and Cover (1998) study the exp-concave log-loss in universal portfolios. In standard online linear optimization where both the adversary and the learner operate with purely linear losses, an $\Omega(\sqrt{T})$ regret lower bound is unavoidable (Auer et al., 2002; Cesa-Bianchi and Lugosi, 2006). In our setting, the adversary chooses reward vectors $r_t \in \mathbb{R}^K$, which induce the linear maps $\pi \mapsto \langle r_t, \pi\rangle$; the curvature comes from the *fixed*, known regularizer $\psi(\cdot)$. While none of the aforementioned $\Omega(\log T)$ bounds operate strictly within a composite optimization framework, they serve as the closest available theoretical analogues.

Under bandit feedback, recent analyses of regularized regret have primarily studied the reverse KL-regularizer (Ji et al., 2026; Tiapkin et al., 2023, 2024; Xiong et al., 2024; Zhao et al., 2025a,b) in stochastic environments, e.g., when $r_t$ is chosen in an $i.i.d.$ manner. Quite recently, Ji et al. (2026) established a high-probability $\widetilde{\mathcal{O}}(\eta K(\log T)^2)$ regret upper bound and an in-expectation regret lower bound of $\Omega(\eta K \log T)$ for the uniform reference policy, in the strongly regularized regime. For contextual bandits, Wang et al. (2026) further establish polylogarithmic regret for greedy sampling under reward and preference feedback without explicit eluder-dimension dependence. The KL-specific analyses exploit the Gibbs form of the optimal policy (Ji et al., 2026; Wu et al., 2025a; Zhao et al., 2025a,b). More recent works extend fast-rate guarantees beyond KL, but studies different learning protocols or lower-bound quantifiers. Such results include online RLHF under generalized bilinear preferences (Lee et al., 2026), online RLHF with $f$-divergence regularization (Wu et al., 2026), and offline contextual bandits (Zhao et al., 2026). In particular, Zhao et al. (2026) establish offline sample-complexity upper and lower bounds, using a dual-Bregman argument for strongly

convex $f$ that does not require a closed-form optimal policy. These results do not characterize minimax regularized regret against adversarial reward sequences for each fixed regularizer, under both full-information and bandit feedback. Our goal is to provide such a characterization through the curvature of $\psi$, without requiring an explicit formula for its optimal policy.

**$\psi$-minimax regularized regret.** We first define ***$\psi$-minimax regularized regret*** as follows:

$$\mathfrak{R}_\eta(T;\psi) \coloneqq \inf_{\mathcal{A}} \sup_{r_1,\cdots,r_T\subset[-R,R]^K} \mathbb{E}^{\mathcal{A}}\left[\mathrm{Reg}_\eta(T;\{r_t\})\right],$$

where $[-R,R]^K$ is the adversary's domain. Note that the regularizer $\psi(\cdot)$ is *fixed* before taking the infimum and supremum, i.e., the minimax regret is defined *per* regularizer. We write $\mathfrak{R}_\eta^{\mathrm{FI}}$ and $\mathfrak{R}_\eta^{\mathrm{B}}$ when distinguishing the two feedback models (full-information and bandit, respectively); otherwise, the feedback model is understood from context. For instance, the logarithmic lower bound of Ji et al. (2026) has scale $\Omega(\eta K\log(T/(\eta^2K)))$ in its stated regime and uses $D_{\mathrm{KL}}(\cdot,\mathrm{Unif}([K]))$ in a stochastic sub-Gaussian-noise model. It does not provide a lower bound for $D_{\mathrm{KL}}(\cdot,\rho)$ for other choices of $\rho$, nor a lower bound for other choices of $\psi(\cdot)$ in general. Considering the aforementioned long line of work motivating non-KL regularizers, this constitutes a significant gap in the literature.

This motivates the central question of the paper:

*Can we characterize $\mathfrak{R}_\eta(T;\psi)$ under full-information and bandit feedback, for **any** regularizer $\psi$?*

**Contributions.** In this work, we take a step towards settling the above question by developing a mirror-geometric framework of regularized regret. We consider regularizers that are relatively strongly convex/smooth (Assumption 2), Legendre (Assumption 3), and coordinate-wise separable (Assumption 4). Such assumptions are satisfied by various regularizers, as shown later in Section 5.

- **Full-info.** We establish an $\Omega(\eta R^2\log T)$ lower bound for ***any*** $\psi$ within this class, with $K,R,\eta,\psi$ fixed as $T$ grows (**Theorem 3.1**). The proof combines quotient Bregman geometry with the ***van Trees inequality*** (Gill and Levit, 1995; van Trees, 1968), which highlights our *mirror-geometric perspective*. Together with **C-FTRL** (**Theorem 3.2**), this gives matching logarithmic dependence on $T$ and $\Theta(R\sqrt{T\log K})$ regret under the stated large-$\eta$ condition.
- **Bandit.** We extend the same geometric reduction to bandit feedback by accounting for the restricted Fisher information (**Theorem 4.1**). Combined with **IW-C-FTRL** (**Theorem 4.2**), this again gives matching logarithmic dependence on $T$ for fixed parameters and $\Theta(R\sqrt{KT})$ regret under the corresponding large-$\eta$ condition. The effective dimensions (Eqn. (2)) describe the additional cost of observing only one coordinate per round.
- By instantiating our bounds for Shannon, Tsallis, and BIS entropy, we obtain matching polynomial dependence on $K$ in the small-$\eta$ regime (**Section 5**). These results complement prior work on online composite optimization (Duchi et al., 2010) and KL-regularized bandits (Ji et al., 2026; Zhao et al., 2025b).

## 2 Preliminaries

**Assumptions.** We first describe the assumptions that we impose throughout. We emphasize beforehand that such assumptions are satisfied by many common regularizers, as we show in **Section 5**.

First is the boundedness of the reward vectors, standard in the online learning literature (Cesa-Bianchi and Lugosi, 2006; Lattimore and Szepesvári, 2020):

**Assumption 1** (Bounded Rewards)**.** *For all $t\ge 1$, $r_t\in[-R,R]^K$ for some known $R>0$.*

We now describe the assumptions on the regularizer $\psi:\Pi\to\mathbb{R}\cup\{+\infty\}$ with $\Pi=\Delta_K$ for $K\ge 2$. Allowing $+\infty$ on the boundary includes Burg entropy; the regularizer is finite on $\mathrm{relint}(\Pi)$. Adding the same constant to every reward coordinate leaves the optimal policy unchanged. The following spaces separate this unidentifiable reward direction from feasible policy variations. We define the ***quotient space*** $\mathcal{Q}$, ***tangent space*** $\mathcal{T}$, and ***projection onto*** $\mathcal{T}$ $P_{\mathcal{T}}$ as

$$\mathcal{Q}\coloneqq\mathbb{R}^K/\mathrm{span}(\{\mathbf{1}\}),\quad \mathcal{T}\coloneqq\{v\in\mathbb{R}^K:\langle v,\mathbf{1}\rangle=0\}=\mathbf{1}^\perp,\quad P_{\mathcal{T}}\coloneqq \boldsymbol{I}_K-\frac{1}{K}\mathbf{1}\mathbf{1}^\top. \tag{1}$$

Derivatives and Hessians are understood relative to the affine hull of $\Pi$, $\mathrm{aff}(\Pi)$. In particular, the Löwner inequalities below are required only along directions in $\mathcal{T}$: $A \succeq B$ means $v^\top A v \geq v^\top B v$ for every $v \in \mathcal{T}$. We describe in detail the dual geometry in Appendix A.

Inspired by the notions of relative smoothness (Lu et al., 2018) and differential consistency of Abernethy et al. (2015, Definition 2.2), we consider the following convex-smooth assumption on $\psi$:

**Assumption 2** (Relative Strong Convexity/Relative Smoothness)**.** *The regularizer $\psi(\cdot)$ is proper, lower-semicontinuous, and is twice continuously differentiable on* $\mathrm{relint}(\Pi)$. *Also, there exist $p \geq 1$ and $0 < c_\psi \leq C_\psi$ such that $\psi$ is*

(i) $(p, c_\psi)$***-relatively strongly convex (RSC)****:* $\nabla^2\psi(\pi) \succeq c_\psi \mathrm{diag}(\pi)^{-p}$, $\forall \pi \in \mathrm{relint}(\Pi)$;

(ii) $(p, C_\psi)$***-relatively smooth (RS)****:*[a] $\nabla^2\psi(\pi) \preceq C_\psi \mathrm{diag}(\pi)^{-p}$, $\forall \pi \in \mathrm{relint}(\Pi)$.

[a] Notation-wise, $\gamma$ and $\tilde{\Psi}$ in Abernethy et al. (2015) correspond to our $p$ and $\psi^*$, respectively.

We also assume that the regularizer is Legendre, a standard and useful assumption in online learning for utilizing potential-based analysis (Cesa-Bianchi and Lugosi, 2006, Chapter 11):

**Assumption 3** (Legendre)**.** *The regularizer $\psi(\cdot)$ is* ***Legendre*** *(Rockafellar, 1997, Theorem 26.1), i.e., strictly convex and essentially smooth relative to* $\mathrm{aff}(\Pi)$, *with* $\lim_{n\to\infty} \|P_{\mathcal{T}} \nabla\psi(\pi_n)\| = \infty$ *for any sequence* $(\pi_n)_{n\geq 1} \subset \mathrm{relint}(\Pi)$ *with* $\lim_{n\to\infty} \pi_n \in \partial_{\mathrm{rel}}\Pi \triangleq \Pi \setminus \mathrm{relint}(\Pi)$.

Lastly, we assume coordinate-wise separability:

**Assumption 4** (Separability)**.** *The regularizer satisfies $\psi(\pi) = \sum_{i=1}^K \phi_i(\pi_i)$, where $\phi_i \in C^2((0,1))$ and $c_\psi x^{-p} \leq \phi_i''(x) \leq C_\psi x^{-p}$, $\forall i \in [K], \forall x \in (0,1)$.*

**Mirror Geometry.** Our viewpoint exploits the classical ***mirror duality*** between reward vectors and policies induced by the regularizer $\psi$ (Beck and Teboulle, 2003; Rockafellar, 1997). This relation also underlies regularized policy optimization (Geist et al., 2019). Indeed, the maximum regularized objective is a scaled ***Fenchel conjugate*** of $\psi$:

$$r_t \mapsto \max_{\pi\in\Pi} \left\{\langle r_t, \pi\rangle - \eta^{-1}\psi(\pi)\right\} = \eta^{-1} \left\{\max_{\pi\in\Pi} \left\{\langle \eta r_t, \pi\rangle - \psi(\pi)\right\}\right\} \triangleq \eta^{-1}\psi^*(\eta r_t).$$

As we consider bounded reward vectors, we consider the following subset of $\mathrm{relint}(\Pi)$:

**Definition 2.1.** *We define the* ***inverse mirror map*** $\nabla\psi^* : \mathbb{R}^K \to \mathrm{relint}(\Pi)$ *as*

$$\nabla\psi^*(u) \coloneqq \arg\max_{\pi\in\Pi} \{\langle u, \pi\rangle - \psi(\pi)\}, \quad u \in \mathbb{R}^K$$

*For each $U \geq 0$, the* ***dual-constrained policy class*** *is defined as*

$$\Pi_{U,\psi} \coloneqq \left\{\nabla\psi^*(\tilde{r}) : \tilde{r} \in [-U, U]^K\right\} \subset \mathrm{relint}(\Pi),$$

***dual-diameter*** *as $M_\psi(U) \coloneqq \max_{\pi\in\Pi_{U,\psi}} \psi(\pi) - \min_{\pi\in\Pi} \psi(\pi)$, and denote $M_\psi \coloneqq M_\psi(\eta R)$.*

Since $\nabla\psi^*(u + c\mathbf{1}) = \nabla\psi^*(u)$, the induced policy depends only on reward differences; this is where $\mathcal{Q}$ and $\mathcal{T}$ in Eqn. (1) play a critical role.

We also recall the Bregman divergence:

**Definition 2.2** (Bregman divergence, Bregman (1967))**.** *The* ***Bregman divergence*** *generated by $\psi$, $B_\psi : \Pi \times \mathrm{relint}(\Pi) \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$, is defined as follows:*

$$B_\psi(\pi, \pi') \coloneqq \psi(\pi) - \psi(\pi') - \langle \nabla\psi(\pi'), \pi - \pi'\rangle, \quad \pi \in \Pi,\ \pi' \in \mathrm{relint}(\Pi).$$

Rewards differing by a constant induce the same policy, so dual metrics must ignore this constant direction. The matrix $L_p(\pi)$ below measures the remaining reward differences near a policy $\pi$. Its two effective dimensions, $d_{\mathrm{FI}}(\pi)$ and $d_{\mathrm{B}}(\pi)$, quantify the local estimation difficulty under full-information and bandit feedback, respectively. Writing $\pi^p = (\pi_i^p)_{i\in[K]}$, we define

$$L_p(\pi) \coloneqq \mathrm{diag}(\pi^p) - \frac{\pi^p(\pi^p)^\top}{\sum_i \pi_i^p}, \quad d_{\mathrm{FI}}(\pi) \coloneqq \mathrm{tr}(L_p(\pi)), \quad d_{\mathrm{B}}(\pi) \coloneqq \left(\sum_i \sqrt{(L_p(\pi))_{ii}}\right)^2.$$

The corresponding uppercase quantities, referred to as ***effective dimensions***, select the largest local difficulty among policies reachable at a given dual radius $U \geq 0$:

$$D_{\mathrm{FI}}(U;\psi) \coloneqq \max_{\pi\in\Pi_{U,\psi}} d_{\mathrm{FI}}(\pi), \quad D_{\mathrm{B}}(U;\psi) \coloneqq \max_{\pi\in\Pi_{U,\psi}} d_{\mathrm{B}}(\pi). \tag{2}$$

Lastly, we define the following function, which measures the average growth of the Fenchel conjugate $\psi^*$ along the $K$ coordinate directions and will appear in our $\Omega(\sqrt{T})$ lower bounds:

$$F_\psi(x) \coloneqq \frac{1}{K}\sum_{i=1}^{K}\{\psi^*(xe_i) - \psi^*(0)\}, \quad x \geq 0.$$

## 3 Regularized Online Learning with Full-Information Feedback

We first consider full-information feedback, where the geometric reduction underlying both feedback models is most transparent. The bandit proof in Section 4 follows from similar reasoning.

### 3.1 Minimax-Optimal Regularized Regret Bounds

We first establish a minimax lower bound for the regularized regret, for *each* $\psi(\cdot)$:

**Theorem 3.1** (Regret Lower Bound, Full-Information)**.** *Suppose Assumptions 1 to 4 hold. For every $T \geq 1$ and $\eta > 0$,*

$$\mathfrak{R}_\eta^{\mathrm{FI}}(T;\psi) \geq \frac{7\exp(-2p/c_\psi)}{32C_\psi}\eta R^2 D_{\mathrm{FI}}(\eta R/2;\psi)\log\left(1 + \frac{T}{7\pi^2\max\{1,(\eta R/4)^2\}}\right).$$

*Furthermore suppose that $T \geq \log K$ and $F_\psi\left((\eta R/4)\sqrt{\log K/T}\right) \geq \frac{7\eta R}{32}\sqrt{\log K/T}$. Then,*

$$\mathfrak{R}_\eta^{\mathrm{FI}}(T;\psi) \geq \frac{1}{32}R\sqrt{T\log K}.$$

*Proof Sketch.* The full proof is provided in Appendix B. We focus here on the logarithmic lower bound. Using Yao's minimax principle (Yao, 1977), we construct a Bayesian adversary and use mirror geometry to reduce regret to a sequential estimation problem. We then lower-bound the resulting estimation risk using the ***matrix van Trees inequality*** (Gassiat and Stoltz, 2024; Gill and Levit, 1995; Letac, 2022; van Trees, 1968).

**Step 1: Bayesian Adversary and Bregman Representation.** Fix an anchor $\pi^\circ = \nabla\psi^*(u^\circ)$ with $u^\circ \in [-\eta R/2, \eta R/2]^K$. Draw a mean reward vector $w$ from a product cosine-squared prior $\mu_w$, centered at $u^\circ/\eta$ with half-width $\Delta = \min\{R/4, 1/\eta\}$. Conditional on $w$, sample the rewards independently across rounds and coordinates, with $r_{t,i} \in \{-R, R\}$ and $\mathbb{E}[r_{t,i} \mid w] = w_i$.

The mean-reward oracle is $\pi^\star(w) = \nabla\psi^*(\eta w)$. Since the best policy in hindsight performs at least as well as this fixed comparator,

$$\mathbb{E}[\mathrm{Reg}_\eta(T;\{r_t\})] \geq \frac{1}{\eta}\sum_{t=1}^{T}\mathbb{E}\left[B_\psi(\pi_t, \pi^\star(w))\right]. \tag{3}$$

The Bregman divergence here is precisely the Fenchel–Young loss of Blondel et al. (2020), $\psi^*(\eta w) + \psi(\pi_t) - \langle \eta w, \pi_t\rangle$.

**Step 2: Posterior Risk and Quotient Duality.** Let $\widehat{w}_t = \mathbb{E}[w \mid \mathcal{H}_{t-1}]$, where $\mathcal{H}_{t-1}$ is the learner's history, including its independent randomization. The policy $\bar{\pi}_t = \nabla\psi^*(\eta\widehat{w}_t)$ minimizes the conditional expected divergence in (3). Therefore, we may lower-bound the learner's risk by the risk of $\bar{\pi}_t$.

Because adding a constant reward vector leaves the optimal policy unchanged, dual Bregman divergences are naturally defined on $\mathcal{Q} = \mathbb{R}^K / \operatorname{span}\{\mathbf{1}\}$. By Lemma A.1,

$$B_\psi(\bar{\pi}_t, \pi^\star(w)) = B_{\psi^*}([\eta w], [\eta\widehat{w}_t]).$$

Both $w$ and $\widehat{w}_t$ lie in the prior box. This is a comparison of conditional risks; the learner's actual actions and observations remain unchanged.

**Step 3: Dual Curvature and Local Metric Comparison.** The curvature assumptions imply

$$C_\psi^{-1} L_p(\nabla\psi^*(u)) \preceq \nabla^2\psi^*(u) \preceq c_\psi^{-1} L_p(\nabla\psi^*(u)),$$

as shown in Lemma A.2. Every point on the segment between $\eta\widehat{w}_t$ and $\eta w$ is within $\eta\Delta \le 1$ of $u^\circ$ in $\ell_\infty$-norm. Separability and the coordinate curvature lower bound therefore give

$$L_p(\nabla\psi^*(u)) \succeq a_\psi L_p(\pi^\circ), \qquad a_\psi \coloneqq e^{-2p/c_\psi},$$

along this segment, by Lemma A.3. Integrating the lower Hessian bound yields

$$\mathbb{E}[\mathrm{Reg}_\eta(T; \{r_t\})] \ge \frac{a_\psi \eta}{2C_\psi} \sum_{t=1}^{T} \mathbb{E}\left[(w - \widehat{w}_t)^\top L_p(\pi^\circ)(w - \widehat{w}_t)\right]. \tag{4}$$

Retaining the anchor matrix $L_p(\pi^\circ)$ preserves the regularizer-specific dimension dependence.

**Step 4: Matrix van Trees Inequality.** Let $\mathcal{I}_t(w)$ denote the Fisher information in the observations available before round $t$, and let $\mathcal{J}(\mu_w)$ denote the prior Fisher information. The matrix van Trees inequality gives

$$\mathbb{E}\left[(w - \widehat{w}_t)(w - \widehat{w}_t)^\top\right] \succeq \left(\mathbb{E}[\mathcal{I}_t(w)] + \mathcal{J}(\mu_w)\right)^{-1}.$$

The prior support satisfies $|w_i| \le 3R/4$. Since every reward coordinate is observed,

$$\mathcal{I}_t(w) \preceq \frac{16(t-1)}{7R^2} I_K, \qquad \mathcal{J}(\mu_w) = \frac{\pi^2}{\Delta^2} I_K.$$

Substituting these bounds into (4) gives

$$\mathbb{E}[\mathrm{Reg}_\eta(T; \{r_t\})] \ge \frac{a_\psi \eta}{2C_\psi} d_{\mathrm{FI}}(\pi^\circ) \sum_{t=1}^{T} \frac{1}{16(t-1)/(7R^2) + \pi^2/\Delta^2}.$$

An integral comparison bounds the sum from below by

$$\frac{7R^2}{16} \log\left(1 + \frac{T}{7\pi^2 \max\{1, (\eta R/4)^2\}}\right).$$

Choosing the anchor to maximize $d_{\mathrm{FI}}$ over $\Pi_{\eta R/2, \psi}$ proves the logarithmic lower bound. □

For completeness, we complement our lower bound with a matching regret upper bound. The algorithm is a simple instantiation of Follow-The-Regularized-Leader (FTRL), with an additional Shannon entropy term $H(\pi) \coloneqq \sum_{i=1}^K \pi_i \log \pi_i$.

**Theorem 3.2** (Regret Upper Bound, Full-Information)**.** *Suppose Assumptions 1 and 3 and Assumption 2(i) hold, and $T$ is known. Then, for every $\eta > 0$, we have*

$$\mathfrak{R}_\eta(T; \psi) \le \min\left\{\eta c_\psi^{-1} R^2 (3 + \log T),\ R\sqrt{2T \log K} + 2R\right\}.$$

*The bound is attained by* ***C-FTRL*** *(Algorithm 1),*

$$\pi_t \in \arg\max_{\pi \in \Pi} \left\{\sum_{s=1}^{t-1} \langle \pi, r_s \rangle - \lambda_t \psi(\pi) - \xi H(\pi)\right\},$$

*with*

$$\lambda_t = \frac{t}{\eta}, \qquad \xi = \min\left\{\frac{\eta R^2}{c_\psi \log K},\ R\sqrt{\frac{T}{2\log K}}\right\}. \tag{5}$$

A refinement retaining the full-information effective dimension is given in Appendix F.1.2.

*Proof Sketch.* We defer the full proof to Appendix C, which is a standard application of FTRL analysis for composite objectives, with a careful choice of the coefficient $\xi$; see Orabona (2025, Chapter 7). The difference $\lambda_{t+1} - \lambda_t = \eta^{-1}$ makes the accumulated objective penalties telescope. Writing $\bar{\psi} = \psi - \min_{\pi\in\Pi}\psi(\pi)$, we obtain a bound of the form

$$\mathrm{Reg}_\eta(T;\{r_t\}) \le \frac{\bar{\psi}(\pi^\star)}{\eta} + \xi\log K + \eta c_\psi^{-1} R^2 \sum_{t=1}^{T} \frac{1}{t + 2\eta c_\psi^{-1}\xi}.$$

The comparator term $\bar{\psi}(\pi^\star)/\eta$ is at most $\min\{\eta c_\psi^{-1}R^2, 2R\}$, and the choice in (5) gives both $\eta\log T$ and $\sqrt{T}$ rates simultaneously. □

The update follows the composite online optimization framework of Duchi and Singer (2009); Duchi et al. (2010); Xiao (2010) and its FTRL treatments (Joulani et al., 2020; McMahan, 2017). Here $\psi/\eta$ is part of each round's objective and is accumulated through $\lambda_t = t/\eta$, while $\xi H$ supplies additional algorithmic regularization. This separation allows the curvature of $\psi$ to give logarithmic regret and the Shannon term to give the standard full-information square-root bound.

### 3.2 Discussions

**Tightness of Regret Bounds.** For fixed $K$, $R > 0$, $\eta > 0$, and $\psi$, Theorems 3.1 and 3.2 give matching logarithmic dependence on $T$ under the assumptions of Theorem 3.1. The comparison of the effective dimensions and dual radii in the lower and refined upper bounds is detailed in Appendix F.1.2. For Shannon entropy, $D_{\mathrm{FI}}(U;H) = 1 - 1/K$ at every radius, so the bounds give $\Theta(\eta R^2 \log T)$ as $T$ grows. Section 5 compares the bounds for other regularizers. When $\eta$ varies with $T$, the logarithmic factors also depend differently on $\eta R$. The lower bound contains a logarithm of order $\log(1+T/\max\{1,(\eta R)^2\})$, while the logarithmic upper bound scales with $1+\log T$. C-FTRL also guarantees $R\sqrt{2T\log K} + 2R$ for every $\eta > 0$. Under the square-root lower-bound condition of Theorem 3.1, this yields the full-information minimax rate $\Theta(R\sqrt{T\log K})$.

**Use of the van Trees inequality.** Other than the Bregman divergence representation of the regularized regret, our main technical tool is the ***van Trees inequality (VTI)*** (Gill and Levit, 1995; van Trees, 1968), also known as the *Bayesian Cramér-Rao Bound* (Bobrovsky et al., 1987), extended to matrices (Gassiat and Stoltz, 2024; Letac, 2022). In rough terms, VTI lower bounds the Bayes $\ell_2$-error of estimating a parameter (over some given prior) by the inverse Fisher information.

Despite its "relative simplicity" compared to other lower bound techniques (Tsybakov, 2009), the use of ***VTI*** for deriving lower bounds remains rather scattered and is not well known in the machine learning community, yet it has been utilized effectively. For instance, it has been used to establish lower bounds for distribution estimation under communication constraints (Barnes et al., 2020; Chen et al., 2021), and in the statistical learning theory of control (Lee et al., 2023; Tsiamis et al., 2023; Wagenmaker et al., 2021; Ziemann and Sandberg, 2021, 2025). Our analysis contributes to this line of work by demonstrating that ***VTI*** is a versatile tool capable of leveraging strongly convex geometry to obtain fast $\Omega(\log T)$ regret lower bounds.

**Logarithmic Regret Lower Bound.** In online learning, logarithmic regret *upper* bounds are well-established for losses exhibiting sufficient curvature, such as exp-concavity (Hazan et al., 2007) and stochastic mixability (van Erven et al., 2015). However, as noted in Section 1, corresponding *lower* bounds are rare because they typically require the adversary to provide the curvature (e.g., Strongly Convex Games (Abernethy et al., 2008)), or are restricted to specific forms of losses such as the squared loss (Gaillard et al., 2019; Takimoto and Warmuth, 2000; Vovk, 2001) or exp-concave log-loss (Ordentlich and Cover, 1998).

Crucially, from a methodological standpoint, almost all of these existing logarithmic lower bounds proceed by computing the *minimax regret* explicitly in closed form. Although Gaillard et al. (2019, Theorem 4) prove a similar logarithmic lower bound for fixed-design linear regression using the multivariate van Trees inequality (Gill and Levit, 1995) without relying on explicitly computing the minimax regret, their analysis is still strictly confined to $\ell_2$-losses. Thus, to our knowledge, it is generally unclear how to extend such proof techniques beyond highly specific domains where the minimax strategy and game value can be explicitly solved.

Our main novelty here is reducing the lower bound for a linear online learning problem to a Bayesian estimation task on a bounded quotient space. This cleanly decouples the curvature requirement from the adversary, proving that the learner's fixed regularizer alone is sufficient to enforce the information-theoretic $\Omega(\log T)$ limit, without the need to solve the minimax game in a closed form.

## 4 Regularized Adversarial Multi-Armed Bandits

Under bandit feedback, the learner draws $a_t \sim \pi_t$ and observes only the reward $r_{t,a_t}$.

### 4.1 Minimax-Optimal Regularized Regret Bounds

Similarly, we first present the lower bound:

**Theorem 4.1** (Regret Lower Bound, Bandit)**.** *Let $K \geq 2$, and suppose Assumptions 1 to 4 hold. For every $T \geq 1$ and $\eta > 0$,*

$$\mathfrak{R}_\eta^{\mathrm{B}}(T;\psi) \geq \frac{7e^{-2p/c_\psi}}{32C_\psi}\eta R^2 D_{\mathrm{B}}(\eta R/2;\psi)\log\left(1+\frac{T}{7\pi^2 K\max\{1,(\eta R/4)^2\}}\right). \tag{6}$$

*Furthermore, suppose that $T \geq K$ and $F_\psi\left(\frac{\eta R}{4}\sqrt{\frac{K}{T}}\right) \geq \frac{7\eta R}{32}\sqrt{\frac{K}{T}}$. Then,*

$$\mathfrak{R}_\eta^{\mathrm{B}}(T;\psi) \geq \frac{1}{32}R\sqrt{KT}.$$

*Proof Sketch.* The full proof is provided in Appendix D. We focus on the logarithmic lower bound. We reuse the Bayesian adversary and geometric reduction from the proof of Theorem 3.1, with $\Delta = \min\{R/4, 1/\eta\}$ and $a_\psi = e^{-2p/c_\psi}$. In particular, (4) remains valid when $\widehat{w}_t$ is the posterior mean given the learner's actual bandit history. The critical difference is the available Fisher information.

Write

$$n_{t,i} := \mathbb{E}\left[\sum_{s=1}^{t-1}\mathbf{1}\{I_s = i\}\right], \qquad \sum_{i=1}^{K} n_{t,i} = t-1,$$

where the expectation is over the prior, rewards, and learner's randomization. The action-selection rule has no direct dependence on $w$, and each observation of arm $i$ contributes Fisher information $1/(R^2 - w_i^2)$ in coordinate $i$. Since $|w_i| \leq 3R/4$, the prior-averaged data information satisfies $\mathbb{E}[\mathcal{I}_t(w)] \preceq \frac{16}{7R^2}\mathrm{diag}(n_{t,1},\ldots,n_{t,K})$, and the prior information remains $\mathcal{J}(\mu_w) = (\pi^2/\Delta^2)I_K$.

Fix an anchor $\pi^\circ \in \Pi_{\eta R/2,\psi}$ and write $\ell_i = (L_p(\pi^\circ))_{ii}$. Applying the matrix van Trees inequality in (4) gives

$$\mathbb{E}[\mathrm{Reg}_\eta(T;\{r_t\})] \geq \frac{a_\psi \eta}{2C_\psi}\sum_{t=1}^{T}\sum_{i=1}^{K}\frac{\ell_i}{16n_{t,i}/(7R^2)+\pi^2/\Delta^2}.$$

The prior information keeps every denominator positive, including for arms that have never been sampled. By Cauchy–Schwarz,

$$\sum_{i=1}^{K}\frac{\ell_i}{16n_{t,i}/(7R^2)+\pi^2/\Delta^2} \geq \frac{\left(\sum_i\sqrt{\ell_i}\right)^2}{\frac{16}{7R^2}\sum_i n_{t,i}+K\pi^2/\Delta^2} = \frac{d_{\mathrm{B}}(\pi^\circ)}{16(t-1)/(7R^2)+K\pi^2/\Delta^2}.$$

Thus the restricted observation budget produces the bandit effective dimension $d_{\mathrm{B}}$. Summing over $t$ using the same integral comparison as in the full-information proof, and maximizing over $\pi^\circ \in \Pi_{\eta R/2,\psi}$, proves (6). □

We obtain the upper bound using C-FTRL with an importance-weighted (IW) estimator (Auer et al., 2002). For $U \geq 0$, let $m_\psi(U) := \max_{\pi \in \Pi_{U,\psi}} \|\pi\|_\infty$ and define the convex cap and its variance profile by

$$\mathcal{C}_U := \{\pi \in \Pi : \|\pi\|_\infty \leq m_\psi(U)\}, \qquad V_{\mathrm{B}}(U;\psi) := \max_{\pi \in \mathcal{C}_U} \sum_i \pi_i^{p-1}, \tag{7}$$

with the convention $0^0 = 1$ for $p = 1$. By Lemma C.2, the hindsight comparator satisfies $\pi^\star \in \Pi_{\eta R,\psi} \subseteq \mathcal{C}_{\eta R}$, so we may use $\mathcal{C}_{\eta R}$ as the feasible set.

**Theorem 4.2** (Regret Upper Bound, Bandit)**.** *Suppose Assumptions 1 and 3 and Assumption 2(i) hold, and $T$ is known. Then, for every $\eta > 0$, we have*

$$\mathfrak{R}_\eta(T;\psi) \leq \min\left\{4\eta c_\psi^{-1} R^2 V_{\mathrm{B}}(\eta R;\psi)(1+\log T),\ 4R\sqrt{2KT}\right\}. \tag{8}$$

*The bound is attained by **IW-C-FTRL** with the update*

$$\pi_t \in \arg\max_{\pi \in \mathcal{C}_{\eta R}} \left\{\sum_{s=1}^{t-1} \langle \pi, \widehat{r}_s\rangle - \lambda_t \psi(\pi) - \xi S_{1/2}(\pi)\right\},$$

*where $S_{1/2}(\pi) := 2(1 - \sum_i \sqrt{\pi_i})$ is the negative $1/2$-Tsallis entropy and*

$$\lambda_t = \frac{t}{\eta}, \qquad \xi = \min\left\{\frac{\eta R^2 V_{\mathrm{B}}(\eta R;\psi)(1+\log T)}{c_\psi \sqrt{K}},\ R\sqrt{2T}\right\}. \tag{9}$$

*After drawing $a_t \sim \pi_t$, the learner forms the shifted loss estimate $\widehat{\ell}_t = (R - r_{t,a_t})e_{a_t}/\pi_{t,a_t}$ as in Algorithm 2 and sets $\widehat{r}_t = R\mathbf{1} - \widehat{\ell}_t$.*

*Proof Sketch.* We defer the full proof to Appendix E.1. The update includes the current penalty $\psi/\eta$ before $a_t$ is drawn. In the equivalent loss formulation, we telescope the minima of the cumulative objectives and use the unbiasedness of $\widehat{\ell}_t$ to bound the expected regret by the range of $\xi S_{1/2}$, at most $2\xi\sqrt{K}$, plus the expected cumulative stability. When $a_t = i$, $\widehat{\ell}_t$ is nonnegative and supported on $i$, so only candidates with $\pi_i < \pi_{t,i}$ can contribute positively to stability. The $i$-th coordinate is at most $\pi_{t,i}$ along the segment from $\pi_t$ to any such candidate. Applying RSC and the curvature $1/(2x^{3/2})$ of $-2\sqrt{x}$ along this segment and averaging over $a_t$ bounds the conditional expected stability by

$$\min\left\{\frac{2\eta R^2}{c_\psi t}\sum_i \pi_{t,i}^{p-1},\ \frac{4R^2}{\xi}\sum_i \sqrt{\pi_{t,i}}\right\}.$$

The two sums are at most $V_{\mathrm{B}}(\eta R;\psi)$ and $\sqrt{K}$, respectively. Summing over $t$, using $\sum_{t=1}^T t^{-1} \leq 1 + \log T$, and substituting (9) gives (8). □

### 4.2 Discussions

**Tightness and Geometry.** For fixed $K, R, \eta$ and $\psi$ satisfying the assumptions of Theorem 4.1, Theorems 4.1 and 4.2 give matching logarithmic dependence on $T$. The upper and lower bounds involve $V_{\mathrm{B}}(\eta R;\psi)$ and $D_{\mathrm{B}}(\eta R/2;\psi)$, respectively. For $1 \leq p \leq 2$, the uniform policy attains the maximum in (7), giving $V_{\mathrm{B}}(\eta R;\psi) = K^{2-p}$. If $\mathbf{1}/K \in \Pi_{\eta R/2,\psi}$, then Lemma F.2 gives $D_{\mathrm{B}}(\eta R/2;\psi) = (1 - 1/K)K^{2-p}$. In this case, their polynomial dependence on $K$ matches, with the curvature constants and logarithmic factors retained in the theorem statements. For every $\eta > 0$, **IW-C-FTRL** also guarantees $\mathcal{O}(R\sqrt{KT})$ regret. When $T \geq K$ and the $F_\psi$ condition in Theorem 4.1 holds, this yields the minimax rate $\Theta(R\sqrt{KT})$.

The lower-bound proof uses the same geometric reduction under both feedback models. Full information reveals every coordinate at each round, leading to the trace in $d_{\mathrm{FI}}$. Bandit feedback allocates one observation per round across coordinates, leading instead to the squared sum in $d_{\mathrm{B}}$ through the restricted Fisher information.

**Beyond KL.** For KL-regularized contextual bandits, Zhao et al. (2025a) relate policy suboptimality to squared reward-estimation error to study sample complexity, while Zhao et al. (2025b) use a related bound to establish logarithmic regret. Wu et al. (2025a) study greedy sampling for KL-regularized RLHF with preference feedback. For stochastic KL-regularized MABs, Ji et al. (2026) establish a high-probability $\widetilde{\mathcal{O}}(\eta K \log^2 T)$ upper bound in the strongly regularized regime and logarithmic lower bounds constructed with a uniform reference policy. Our guarantees concern expected regret under bounded adversarial rewards and extend beyond KL: for fixed $\eta > 0$, Theorem 4.2 gives logarithmic regret for every Legendre regularizer satisfying $(p, c_\psi)$-RSC.

## 5 Instantiating the Regret Bounds

We instantiate our regret bounds for four standard regularizers: **negative Shannon entropy** $H(\pi) := \sum_i \pi_i \log \pi_i$, **reverse KL-divergence** $D_{\mathrm{KL}}(\pi, \rho) := \sum_i \pi_i \log(\pi_i/\rho_i)$ for a fixed reference $\rho \in \mathrm{relint}(\Pi)$, **negative $\alpha$-Tsallis entropy** $S_\alpha(\pi) := (1 - \sum_i \pi_i^\alpha)/(1-\alpha)$ for $\alpha \in (0,1)$, and **Burg–Itakura–Saito (BIS) entropy** $\psi_{\mathrm{BIS}}(\pi) := -\sum_i \log \pi_i$. Throughout this section, we set $R = 1$ for simplicity. Table 2 in Appendix F summarizes the detailed bounds and their conditions.

**Shannon, Tsallis, and BIS Entropy.** These regularizers are minimized at the uniform policy and have curvature exponents $p = 1$, $2 - \alpha$, and 2, respectively. For $0 < \eta \le 1$, our full-information bounds match at $\Theta(\eta K^{1-p} \log T)$. Under bandit feedback, the corresponding bounds are $\Omega(\eta K^{2-p} \log(T/K))$ and $\mathcal{O}(\eta K^{2-p} \log T)$. The implicit constants are independent of $K, T, \eta$, but may depend on the fixed Tsallis parameter $\alpha$. Thus, the full-information factors are $1$, $K^{\alpha-1}$, and $K^{-1}$, whereas the bandit factors are $K$, $K^\alpha$, and $1$. In particular, BIS entropy admits a logarithmic bandit upper bound independent of $K$.

**KL-Regularized Bandit.** For stochastic KL-regularized MABs, Ji et al. (2026, Theorems 5.1 and 5.3) established an $\Omega(\sqrt{KT})$ lower bound for $K \ge 9$ and $\eta \ge \sqrt{T/K} \log K$, and an $\Omega(\eta K \log(T/(\eta^2 K)))$ lower bound for $\eta \le \sqrt{T/K}$, both using $\rho = \mathrm{Unif}([K])$ in their constructions. In contrast, our bounds hold under bounded adversarial rewards for *any* reference policy $\rho$, with the logarithmic lower bound retaining the effective dimension $D_{\mathrm{B}}(\eta/2; D_{\mathrm{KL}}(\cdot, \rho))$. For $\rho = \mathrm{Unif}([K])$, this quantity equals $K - 1$, recovering the $\eta K \log T$ scaling for fixed $K, \eta$ as $T$ grows. Our $\Omega(\sqrt{KT})$ lower bound also holds for $\eta \gtrsim \sqrt{T/K} \log K$ when $T \ge K$.

For a general reference policy and $T \ge K$, a sufficient condition for our $\sqrt{KT}$ lower bound is $\eta \gtrsim \sqrt{T/K}\, K^{-1} \sum_i \log(1/\rho_i)$. Meanwhile, Ji et al. (2026, Theorem 4.2) give a high-probability $\widetilde{\mathcal{O}}(\sqrt{KT} \log T)$ upper bound for $\eta \ge \sqrt{T/K}$, whereas **IW-C-FTRL** guarantees $\mathcal{O}(\sqrt{KT})$ expected regret for every $\eta > 0$. The reference-policy dependence in our lower bounds remains relevant: a concentrated $\rho$ can restrict the reachable effective dimensions. Determining the sharp dependence of the transition on $\rho$ remains an interesting direction for future work.

## 6 Conclusion

In this work, we established lower and upper bounds on the ***$\psi$-minimax regularized regret*** for a broad class of strongly convex regularizers under both full-information and bandit feedback. By formulating the dual geometry over the quotient space $\mathcal{Q}$ and applying the matrix-variate van Trees inequality, we derived logarithmic lower bounds directly from the curvature of the learner's fixed regularizer. This approach bypasses the need for curved adversarial losses, solving the minimax game explicitly, or closed-form solutions unique to the reverse KL-divergence. Our bounds match in their logarithmic dependence on $T$ for fixed parameters and recover the unregularized square-root rates under explicit large-$\eta$ conditions. The effective dimensions describe how the regularizer's curvature and reachable policies determine the dependence on $K$.

## Acknowledgments and Disclosure of Funding

In the initial submission phrase, J. Lee acknowledges support from Gemini 3.1 Pro and ChatGPT 5.5 Pro (Extended Thinking) for help with the proof and Cursor (GPT-5.3 Codex) for help with the

LaTeX. For the camera-ready revision, we used ChatGPT 6.0 Astra to further polish the writing and proofs, with some improvements.

We thank the anonymous reviewers for helpful comments and suggestions that helped us improve the manuscript. This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2019-II190075 Artificial Intelligence Graduate School Support Program (KAIST), 5%; No. RS-2024-00457882, AI Research Hub Project, 5%; No. RS-2022-II220311 Development of Goal-Oriented Reinforcement Learning Techniques for Contact-Rich Robotic Manipulation of Everyday Objects, 45%; No. RS-2022-II220184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics, 45%).

## References

Jacob Abernethy, Peter L. Bartlett, Alexander Rakhlin, and Ambuj Tewari. Optimal Strategies and Minimax Lower Bounds for Online Convex Games. In *COLT 2008 - The 21st Annual Conference on Learning Theory*, 2008. URL https://www.learningtheory.org/colt2008/papers/111-Abernethy.pdf.

Jacob Abernethy, Alekh Agarwal, Peter L. Bartlett, and Alexander Rakhlin. A Stochastic View of Optimal Regret through Minimax Duality. In *COLT 2009 - The 22nd Conference on Learning Theory*, 2009. URL https://www.learningtheory.org/colt2009/papers/026.pdf.

Jacob D. Abernethy, Elad Hazan, and Alexander Rakhlin. Interior-Point Methods for Full-Information and Bandit Online Learning. *IEEE Transactions on Information Theory*, 58(7):4164–4175, 2012. doi:10.1109/TIT.2012.2192096.

Jacob D Abernethy, Chansoo Lee, and Ambuj Tewari. Fighting Bandits with a New Kind of Smoothness. In *Advances in Neural Information Processing Systems*, volume 28, pages 2197–2205. Curran Associates, Inc., 2015. URL https://arxiv.org/abs/1512.04152.

Shun-ichi Amari. *Information Geometry and Its Applications*. Applied Mathematical Sciences. Springer Tokyo, 2016.

Gholamali Aminian, Amir R. Asadi, Idan Shenfeld, and Youssef Mroueh. KL-Regularized RLHF with Multiple Reference Models: Exact Solutions and Sample Complexity. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. URL https://openreview.net/forum?id=j8XnFfTvXF.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The Non-stochastic Multiarmed Bandit Problem. *SIAM Journal on Computing*, 32(1):48–77, 2002. doi:10.1137/S0097539701398375.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback. *arXiv preprint arXiv:2204.05862*, 2022. URL https://arxiv.org/abs/2204.05862.

Leighton Pate Barnes, Yanjun Han, and Ayfer Özgür. Lower bounds for learning distributions under communication constraints via fisher information. *Journal of Machine Learning Research*, 21 (236):1–30, 2020. URL http://jmlr.org/papers/v21/19-737.html.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003. doi:10.1016/S0167-6377(02)00231-6.

Mathieu Blondel, André F.T. Martins, and Vlad Niculae. Learning with Fenchel-Young losses. *Journal of Machine Learning Research*, 21(35):1–69, 2020. URL http://jmlr.org/papers/v21/19-021.html.

B. Z. Bobrovsky, E. Mayer-Wolf, and M. Zakai. Some Classes of Global Cramer-Rao Bounds. *The Annals of Statistics*, 15(4):1421 – 1438, 1987. doi:10.1214/aos/1176350602.

L.M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967. doi:10.1016/0041-5553(67)90040-7.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomek Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphael Segerie, Micah Carroll, Andi Peng, Phillip J.K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=bx24KpJ4Eb.

Shicong Cen, Jincheng Mei, Katayoon Goshvadi, Hanjun Dai, Tong Yang, Sherry Yang, Dale Schuurmans, Yuejie Chi, and Bo Dai. Value-Incentivized Preference Optimization: A Unified Approach to Online and Offline RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=SQnitDuow6.

Nicoló Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning and Games*. Cambridge University Press, 2006.

Wei-Ning Chen, Peter Kairouz, and Ayfer Özgür. Pointwise Bounds for Distribution Estimation under Communication Constraints. In *Advances in Neural Information Processing Systems*, volume 34, pages 24593–24603. Curran Associates, Inc., 2021. URL https://openreview.net/forum?id=l41jc6kUfKr.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, volume 30, pages 4302–4310. Curran Associates, Inc., 2017. URL https://arxiv.org/abs/1706.03741.

John Duchi and Yoram Singer. Efficient Online and Batch Learning Using Forward Backward Splitting. *Journal of Machine Learning Research*, 10(99):2899–2934, 2009. URL http://jmlr.org/papers/v10/duchi09a.html.

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. URL http://jmlr.org/papers/v12/duchi11a.html.

John C. Duchi, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari. Composite Objective Mirror Descent. In *COLT 2010 - The 23rd Conference on Learning Theory*, pages 14–26, 2010. URL https://www.learningtheory.org/colt2010/papers/057Duchi.pdf.

Pierre Gaillard, Sébastien Gerchinovitz, Malo Huard, and Gilles Stoltz. Uniform regret bounds over $\mathbb{R}^d$ for the sequential linear regression problem with the square loss. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 404–432. PMLR, 22–24 Mar 2019. URL https://proceedings.mlr.press/v98/gaillard19a.html.

Leo Gao, John Schulman, and Jacob Hilton. Scaling laws for reward model overoptimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10835–10866. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/gao23h.html.

Elisabeth Gassiat and Gilles Stoltz. The van Trees Inequality in the Spirit of Hájek and Le Cam. *Statistical Science*, 39(4):644 – 653, 2024. doi:10.1214/24-STS941.

Matthieu Geist, Bruno Scherrer, and Olivier Pietquin. A Theory of Regularized Markov Decision Processes. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2160–2169. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/geist19a.html.

Richard D. Gill and Boris Y. Levit. Applications of the van Trees inequality: a Bayesian Cramér-Rao bound. *Bernoulli*, 1(1-2):59 – 79, 1995. doi:10.2307/3318681.

Dongyoung Go, Tomasz Korbak, Germàn Kruszewski, Jos Rozen, Nahyeon Ryu, and Marc Dymetman. Aligning Language Models with Preferences through $f$-divergence Minimization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 11546–11583. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/go23a.html.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

Jiaqi Han, Mingjian Jiang, Yuxuan Song, Stefano Ermon, and Minkai Xu. $f$-PO: Generalizing Preference Optimization with $f$-divergence Minimization. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 1144–1152. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/han25a.html.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007. doi:10.1007/s10994-007-5016-8.

Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason D. Lee, Wen Sun, Akshay Krishnamurthy, and Dylan J Foster. Correcting the Mythos of KL-Regularization: Direct Alignment without Overoptimization via Chi-Squared Preference Optimization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=hXm0Wu2U9K.

Athul Paul Jacob, David J Wu, Gabriele Farina, Adam Lerer, Hengyuan Hu, Anton Bakhtin, Jacob Andreas, and Noam Brown. Modeling Strong and Human-Like Gameplay with KL-Regularized Search. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9695–9728. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/jacob22a.html.

Kaixuan Ji, Qingyue Zhao, Heyang Zhao, Qiwei Di, and Quanquan Gu. Near-Optimal Regret for KL-Regularized Multi-Armed Bandits. In *Proceedings of the 43rd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 06–11 Jul 2026. URL https://openreview.net/forum?id=XarOZG8Un0.

Pooria Joulani, András György, and Csaba Szepesvári. A modular analysis of adaptive (non-)convex optimization: Optimism, composite objectives, variance reduction, and variational bounds. *Theoretical Computer Science*, 808:108–138, 2020. URL https://www.sciencedirect.com/science/article/pii/S0304397519307273. Special Issue on Algorithmic Learning Theory.

Masahiro Kato and Shinji Ito. LC-Tsallis-INF: Generalized Best-of-Both-Worlds Linear Contextual Bandits. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 3655–3663. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/kato25a.html.

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

Hung Le, Quan Hung Tran, Dung Nguyen, Kien Do, Saloni Mittal, Kelechi Ogueji, and Svetha Venkatesh. Multi-Reference Preference Optimization for Large Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(23):24375–24383, Apr. 2025. doi:10.1609/aaai.v39i23.34615.

Bruce D. Lee, Ingvar Ziemann, Anastasios Tsiamis, Henrik Sandberg, and Nikolai Matni. The fundamental limitations of learning linear-quadratic regulators. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 4053–4060, 2023. doi:10.1109/CDC49753.2023.10383608.

Junghyun Lee, Minju Hong, Kwang-Sung Jun, Chulhee Yun, and Se-Young Yun. Provably Efficient Regularized Online RLHF with Generalized Bilinear Preferences. *arXiv preprint arXiv:2602.23116*, 2026. URL https://arxiv.org/abs/2602.23116.

Gérard G Letac. The randomization by Wishart laws and the Fisher information. *arXiv preprint arXiv:2211.14137*, 2022. URL https://arxiv.org/abs/2211.14137.

Haihao Lu, Robert M. Freund, and Yurii Nesterov. Relatively Smooth Convex Optimization by First-Order Methods, and Applications. *SIAM Journal on Optimization*, 28(1):333–354, 2018. doi:10.1137/16M1099546.

H. Brendan McMahan. A Unified View of Regularized Dual Averaging and Mirror Descent with Implicit Updates. *arXiv preprint arXiv:1009.3240*, 2010. URL https://arxiv.org/abs/1009.3240.

H. Brendan McMahan. A survey of Algorithms and Analysis for Adaptive Online Learning. *Journal of Machine Learning Research*, 18(90):1–50, 2017. URL http://jmlr.org/papers/v18/14-428.html.

Gergely Neu, Anders Jonsson, and Vicenç Gómez. A unified view of entropy-regularized Markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017. URL https://arxiv.org/abs/1705.07798.

Frank Nielsen. An Elementary Introduction to Information Geometry. *Entropy*, 22(10), 2020. doi:10.3390/e22101100.

Francesco Orabona. A Modern Introduction to Online Learning. *arXiv preprint arXiv:1912.13213*, 2025. URL https://arxiv.org/abs/1912.13213.

Erik Ordentlich and Thomas M. Cover. The Cost of Achieving the Best Portfolio in Hindsight. *Mathematics of Operations Research*, 23(4):960–982, 1998. doi:10.1287/moor.23.4.960.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022. URL https://openreview.net/forum?id=TG8KACxEON.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. In *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc., 2023. URL https://openreview.net/forum?id=HPuSIXJaa9.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton Landmarks in Mathematics. Princeton University Press, Princeton, NJ, 1997.

Eiji Takimoto and Manfred Warmuth. The Minimax Strategy for Gaussian Density Estimation. In *COLT 2000 - The 13nd Conference on Learning Theory*, pages 100–106, 2000. URL https://www.learningtheory.org/colt2000/papers/TakimotoWarmuth.pdf.

Daniil Tiapkin, Denis Belomestny, Daniele Calandriello, Eric Moulines, Remi Munos, Alexey Naumov, Pierre Perrault, Yunhao Tang, Michal Valko, and Pierre Menard. Fast Rates for Maximum Entropy Exploration. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 34161–34221. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/tiapkin23a.html.

Daniil Tiapkin, Denis Belomestny, Daniele Calandriello, Eric Moulines, Alexey Naumov, Pierre Perrault, Michal Valko, and Pierre Menard. Demonstration-Regularized RL. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=lF2aip4Scn.

Anastasios Tsiamis, Ingvar Ziemann, Nikolai Matni, and George J. Pappas. Statistical Learning Theory for Control: A Finite-Sample Perspective. *IEEE Control Systems Magazine*, 43(6):67–97, 2023. doi:10.1109/MCS.2023.3310345.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Series in Statistics. Springer New York, 2009.

Tim van Erven, Peter D. Grünwald, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. Fast Rates in Statistical and Online Learning. *Journal of Machine Learning Research*, 16(54): 1793–1861, 2015. URL http://jmlr.org/papers/v16/vanerven15a.html.

Harry L. van Trees. *Detection, Estimation, and Modulation Theory, Part I: Detection, Estimation, and Linear Modulation Theory*. John Wiley & Sons, New York, 1968. ISBN 978-0471899556.

Volodya Vovk. Competitive On-line Statistics. *International Statistical Review*, 69(2):213–248, 2001. doi:10.1111/j.1751-5823.2001.tb00457.x.

Andrew J Wagenmaker, Max Simchowitz, and Kevin Jamieson. Task-Optimal Exploration in Linear Dynamical Systems. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10641–10652. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/wagenmaker21a.html.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond Reverse KL: Generalizing Direct Preference Optimization with Diverse Divergence Constraints. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=2cRzmWXK9N.

Yibo Wang, Sijia Chen, Wei Jiang, Wenhao Yang, Yuanyu Wan, and Lijun Zhang. Online Composite Optimization Between Stochastic and Adversarial Environments. In *Advances in Neural Information Processing Systems*, volume 37, pages 94808–94850. Curran Associates, Inc., 2024b. doi:10.52202/079017-3005. URL https://openreview.net/forum?id=MbEB5aKmMK.

Zichen Wang, Haoyang Hong, and Huazheng Wang. When Greedy Sampling Explores: KL-Regularized Contextual Bandits without Eluder-Dimension Dependence. *arXiv preprint arXiv:2609.13564*, 2026. URL https://arxiv.org/abs/2609.13564.

Di Wu, Chengshuai Shi, Jing Yang, and Cong Shen. Greedy Sampling Is Provably Efficient For RLHF. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025a. URL https://openreview.net/forum?id=jQH0gdwsuT.

Di Wu, Chengshuai Shi, Jing Yang, and Cong Shen. $f$-Divergence Regularized RLHF: Two Tales of Sampling and Unified Analyses. In *Proceedings of the 43rd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 06–11 Jul 2026. URL https://openreview.net/forum?id=XrtiZYwcU4.

Yulian Wu, Rushil Thareja, Praneeth Vepakomma, and Francesco Orabona. Offline and Online KL-Regularized RLHF under Differential Privacy. In *Socially Responsible and Trustworthy Foundation Models at NeurIPS 2025*, 2025b. URL https://openreview.net/forum?id=oFRYuSvlAe.

Lin Xiao. Dual Averaging Methods for Regularized Stochastic Learning and Online Optimization. *Journal of Machine Learning Research*, 11(88):2543–2596, 2010. URL http://jmlr.org/papers/v11/xiao10a.html.

Tengyang Xie, Dylan J Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Hassan Awadallah, and Alexander Rakhlin. Exploratory Preference Optimization: Harnessing Implicit Q*-Approximation for Sample-Efficient RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=QYigQ6gXNw.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative Preference Learning from Human Feedback: Bridging Theory and Practice for RLHF under KL-constraint. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 54715–54754. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/xiong24a.html.

Wenhao Yang, Xiang Li, and Zhihua Zhang. A Regularized Approach to Sparse Optimal Policy in Reinforcement Learning. In *Advances in Neural Information Processing Systems*, volume 32, pages 5940–5950. Curran Associates, Inc., 2019. URL https://arxiv.org/abs/1903.00725.

Andrew Chi-Chin Yao. Probabilistic computations: Toward a unified measure of complexity. In *18th Annual Symposium on Foundations of Computer Science (sfcs 1977)*, pages 222–227, 1977. doi:10.1109/SFCS.1977.24.

Deming Yuan, Yiguang Hong, Daniel W. C. Ho, and Shengyuan Xu. Distributed Mirror Descent for Online Composite Optimization. *IEEE Transactions on Automatic Control*, 66(2):714–729, 2021. doi:10.1109/TAC.2020.2987379.

Huiming Zhang, Binghan Li, Wan Tian, and Qiang Sun. Tail-Aware Information-Theoretic Generalization for RLHF and SGLD. *arXiv preprint arXiv:2604.10727*, 2026. URL https://arxiv.org/abs/2604.10727.

Heyang Zhao, Chenlu Ye, Quanquan Gu, and Tong Zhang. Sharp Analysis for KL-Regularized Contextual Bandits and RLHF. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025a. URL https://openreview.net/forum?id=TE63KPCXWt.

Heyang Zhao, Chenlu Ye, Wei Xiong, Quanquan Gu, and Tong Zhang. Logarithmic Regret for Online KL-Regularized Reinforcement Learning. In *Proceedings of the 42nd International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 13–19 Jul 2025b. URL https://openreview.net/forum?id=6QH9IB53uy.

Qingyue Zhao, Kaixuan Ji, Heyang Zhao, Tong Zhang, and Quanquan Gu. Towards a Sharp Analysis of Offline Policy Learning for $f$-Divergence-Regularized Contextual Bandits. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ly6MB2Cfx2.

Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum Entropy Inverse Reinforcement Learning. In *Proceedings of the 23rd National Conference on Artificial Intelligence - Volume 3*, AAAI'08, page 1433–1438. AAAI Press, 2008. URL https://cdn.aaai.org/AAAI/2008/AAAI08-227.pdf.

Ingvar Ziemann and Henrik Sandberg. On Uninformative Optimal Policies in Adaptive LQR with Unknown B-Matrix. In *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, volume 144 of *Proceedings of Machine Learning Research*, pages 213–226. PMLR, 07 – 08 June 2021. URL https://proceedings.mlr.press/v144/ziemann21a.html.

Ingvar Ziemann and Henrik Sandberg. Regret Lower Bounds for Learning Linear Quadratic Gaussian Systems. *IEEE Transactions on Automatic Control*, 70(1):159–173, 2025. doi:10.1109/TAC.2024.3439132.

Julian Zimmert and Yevgeny Seldin. An Optimal Algorithm for Stochastic and Adversarial Bandits. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 467–475. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/zimmert19a.html.

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. URL http://jmlr.org/papers/v22/19-753.html.

## Contents


**1 Introduction** **1**

**2 Preliminaries** **3**

**3 Regularized Online Learning with Full-Information Feedback** **5**

3.1 Minimax-Optimal Regularized Regret Bounds . . . . . 5

3.2 Discussions . . . . . 7

**4 Regularized Adversarial Multi-Armed Bandits** **8**

4.1 Minimax-Optimal Regularized Regret Bounds . . . . . 8

4.2 Discussions . . . . . 9

**5 Instantiating the Regret Bounds** **10**

**6 Conclusion** **10**

**A Quotient-Space Formulation of the Dual Geometry** **19**

A.1 Motivation: A Non-Identifiability Issue . . . . . 19

A.2 Local (Riemannian) Geometry of Quotient-Tangent Spaces . . . . . 19

A.3 Mirror Maps and Bregman Divergences on the Quotient Space . . . . . 20

A.4 Quotient Versions of Convex-Analytic Lemmas . . . . . 21

**B Proof of Theorem 3.1: Minimax Lower Bound (Full Information)** **25**

B.1 Part I. Logarithmic Lower Bound . . . . . 25

B.2 Part II. Square-Root Lower Bound . . . . . 29

**C Proof of Theorem 3.2: Regret Upper Bound (Full Information) via C-FTRL** **32**

C.1 Generic Full Information Upper Bound . . . . . 32

C.2 Supporting Lemmas and the Effective-Dimension Refinement . . . . . 33

**D Proof of Theorem 4.1: Minimax Lower Bound (Bandit)** **37**

D.1 Part I. Logarithmic Lower Bound . . . . . 37

D.2 Part II. Square-Root Lower Bound . . . . . 39

**E Proof of Theorem 4.2: Regret Upper Bound (Bandit) via IW-C-FTRL** **41**

E.1 Generic Bandit Upper Bound . . . . . 41

**F Instantiations and Geometric Quantities for Standard Regularizers** **43**

F.1 Results . . . . . 43

F.1.1 Curvature and Dual-Constrained Diameter . . . . . 43

F.1.2 Full-Information Effective-Dimension Refinement . . . . . 43

F.1.3 Effective Dimensions . . . . . 44

F.1.4 Sufficient Transition Conditions . . . . . 45
F.1.5 Instantiated Regret Bounds . . . . . 45
F.2 Proofs and Computations . . . . . 46

# A Quotient-Space Formulation of the Dual Geometry

In this Appendix, we formalize the *quotient-space formulation* of the dual geometry, which provides the theoretical basis for our subsequent lower bound proofs. Along the way, we state several useful convex-analytic and information-geometric lemmas adapted to our quotient space; see Amari (2016); Nielsen (2020); Rockafellar (1997) for background. Throughout, let $\psi$ satisfy Assumptions 2 and 3. We impose coordinate-wise separability only when explicitly stated.

## A.1 Motivation: A Non-Identifiability Issue

We begin by observing that, for a fixed policy sequence, the regret is invariant under adding the same constant to every reward coordinate. Indeed, for any $c_t \in \mathbb{R}$ and $\pi_t \in \Pi = \Delta_K$,

$$\langle r_t + c_t \mathbf{1}, \pi_t \rangle = \langle r_t, \pi_t \rangle + c_t.$$

The same additive term appears for every comparator policy, and hence cancels from the regret. This is an identity for a fixed policy sequence; a learner interacting with shifted feedback need not choose the same policies.

To account for this non-identifiability, recall the ***quotient space*** $\mathcal{Q}$ and the ***tangent space*** $\mathcal{T}$:

$$\mathcal{Q} := \mathbb{R}^K / \mathrm{span}\{\mathbf{1}\}, \qquad \mathcal{T} := \left\{ x \in \mathbb{R}^K : \langle x, \mathbf{1} \rangle = 0 \right\}.$$

For each $u \in \mathbb{R}^K$, its equivalence class is

$$[u] := \{ u + c\mathbf{1} : c \in \mathbb{R} \}.$$

The dual pairing between $\mathcal{Q}$ and $\mathcal{T}$ is defined by

$$\langle [u], x \rangle := \langle u, x \rangle, \qquad x \in \mathcal{T}.$$

This is independent of the representative $u$, since $\langle \mathbf{1}, x \rangle = 0$.

## A.2 Local (Riemannian) Geometry of Quotient-Tangent Spaces

We first describe the local norms associated with the relative curvature assumptions. For this subsection, the definitions also make sense for $p = 0$.

**Definition A.1** (Local Norms)**.** *Let $p \geq 0$ and $\tilde{\pi} \in \mathrm{relint}(\Pi)$. For $x \in \mathcal{T}$ and $[u] \in \mathcal{Q}$, define*

$$\textit{Primal } p\textit{-local norm:} \quad \|x\|_{\tilde{\pi},p}^2 := \sum_{i=1}^K \frac{x_i^2}{\tilde{\pi}_i^p},$$

$$\textit{Dual } p\textit{-local norm:} \quad \|[u]\|_{*,\tilde{\pi},p}^2 := \min_{c \in \mathbb{R}} \sum_{i=1}^K \tilde{\pi}_i^p (u_i - c)^2.$$

The minimization over $c$ makes the dual norm independent of the representative of $[u]$. The minimizing scalar and the resulting quadratic form are

$$c^* = \frac{\sum_i \tilde{\pi}_i^p u_i}{\sum_i \tilde{\pi}_i^p}, \qquad \|[u]\|_{*,\tilde{\pi},p}^2 = u^\top L_p(\tilde{\pi}) u, \tag{10}$$

where

$$L_p(\tilde{\pi}) = \mathrm{diag}(\tilde{\pi}^p) - \frac{\tilde{\pi}^p (\tilde{\pi}^p)^\top}{\sum_i \tilde{\pi}_i^p}.$$

In particular, $L_p(\tilde{\pi})$ is positive semidefinite and has kernel $\mathrm{span}\{\mathbf{1}\}$. Thus, it defines a positive-definite quadratic form on $\mathcal{Q}$.

For $p = 0$, this reduces to

$$\|[u]\|_{*,\tilde{\pi},0}^2 = \|P_\perp u\|_2^2, \qquad P_\perp := I_K - \frac{1}{K}\mathbf{1}\mathbf{1}^\top = P_{\mathcal{T}}.$$

As the names suggest, the two local norms are dual to each other.

**Proposition A.1.** *For $p \geq 0$ and $\tilde{\pi} \in \operatorname{relint}(\Pi)$,*

$$\|[u]\|_{*,\tilde{\pi},p} = \sup_{x \in \mathcal{T}:\, \|x\|_{\tilde{\pi},p} \leq 1} \langle [u], x \rangle.$$

*Proof.* Let $c^*$ be the minimizing scalar in (10). For every $x \in \mathcal{T}$, the weighted Cauchy–Schwarz inequality gives

$$\begin{aligned}
\langle [u], x \rangle &= \sum_i (u_i - c^*) x_i \\
&\leq \left( \sum_i \tilde{\pi}_i^p (u_i - c^*)^2 \right)^{1/2} \left( \sum_i \frac{x_i^2}{\tilde{\pi}_i^p} \right)^{1/2} \\
&= \|[u]\|_{*,\tilde{\pi},p} \, \|x\|_{\tilde{\pi},p} \, .
\end{aligned}$$

This proves the upper bound on the supremum.

If $[u] = [0]$, both sides vanish. Otherwise, set

$$x_i = \frac{\tilde{\pi}_i^p (u_i - c^*)}{\|[u]\|_{*,\tilde{\pi},p}}.$$

The definition of $c^*$ gives $\sum_i x_i = 0$, so $x \in \mathcal{T}$. Moreover, $\|x\|_{\tilde{\pi},p} = 1$, and

$$\langle [u], x \rangle = \|[u]\|_{*,\tilde{\pi},p} \, .$$

Hence, this choice attains the supremum. □

### A.3 Mirror Maps and Bregman Divergences on the Quotient Space

We retain the ordinary Fenchel conjugate from the main text:

$$\psi^*(u) \coloneqq \max_{\pi \in \Pi} \{ \langle u, \pi \rangle - \psi(\pi) \}, \qquad u \in \mathbb{R}^K.$$

The maximum exists because $\Pi$ is compact and $\psi$ is proper and lower-semicontinuous. The Legendre assumption ensures that the maximizer lies in $\operatorname{relint}(\Pi)$: a finite boundary maximizer would give a relative subgradient of $\psi$ at the boundary, contradicting essential smoothness. Strict convexity then gives uniqueness. Consequently,

$$\nabla \psi^*(u) = \arg\max_{\pi \in \Pi} \{ \langle u, \pi \rangle - \psi(\pi) \} \in \operatorname{relint}(\Pi).$$

The conjugate itself is not translation-invariant:

$$\psi^*(u + c\mathbf{1}) = \psi^*(u) + c.$$

Its gradient, however, satisfies

$$\nabla \psi^*(u + c\mathbf{1}) = \nabla \psi^*(u).$$

Thus, the inverse mirror map depends only on $[u]$.

Since $\psi$ is defined on the simplex, its differential at an interior policy acts on tangent directions. We use $\nabla \psi(\pi) \in \mathbb{R}^K$ to denote any representative satisfying

$$d\psi(\pi)[x] = \langle \nabla \psi(\pi), x \rangle, \qquad x \in \mathcal{T}.$$

Any two such representatives differ by a multiple of $\mathbf{1}$. Hence, the **mirror map**

$$\pi \longmapsto [\nabla \psi(\pi)]$$

is well-defined. The first-order optimality condition gives

$$[\nabla \psi(\pi)] = [u] \quad \Longleftrightarrow \quad \pi = \nabla \psi^*(u).$$

Therefore, the mirror map is a bijection from $\operatorname{relint}(\Pi)$ to $\mathcal{Q}$, with inverse $[u] \mapsto \nabla \psi^*(u)$. When convenient, we write $\nabla \psi^*([u]) \coloneqq \nabla \psi^*(u)$ for this policy-valued inverse.

For completeness, a scalar conjugate on the quotient space is obtained by centering the simplex:

$$\begin{aligned}\bar{\psi}^*([u]) &:= \max_{\pi\in\Pi}\left\{\left\langle u, \pi - \frac{1}{K}\mathbf{1}\right\rangle - \psi(\pi)\right\} \\ &= \psi^*(u) - \left\langle u, \frac{1}{K}\mathbf{1}\right\rangle.\end{aligned}$$

This expression is independent of the representative $u$. Its gradient, viewed as an element of $\mathcal{T}$, is $\nabla\psi^*(u) - \mathbf{1}/K$. We continue to use the ordinary $\psi^*$ for expressions evaluated at reward vectors, including $F_\psi$ in the main text.

We recall the primal Bregman divergence and define its quotient dual counterpart under these conventions:

**Definition A.2** (Primal and Quotient Dual Bregman Divergences)**.** *For $\pi \in \Pi$ and $\pi' \in \mathrm{relint}(\Pi)$, the primal Bregman divergence is*

$$B_\psi(\pi, \pi') := \psi(\pi) - \psi(\pi') - \langle \nabla\psi(\pi'), \pi - \pi'\rangle.$$

*For $[u], [v] \in \mathcal{Q}$, define the quotient dual Bregman divergence by*

$$B_{\psi^*}([u], [v]) := \psi^*(u) - \psi^*(v) - \langle u - v, \nabla\psi^*(v)\rangle. \tag{11}$$

The primal divergence is independent of the gradient representative, since $\pi - \pi' \in \mathcal{T}$. Similarly, replacing $u$ and $v$ by $u + a\mathbf{1}$ and $v + b\mathbf{1}$ adds $a - b$ to the conjugate difference and to the linear term, so the dual divergence is well-defined on $\mathcal{Q}$. It is also precisely the Bregman divergence generated by $\bar{\psi}^*$. The primal divergence may equal $+\infty$ at a boundary policy, whereas the dual divergence is finite for every pair of quotient elements.

Lastly, the dual-constrained policy class has the following equivalent quotient representation.

**Definition A.3** (Dual-Constrained Policy Class)**.** *For $U \geq 0$, define*

$$\mathcal{W}_U := \{[u] \in \mathcal{Q} : \mathrm{osc}([u]) \leq 2U\}, \qquad \mathrm{osc}([u]) := \max_i u_i - \min_i u_i.$$

*Then,*

$$\Pi_{U,\psi} = \{\nabla\psi^*(u) : [u] \in \mathcal{W}_U\} = \left\{\nabla\psi^*(u) : u \in [-U, U]^K\right\}.$$

*As in Definition 2.1, define*

$$M_\psi(U) := \max_{\pi\in\Pi_{U,\psi}} \psi(\pi) - \min_{\pi\in\Pi} \psi(\pi), \qquad M_\psi := M_\psi(\eta R).$$

The two representations agree because a quotient element has oscillation at most $2U$ if and only if it has a representative in $[-U, U]^K$. Moreover, $\Pi_{U,\psi}$ is a compact subset of $\mathrm{relint}(\Pi)$, being the continuous image of $[-U, U]^K$ under $\nabla\psi^*$. Thus, the maximum defining $M_\psi(U)$ is finite and attained.

### A.4 Quotient Versions of Convex-Analytic Lemmas

We first verify Bregman duality under the preceding conventions.

**Lemma A.1** (Bregman Duality)**.** *For $\pi, \pi' \in \mathrm{relint}(\Pi)$,*

$$B_\psi(\pi, \pi') = B_{\psi^*}\left([\nabla\psi(\pi')], [\nabla\psi(\pi)]\right).$$

*Proof.* Choose representatives $u = \nabla\psi(\pi')$ and $v = \nabla\psi(\pi)$. Then $\nabla\psi^*(u) = \pi'$ and $\nabla\psi^*(v) = \pi$. Fenchel equality gives

$$\psi^*(u) = \langle u, \pi'\rangle - \psi(\pi'), \qquad \psi^*(v) = \langle v, \pi\rangle - \psi(\pi).$$

Substituting into (11),

$$\begin{aligned} B_{\psi^*}([u],[v]) &= \langle u, \pi' \rangle - \psi(\pi') - \langle v, \pi \rangle + \psi(\pi) - \langle u - v, \pi \rangle \\ &= \psi(\pi) - \psi(\pi') - \langle u, \pi - \pi' \rangle \\ &= B_\psi(\pi, \pi'). \end{aligned}$$

□

The next lemma converts the relative curvature bounds for $\psi$ into bounds for the dual Hessian. The inversion must be performed on $\mathcal{T}$, since the dual Hessian vanishes in the constant direction.

**Lemma A.2** (Convex-Smooth Duality)**.** *Suppose Assumptions 2 and 3 hold. Then $\psi^*$ is twice continuously differentiable on $\mathbb{R}^K$. For every $u \in \mathbb{R}^K$, writing $\pi = \nabla\psi^*(u)$,*

$$\frac{1}{C_\psi} L_p(\pi) \preceq \nabla^2\psi^*(u) \preceq \frac{1}{c_\psi} L_p(\pi). \tag{12}$$

*Here the inequalities hold as quadratic forms on $\mathbb{R}^K$, and all three matrices annihilate $\mathbf{1}$. Consequently, for $[u], [v] \in \mathcal{Q}$ and $\pi_z := \nabla\psi^*(v + z(u - v))$,*

$$B_{\psi^*}([u],[v]) \geq \frac{1}{C_\psi} \int_0^1 (1 - z) \, \|[u - v]\|^2_{*,\pi_z,p} \, dz.$$

*Proof.* Let $Q \in \mathbb{R}^{K \times (K-1)}$ have orthonormal columns spanning $\mathcal{T}$. At $\pi = \nabla\psi^*(u)$, define

$$A := Q^\top \nabla^2\psi(\pi) Q, \qquad G := Q^\top \mathrm{diag}(\pi)^{-p} Q.$$

Both matrices are positive definite, and the relative curvature assumptions give

$$c_\psi G \preceq A \preceq C_\psi G. \tag{13}$$

The first-order optimality condition is

$$Q^\top \nabla\psi(\nabla\psi^*(u)) = Q^\top u.$$

Since the primal Hessian is continuous and positive definite on $\mathcal{T}$, the inverse function theorem in tangent coordinates shows that the inverse mirror map is continuously differentiable. Differentiating the optimality condition therefore yields

$$\nabla^2\psi^*(u) = Q A^{-1} Q^\top.$$

In particular, $\psi^*$ is twice continuously differentiable.

We next verify the identity

$$Q G^{-1} Q^\top = L_p(\pi). \tag{14}$$

Indeed, for each $v \in \mathbb{R}^K$, both sides map $v$ to the unique $x \in \mathcal{T}$ satisfying

$$Q^\top \mathrm{diag}(\pi)^{-p} x = Q^\top v.$$

For $x = L_p(\pi)v$, this follows from

$$\mathrm{diag}(\pi)^{-p} L_p(\pi) v = v - \frac{\sum_i \pi_i^p v_i}{\sum_i \pi_i^p} \mathbf{1},$$

together with $L_p(\pi)v \in \mathcal{T}$.

Inverting (13) reverses the inequalities:

$$\frac{1}{C_\psi} G^{-1} \preceq A^{-1} \preceq \frac{1}{c_\psi} G^{-1}.$$

Multiplying by $Q$ and $Q^\top$ and using (14) proves (12).

Finally, Taylor's theorem with integral remainder gives

$$\begin{aligned} B_{\psi^*}([u],[v]) &= \int_0^1 (1-z)(u-v)^\top \nabla^2 \psi^*(v+z(u-v))(u-v)\, dz \\ &\geq \frac{1}{C_\psi} \int_0^1 (1-z)(u-v)^\top L_p(\pi_z)(u-v)\, dz \\ &= \frac{1}{C_\psi} \int_0^1 (1-z)\, \|[u-v]\|^2_{*,\pi_z,p}\, dz. \end{aligned}$$

□

The preceding lemma does not require coordinate-wise separability. We now use separability to compare the local metric at nearby dual vectors, which is the remaining geometric ingredient in the logarithmic lower bounds.

**Lemma A.3** (Local Metric Comparison)**.** *Suppose Assumptions 2 to 4 hold. Let $u, u^\circ \in \mathbb{R}^K$, and write*

$$\pi = \nabla\psi^*(u), \qquad \pi^\circ = \nabla\psi^*(u^\circ), \qquad \delta = \|u - u^\circ\|_\infty.$$

*Then, for every $i \in [K]$,*

$$\exp\left(-\frac{2\delta}{c_\psi}\right) \leq \frac{\pi_i}{\pi_i^\circ} \leq \exp\left(\frac{2\delta}{c_\psi}\right). \tag{15}$$

*Consequently,*

$$\exp\left(-\frac{2p\delta}{c_\psi}\right) L_p(\pi^\circ) \preceq L_p(\pi) \preceq \exp\left(\frac{2p\delta}{c_\psi}\right) L_p(\pi^\circ). \tag{16}$$

*In particular, if $\|u - u^\circ\|_\infty \leq 1$, then*

$$L_p(\pi) \succeq \exp\left(-\frac{2p}{c_\psi}\right) L_p(\pi^\circ).$$

*Proof.* By separability, write $\psi(\pi) = \sum_i \phi_i(\pi_i)$. Since both policies lie in $\operatorname{relint}(\Pi)$, their optimality conditions give scalars $\lambda, \lambda^\circ \in \mathbb{R}$ such that

$$\phi_i'(\pi_i) = u_i - \lambda, \qquad \phi_i'(\pi_i^\circ) = u_i^\circ - \lambda^\circ, \qquad i \in [K].$$

Let $h = u - u^\circ$. We first show that

$$\min_i h_i \leq \lambda - \lambda^\circ \leq \max_i h_i. \tag{17}$$

If $\lambda - \lambda^\circ > \max_i h_i$, then $\phi_i'(\pi_i) < \phi_i'(\pi_i^\circ)$ for every $i$. Since each $\phi_i'$ is strictly increasing, this implies $\pi_i < \pi_i^\circ$ for every $i$, contradicting that both policies sum to one. The other inequality follows by the same argument with the roles reversed.

It follows from (17) that

$$|\phi_i'(\pi_i) - \phi_i'(\pi_i^\circ)| = |h_i - (\lambda - \lambda^\circ)| \leq \max_j h_j - \min_j h_j \leq 2\delta.$$

On the other hand, the coordinate curvature bound in Assumption 4 and $p \geq 1$ imply

$$\phi_i''(x) \geq c_\psi x^{-p} \geq c_\psi x^{-1}, \qquad x \in (0,1).$$

Integrating between $\pi_i^\circ$ and $\pi_i$ therefore gives

$$c_\psi \left|\log \frac{\pi_i}{\pi_i^\circ}\right| \leq |\phi_i'(\pi_i) - \phi_i'(\pi_i^\circ)| \leq 2\delta.$$

Exponentiating proves (15).

Raising the coordinate bounds to the power $p$, for every $v \in \mathbb{R}^K$ and $c \in \mathbb{R}$, we obtain

$$\exp\left(-\frac{2p\delta}{c_\psi}\right) \sum_i (\pi_i^\circ)^p (v_i - c)^2 \leq \sum_i \pi_i^p (v_i - c)^2$$

$$\leq \exp\left(\frac{2p\delta}{c_\psi}\right) \sum_i (\pi_i^\circ)^p (v_i - c)^2.$$

Taking the minimum over $c$ in each expression and using (10) proves (16). □

The local comparison uses the coordinate curvature lower bound and separability, but not the discrepancy parameter $\omega$.

## B Proof of Theorem 3.1: Minimax Lower Bound (Full Information)

We divide the proof into two parts. The first establishes the logarithmic lower bound with effective dimension $D_{\mathrm{FI}}(\eta R/2; \psi)$. The second establishes a lower bound of order $R\min\{T, \sqrt{T\log K}\}$ under a sufficient condition on $\eta$.

Throughout, suppose Assumptions 1 to 4 hold, with $K \geq 2$, $T \geq 1$, and $\eta > 0$. All adversarial constructions below are distributions over reward sequences fixed before the interaction. For every learner, its worst-case expected regret is at least its average under any such distribution. Thus, it suffices to establish Bayesian lower bounds uniformly over learners.

We first recall a version of the matrix van Trees inequality (Gassiat and Stoltz, 2024; Gill and Levit, 1995; van Trees, 1968), together with a proof and sufficient regularity conditions for our application.

**Lemma B.1** (Matrix van Trees Inequality)**.** *Let $\theta$ have a continuously differentiable prior density $\mu$ on a bounded rectangle $\Theta \subset \mathbb{R}^K$ with nonempty interior. Suppose $\mu$ is positive in the interior, vanishes on $\partial\Theta$, and has finite prior Fisher information.*
*Suppose the conditional law of $X$ has a density $p(x \mid \theta)$ with respect to a common measure, with parameter-independent support. For almost every $x$, assume that $p(x \mid \theta)$ is continuously differentiable on a neighborhood of $\Theta$, and that the density and its first derivatives are dominated on $\Theta$ by integrable functions of $x$.*
*Define*

$$\mathcal{I}(\theta) := \mathbb{E}_{X|\theta}\left[\nabla_\theta \log p(X \mid \theta) \nabla_\theta \log p(X \mid \theta)^\top\right],$$
$$\mathcal{J}(\mu) := \mathbb{E}_\theta\left[\nabla \log \mu(\theta) \nabla \log \mu(\theta)^\top\right].$$

*If $M := \mathbb{E}_\theta[\mathcal{I}(\theta)] + \mathcal{J}(\mu)$ is finite and positive definite, then every bounded estimator $\widehat{\theta}(X)$ satisfies*

$$\mathbb{E}\left[(\widehat{\theta} - \theta)(\widehat{\theta} - \theta)^\top\right] \succeq M^{-1}.$$

*Independent auxiliary randomization may be included in $X$.*

*Proof.* Let

$$s := \nabla_\theta \log\big(p(X \mid \theta)\mu(\theta)\big)$$

be the joint score. Differentiating the normalization of the likelihood gives

$$\mathbb{E}_{X|\theta}[\nabla_\theta \log p(X \mid \theta)] = 0.$$

Hence the likelihood and prior score cross-terms vanish, and $\mathbb{E}[ss^\top] = M$.

Integration by parts in each parameter coordinate gives

$$\mathbb{E}[(\widehat{\theta} - \theta)s^\top] = I_K.$$

Indeed, the estimator is a function of $X$ alone, and the boundary term vanishes because $\mu$ vanishes on $\partial\Theta$. The stated domination assumptions justify interchanging the integrations.

Consequently,

$$\begin{aligned} 0 &\preceq \mathbb{E}\left[\big((\widehat{\theta} - \theta) - M^{-1}s\big)\big((\widehat{\theta} - \theta) - M^{-1}s\big)^\top\right] \\ &= \mathbb{E}\left[(\widehat{\theta} - \theta)(\widehat{\theta} - \theta)^\top\right] - M^{-1}, \end{aligned}$$

proving the claim. □

### B.1 Part I. Logarithmic Lower Bound

The main idea is to center a Bayesian adversary at an arbitrary reachable oracle policy and measure its estimation error through the local metric $L_p$. Maximizing over the anchor policy then gives the effective dimension.

**(i) Constructing the Bayesian adversary.** Fix any $\pi^\circ \in \Pi_{\eta R/2,\psi}$. Choose $u^\circ \in [-\eta R/2, \eta R/2]^K$ such that $\nabla\psi^*(u^\circ) = \pi^\circ$, and set $w^\circ = u^\circ/\eta$. Define

$$\Delta \coloneqq \min\left\{\frac{R}{4}, \frac{1}{\eta}\right\}, \qquad a_\psi \coloneqq \exp\left(-\frac{2p}{c_\psi}\right). \tag{18}$$

Draw the coordinates of $w - w^\circ$ independently from the cosine-squared density

$$\mu(z) = \frac{1}{\Delta}\cos^2\left(\frac{\pi z}{2\Delta}\right)\mathbb{1}[-\Delta \le z \le \Delta].$$

Denote the resulting product prior on $w$ by $\mu_w$. Conditional on $w$, draw all rewards independently over coordinates and rounds according to

$$\mathbb{P}(r_{t,i} = R \mid w) = \frac{1}{2} + \frac{w_i}{2R}, \qquad \mathbb{P}(r_{t,i} = -R \mid w) = \frac{1}{2} - \frac{w_i}{2R}.$$

Since $|w_i| \le R/2 + \Delta \le 3R/4$, these probabilities belong to $[1/8, 7/8]$. In particular, the rewards are admissible and satisfy $\mathbb{E}[r_t \mid w] = w$.

The construction is an oblivious distribution over reward sequences. We remark in advance that the following risk reduction applies under either feedback model; only the information calculation in **part (iv)** uses full-information feedback.

**(ii) Comparing with the mean-reward oracle.** For each $w$, define the regularized mean-reward oracle

$$\pi^\star(w) \coloneqq \nabla\psi^*(\eta w) = \arg\max_{\pi\in\Pi}\left\{\langle \eta w, \pi\rangle - \psi(\pi)\right\}.$$

The next lemma compares the learner with this oracle and then with the posterior Bayes act. The resulting estimator is bounded, while the learner's actual policies and observations are preserved.

**Lemma B.2** (Bayesian Regret as a Dual Estimation Risk)**.** *Suppose the adversary first draws a mean vector $w$ from a bounded prior and then draws rewards in $[-R, R]^K$ independently over time conditional on $w$, with $\mathbb{E}[r_t \mid w] = w$.*
*Let $\mathcal{H}_{t-1}$ be the learner's information before choosing its round-$t$ action, including the private randomization used to select $\pi_t$. The private randomization is generated independently of $w$ and the reward sequence. Define*

$$\widehat{w}_t \coloneqq \mathbb{E}[w \mid \mathcal{H}_{t-1}].$$

*Under either feedback model,*

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}_\eta(T)] &\ge \frac{1}{\eta}\sum_{t=1}^T \mathbb{E}[B_\psi(\pi_t, \pi^\star(w))] \\ &\ge \frac{1}{\eta}\sum_{t=1}^T \mathbb{E}[B_{\psi^*}([\eta w], [\eta\widehat{w}_t])].\end{aligned} \tag{19}$$

*If the prior is supported on a closed convex set, then $\widehat{w}_t$ belongs to that set almost surely.*

*Proof.* If the learner has infinite expected regret under this experiment, the claim is immediate. Otherwise, the expectations below are finite, since rewards are bounded and $\psi$ is bounded below on $\Pi$.

On every realized reward sequence, the best comparator in hindsight performs at least as well as $\pi^\star(w)$. Conditional on $w$ and $\mathcal{H}_{t-1}$, the next reward vector has mean $w$, while $\pi_t$ is fixed. Therefore,

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}_\eta(T)] &\ge \sum_{t=1}^T \mathbb{E}\left[J_\eta(\pi^\star(w); w) - J_\eta(\pi_t; w)\right] \\ &= \frac{1}{\eta}\sum_{t=1}^T \mathbb{E}\left[\psi(\pi_t) + \psi^*(\eta w) - \langle \eta w, \pi_t\rangle\right]\end{aligned}$$

$$= \frac{1}{\eta} \sum_{t=1}^{T} \mathbb{E}[B_\psi(\pi_t, \pi^\star(w))].$$

The identities follow from Fenchel equality at $\pi^\star(w)$. They also apply to boundary policies with finite regularizer value.

Next, conditional on $\mathcal{H}_{t-1}$,

$$\begin{aligned}
&\mathbb{E}[B_\psi(\pi_t, \pi^\star(w)) \mid \mathcal{H}_{t-1}] \\
&\quad = \psi(\pi_t) - \langle \eta\widehat{w}_t, \pi_t \rangle + \mathbb{E}[\psi^*(\eta w) \mid \mathcal{H}_{t-1}] \\
&\quad \geq \mathbb{E}[\psi^*(\eta w) \mid \mathcal{H}_{t-1}] - \psi^*(\eta\widehat{w}_t).
\end{aligned}$$

The minimum over the policy is attained by $\nabla\psi^*(\eta\widehat{w}_t)$. This compares conditional risks without replacing the learner's actions or the history generated by those actions.

Moreover,

$$\mathbb{E}\left[\langle w - \widehat{w}_t, \nabla\psi^*(\eta\widehat{w}_t)\rangle \,|\, \mathcal{H}_{t-1}\right] = 0.$$

Hence,

$$\begin{aligned}
&\mathbb{E}[B_{\psi^*}([\eta w], [\eta\widehat{w}_t]) \mid \mathcal{H}_{t-1}] \\
&\quad = \mathbb{E}[\psi^*(\eta w) \mid \mathcal{H}_{t-1}] - \psi^*(\eta\widehat{w}_t).
\end{aligned}$$

Taking expectations and summing proves the second inequality in (19). The assertion about $\widehat{w}_t$ follows from convexity of the prior support. □

**(iii) Comparing the local metric with the anchor.** Apply Lemma B.2 to the preceding Bayesian construction. Both $w$ and $\widehat{w}_t$ belong to $w^\circ + [-\Delta, \Delta]^K$, as does every point on the segment joining them. For $z \in [0, 1]$, define

$$u_{t,z} \coloneqq \eta\widehat{w}_t + z\eta(w - \widehat{w}_t), \qquad \pi_{t,z} \coloneqq \nabla\psi^*(u_{t,z}).$$

Since

$$\|u_{t,z} - u^\circ\|_\infty \leq \eta\Delta \leq 1,$$

Lemma A.3 gives

$$L_p(\pi_{t,z}) \succeq a_\psi L_p(\pi^\circ), \qquad z \in [0, 1].$$

Combining this with Lemma A.2,

$$\begin{aligned}
B_{\psi^*}([\eta w], [\eta\widehat{w}_t]) &\geq \frac{\eta^2}{C_\psi} \int_0^1 (1 - z)(w - \widehat{w}_t)^\top L_p(\pi_{t,z})(w - \widehat{w}_t)\, dz \\
&\geq \frac{a_\psi \eta^2}{2C_\psi} (w - \widehat{w}_t)^\top L_p(\pi^\circ)(w - \widehat{w}_t).
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}_\eta(T)] &\geq \frac{a_\psi \eta}{2C_\psi} \sum_{t=1}^{T} \mathbb{E}\left[(w - \widehat{w}_t)^\top L_p(\pi^\circ)(w - \widehat{w}_t)\right] \\
&= \frac{a_\psi \eta}{2C_\psi} \sum_{t=1}^{T} \mathrm{tr}\left(L_p(\pi^\circ)\mathbb{E}[(w - \widehat{w}_t)(w - \widehat{w}_t)^\top]\right). \qquad (20)
\end{aligned}$$

The Bregman-to-quadratic conversion uses relative smoothness. The comparison with the anchor uses separability and the coordinate curvature lower bound. In particular, (20) applies under both feedback models.

**(iv) Applying the van Trees inequality.** We now specialize to full-information feedback. Let

$$X_t = (r_1, \dots, r_{t-1}, \xi),$$

where $\xi$ collects the learner's private randomization, generated independently of $w$ and the rewards. The estimator $\widehat{w}_t$ is a measurable function of $X_t$. Relative to counting measure on the reward vectors and the law of $\xi$, the likelihood is

$$p_t(X_t \mid w) = \prod_{s=1}^{t-1} \prod_{i=1}^{K} \frac{1 + r_{s,i} w_i / R^2}{2}.$$

The support is parameter-independent, and the likelihood is continuously differentiable on a neighborhood of the prior rectangle. Its reward support is finite, and $|w_i| \le 3R/4$. Thus, the domination and differentiation requirements in Lemma B.1 hold. The private randomization contributes no Fisher information.

We first compute the prior Fisher information. For each coordinate,

$$\begin{aligned} \int_{-\Delta}^{\Delta} \mu(z) \left( \frac{d}{dz} \log \mu(z) \right)^2 dz &= \frac{\pi^2}{\Delta^3} \int_{-\Delta}^{\Delta} \sin^2 \left( \frac{\pi z}{2\Delta} \right) dz \\ &= \frac{\pi^2}{\Delta^2}. \end{aligned} \tag{21}$$

Translation by $w^\circ$ does not change this information. Independence of the prior coordinates and the zero mean of their scores give

$$\mathcal{J}(\mu_w) = \frac{\pi^2}{\Delta^2} I_K.$$

For the data information, one reward coordinate has Fisher information

$$\mathbb{E}_{r_{s,i} \mid w_i} \left[ \left( \frac{\partial}{\partial w_i} \log p(r_{s,i} \mid w_i) \right)^2 \right] = \frac{1}{R^2 - w_i^2} \le \frac{16}{7R^2}. \tag{22}$$

Conditional independence across coordinates and rounds makes the score cross-terms vanish. Consequently,

$$\mathcal{I}_t(w) = (t-1) \mathrm{diag} \left( \left\{ \frac{1}{R^2 - w_i^2} \right\}_{i \in [K]} \right) \preceq \frac{16(t-1)}{7R^2} I_K.$$

The posterior estimator is bounded by the prior rectangle. Applying Lemma B.1 and then the preceding information bounds,

$$\begin{aligned} \mathbb{E}[(w - \widehat{w}_t)(w - \widehat{w}_t)^\top] &\succeq \left( \mathbb{E}_w[\mathcal{I}_t(w)] + \mathcal{J}(\mu_w) \right)^{-1} \\ &\succeq \frac{1}{16(t-1)/(7R^2) + \pi^2/\Delta^2} I_K. \end{aligned}$$

Substituting into (20),

$$\begin{aligned} \mathbb{E}[\mathrm{Reg}_\eta(T)] &\ge \frac{a_\psi \eta}{2C_\psi} d_{\mathrm{FI}}(\pi^\circ) \sum_{t=1}^{T} \frac{1}{16(t-1)/(7R^2) + \pi^2/\Delta^2} \\ &\ge \frac{7 a_\psi \eta R^2}{32 C_\psi} d_{\mathrm{FI}}(\pi^\circ) \log \left( 1 + \frac{16 \Delta^2 T}{7 \pi^2 R^2} \right). \end{aligned}$$

The last inequality integrates the positive decreasing summand over $[0, T]$.

This lower bound holds uniformly over learners for every fixed anchor. Maximizing over the compact class $\Pi_{\eta R/2, \psi}$ and recalling (18), we obtain

$$\mathfrak{R}_\eta^{\mathrm{FI}}(T; \psi) \ge \frac{7 a_\psi}{32 C_\psi} \eta R^2 D_{\mathrm{FI}}(\eta R/2; \psi) \log \left( 1 + \frac{T}{7\pi^2 \max\{1, (\eta R/4)^2\}} \right). \tag{23}$$

This proves the logarithmic part of Theorem 3.1.

**(v) Uniform and two-arm anchors.** We record two direct evaluations of the effective dimensions. They will also be useful under bandit feedback.

**Proposition B.1** (Uniform Anchor). *At the uniform policy,*

$$L_p(\mathbf{1}/K) = K^{-p}P_\perp, \qquad d_{\mathrm{FI}}(\mathbf{1}/K) = (K-1)K^{-p}, \qquad d_{\mathrm{B}}(\mathbf{1}/K) = (K-1)K^{1-p}.$$

*Consequently, whenever* $\mathbf{1}/K \in \Pi_{U,\psi}$,

$$D_{\mathrm{FI}}(U;\psi) \geq (K-1)K^{-p}, \qquad D_{\mathrm{B}}(U;\psi) \geq (K-1)K^{1-p}.$$

*In particular, this holds for every* $U \geq 0$ *if the uniform policy minimizes* $\psi$.

*Proof.* Substituting $\pi_i = 1/K$ into the definition of $L_p$ gives $L_p(\mathbf{1}/K) = K^{-p}P_\perp$. Its trace is $(K-1)K^{-p}$, and each diagonal entry equals $(K-1)K^{-p-1}$. Therefore,

$$d_{\mathrm{B}}(\mathbf{1}/K) = \left(K\sqrt{(K-1)K^{-p-1}}\right)^2 = (K-1)K^{1-p}.$$

The bounds for $D_{\mathrm{FI}}$ and $D_{\mathrm{B}}$ follow by maximization over the reachable policies. If the uniform policy minimizes $\psi$, it equals $\nabla\psi^*(0)$ and is reachable at every radius. □

**Proposition B.2** (Two-Arm Anchor). *Let* $K \geq 3$, *and define*

$$\pi^\circ := \left(\frac{3}{8}, \frac{3}{8}, \underbrace{\frac{1}{4(K-2)}, \ldots, \frac{1}{4(K-2)}}_{K-2 \text{ coordinates}}\right).$$

*Write* $\psi(\pi) = \sum_{i=1}^K \phi_i(\pi_i)$. *If*

$$\eta R \geq \max_{i\in[K]} \phi_i'(\pi_i^\circ) - \min_{i\in[K]} \phi_i'(\pi_i^\circ), \tag{24}$$

*then*

$$D_{\mathrm{FI}}(\eta R/2;\psi) \geq \left(\frac{3}{8}\right)^p, \qquad D_{\mathrm{B}}(\eta R/2;\psi) \geq 2\left(\frac{3}{8}\right)^p.$$

*Proof.* By the mirror-map optimality condition, $\pi^\circ \in \Pi_{\eta R/2,\psi}$ if and only if there exists $\lambda \in \mathbb{R}$ such that

$$\phi_i'(\pi_i^\circ) - \lambda \in [-\eta R/2, \eta R/2], \qquad i \in [K].$$

Subtracting the midpoint of the largest and smallest coordinate derivatives shows that this is equivalent to (24). Its right-hand side is finite because $\pi^\circ \in \mathrm{relint}(\Pi)$ and each $\phi_i \in C^2((0,1))$.

Let $b = (3/8)^p$ and $S = \sum_i (\pi_i^\circ)^p$. Since $S \geq 2b$,

$$(L_p(\pi^\circ))_{11} = (L_p(\pi^\circ))_{22} = b - \frac{b^2}{S} \geq \frac{b}{2}.$$

Therefore,

$$d_{\mathrm{FI}}(\pi^\circ) \geq b, \qquad d_{\mathrm{B}}(\pi^\circ) \geq \left(2\sqrt{b/2}\right)^2 = 2b.$$

Maximizing over the reachable policies proves the claim. □

For $K = 2$, the uniform-anchor calculation already gives dimension coefficients depending only on $p$, whenever that anchor is reachable. For $K \geq 3$, (24) supplies a sufficient condition for a $K$-independent coefficient in the full-information logarithmic lower bound.

### B.2 Part II. Square-Root Lower Bound

We again compare with the regularized mean-reward oracle, now under an adversary with a hidden slightly better arm. We first state a comparison that applies to both feedback models.

**Lemma B.3** (Comparison with the Regularized Oracle)**.** *Draw $I$ uniformly from $[K]$. Conditional on $I$, draw independent reward vectors in $[-R, R]^K$ with mean $\Delta e_I$, where $0 < \Delta \le R$. Let $x = \eta\Delta$ and recall*

$$F_\psi(x) = \frac{1}{K}\sum_{i=1}^{K}\left\{\psi^*(xe_i) - \psi^*(0)\right\}.$$

*Every learner under either feedback model satisfies*

$$\mathbb{E}[\mathrm{Reg}_\eta(T)] \ge \frac{T}{\eta}F_\psi(x) - \Delta\sum_{t=1}^{T}\mathbb{E}[\pi_{t,I}]. \tag{25}$$

*Proof.* The regularized mean-reward oracle is $\pi_I^\star = \nabla\psi^*(xe_I)$. Let $\bar\psi(\pi) \coloneqq \psi(\pi) - \min_{\pi'\in\Pi}\psi(\pi')$. Since $\psi^*(0) = -\min_{\pi'\in\Pi}\psi(\pi')$, Fenchel equality gives

$$B_\psi(\pi, \pi_I^\star) = \bar\psi(\pi) + \psi^*(xe_I) - \psi^*(0) - x\pi_I.$$

The first inequality of Lemma B.2 applies with $w = \Delta e_I$. Averaging the displayed identity over $I$ and the learner's history, summing over $t$, and using $\bar\psi \ge 0$ yields (25). □

We next bound the mass that a full-information learner can place on the hidden arm.

**Lemma B.4** (Full-Information Hidden-Arm Testing)**.** *Let*

$$\Delta = \frac{R}{4}\min\left\{1, \sqrt{\frac{\log K}{T}}\right\}.$$

*Draw $I$ uniformly from $[K]$. Conditional on $I$, draw rewards independently over coordinates and rounds, with values in $\{-R, R\}$ and success probability $1/2 + \Delta/(2R)$ on arm $I$, and $1/2$ on every other arm. Then every full-information learner satisfies*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\pi_{t,I}] \le \frac{3}{4}.$$

*Proof.* Let $P_i$ be the joint law of the full reward sequence and the learner's independent private randomization when $I = i$. Let $P_0$ be the corresponding law when every reward coordinate is unbiased. Define

$$L_i \coloneqq \frac{dP_i}{dP_0}, \qquad \bar\pi_i \coloneqq \frac{1}{T}\sum_{t=1}^{T}\pi_{t,i}.$$

For each realized reward sequence and private randomization, $\bar\pi \in \Pi$. In particular, $\sum_i \bar\pi_i^2 \le 1$.

Under $P_0$, the rewards are independent unbiased signs multiplied by $R$. The private randomization has the same law under every instance, so it cancels from the likelihood ratio. Therefore,

$$L_i = \prod_{t=1}^{T}\left(1 + \frac{\Delta r_{t,i}}{R^2}\right), \qquad \mathbb{E}_0[L_i^2] = \left(1 + \frac{\Delta^2}{R^2}\right)^T.$$

Applying Cauchy–Schwarz jointly over the instance index and the observation space,

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\pi_{t,I}] &= \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_i[\bar\pi_i] \\ &= \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_0[L_i\bar\pi_i]\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{K}\left(\sum_{i=1}^{K}\mathbb{E}_0[L_i^2]\right)^{1/2}\left(\mathbb{E}_0\left[\sum_{i=1}^{K}\bar{\pi}_i^2\right]\right)^{1/2} && \text{(Cauchy–Schwarz inequality)}\\
&\leq \frac{1}{\sqrt{K}}\left(1+\frac{\Delta^2}{R^2}\right)^{T/2}\\
&\leq \frac{1}{\sqrt{K}}\exp\left(\frac{T\Delta^2}{2R^2}\right)\\
&\leq K^{-15/32} \leq 2^{-15/32} < \frac{3}{4}.
\end{aligned}
$$

The penultimate bound follows from

$$
\frac{T\Delta^2}{R^2} = \frac{1}{16}\min\{T, \log K\} \leq \frac{\log K}{16}.
$$

This proves the claim for every $K \geq 2$. □

**Completing the square-root lower bound.** Use the preceding hidden-arm experiment with

$$
\Delta = \frac{R}{4}\min\left\{1, \sqrt{\frac{\log K}{T}}\right\}.
$$

Combining Lemmas B.3 and B.4 and then taking the infimum over learners gives

$$
\mathfrak{R}_\eta^{\mathrm{FI}}(T;\psi) \geq \frac{T}{\eta}F_\psi(\eta\Delta) - \frac{3}{4}\Delta T.
$$

Consequently,

$$
F_\psi(\eta\Delta) \geq \frac{7}{8}\eta\Delta \quad\Longrightarrow\quad \mathfrak{R}_\eta^{\mathrm{FI}}(T;\psi) \geq \frac{\Delta T}{8} = \frac{R}{32}\min\{T, \sqrt{T\log K}\}. \tag{26}
$$

In particular, if $T \geq \log K$, the sufficient condition becomes

$$
F_\psi\left(\frac{\eta R}{4}\sqrt{\frac{\log K}{T}}\right) \geq \frac{7\eta R}{32}\sqrt{\frac{\log K}{T}},
$$

and the resulting lower bound is

$$
\mathfrak{R}_\eta^{\mathrm{FI}}(T;\psi) \geq \frac{R}{32}\sqrt{T\log K}.
$$

For $T < \log K$, the same argument gives $\mathfrak{R}_\eta^{\mathrm{FI}}(T;\psi) \geq RT/32$ whenever $F_\psi(\eta R/4) \geq 7\eta R/32$.

**A finite sufficient threshold.** The condition in (26) holds for sufficiently large $\eta$ for every fixed regularizer in our class. To see this directly, define

$$
q^{(i)} \coloneqq \frac{15}{16}e_i + \frac{1}{16K}\mathbf{1}, \qquad C_0 \coloneqq \frac{1}{K}\sum_{i=1}^{K}\left\{\psi(q^{(i)}) - \min_{\pi\in\Pi}\psi(\pi)\right\}.
$$

Every $q^{(i)}$ is an interior policy, so $0 \leq C_0 < \infty$. Evaluating each conjugate at $q^{(i)}$ gives

$$
\begin{aligned}
F_\psi(x) &\geq \frac{1}{K}\sum_{i=1}^{K}\left\{xq_i^{(i)} - \psi(q^{(i)}) + \min_{\pi\in\Pi}\psi(\pi)\right\}\\
&\geq \frac{15}{16}x - C_0.
\end{aligned}
$$

Thus, $x \geq 16C_0$ implies $F_\psi(x) \geq 7x/8$. Taking $x = \eta\Delta$ shows that the sufficient condition is non-vacuous, including for regularizers that diverge at the boundary.

# C Proof of Theorem 3.2: Regret Upper Bound (Full Information) via C-FTRL

We prove the full-information regret bound for C-FTRL (Algorithm 1).

**Algorithm 1** Composite Follow-The-Regularized-Leader (C-FTRL)

**Require:** Coefficient $\xi > 0$ and convex feasible set $\mathcal{C}_\psi$; by default, $\mathcal{C}_\psi = \Pi$.

1: For each $t \geq 1$, set $\lambda_t = t/\eta$ and

$$\pi_t \in \arg\max_{\pi \in \mathcal{C}_\psi} \left\{ \sum_{s=1}^{t-1} \langle \pi, r_s \rangle - \lambda_t \psi(\pi) - \xi H(\pi) \right\}.$$

Appendix C.1 proves Theorem 3.2. The supporting lemmas in Appendix C.2 also yield the effective-dimension refinement in Proposition F.3.

## C.1 Generic Full Information Upper Bound

We first prove a slightly more general statement than the one in Theorem 3.2. For any $\xi > 0$, consider **C-FTRL** over $\Pi$ with $\lambda_t = t/\eta$. We claim that

$$\mathrm{Reg}_\eta(T; \{r_t\}) \leq \min\{\eta c_\psi^{-1} R^2, 2R\} + \xi \log K + \eta c_\psi^{-1} R^2 \sum_{t=1}^{T} \frac{1}{t + 2\eta c_\psi^{-1} \xi}. \tag{27}$$

We first prove the claim, and later deduce how the claim reduces to the theorem statement. Define the shifted regularizer

$$\bar{\psi}(\pi) \coloneqq \psi(\pi) - \min_{\pi' \in \Pi} \psi(\pi').$$

This changes neither the updates nor the regret, and ensures that $\bar{\psi} \geq 0$. Denoting $\ell_s(\pi) \coloneqq \langle -r_s, \pi \rangle$ and utilizing the composite FTRL regularizer

$$\bar{\psi}_t(\pi) \coloneqq \lambda_t \bar{\psi}(\pi) + \xi \left( H(\pi) + \log K \right), \tag{28}$$

we can write

$$\pi_t \in \arg\min_{\pi \in \Pi} F_t(\pi), \qquad F_t(\pi) \coloneqq \bar{\psi}_t(\pi) + \sum_{s=1}^{t-1} \ell_s(\pi).$$

The Legendre property ensures that these minimizers lie in $\mathrm{relint}(\Pi)$. First, we invoke the FTRL stability lemma, which we restate in Lemma C.1 using notation consistent with ours. Adding $\eta^{-1} \sum_{t=1}^{T} \{\bar{\psi}(\pi_t) - \bar{\psi}(\pi^\star)\}$ to its linear-regret bound gives

$$\begin{aligned} \mathrm{Reg}_\eta(T; \{r_t\}) \leq \bar{\psi}_{T+1}(\pi^\star) - \frac{T}{\eta} \bar{\psi}(\pi^\star) - \min_{\pi \in \Pi} \bar{\psi}_1(\pi) \\ + \sum_{t=1}^{T} \left\{ \bar{\psi}_t(\pi_{t+1}) - \bar{\psi}_{t+1}(\pi_{t+1}) + \eta^{-1} \bar{\psi}(\pi_t) \right\} \\ + \sum_{t=1}^{T} \sup_{\pi \in \Pi} \left\{ \langle r_t, \pi - \pi_t \rangle - B_{\bar{\psi}_t}(\pi, \pi_t) \right\}. \end{aligned}$$

Since $\lambda_{t+1} - \lambda_t = \eta^{-1}$ and $\xi$ is constant, the middle sum telescopes to $\eta^{-1}\{\bar{\psi}(\pi_1) - \bar{\psi}(\pi_{T+1})\}$. Moreover, the initialization gives

$$\min_{\pi \in \Pi} \bar{\psi}_1(\pi) = \eta^{-1} \bar{\psi}(\pi_1) + \xi \left( H(\pi_1) + \log K \right).$$

Thus the initial term cancels even when $\pi_1$ does not separately minimize $\psi$. Since $0 \leq H(\pi) + \log K \leq \log K$, we obtain

$$\mathrm{Reg}_\eta(T; \{r_t\}) \leq \frac{\bar{\psi}(\pi^\star)}{\eta} + \xi \log K + \sum_{t=1}^{T} \sup_{\pi \in \Pi} \left\{ \langle r_t, \pi - \pi_t \rangle - B_{\bar{\psi}_t}(\pi, \pi_t) \right\}. \tag{29}$$

We now control stability. As $p \geq 1$, relative strong convexity and Cauchy–Schwarz give, for every $v \in \mathcal{T}$,

$$v^\top \nabla^2 \psi(\pi) v \geq c_\psi \sum_i \frac{v_i^2}{\pi_i^p} \geq c_\psi \|v\|_1^2, \qquad v^\top \nabla^2 H(\pi) v \geq \|v\|_1^2.$$

Taylor's integral formula therefore yields

$$B_{\bar{\psi}_t}(\pi, \pi_t) \geq \frac{1}{2}(c_\psi \lambda_t + \xi)\|\pi - \pi_t\|_1^2.$$

For boundary points with finite regularizer value, the same inequality follows by an interior limit. Using $\|r_t\|_\infty \leq R$ and maximizing the resulting scalar quadratic, we obtain

$$\sup_{\pi \in \Pi} \left\{ \langle r_t, \pi - \pi_t \rangle - B_{\bar{\psi}_t}(\pi, \pi_t) \right\} \leq \frac{R^2}{2(c_\psi \lambda_t + \xi)} \leq \frac{\eta c_\psi^{-1} R^2}{t + 2\eta c_\psi^{-1} \xi}. \tag{30}$$

Lemma C.3 gives $\bar{\psi}(\pi^\star)/\eta \leq \min\{\eta c_\psi^{-1} R^2, 2R\}$. Substituting this and (30) into (29) proves (27).

We now set $\xi$ as in (5). Since $\xi \log K \leq \eta c_\psi^{-1} R^2$, the harmonic-sum bound gives

$$\mathrm{Reg}_\eta(T; \{r_t\}) \leq \eta c_\psi^{-1} R^2 (3 + \log T).$$

To obtain the square-root bound, consider the two choices in the minimum. If $\xi = R\sqrt{T/(2\log K)}$, the entropy cost and the stability sum in (27) are each at most $R\sqrt{T \log K/2}$. If $\xi = \eta c_\psi^{-1} R^2 / \log K$, the entropy cost is again at most this quantity, and integration gives

$$\begin{aligned} \eta c_\psi^{-1} R^2 \sum_{t=1}^T \frac{1}{t + 2\eta c_\psi^{-1} \xi} &\leq \eta c_\psi^{-1} R^2 \log\left(1 + \frac{c_\psi^2 T \log K}{2\eta^2 R^2}\right) \\ &\leq R\sqrt{\frac{T \log K}{2}}, \end{aligned}$$

where $\log(1 + x^2) \leq x$ for $x \geq 0$. Using $\bar{\psi}(\pi^\star)/\eta \leq 2R$ in both cases proves

$$\mathrm{Reg}_\eta(T; \{r_t\}) \leq R\sqrt{2T \log K} + 2R.$$

Since both bounds hold for the same run and every reward sequence, they prove Theorem 3.2. □

### C.2 Supporting Lemmas and the Effective-Dimension Refinement

We first state the standard FTRL decomposition in a form that keeps the stability term explicit; see Orabona (2025, Chapter 7) and McMahan (2017, Section 5).

**Lemma C.1** (FTRL stability). *Let $\mathcal{C}_\psi \subseteq \Pi$ be nonempty, compact, and convex. Let $\bar{\psi}_1, \ldots, \bar{\psi}_{T+1}$ be proper, lower-semicontinuous convex regularizers, differentiable on* $\mathrm{relint}(\Pi)$*, and let $\ell_t(\pi) = \langle g_t, \pi \rangle$. Define*

$$F_t(\pi) := \bar{\psi}_t(\pi) + \sum_{s=1}^{t-1} \ell_s(\pi), \qquad \pi_t \in \arg\min_{\pi \in \mathcal{C}_\psi} F_t(\pi).$$

*Assume these minima are attained in* $\mathrm{relint}(\Pi)$*. Then, for every $\pi^\star \in \mathcal{C}_\psi$ with finite regularizer values,*

$$\begin{aligned} \sum_{t=1}^T \{\ell_t(\pi_t) - \ell_t(\pi^\star)\} \leq \bar{\psi}_{T+1}(\pi^\star) - \min_{\pi \in \mathcal{C}_\psi} \bar{\psi}_1(\pi) \\ + \sum_{t=1}^T \{\bar{\psi}_t(\pi_{t+1}) - \bar{\psi}_{t+1}(\pi_{t+1})\} \end{aligned}$$

$$+\sum_{t=1}^{T}\sup_{\pi\in\mathcal{C}_\psi}\left\{\langle g_t,\pi_t-\pi\rangle-B_{\bar{\psi}_t}(\pi,\pi_t)\right\}.$$

*Proof.* By constrained first-order optimality at $\pi_t$,

$$F_t(\pi)-F_t(\pi_t)\geq B_{\bar{\psi}_t}(\pi,\pi_t),\qquad \pi\in\mathcal{C}_\psi.$$

Applying this at $\pi_{t+1}$ and rearranging gives

$$\begin{aligned}\ell_t(\pi_t)\leq F_{t+1}(\pi_{t+1})-F_t(\pi_t)+\bar{\psi}_t(\pi_{t+1})-\bar{\psi}_{t+1}(\pi_{t+1})\\+\sup_{\pi\in\mathcal{C}_\psi}\left\{\langle g_t,\pi_t-\pi\rangle-B_{\bar{\psi}_t}(\pi,\pi_t)\right\}.\end{aligned}$$

Summing over $t$, using $F_{T+1}(\pi_{T+1})\leq F_{T+1}(\pi^\star)$, and subtracting $\sum_t \ell_t(\pi^\star)$ proves the claim. □

**Lemma C.2** (Dual domain restriction)**.** *The optimal regularized policy in hindsight satisfies $\pi^\star\in\Pi_{\eta R,\psi}$. Suppose additionally that $\psi$ is separable as in Assumption 4.*

1. *If $\psi$ is minimized at $\mathbf{1}/K$, then every policy*

$$z_t(s)\coloneqq\operatorname*{arg\,max}_{\pi\in\Pi}\left\{\left\langle\sum_{u<t}r_u+sr_t,\pi\right\rangle-\lambda_t\psi(\pi)-\xi H(\pi)\right\},\qquad 0\leq s\leq 1,$$

   *belongs to $\Pi_{\eta R,\psi}$, where $\lambda_t=t/\eta$.*
2. *For $\psi=D_{\mathrm{KL}}(\cdot,\rho)$, $\Pi_{\eta R,\psi}$ is convex.*
3. *In general,* $\operatorname{conv}\Pi_{\eta R,\psi}\subseteq\Pi_{2\eta R,\psi}$.

*All three feasible sets used in the refinements contain $\pi^\star$ and the minimizer of $\psi$.*

*Proof.* The comparator can be written as

$$\pi^\star=\nabla\psi^*\left(\frac{\eta}{T}\sum_{t=1}^{T}r_t\right).$$

The argument has $\ell_\infty$-norm at most $\eta R$, proving the first assertion. Write $\psi(\pi)=\sum_i\phi_i(\pi_i)$. By the first-order conditions, membership in $\Pi_{U,\psi}$ is equivalent to the coordinate derivatives $\phi_i'(\pi_i)$ having range at most $2U$. For part 1, their values at $1/K$ agree. Choose $i,j$ attaining the largest and smallest coordinate derivatives at $z_t(s)$. Since the derivatives increase and the coordinates sum to one, $z_{t,i}(s)\geq 1/K\geq z_{t,j}(s)$. The optimality equations give

$$\begin{aligned}\lambda_t\{\phi_i'(z_{t,i}(s))-\phi_j'(z_{t,j}(s))\}&=\sum_{u<t}(r_{u,i}-r_{u,j})+s(r_{t,i}-r_{t,j})-\xi\log\frac{z_{t,i}(s)}{z_{t,j}(s)}\\&\leq 2tR.\end{aligned}$$

Since $\lambda_t=t/\eta$, the derivative range is at most $2\eta R$. For part 2, membership is equivalent to the linear inequalities

$$\frac{\pi_i}{\rho_i}\leq e^{2\eta R}\frac{\pi_j}{\rho_j},\qquad i,j\in[K].$$

These inequalities and the simplex constraint also force positivity, so they describe exactly $\Pi_{\eta R,\psi}$. For part 3, let $\pi_\psi$ minimize $\psi$. The derivatives $\phi_i'(\pi_{\psi,i})$ have a common value. At any $\pi\in\Pi_{\eta R,\psi}$, the largest derivative is at least this value and the smallest is at most it; otherwise every coordinate would be larger, or every coordinate smaller, than that of $\pi_\psi$. Thus each derivative lies within $2\eta R$ of the common value. Coordinate-wise monotonicity preserves these bounds under convex combinations, giving derivative range at most $4\eta R$. Finally, $\pi_\psi=\nabla\psi^*(0)$ belongs to every dual-constrained class, as does the comparator at radius $\eta R$. □

**Lemma C.3** (Regularizer gap of the hindsight comparator)**.** *Let $\psi$ be $(p, c_\psi)$-RSC and Legendre, $\pi^\star = \arg\max_{\pi\in\Pi} \sum_{t=1}^T \{\langle r_t, \pi\rangle - \eta^{-1}\psi(\pi)\}$ be the best policy in hindsight, and $\pi_\psi = \arg\min_{\pi\in\Pi} \psi(\pi)$ be the minimizer of the regularizer. Then, we have*

$$\psi(\pi^\star) - \psi(\pi_\psi) \le \min\left\{c_\psi^{-1}\eta^2 R^2 D_{\mathrm{FI}}(\eta R; \psi),\ 2\eta R,\ M_\psi\right\}.$$

*In particular, $D_{\mathrm{FI}}(\eta R; \psi) \le 1$ gives $\bar\psi(\pi^\star)/\eta \le \min\{\eta c_\psi^{-1} R^2, 2R\}$.*

*Proof.* We first record the inverse-curvature estimate. For $\pi = \nabla\psi^*(u)$ and any $v \in \mathbb{R}^K$,

$$\begin{aligned} v^\top \nabla^2\psi^*(u) v &= \sup_{d\in\mathcal{T}} \left\{2\langle v, d\rangle - d^\top \nabla^2\psi(\pi) d\right\} \\ &\le c_\psi^{-1} \min_{c\in\mathbb{R}} \sum_i \pi_i^p (v_i - c)^2 = c_\psi^{-1} v^\top L_p(\pi) v. \end{aligned} \tag{31}$$

The inverse is taken on the tangent space $\mathcal{T}$; the ambient conjugate removes the constant reward direction. The inequality follows from RSC and the dual-norm identity in Appendix A. For $\|v\|_\infty \le U$, the definition of $L_p$ gives

$$v^\top L_p(\pi) v = \frac{\sum_{i<j} \pi_i^p \pi_j^p (v_i - v_j)^2}{\sum_i \pi_i^p} \le 2U^2 \operatorname{tr} L_p(\pi). \tag{32}$$

Let $u = (\eta/T)\sum_t r_t$ and $\pi(s) = \nabla\psi^*(su)$. Then $\pi(0) = \pi_\psi$, $\pi(1) = \pi^\star$, and $\pi(s) \in \Pi_{\eta R,\psi}$. Since $\nabla\psi(\pi(s))$ equals $su$ up to a multiple of $\mathbf{1}$, while $\mathbf{1}^\top \pi'(s) = 0$, we obtain

$$\begin{aligned} \psi(\pi^\star) - \psi(\pi_\psi) &= \int_0^1 s\, u^\top \nabla^2\psi^*(su) u\, ds \\ &\le c_\psi^{-1}\eta^2 R^2 D_{\mathrm{FI}}(\eta R; \psi). \end{aligned}$$

For the second bound, comparator optimality gives

$$\psi(\pi^\star) - \psi(\pi_\psi) \le \frac{\eta}{T}\sum_{t=1}^T \langle r_t, \pi^\star - \pi_\psi\rangle \le 2\eta R.$$

The remaining bounds follow from Lemma C.2, the definition of $M_\psi$, and $\operatorname{tr} L_p(\pi) \le \sum_i \pi_i^p \le 1$. □

**Refined full-information bound.** We now retain the dimension factor in the argument above. Let $U = \eta R$ for a uniform minimizer or KL regularization, and $U = 2\eta R$ for a general separable regularizer. Use the corresponding feasible set in Lemma C.2. In (28), replace $H + \log K$ by $H - \min_{\mathcal{C}_\psi} H$; its range is still at most $\log K$. At any relevant policy $\pi \in \Pi_{U,\psi}$, (31)–(32) give

$$r_t^\top \left(\nabla^2\psi(\pi)|_{\mathcal{T}}\right)^{-1} r_t \le 2c_\psi^{-1} R^2 D_{\mathrm{FI}}(U; \psi).$$

Here and below projections onto $\mathcal{T}$ are understood. For Shannon entropy, the same inverse quadratic form is the variance of the coordinates of $r_t$ under $\pi$, and is at most $R^2$. Cauchy–Schwarz therefore implies, for all $d \in \mathcal{T}$,

$$d^\top \nabla^2\bar\psi_t(\pi) d \ge \left(\frac{c_\psi \lambda_t}{2D_{\mathrm{FI}}(U; \psi)} + \xi\right)\frac{\langle r_t, d\rangle^2}{R^2}. \tag{33}$$

For the two constrained domains, every segment in $\mathcal{C}_\psi$ lies in $\Pi_{U,\psi}$ by Lemma C.2. Integrating (33) and maximizing a scalar quadratic gives

$$\sup_{\pi\in\mathcal{C}_\psi} \left\{\langle r_t, \pi - \pi_t\rangle - B_{\bar\psi_t}(\pi, \pi_t)\right\} \le \frac{c_\psi^{-1} R^2 D_{\mathrm{FI}}(U; \psi)}{\lambda_t + 2c_\psi^{-1}\xi D_{\mathrm{FI}}(U; \psi)}. \tag{34}$$

For the uniform-minimizer case, use the path $z_t(s)$ of Lemma C.2, which stays in $\Pi_{\eta R,\psi}$. The supremum in (34) equals

$$\int_0^1 (1-s)\, r_t^\top \left(\nabla^2\bar\psi_t(z_t(s))|_{\mathcal{T}}\right)^{-1} r_t\, ds.$$

This identity follows by differentiating the maximized objective along the reward segment and applying Taylor's integral formula. Inequality (33) gives the same bound. All paths and inverse forms are interior by the Legendre and RSC assumptions. The constrained sets are compact subsets of $\operatorname{relint}(\Pi)$, so their updates are also well-defined. Finally, Lemma C.3 bounds the comparator cost by $\min\{\eta c_\psi^{-1} R^2 D_{\mathrm{FI}}(U;\psi), 2R\}$. Thus (27) holds with each occurrence of $c_\psi^{-1}$ replaced by $c_\psi^{-1} D_{\mathrm{FI}}(U;\psi)$. Choosing $\xi$ as in Proposition F.3 and summing as in Appendix C.1 gives the stated bound for each of the three update domains.

## D Proof of Theorem 4.1: Minimax Lower Bound (Bandit)

We reuse the Bayesian constructions and geometric reductions from Appendix B. For the logarithmic bound, the only change is the data Fisher information available under bandit feedback. For the square-root bound, we reuse the regularized oracle comparison and establish the corresponding bandit testing bound.

Throughout, suppose Assumptions 1 to 4 hold, with $K \geq 2$, $T \geq 1$, and $\eta > 0$.

### D.1 Part I. Logarithmic Lower Bound

Fix any anchor $\pi^\circ \in \Pi_{\eta R/2,\psi}$ and use exactly the prior and reward construction from Appendix B.1. In particular, retain $\Delta$ and $a_\psi$ from (18). The posterior estimator $\widehat{w}_t = \mathbb{E}[w \mid \mathcal{H}_{t-1}]$ is defined using the learner's actual bandit history. By Lemma B.2 and the local metric comparison, the shared risk bound (20) applies without modification, which we recall here:

$$\mathbb{E}[\mathrm{Reg}_\eta(T)] \geq \frac{a_\psi \eta}{2C_\psi} \sum_{t=1}^{T} \mathrm{tr}\left(L_p(\pi^\circ)\mathbb{E}[(w-\widehat{w}_t)(w-\widehat{w}_t)^\top]\right).$$

**(i) Fisher information under adaptive sampling.** Let $i_s$ be the arm sampled at round $s$, and let $Y_s = r_{s,i_s}$ be the observed reward. Represent all private randomization of the learner, including arm sampling, by a seed $\xi$ independent of $w$ and the rewards. Conditional on $\xi$, the learner's policies and sampled arms are deterministic functions of its preceding observations. Consider the statistical data

$$X_t := \left(\xi, (i_s, Y_s)_{s=1}^{t-1}\right).$$

The posterior estimator $\widehat{w}_t$ remains the one defined from the learner's actual information; it is a measurable function of $X_t$, which suffices for applying Lemma B.1.

Relative to the law of $\xi$ and counting measure on the finite action and reward histories, the likelihood is

$$p_t(X_t \mid w) = \Gamma_t(X_t) \prod_{s=1}^{t-1} \frac{1 + Y_s w_{i_s}/R^2}{2}, \tag{35}$$

where $\Gamma_t(X_t) \in \{0,1\}$ indicates compatibility with the learner's decision rules. Crucially, $\Gamma_t$ does not depend on $w$. Since $|w_i| \leq 3R/4$, all reward probabilities are positive on the prior rectangle. Thus, the likelihood has parameter-independent support, including when some arms have zero sampling probability. Its density and first derivatives satisfy the domination conditions of Lemma B.1, since the horizon and the action and reward alphabets are finite.

On the support of the likelihood, the score is

$$\nabla_w \log p_t(X_t \mid w) = \sum_{s=1}^{t-1} Z_s, \qquad Z_s := e_{i_s} \frac{Y_s}{R^2 + Y_s w_{i_s}}.$$

Conditional on $w$, $\xi$, and the observations preceding round $s$, the sampled arm $i_s$ is fixed. The single-observation calculation in (22) gives

$$\mathbb{E}[Z_s \mid w, \xi, (i_j, Y_j)_{j<s}] = 0,$$

$$\mathbb{E}[Z_s Z_s^\top \mid w, \xi, (i_j, Y_j)_{j<s}] = \frac{e_{i_s} e_{i_s}^\top}{R^2 - w_{i_s}^2}.$$

Therefore, the score cross-terms from different rounds vanish.

Define the random pull counts and their prior-averaged expectations by

$$N_i(t-1) := \sum_{s=1}^{t-1} \mathbb{1}[i_s = i], \qquad n_i(t-1) := \mathbb{E}[N_i(t-1)].$$

Here the latter expectation is jointly over the prior and the interaction. The data Fisher information is

$$\mathcal{I}_t(w) = \mathrm{diag}\left(\left\{\frac{\mathbb{E}[N_i(t-1) \mid w]}{R^2 - w_i^2}\right\}_{i \in [K]}\right).$$

Averaging over the prior and using $|w_i| \le 3R/4$ yields

$$\mathbb{E}_w[\mathcal{I}_t(w)] \preceq \frac{16}{7R^2}\mathrm{diag}\big((n_i(t-1))_{i\in[K]}\big), \qquad \sum_{i=1}^{K} n_i(t-1) = t-1. \tag{36}$$

The order of these steps matters: the conditional pull counts may depend on $w$, whereas $n_i(t-1)$ averages over both the prior and the interaction.

**(ii) Applying the van Trees inequality.** The prior is unchanged, so its Fisher information is $\mathcal{J}(\mu_w) = (\pi^2/\Delta^2)I_K$, as calculated in (21). Combining Lemma B.1 with (36) gives

$$\mathbb{E}[(w-\widehat{w}_t)(w-\widehat{w}_t)^\top] \succeq \mathrm{diag}\left(\left\{\frac{1}{16n_i(t-1)/(7R^2)+\pi^2/\Delta^2}\right\}_{i\in[K]}\right).$$

Writing $\ell_i \coloneqq (L_p(\pi^\circ))_{ii}$, the shared risk bound (20) therefore implies

$$\mathbb{E}[\mathrm{Reg}_\eta(T)] \ge \frac{a_\psi \eta}{2C_\psi}\sum_{t=1}^{T}\sum_{i=1}^{K}\frac{\ell_i}{16n_i(t-1)/(7R^2)+\pi^2/\Delta^2}.$$

By the weighted Cauchy–Schwarz inequality,

$$\begin{aligned}\sum_{i=1}^{K}\frac{\ell_i}{16n_i(t-1)/(7R^2)+\pi^2/\Delta^2} &\ge \frac{\left(\sum_i \sqrt{\ell_i}\right)^2}{\sum_i\left(16n_i(t-1)/(7R^2)+\pi^2/\Delta^2\right)} \\ &= \frac{d_{\mathrm{B}}(\pi^\circ)}{16(t-1)/(7R^2)+K\pi^2/\Delta^2}.\end{aligned}$$

This is where the bandit effective dimension and the factor $K$ multiplying the prior information arise.

Integrating the decreasing reciprocal summand as in the full-information proof,

$$\begin{aligned}\mathbb{E}[\mathrm{Reg}_\eta(T)] &\ge \frac{a_\psi\eta}{2C_\psi}d_{\mathrm{B}}(\pi^\circ)\sum_{t=1}^{T}\frac{1}{16(t-1)/(7R^2)+K\pi^2/\Delta^2} \\ &\ge \frac{7a_\psi\eta R^2}{32C_\psi}d_{\mathrm{B}}(\pi^\circ)\log\left(1+\frac{16\Delta^2 T}{7\pi^2 K R^2}\right).\end{aligned}$$

The bound holds uniformly over learners for every fixed anchor. Maximizing over $\Pi_{\eta R/2,\psi}$ and substituting $\Delta = \min\{R/4, 1/\eta\}$ gives

$$\mathfrak{R}^{\mathrm{B}}_\eta(T;\psi) \ge \frac{7a_\psi}{32C_\psi}\eta R^2 D_{\mathrm{B}}(\eta R/2;\psi)\log\left(1+\frac{T}{7\pi^2 K\max\{1,(\eta R/4)^2\}}\right). \tag{37}$$

**(iii) Uniform and two-arm anchors.** Whenever the uniform policy belongs to $\Pi_{\eta R/2,\psi}$, Proposition B.1 gives

$$D_{\mathrm{B}}(\eta R/2;\psi) \ge (K-1)K^{1-p}.$$

Thus, (37) recovers the $K^{2-p}$ coefficient in the logarithmic bandit lower bound.

The two-arm regime can be obtained directly from the full-information result. Indeed, a full-information learner can simulate any bandit learner by retaining only the reward of its sampled arm. Therefore,

$$\mathfrak{R}^{\mathrm{B}}_\eta(T;\psi) \ge \mathfrak{R}^{\mathrm{FI}}_\eta(T;\psi).$$

For $K \ge 3$, under the sufficient condition (24), Proposition B.2 and (23) consequently give

$$\mathfrak{R}^{\mathrm{B}}_\eta(T;\psi) \ge \frac{7a_\psi}{32C_\psi}\left(\frac{3}{8}\right)^p \eta R^2\log\left(1+\frac{T}{7\pi^2\max\{1,(\eta R/4)^2\}}\right).$$

This inherits both the coefficient and the logarithmic factor from the full-information bound, without a separate bandit construction.

### D.2 Part II. Square-Root Lower Bound

Use the same hidden-arm reward construction as in Lemma B.4, now with

$$\Delta := \frac{R}{4}\min\left\{1, \sqrt{\frac{K}{T}}\right\}, \qquad \delta := \frac{\Delta}{R} \le \frac{1}{4}.$$

Thus, the hidden arm $I$ is uniform on $[K]$, its mean reward is $\Delta$, and all other arms have mean zero. The rewards remain independent across coordinates and rounds conditional on $I$.

By Lemma B.3, it suffices to bound the expected mass placed on the hidden arm. The full-information testing calculation does not directly give the required dependence on $K$; we instead compare the observed bandit histories.

**(i) Testing the hidden arm under bandit feedback.** Fix a learner. Let $P_i$ be the law of its complete observed history and independent private randomization under hidden arm $i$. Let $P_0$ be the corresponding law when every reward coordinate is unbiased. These laws have common support, by the same argument used for (35).

The learner's decision rules cancel in the likelihood ratio. Only observations of arm $i$ distinguish $P_0$ from $P_i$, so

$$D_{\mathrm{KL}}(P_0, P_i) = \mathbb{E}_0[N_i(T)]\mathrm{kl}\left(\frac{1}{2}, \frac{1}{2}+\frac{\delta}{2}\right).$$

Moreover,

$$\mathrm{kl}\left(\frac{1}{2}, \frac{1}{2}+\frac{\delta}{2}\right) = -\frac{1}{2}\log(1-\delta^2) \le \frac{\delta^2}{2(1-\delta^2)} \le \delta^2,$$

where the last inequality uses $\delta \le 1/4$. Consequently,

$$D_{\mathrm{KL}}(P_0, P_i) \le \delta^2 \mathbb{E}_0[N_i(T)].$$

Define

$$S_i := \sum_{t=1}^{T} \pi_{t,i}.$$

For the fixed learner, $S_i$ is the same measurable function of the observed history and private randomization under every instance. It satisfies $0 \le S_i \le T$ and $\sum_i S_i = T$. Thus,

$$\begin{aligned}
\frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_i[S_i] &\le \frac{1}{K}\sum_{i=1}^{K}\mathbb{E}_0[S_i] + \frac{T}{K}\sum_{i=1}^{K}\mathrm{TV}(P_0, P_i) \\
&\le \frac{T}{K} + \frac{T}{K}\sum_{i=1}^{K}\sqrt{\frac{1}{2}D_{\mathrm{KL}}(P_0, P_i)} && \text{(Pinsker's inequality)} \\
&\le \frac{T}{K} + \frac{T\delta}{K}\sum_{i=1}^{K}\sqrt{\frac{\mathbb{E}_0[N_i(T)]}{2}} \\
&\le \frac{T}{K} + T\delta\sqrt{\frac{T}{2K}} && \text{(Jensen's inequality)} \\
&\le T\left(\frac{1}{2} + \frac{1}{4\sqrt{2}}\right) < \frac{3}{4}T.
\end{aligned}$$

Here $\sum_i \mathbb{E}_0[N_i(T)] = T$, and the last line uses $K \ge 2$ and $\delta = \frac{1}{4}\min\{1, \sqrt{K/T}\}$. Since $I$ is uniform,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\pi_{t,I}] = \frac{1}{KT}\sum_{i=1}^{K}\mathbb{E}_i[S_i] \le \frac{3}{4}. \tag{38}$$

**(ii) Applying the shared oracle comparison.** Combining Lemma B.3 with (38) and taking the infimum over learners,

$$\mathfrak{R}_{\eta}^{\mathrm{B}}(T;\psi) \geq \frac{T}{\eta} F_{\psi}(\eta\Delta) - \frac{3}{4}\Delta T.$$

Therefore,

$$F_{\psi}(\eta\Delta) \geq \frac{7}{8}\eta\Delta \quad \Longrightarrow \quad \mathfrak{R}_{\eta}^{\mathrm{B}}(T;\psi) \geq \frac{\Delta T}{8} = \frac{R}{32}\min\{T, \sqrt{KT}\}. \tag{39}$$

In particular, for $T \geq K$, the sufficient condition is

$$F_{\psi}\left(\frac{\eta R}{4}\sqrt{\frac{K}{T}}\right) \geq \frac{7\eta R}{32}\sqrt{\frac{K}{T}},$$

and the resulting lower bound is

$$\mathfrak{R}_{\eta}^{\mathrm{B}}(T;\psi) \geq \frac{R}{32}\sqrt{KT}.$$

For $T < K$, the same argument gives the saturated lower bound $RT/32$ whenever $F_{\psi}(\eta R/4) \geq 7\eta R/32$.

The finite sufficient-threshold argument at the end of Appendix B.2 concerns only $F_{\psi}$ and applies here unchanged. Thus, the condition in (39) holds for sufficiently large $\eta$ for every fixed regularizer in our class.

# E Proof of Theorem 4.2: Regret Upper Bound (Bandit) via IW-C-FTRL

In this section, we prove Theorem 4.2 for the importance-weighted version of C-FTRL in Algorithm 2. As in Appendix C.1, we first bound the regret for any fixed $\xi > 0$ and then choose $\xi$ as in (9).

**Algorithm 2** Importance-Weighted Composite Follow-The-Regularized-Leader (IW-C-FTRL)

**Require:** Coefficient $\xi > 0$.

1: **for** $t = 1, \dots, T$ **do**

2: Set $\lambda_t = t/\eta$ and choose

$$\pi_t \in \arg\max_{\pi \in \mathcal{C}_{\eta R}} \left\{ \sum_{s=1}^{t-1} \langle \pi, \widehat{r}_s \rangle - \lambda_t \psi(\pi) - \xi S_{1/2}(\pi) \right\}.$$

3: Draw $a_t \sim \pi_t$ and observe $r_{t,a_t}$.

4: Set $\widehat{\ell}_t = (R - r_{t,a_t}) e_{a_t} / \pi_{t,a_t}$ and $\widehat{r}_t = R\mathbf{1} - \widehat{\ell}_t$.

5: **end for**

## E.1 Generic Bandit Upper Bound

Fix a reward sequence $\{r_t\}_{t=1}^T$ and let $\pi^\star$ be its optimal regularized policy in hindsight. By Lemma C.2 and (7), $\pi^\star \in \Pi_{\eta R, \psi} \subseteq \mathcal{C}_{\eta R}$. We claim that, for any $\xi > 0$,

$$\mathbb{E}\big[\mathrm{Reg}_\eta(T; \{r_t\})\big] \le 2\xi\sqrt{K} + \min\left\{ \frac{2\eta R^2}{c_\psi} V_{\mathrm{B}}(\eta R; \psi)(1 + \log T),\ \frac{4R^2 T\sqrt{K}}{\xi} \right\}. \tag{40}$$

Write $\ell_t = R\mathbf{1} - r_t \in [0, 2R]^K$. Since $\langle \pi, \widehat{r}_t \rangle = R - \langle \pi, \widehat{\ell}_t \rangle$ on the simplex, we can write the update in terms of the cumulative estimated losses. With $F_0 = \xi S_{1/2}$, we have

$$F_t(\pi) \coloneqq F_{t-1}(\pi) + \eta^{-1}\psi(\pi) + \langle \pi, \widehat{\ell}_t \rangle,$$
$$\pi_t \in \arg\min_{\pi \in \mathcal{C}_{\eta R}} \left\{ F_{t-1}(\pi) + \eta^{-1}\psi(\pi) \right\}.$$

Each objective is lower-semicontinuous and finite at $\mathbf{1}/K$, so it attains its minimum on the compact cap. At any boundary point of $\Pi$ with finite objective, the directional derivative of $S_{1/2}$ toward $\mathbf{1}/K$ is $-\infty$, while convexity bounds the directional derivative of the remaining terms from above by their finite difference to $\mathbf{1}/K$. Thus each $\pi_t$ has positive coordinates, even if a cap constraint is active.

Let $\mathbb{E}_t$ denote expectation conditional on the history before drawing $a_t$. Since $a_t \sim \pi_t$, we have $\mathbb{E}_t[\widehat{\ell}_t] = \ell_t$. The reward sequence is fixed, so $\pi^\star$ is deterministic and $\pi_t$ is known before the draw. Hence

$$\mathbb{E}\big[\mathrm{Reg}_\eta(T; \{r_t\})\big] = \mathbb{E} \sum_{t=1}^T \left\{ \langle \widehat{\ell}_t, \pi_t - \pi^\star \rangle + \eta^{-1}\big(\psi(\pi_t) - \psi(\pi^\star)\big) \right\}. \tag{41}$$

We now telescope the minimum values of $F_t$ over $\mathcal{C}_{\eta R}$. As in the proof of Lemma C.1, constrained first-order optimality at $\pi_t$ gives, for every feasible $\pi$ of finite objective,

$$F_{t-1}(\pi) + \eta^{-1}\psi(\pi) \ge F_{t-1}(\pi_t) + \eta^{-1}\psi(\pi_t) + B_{\lambda_t \psi + \xi S_{1/2}}(\pi, \pi_t).$$

Define the stability term

$$\Delta_t \coloneqq \sup_{\pi \in \mathcal{C}_{\eta R}} \left\{ \langle \widehat{\ell}_t, \pi_t - \pi \rangle - B_{\lambda_t \psi + \xi S_{1/2}}(\pi, \pi_t) \right\}.$$

Adding $\langle \pi, \widehat{\ell}_t \rangle$ to the optimality inequality and using $F_{t-1}(\pi_t) \ge \min F_{t-1}$ yields

$$\langle \pi_t, \widehat{\ell}_t \rangle + \eta^{-1}\psi(\pi_t) \le \min F_t - \min F_{t-1} + \Delta_t.$$

When we sum over $t$ and use $\min F_T \le F_T(\pi^\star)$, the term $T\psi(\pi^\star)/\eta$ in $F_T(\pi^\star)$ cancels the comparator penalty in (41). Since $2(1-\sqrt{K}) \le S_{1/2}(\pi) \le 0$ on $\Pi$, we obtain

$$\begin{aligned}\mathbb{E}\big[\mathrm{Reg}_\eta(T;\{r_t\})\big] &\le \xi\big(S_{1/2}(\pi^\star) - \min_{\pi\in\mathcal{C}_{\eta R}} S_{1/2}(\pi)\big) + \sum_{t=1}^{T}\mathbb{E}[\Delta_t] \\ &\le 2\xi\sqrt{K} + \sum_{t=1}^{T}\mathbb{E}[\Delta_t]. \end{aligned} \tag{42}$$

It remains to bound $\mathbb{E}_t[\Delta_t]$. On the event $a_t = i$, we have $\widehat{\ell}_t = (\ell_{t,i}/\pi_{t,i})e_i$ with $\ell_{t,i} \ge 0$. Thus a candidate with $\pi_i \ge \pi_{t,i}$ contributes at most zero to the supremum. For any remaining candidate, the $i$-th coordinate along the segment from $\pi_t$ to $\pi$ is at most $\pi_{t,i}$. Using Assumption 2(i) in the direction $\pi - \pi_t \in \mathcal{T}$ and $(-2\sqrt{x})'' = 1/(2x^{3/2})$, we obtain from Taylor's integral formula

$$B_{\lambda_t\psi+\xi S_{1/2}}(\pi,\pi_t) \ge \frac{(\pi_{t,i}-\pi_i)^2}{2}\left(\frac{c_\psi\lambda_t}{\pi_{t,i}^{p}} + \frac{\xi}{2\pi_{t,i}^{3/2}}\right).$$

For boundary candidates of finite objective, the same inequality follows by taking a limit along the segment. Maximizing the resulting scalar quadratic in $\pi_{t,i} - \pi_i$ yields

$$\Delta_t \le \min\left\{\frac{\ell_{t,i}^2\pi_{t,i}^{p-2}}{2c_\psi\lambda_t},\ \frac{\ell_{t,i}^2}{\xi\sqrt{\pi_{t,i}}}\right\}.$$

Averaging over $a_t$ with probabilities $\pi_{t,i}$, and using $\ell_{t,i} \le 2R$ and $\lambda_t = t/\eta$, we obtain

$$\mathbb{E}_t[\Delta_t] \le \min\left\{\frac{2\eta R^2}{c_\psi t}\sum_i \pi_{t,i}^{p-1},\ \frac{4R^2}{\xi}\sum_i\sqrt{\pi_{t,i}}\right\}. \tag{43}$$

The first sum is at most $V_{\mathrm{B}}(\eta R;\psi)$ by (7), and the second is at most $\sqrt{K}$ by Cauchy–Schwarz. Substituting (43) into (42) and using $\sum_{t=1}^{T} t^{-1} \le 1 + \log T$ proves (40).

With $\xi$ chosen as in (9), the minimum in (40) equals $2\xi\sqrt{K}$. The same run therefore has expected regret at most $4\xi\sqrt{K}$ for every reward sequence, which proves (8).

For updates over $\Pi$, the same argument replaces $V_{\mathrm{B}}(\eta R;\psi)$ in the bound and in the choice of $\xi$ by $\max_{\pi\in\Pi}\sum_i \pi_i^{p-1} = K^{(2-p)_+}$. For $p = 1$ the sum is $K$ by convention. For $1 < p \le 2$, concavity gives the maximum $K^{2-p}$ at $\mathbf{1}/K$. For $p > 2$, $\pi_i^{p-1} \le \pi_i$ gives the maximum 1 at a vertex. □

# F Instantiations and Geometric Quantities for Standard Regularizers

We instantiate our regret bounds for the four regularizers introduced in Section 5. We first present the geometric quantities and resulting regret bounds, and then provide their proofs and computations. Throughout, $K \geq 2$, $T \geq 1$, $\eta, R > 0$, $U \geq 0$, $\alpha \in (0,1)$, and $\rho \in \operatorname{relint}(\Pi)$. Recall that $M_\psi = M_\psi(\eta R)$. We set $R = 1$ only in the subsection on instantiated regret bounds.

## F.1 Results

### F.1.1 Curvature and Dual-Constrained Diameter

**Proposition F.1** (Curvature and Legendre Properties)**.** *All four regularizers, with their lower-semicontinuous extensions to $\Pi$, are separable and Legendre. They satisfy the coordinate curvature bounds in Assumption 4 with the constants in Table 1. Their minimizers are $\mathbf{1}/K$ for $H$, $S_\alpha$, and $\psi_{\mathrm{BIS}}$, and $\rho$ for $D_{\mathrm{KL}}(\cdot, \rho)$.*

**Proposition F.2** (Dual-Constrained Diameter)**.** *For every $U \geq 0$, the upper bounds on $M_\psi(U)$ in Table 1 hold.*

Table 1: Curvature constants and upper bounds on the dual-constrained diameter of standard regularizers.

| **Regularizer** | $p$ | $c_\psi = C_\psi$ | **Upper bound on** $M_\psi(U)$ |
|---|---|---|---|
| $H$ | 1 | 1 | $\log K$ |
| $D_{\mathrm{KL}}(\cdot, \rho)$ | 1 | 1 | $\max_i \log(1/\rho_i)$ |
| $S_\alpha$ | $2-\alpha$ | $\alpha$ | $(K^{1-\alpha}-1)/(1-\alpha)$ |
| $\psi_{\mathrm{BIS}}$ | 2 | 1 | $(K-1)\log(1+2U/K)$ |

### F.1.2 Full-Information Effective-Dimension Refinement

The logarithmic upper bound in Theorem 3.2 can be refined by retaining the full-information effective dimension. The appropriate dual radius depends on the update domain.

**Proposition F.3** (Effective-Dimension Refinement)**.** *Suppose Assumptions 1, 3 and 4 and Assumption 2(i) hold, and $T$ is known. Consider the following choices of update domain $\mathcal{C}$ and radius $U$:*

1. *If $\psi$ is minimized at $\mathbf{1}/K$, take $\mathcal{C} = \Pi$ and $U = \eta R$.*
2. *If $\psi = D_{\mathrm{KL}}(\cdot, \rho)$, take $\mathcal{C} = \Pi_{\eta R, \psi}$ and $U = \eta R$.*
3. *For a general separable regularizer, take $\mathcal{C} = \operatorname{conv} \Pi_{\eta R, \psi}$ and $U = 2\eta R$.*

*Run the **C-FTRL** update of Theorem 3.2 over $\mathcal{C}$, with*

$$\lambda_t = \frac{t}{\eta}, \qquad \xi = \min\left\{ \frac{\eta R^2 D_{\mathrm{FI}}(U;\psi)}{c_\psi \log K}, R\sqrt{\frac{T}{2\log K}} \right\}.$$

*Then, for every bounded reward sequence,*

$$\mathrm{Reg}_\eta(T; \{r_t\}) \leq \min\left\{ \eta c_\psi^{-1} R^2 D_{\mathrm{FI}}(U;\psi)(3+\log T), R\sqrt{2T\log K} + 2R \right\}. \tag{44}$$

*Both bounds hold for the same sequence of policies.*

The lower bound in Theorem 3.1 involves $D_{\mathrm{FI}}(\eta R/2; \psi)$. Thus, comparison with (44) depends on the effective dimensions at the corresponding dual radii, the curvature constants, and the logarithmic factors. For all four standard regularizers, the refined upper bound uses $U = \eta R$.

### F.1.3 Effective Dimensions

We next collect bounds on the full-information and bandit effective dimensions, and then instantiate them for the four regularizers.

**Lemma F.1** (Full-Information Effective-Dimension Bounds). *For every $p \geq 1$,*

$$d_{\mathrm{FI}}(\pi) \leq 1 - \sum_i \pi_i^2 \leq 1 - \frac{1}{K}, \qquad \pi \in \mathrm{relint}(\Pi).$$

*Consequently, $D_{\mathrm{FI}}(U;\psi) \leq 1 - 1/K$. If $p = 1$ and $\mathbf{1}/K \in \Pi_{U,\psi}$, then*

$$D_{\mathrm{FI}}(U;\psi) = 1 - \frac{1}{K}.$$

*Suppose additionally that Assumptions 2 to 4 hold and $\psi$ is minimized at $\mathbf{1}/K$. Then, for every $U \geq 0$,*

$$(K-1)K^{-p} \leq D_{\mathrm{FI}}(U;\psi) \leq \min\left\{1 - \frac{1}{K},\, e^{2pU/c_\psi}(K-1)K^{-p}\right\}. \tag{45}$$

For $p > 1$, the uniform policy need not maximize $d_{\mathrm{FI}}$. The bandit dimension, however, admits the following uniform-maximizer identity throughout $1 \leq p \leq 2$.

**Lemma F.2** (Maximum Bandit Effective Dimension). *For $1 \leq p \leq 2$,*

$$d_{\mathrm{B}}(\pi) \leq (K-1)K^{1-p}, \qquad \pi \in \mathrm{relint}(\Pi),$$

*with equality at $\pi = \mathbf{1}/K$. Consequently, whenever $\mathbf{1}/K \in \Pi_{U,\psi}$,*

$$D_{\mathrm{B}}(U;\psi) = (K-1)K^{1-p}.$$

*Moreover, the cap in (7) always contains $\mathbf{1}/K$, and*

$$V_{\mathrm{B}}(U;\psi) = K^{2-p}.$$

For completeness, when $p > 2$, the cap instead gives $V_{\mathrm{B}}(U;\psi) \leq m_\psi(U)^{p-2} < 1$. The four regularizers considered here all have $1 \leq p \leq 2$.

**Proposition F.4** (Effective Dimensions of Standard Regularizers). *For every $U \geq 0$,*

$$D_{\mathrm{B}}(U;H) = K - 1,$$
$$D_{\mathrm{B}}(U;S_\alpha) = (K-1)K^{\alpha-1},$$
$$D_{\mathrm{B}}(U;\psi_{\mathrm{BIS}}) = \frac{K-1}{K}.$$

*The full-information bounds for these three regularizers follow from Lemma F.1, since each is minimized at $\mathbf{1}/K$.*
*For $\psi = D_{\mathrm{KL}}(\cdot,\rho)$, the uniform policy belongs to $\Pi_{U,\psi}$ if and only if*

$$\log \frac{\max_i \rho_i}{\min_i \rho_i} \leq 2U. \tag{46}$$

*Under this condition, Lemmas F.1 and F.2 give*

$$D_{\mathrm{FI}}(U;\psi) = 1 - \frac{1}{K}, \qquad D_{\mathrm{B}}(U;\psi) = K - 1.$$

*For an arbitrary reference policy,*

$$d_{\mathrm{FI}}(\rho) \leq D_{\mathrm{FI}}(U;\psi) \leq \min\left\{e^{2U} d_{\mathrm{FI}}(\rho),\, 1 - \frac{1}{K}\right\},$$

$$d_{\mathrm{B}}(\rho) \le D_{\mathrm{B}}(U;\psi) \le \min\left\{e^{2U} d_{\mathrm{B}}(\rho),\, K-1\right\},$$

*where*

$$d_{\mathrm{FI}}(\rho) = 1 - \sum_i \rho_i^2, \qquad d_{\mathrm{B}}(\rho) = \left(\sum_i \sqrt{\rho_i(1-\rho_i)}\right)^2.$$

### F.1.4 Sufficient Transition Conditions

We first instantiate the condition for the two-arm anchor of Proposition B.2, and then give sufficient conditions for the square-root lower bounds.

**Proposition F.5** (Two-Arm Anchor Conditions)**.** *Let $K \ge 3$. The condition in (24) holds whenever*

$$\eta R \ge \begin{cases} \log\left(\frac{3(K-2)}{2}\right), & \psi = H, \\ \frac{\alpha}{1-\alpha}\left([4(K-2)]^{1-\alpha} - \left(\frac{8}{3}\right)^{1-\alpha}\right), & \psi = S_\alpha, \\ 4(K-2) - \frac{8}{3}, & \psi = \psi_{\mathrm{BIS}}. \end{cases} \tag{47}$$

*For $\psi = D_{\mathrm{KL}}(\cdot,\rho)$, the exact condition is*

$$\eta R \ge \max_i \log \frac{\pi_i^\circ}{\rho_i} - \min_i \log \frac{\pi_i^\circ}{\rho_i},$$

*where $\pi^\circ$ is the policy in Proposition B.2.*

Under these conditions, Proposition B.2 gives $D_{\mathrm{FI}}(\eta R/2;\psi) \ge (3/8)^p$. The resulting full-information lower bound also applies to bandit feedback, as shown in Appendix D. This inherited bound retains the full-information logarithm; it does not introduce a factor of $K$ in its denominator.

**Proposition F.6** (Sufficient Conditions for the Square-Root Lower Bounds)**.** *Define*

$$x_\psi := \begin{cases} 8\log K, & \psi = H, \\ \frac{8}{K}\sum_i \log\frac{1}{\rho_i}, & \psi = D_{\mathrm{KL}}(\cdot,\rho), \\ 8\frac{K^{1-\alpha}-1}{1-\alpha}, & \psi = S_\alpha, \\ 16(K-1)\log 16, & \psi = \psi_{\mathrm{BIS}}. \end{cases}$$

*Then $F_\psi(x) \ge 7x/8$ for every $x \ge x_\psi$. Consequently,*

$$\eta R \ge 4x_\psi\sqrt{\frac{T}{\log K}}, \quad T \ge \log K \quad \Longrightarrow \quad \mathfrak{R}_\eta^{\mathrm{FI}}(T;\psi) \ge \frac{R}{32}\sqrt{T\log K},$$

$$\eta R \ge 4x_\psi\sqrt{\frac{T}{K}}, \quad T \ge K \quad \Longrightarrow \quad \mathfrak{R}_\eta^{\mathrm{B}}(T;\psi) \ge \frac{R}{32}\sqrt{KT}.$$

These conditions are sufficient; we do not claim necessity. For Shannon entropy and KL divergence, the conjugate-growth functions also admit the exact expressions

$$F_H(x) = \log\frac{e^x + K - 1}{K}, \qquad F_{D_{\mathrm{KL}}(\cdot,\rho)}(x) = \frac{1}{K}\sum_i \log\left(1 + \rho_i(e^x - 1)\right). \tag{48}$$

### F.1.5 Instantiated Regret Bounds

In this subsection, we set $R = 1$, as in Section 5. To retain the finite-horizon dependence, write

$$L_T := 1 + \log T,$$

$$L_{\mathrm{FI}} \coloneqq \log\left(1 + \frac{T}{7\pi^2 \max\{1, (\eta/4)^2\}}\right),$$
$$L_{\mathrm{B}} \coloneqq \log\left(1 + \frac{T}{7\pi^2 K \max\{1, (\eta/4)^2\}}\right). \tag{49}$$

Table 2 summarizes the logarithmic lower and upper bounds. Within each row, $\psi$ denotes the corresponding regularizer. For Tsallis and BIS entropy, the displayed full-information lower bounds use the uniform anchor; retaining $D_{\mathrm{FI}}(\eta/2; \psi)$ can give a stronger bound.

Table 2: Instantiated regret bounds for $R = 1$. The logarithmic factors are defined in (49). Implicit constants are independent of $K, T, \eta, \rho$, but may depend on the fixed Tsallis parameter $\alpha$.

| | **Full Information** | | **Bandit** | |
|---|---|---|---|---|
| **Regularizer** | **Lower bound** | **Upper bound** | **Lower bound** | **Upper bound** |
| $H$ | $\Omega(\eta L_{\mathrm{FI}})$ | $\mathcal{O}(\eta L_T)$ | $\Omega(\eta K L_{\mathrm{B}})$ | $\mathcal{O}(\eta K L_T)$ |
| $D_{\mathrm{KL}}(\cdot, \rho)$ | $\Omega(\eta D_{\mathrm{FI}}(\eta/2; \psi) L_{\mathrm{FI}})$ | $\mathcal{O}(\eta D_{\mathrm{FI}}(\eta; \psi) L_T)$ | $\Omega(\eta D_{\mathrm{B}}(\eta/2; \psi) L_{\mathrm{B}})$ | $\mathcal{O}(\eta K L_T)$ |
| $S_\alpha$ | $\Omega(\eta K^{\alpha-1} L_{\mathrm{FI}})$ | $\mathcal{O}(\eta D_{\mathrm{FI}}(\eta; \psi) L_T)$ | $\Omega(\eta K^{\alpha} L_{\mathrm{B}})$ | $\mathcal{O}(\eta K^{\alpha} L_T)$ |
| $\psi_{\mathrm{BIS}}$ | $\Omega(\eta K^{-1} L_{\mathrm{FI}})$ | $\mathcal{O}(\eta D_{\mathrm{FI}}(\eta; \psi) L_T)$ | $\Omega(\eta L_{\mathrm{B}})$ | $\mathcal{O}(\eta L_T)$ |

Every upper bound in Table 2 can also be intersected with $\mathcal{O}(\sqrt{T \log K})$ under full-information feedback and $\mathcal{O}(\sqrt{KT})$ under bandit feedback. These square-root upper bounds hold for every $\eta > 0$; Proposition F.6 gives sufficient conditions for matching lower bounds.

The following consequence makes the small-$\eta$ dimension dependence explicit.

**Corollary F.1** (Small-$\eta$ Regret Bounds)**.** *Suppose $R = 1$ and $0 < \eta \leq 1$. For $\psi \in \{H, S_\alpha, \psi_{\mathrm{BIS}}\}$, with $p = 1, 2 - \alpha, 2$, respectively,*

$$\mathfrak{R}_\eta^{\mathrm{FI}}(T; \psi) \asymp \eta K^{1-p} \log(1 + T).$$

*Under bandit feedback,*

$$\eta K^{2-p} \log(1 + T/K) \lesssim \mathfrak{R}_\eta^{\mathrm{B}}(T; \psi) \lesssim \eta K^{2-p} \log(1 + T).$$

*The implicit constants are independent of $K, T, \eta$, but may depend on the fixed parameter $\alpha$ for Tsallis entropy.*

For $\rho = \mathbf{1}/K$, KL divergence equals $H + \log K$, so its regularized regret coincides with that of Shannon entropy. For an arbitrary reference policy, the reachable dimensions in Proposition F.4 must be retained.

### F.2 Proofs and Computations

*Proof of Proposition F.1.* The coordinate derivatives and second derivatives are

$$\begin{aligned}
H: &\quad \phi_i'(x) = 1 + \log x, && \phi_i''(x) = x^{-1}, \\
D_{\mathrm{KL}}(\cdot, \rho): &\quad \phi_i'(x) = 1 + \log(x/\rho_i), && \phi_i''(x) = x^{-1}, \\
S_\alpha: &\quad \phi_i'(x) = -\frac{\alpha}{1-\alpha} x^{\alpha-1}, && \phi_i''(x) = \alpha x^{\alpha-2}, \\
\psi_{\mathrm{BIS}}: &\quad \phi_i'(x) = -x^{-1}, && \phi_i''(x) = x^{-2}.
\end{aligned}$$

These identities establish separability, strict convexity, and the stated curvature constants. The usual boundary extensions are proper and lower-semicontinuous.

To verify the relative Legendre condition, consider $\pi_n \in \mathrm{relint}(\Pi)$ converging to the relative boundary. There exist coordinates $i, j$ such that $\pi_{n,i} \to 0$ and $\pi_{n,j} \to \bar{\pi}_j > 0$. For each regularizer,

$$\phi_i'(\pi_{n,i}) \to -\infty, \qquad \phi_j'(\pi_{n,j}) \to \phi_j'(\bar{\pi}_j) \in \mathbb{R},$$

where the derivative at 1, if needed, is understood by its finite left limit. Since $e_i - e_j \in \mathcal{T}$,

$$\sqrt{2}\,\|P_{\mathcal{T}} \nabla\psi(\pi_n)\|_2 \geq |\phi_i'(\pi_{n,i}) - \phi_j'(\pi_{n,j})| \longrightarrow \infty.$$

Thus the required boundary divergence holds on the tangent space.

Symmetry and strict convexity give the uniform minimizer for $H$, $S_\alpha$, and $\psi_{\mathrm{BIS}}$. Nonnegativity of KL divergence gives the minimizer $\rho$ for $D_{\mathrm{KL}}(\cdot, \rho)$. □

*Proof of Proposition F.2.* The global ranges of Shannon and Tsallis entropy satisfy

$$-\log K \leq H(\pi) \leq 0, \qquad -\frac{K^{1-\alpha} - 1}{1-\alpha} \leq S_\alpha(\pi) \leq 0.$$

Also,

$$0 \leq D_{\mathrm{KL}}(\pi, \rho) = H(\pi) + \sum_i \pi_i \log \frac{1}{\rho_i} \leq \max_i \log \frac{1}{\rho_i}.$$

Restricting the maximum to $\Pi_{U,\psi}$ proves the first three bounds.

For BIS entropy, the minimum is $K \log K$. If $\pi \in \Pi_{U,\psi_{\mathrm{BIS}}}$, the optimality condition implies that the derivatives $-1/\pi_i$ have range at most $2U$. Hence,

$$\frac{1}{\pi_{\min}} - \frac{1}{\pi_{\max}} \leq 2U.$$

Since $\pi_{\max} \geq 1/K$, every coordinate satisfies $1/\pi_i \leq K + 2U$, and a maximizing coordinate satisfies $1/\pi_{\max} \leq K$. Therefore,

$$\begin{aligned}\psi_{\mathrm{BIS}}(\pi) - K \log K &\leq \log K + (K-1)\log(K + 2U) - K \log K \\ &= (K-1)\log\left(1 + \frac{2U}{K}\right).\end{aligned}$$

Maximizing over the reachable policies proves the claim. □

*Proof of Proposition F.3.* By Lemma C.2, each update domain contains the hindsight comparator and the minimizer of $\psi$. The relevant stability paths lie in $\Pi_{U,\psi}$. Thus (34) retains the factor $D_{\mathrm{FI}}(U;\psi)$, while Lemma C.3 gives

$$\frac{\psi(\pi^\star) - \min_{\pi \in \Pi} \psi(\pi)}{\eta} \leq \min\left\{\eta c_\psi^{-1} R^2 D_{\mathrm{FI}}(U;\psi), 2R\right\}.$$

Consequently, the generic bound in (27) holds with $c_\psi^{-1}$ replaced by $c_\psi^{-1} D_{\mathrm{FI}}(U;\psi)$. The stated choice of $\xi$ and the summation argument in Appendix C.1 prove the result. □

*Proof of Lemma F.1.* Since $p \geq 1$, we have $\pi_i^p \leq \pi_i$. The weighted-variance representation therefore gives, for every $v \in \mathbb{R}^K$,

$$\begin{aligned}v^\top L_p(\pi) v &= \min_{a \in \mathbb{R}} \sum_i \pi_i^p (v_i - a)^2 \\ &\leq \min_{a \in \mathbb{R}} \sum_i \pi_i (v_i - a)^2 = v^\top L_1(\pi) v.\end{aligned}$$

Taking traces yields

$$d_{\mathrm{FI}}(\pi) \leq \operatorname{tr} L_1(\pi) = 1 - \sum_i \pi_i^2 \leq 1 - \frac{1}{K}.$$

For $p = 1$, the uniform policy attains equality. This proves the corresponding identity whenever it belongs to $\Pi_{U,\psi}$.

If $\psi$ is minimized at $\mathbf{1}/K$, then $\nabla\psi^*(0) = \mathbf{1}/K$. The uniform-anchor calculation in Proposition B.1 gives the lower bound in (45). For the remaining upper bound, apply Lemma A.3 with $u^\circ = 0$:

$$L_p(\nabla\psi^*(u)) \preceq e^{2pU/c_\psi} L_p(\mathbf{1}/K), \qquad u \in [-U, U]^K.$$

Taking traces and maximizing over the dual box proves the claim. □

*Proof of Lemma F.2.* Write $q_i = \pi_i^{p/2}$, $A = \sum_i q_i$, and $S = \sum_i q_i^2$. By Cauchy–Schwarz,

$$\begin{aligned} d_{\mathrm{B}}(\pi) &= \left( \sum_i q_i \sqrt{1 - \frac{q_i^2}{S}} \right)^2 \\ &\le A \left( A - \frac{\sum_i q_i^3}{S} \right). \end{aligned}$$

The identity

$$K \sum_i q_i^3 - AS = \sum_{i<j} (q_i - q_j)^2 (q_i + q_j) \ge 0$$

therefore gives

$$d_{\mathrm{B}}(\pi) \le \left( 1 - \frac{1}{K} \right) A^2.$$

Since $p/2 \le 1$, Jensen's inequality yields $A \le K^{1-p/2}$. This proves the upper bound. Equality at the uniform policy follows from Proposition B.1.

To evaluate $V_{\mathrm{B}}$, note that $m_\psi(U) \ge 1/K$, so the cap contains $\mathbf{1}/K$. For $1 < p \le 2$, concavity gives

$$\sum_i \pi_i^{p-1} \le K^{2-p},$$

with equality at the uniform policy. For $p = 1$, the sum is $K$ under the convention $0^0 = 1$.

Finally, for the stated observation when $p > 2$,

$$\sum_i \pi_i^{p-1} \le \|\pi\|_\infty^{p-2} \sum_i \pi_i \le m_\psi(U)^{p-2}.$$

Compactness of $\Pi_{U,\psi} \subset \mathrm{relint}(\Pi)$ gives $m_\psi(U) < 1$. □

*Proof of Proposition F.4.* **Shannon, Tsallis, and BIS entropy.** Their uniform minimizer belongs to $\Pi_{U,\psi}$ for every $U \ge 0$. The bandit identities therefore follow from Lemma F.2, with $p = 1, 2 - \alpha, 2$, respectively. The full-information bounds follow from Lemma F.1.

**KL divergence.** The inverse mirror map is

$$\big(\nabla \psi^*(u)\big)_i = \frac{\rho_i e^{u_i}}{\sum_j \rho_j e^{u_j}}.$$

It equals $\mathbf{1}/K$ precisely when $u_i = -\log \rho_i + c$ for a common scalar $c$. Such a vector has a representative in $[-U, U]^K$ if and only if (46) holds. The dimension identities then follow from Lemmas F.1 and F.2.

For an arbitrary reference policy, $\nabla \psi^*(0) = \rho$ gives the lower bounds. For $u \in [-U, U]^K$, write $\pi(u) = \nabla \psi^*(u)$. Then $\pi_i(u) \le e^{2U} \rho_i$. The weighted-variance representation gives, for every $v \in \mathbb{R}^K$,

$$\begin{aligned} v^\top L_1(\pi(u)) v &= \min_{a \in \mathbb{R}} \sum_i \pi_i(u)(v_i - a)^2 \\ &\le e^{2U} \min_{a \in \mathbb{R}} \sum_i \rho_i (v_i - a)^2 \\ &= e^{2U} v^\top L_1(\rho) v. \end{aligned}$$

Taking traces proves the full-information comparison. Taking diagonal entries, square roots, and then the squared sum proves the bandit comparison. The remaining upper bounds follow from Lemmas F.1 and F.2. □

*Proof of Proposition F.5.* Recall that the anchor in Proposition B.2 has two coordinates $b = 3/8$ and $K - 2$ coordinates $a = 1/[4(K-2)]$. Since $K \ge 3$, we have $a \le b$. For each symmetric

regularizer, the coordinate derivative is increasing, so its gradient range at this anchor is $\phi'(b) - \phi'(a)$. Substituting the derivatives computed above gives, respectively,

$$\log \frac{b}{a}, \qquad \frac{\alpha}{1-\alpha}\left(a^{\alpha-1} - b^{\alpha-1}\right), \qquad \frac{1}{a} - \frac{1}{b}.$$

These are exactly the expressions in (47). For KL divergence, $\phi_i'(\pi_i^\circ) = 1 + \log(\pi_i^\circ/\rho_i)$, giving the stated range condition. □

*Proof of Proposition F.6.* We use the ordinary conjugate, for which $\psi^*(0) = -\min_{\pi\in\Pi}\psi(\pi)$. For $H$, $D_{\mathrm{KL}}(\cdot,\rho)$, and $S_\alpha$, the regularizer is finite at each simplex vertex. Evaluating $\psi^*(xe_i)$ at $e_i$ gives

$$F_\psi(x) \geq x - \frac{1}{K}\sum_i \left(\psi(e_i) - \min_{\pi\in\Pi}\psi(\pi)\right).$$

The three average regularizer gaps are

$$\log K, \qquad \frac{1}{K}\sum_i \log\frac{1}{\rho_i}, \qquad \frac{K^{1-\alpha}-1}{1-\alpha},$$

respectively. Thus $x \geq x_\psi$ implies $F_\psi(x) \geq 7x/8$ for these regularizers.

For BIS entropy, use the interior policies

$$q^{(i)} = \frac{15}{16}e_i + \frac{1}{16K}\mathbf{1}.$$

Their regularizer gaps satisfy

$$\begin{aligned}\psi_{\mathrm{BIS}}(q^{(i)}) - K\log K &= K\log 16 - \log(15K+1)\\ &\leq (K-1)\log 16.\end{aligned}$$

Since $q_i^{(i)} \geq 15/16$, evaluating the conjugate at $q^{(i)}$ yields

$$F_{\psi_{\mathrm{BIS}}}(x) \geq \frac{15}{16}x - (K-1)\log 16.$$

This is at least $7x/8$ whenever $x \geq 16(K-1)\log 16$.

Finally, substitute $x = (\eta R/4)\sqrt{\log K/T}$ and $x = (\eta R/4)\sqrt{K/T}$ into Theorems 3.1 and 4.1, respectively. □

**Exact conjugate-growth functions.** For Shannon entropy and KL divergence,

$$H^*(u) = \log\sum_i e^{u_i}, \qquad D_{\mathrm{KL}}(\cdot,\rho)^*(u) = \log\sum_i \rho_i e^{u_i}.$$

Evaluating these expressions at $xe_i$ and at 0, and averaging over $i$, proves (48).

**Derivation of Table 2.** Apply Theorems 3.1 and 4.1 with the curvature constants in Table 1. For Shannon entropy, use $D_{\mathrm{FI}} = 1 - 1/K$ and $D_{\mathrm{B}} = K - 1$. For Tsallis and BIS entropy, use $D_{\mathrm{FI}} \geq (K-1)K^{-p}$ and $D_{\mathrm{B}} = (K-1)K^{1-p}$. Since $1/2 \leq 1 - 1/K < 1$, these give the stated lower bounds. For KL divergence with an arbitrary reference policy, retain the effective dimensions in the lower bounds.

For the full-information upper bounds, apply Proposition F.3 with $U = \eta$, which is valid for all four regularizers. For the bandit upper bounds, apply Theorem 4.2 with $V_{\mathrm{B}}(\eta;\psi) = K^{2-p}$. The curvature constants depend only on the fixed $\alpha$ in the Tsallis case. Finally, $3 + \log T \leq 3L_T$, giving the displayed logarithmic upper bounds.

*Proof of Corollary F.1.* Since $0 < \eta \leq 1$,

$$L_{\mathrm{FI}} \asymp \log(1+T), \qquad L_{\mathrm{B}} \asymp \log(1+T/K), \qquad L_T \asymp \log(1+T),$$

with universal comparison constants. By (45),

$$D_{\mathrm{FI}}(\eta/2;\psi) \geq \frac{1}{2}K^{1-p}, \qquad D_{\mathrm{FI}}(\eta;\psi) \leq e^{2p/c_\psi}K^{1-p}.$$

Combining these inequalities with the full-information lower bound and Proposition F.3 proves the full-information claim.

For bandit feedback,

$$D_{\mathrm{B}}(\eta/2;\psi) = \left(1 - \frac{1}{K}\right)K^{2-p}, \qquad V_{\mathrm{B}}(\eta;\psi) = K^{2-p}.$$

Substituting into Theorems 4.1 and 4.2 proves the remaining bounds. □

## 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 main claims made in the abstract and introduction accurately state the paper's theoretical scope: regularized regret minimization under full-information and bandit feedback for adversarial rewards and a broad class of strongly convex regularizers. The claimed lower and upper bounds correspond to Theorems 3.1, 3.2, 4.1, and 4.2, and the instantiated examples are summarized in Section 5.

   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: The paper explicitly states the main assumptions on the rewards and regularizers, including bounded rewards, relative strong convexity/smoothness, Legendre structure, and coordinate-wise separability.

   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 main results are theoretical, and the paper states the relevant assumptions before the theorems and their proofs in the appendices.

   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: [N/A]

   Justification: The paper does not include experimental results.

   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: [N/A]

   Justification: The paper does not include experiments requiring data or code.

   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: [N/A]

   Justification: The paper does not include experiments.

   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: [N/A]

   Justification: The paper does not include experiments

   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: [N/A]

   Justification: The paper does not include computational experiments.

   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: The research is theoretical.

   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: [No]

    Justification: The research is theoretical.

    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 does not release data, models, pretrained systems, scraped datasets, or other artifacts with a high risk of misuse.

    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: The paper does not use existing datasets, codebases, pretrained models, or other external assets.

    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: [N/A]

    Justification: The paper does not introduce or release new datasets, code packages, models, benchmarks, or other assets.

    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 direct crowdsourcing or user studies, although our analysis has a close relationship to such instances as RLHF.

    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 direct crowdsourcing or research with human subjects, so IRB approval or equivalent review is not 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 an important, original, or non-standard component of the paper's core methodology.

    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.