Provably Efficient Regularized Online RLHF with Generalized Bilinear Preferences

Junghyun Lee [1] **Minju Hong** [2] **Kwang-Sung Jun** [3] **Chulhee Yun** [1] **Se-Young Yun** [1]

Abstract

We consider the problem of *regularized* best-response max-regret minimization in online RLHF under general preferences and bandit feedback. While various regularizers are utilized to robustify alignment, known polylogarithmic regret guarantees remain heavily specific to KL. To investigate whether such fast rates extend beyond KL, we adopt the *Generalized Bilinear Preference Model (GBPM)*—capturing intransitive preferences over d-dimensional item-wise features via a rank-$2r$ skew-symmetric matrix—to isolate the impact of generic regularization. Crucially, under GBPM, we prove that the dual gap of any greedy policy is bounded by the *squared* estimation error, derived using *only* strong convexity and skew-symmetry. Under a feature coverage assumption, we establish a *generic* polylogarithmic regret of $\tilde{\mathcal{O}}(\eta d^4 C_{\min}^{-1}(\log T)^2 \wedge d^2 C_{\min}^{-1/2}\sqrt{T})$ with Greedy Sampling, and a dimension-wise improved regret (for well-conditioned arm-sets) of $\tilde{\mathcal{O}}(C_{\min}^{-2}\sqrt{\eta r T} \wedge r^{1/3} C_{\min}^{-4/3} T^{2/3})$ with Explore-Then-Commit, where η^{-1} is the regularization coefficient, T is the time horizon, and $C_{\min}$ is an arm-set dependent quantity. This demonstrates that "fast" regrets are *not* KL-specific, but rather a fundamental consequence of generic strongly convex geometry.

1. Introduction

General Preference Learning. Aligning large language models (LLMs) with human values has emerged as a central challenge in modern AI (Llama Team, 2024; Qwen Team, 2024; OpenAI, 2024). While the common approach to Reinforcement Learning from Human Feedback (RLHF) heavily relies on reward-based models like the Bradley-Terry-Luce (BTL) model (Bradley & Terry, 1952; Christiano et al., 2017), scalar utilities inherently struggle to capture the cyclic, intransitive, and diverse nature of human preferences (May, 1954; Tversky, 1969). This representational bottleneck has motivated a shift toward *General Preference Learning* (or *Nash Learning*), which directly targets the Nash equilibrium (NE) of a preference game (Nash, 1951; von Neumann, 1928; McKelvey & Palfrey, 1995). Recently, this game-theoretic perspective has demonstrated notable empirical promise in LLM alignment (Munos et al., 2024; Ye et al., 2024; Cui et al., 2024; Rosset et al., 2024).

In both practice and theory, solving these preference games online under bandit feedback relies heavily on optimizing *regularized* objectives. In practical RLHF, purely maximizing an unregularized, observed preference often leads to reward hacking, diversity collapse, and hallucinatory text generation (Michaud et al., 2020; Tien et al., 2023; Casper et al., 2023). Embedding regularization directly into the objective prevents large drifts from reference models (e.g., SFT models or expert demonstrations) and robustifies the resulting equilibrium—a concept dating back to the seminal work on *quantal response equilibria* by McKelvey & Palfrey (1995).

Consequently, understanding the statistical efficiency of these regularized games, often measured via ***regularized best-response regret***, has become a primary target in modern RL, game-theoretic alignment, and learning in regularized zero-sum game literature (Munos et al., 2024; Wu et al., 2025; Xiong et al., 2024; Ye et al., 2024; Nayak et al., 2025; Yang et al., 2025). In particular, our analysis focuses on ***regularized max-regret***, which measures suboptimality of the *max* player only (see Section 2.3 for its definition). This notion originates from self-play frameworks for RL in two-player zero-sum games (Bai & Jin, 2020; Bai et al., 2020; Liu et al., 2021; Jin et al., 2022; Xiong et al., 2022), and has become the standard metric in theoretical analyses of online RLHF under general preferences (Ye et al., 2024; Wu et al., 2025). The intuition is that the learner ultimately only cares about obtaining the NE policy for the max-player, the policy actually deployed in practice.

[1]Kim Jaechul Graduate School of AI, KAIST, Seoul, Republic of Korea [2]School of Electrical Engineering, KAIST, Daejeon, Republic of Korea [3]Department of Computer Science and Engineering & Graduate School of AI, POSTECH, Pohang, Republic of Korea. Correspondence to: Junghyun Lee <jh_lee00@kaist.ac.kr>, Se-Young Yun <yunseyoung@kaist.ac.kr>.

Pluralistic Alignment Workshop @ ICML 2026, Seoul, South Korea.

The KL-Centric Theories. The current theoretical landscape for establishing polylogarithmic regularized best-response regret remains overwhelmingly KL-centric. Recent theoretical advances in reward-based RLHF rely heavily on the explicit closed-form Gibbs densities unique to KL-regularized bandits (Xiong et al., 2024; Zhao et al., 2025b;a; Wu et al., 2025; Ji et al., 2026). Consequently, the corresponding literature on regularized best-response regret has followed this same KL-specific trajectory (Wu et al., 2025; Nayak et al., 2025; Yang et al., 2025).

On the other hand, much RL and LLM literature has increasingly explored strongly convex penalties *other than* reverse KL to achieve distinct structural benefits. For example, negative Shannon entropy is used in max-entropy exploration (Ziebart et al., 2008; Neu et al., 2017; Haarnoja et al., 2018); mixtures of reverse KL divergences are used to yield more diverse and less biased alignment (Le et al., 2025; Aminian et al., 2025); the chi-squared divergence is known to provably mitigate overoptimization (Huang et al., 2025a); and Tsallis entropy can encourage sparse policy selection (Lee et al., 2018; Chow et al., 2018). Broader classes of regularizers – such as α-Rényi entropies (Zhang et al., 2026a), Csiszár f-divergences (Wang et al., 2024; Go et al., 2023; Han et al., 2025), and expected or strongly convex regularizers (Yang et al., 2019; Geist et al., 2019) – have been studied as well.

Because prior max-regret analyses rely entirely on the *closed-form solution* of the KL-regularized bandit (Ye et al., 2024; Wu et al., 2025; Nayak et al., 2025), it is unclear whether similar polylogarithmic regret can be attained for regularized games with non-KL regularizers. This motivates our central theoretical question:

Is the fast regularized max-regret achievable under any strongly convex regularizer beyond KL?

The GBPM Abstraction and Theoretical Setup. To answer this, we require a theoretically tractable mathematical abstraction, analogous to the linear BTL model (Bradley & Terry, 1952; Plackett, 1975). To this end, we adopt the ***Generalized Bilinear Preference Model (GBPM)*** (Lee et al., 2026; Zhang et al., 2025b): given *item-wise* features $\boldsymbol{\phi}^1, \boldsymbol{\phi}^2 \in \mathbb{R}^d$, the preference probability is modeled as $P^*(\boldsymbol{\phi}^1 \succ \boldsymbol{\phi}^2) := \mu\left((\boldsymbol{\phi}^1)^\top \boldsymbol{\Theta}_\star \boldsymbol{\phi}^2\right)$, where $\mu(\cdot)$ is a link function satisfying $\mu(z) + \mu(-z) = 1$, and $\boldsymbol{\Theta}_\star \in \mathbb{R}^{d\times d}$ is a *skew-symmetric* matrix of rank at most $2r < d$. By considering such a contextual counterpart to the linear BTL model (Zhu et al., 2023), GBPM allows us to isolate the statistical complexity of online RLHF and rigorously analyze the impact of general regularizers.

Contributions. Under the GBPM framework, we demonstrate that "fast" regret rates are achievable for ***any*** strongly convex regularizer. Our technical contributions are two-fold:

- **Quadratic Bound on Dual Gap.** We show that the dual gap of any greedy NE policy is upper bounded by the *square* of the estimation error of $\boldsymbol{\Theta}_\star$ for ***any*** strongly convex regularizer. Our analysis leverages the skew-symmetry of GBPM, the strong convexity of the regularized game objective, and the integral probability metric representation of the ℓ_1-distance (Müller, 1997) to derive a novel, *self-bounding quadratic inequality* **(Section 3)**.
- **Fast Max-Regrets.** We establish fast regret bounds using two different algorithms tailored to distinct regimes, both critically utilizing the novel quadratic error bound. These results are shown under a feature coverage assumption, which introduces an arm-set dependent quantity $C_{\min} > 0$ **(Section 4.1)**. First, we prove that `Greedy Sampling (GS)` achieves *polylogarithmic* regrets of $\widetilde{\mathcal{O}}(\eta d^4 C_{\min}^{-1}(\log T)^2 \wedge d^2 C_{\min}^{-1/2}\sqrt{T})$ **(Section 4.2)**. Second, to address the high-dimensional regime, we demonstrate that `Explore-Then-Commit (ETC)` with nuclear-norm regularized MLE achieves regrets of $\widetilde{\mathcal{O}}\left(C_{\min}^{-2}\sqrt{\eta r T}\right)$ and $\tilde{\mathcal{O}}(r^{1/3}C_{\min}^{-4/3}T^{2/3})$ **(Section 4.3)**.

2. Problem Setting

We study contextual preference learning with contexts $\boldsymbol{x} \in \mathcal{X}$ and actions $\boldsymbol{a} \in \mathcal{A}$. A policy is a mapping $\pi : \mathcal{X} \to \Delta(\mathcal{A})$, where $\pi(\cdot \mid \boldsymbol{x})$ denotes the conditional distribution over actions given $\boldsymbol{x}$. We denote the policy class by Π. Given two actions $\boldsymbol{a}^1, \boldsymbol{a}^2$ and a context $\boldsymbol{x}$, the event $\boldsymbol{a}^1 \succ \boldsymbol{a}^2 \mid \boldsymbol{x}$ means that response $\boldsymbol{a}^1$ is preferred to response $\boldsymbol{a}^2$ under context $\boldsymbol{x}$.

2.1. Generalized Bilinear Preference Model (GBPM)

We first introduce the low-rank contextual general preference model that we consider in this work. Define $\mathrm{Skew}(d; 2r, S)$ as $\{\boldsymbol{\Theta} \in \mathrm{Skew}(d) : \mathrm{rank}(\boldsymbol{\Theta}) \le 2r, \|\boldsymbol{\Theta}\|_{\mathrm{nuc}} \le S\}$, where $\mathrm{Skew}(d) := \left\{\boldsymbol{\Theta} \in \mathbb{R}^{d\times d} : \boldsymbol{\Theta}^\top = -\boldsymbol{\Theta}\right\}$. We assume a known feature map $\phi : \mathcal{X} \times \mathcal{A} \to \mathcal{B}^d(1) \triangleq \{\boldsymbol{\phi} \in \mathbb{R}^d : \|\boldsymbol{\phi}\|_2 \le 1\}$. Now the definition of ***GBPM*** (Zhang et al., 2025b; Lee et al., 2026):

Definition 2.1 (***Generalized Bilinear Preference Model***)**.** *Let rank $r \le \lfloor d/2 \rfloor$ and norm bound $S > 0$ be known. For any $\boldsymbol{x} \in \mathcal{X}, \boldsymbol{a}^1, \boldsymbol{a}^2 \in \mathcal{A}$, the ground-truth preference under **GBPM** is:*

$$P^*(\boldsymbol{a}^1 \succ \boldsymbol{a}^2 \mid \boldsymbol{x}) := \mu\left(\phi(\boldsymbol{x}, \boldsymbol{a}^1)^\top \boldsymbol{\Theta}_\star \phi(\boldsymbol{x}, \boldsymbol{a}^2)\right), \quad (1)$$

*where $\boldsymbol{\Theta}_\star \in \mathrm{Skew}(d; 2r, S)$ is an unknown **skew-symmetric, low-rank** matrix, and $\mu : \mathbb{R} \to [0, 1]$ is a*

known link satisfying: for some $L_\mu \geq \kappa > 0$,

1. *μ is twice differentiable, monotone increasing, and symmetric ($\mu(z) + \mu(-z) = 1$).*
2. *$\kappa \leq \dot{\mu}\left(\phi^\top \mathbf{\Theta} \phi'\right) \leq L_\mu$, $\left|\ddot{\mu}\left(\phi^\top \mathbf{\Theta} \phi'\right)\right| \leq L_\mu$, $\forall \boldsymbol{\phi}, \boldsymbol{\phi}' \in \mathcal{B}^d(1), \mathbf{\Theta} \in \mathrm{Skew}(d; 2r, S)$.*[a]

[a] We assume the same constant L_μ for the upper bounds of $\dot{\mu}$ and $|\ddot{\mu}|$ for simplicity of the exposition.

The conditions for μ are standard in logistic and generalized linear (dueling) bandits (Faury et al., 2020; Abeille et al., 2021; Lee et al., 2024b;a; 2026; Wu et al., 2024; Bengs et al., 2022). The logistic link $\mu(z) = (1 + e^{-z})^{-1}$ satisfies the above with $L_\mu = \frac{1}{4}$. The linear link $\mu(z) = \frac{1}{2} + z$ is also covered, in which case $L_\mu = 1$ (Gajane et al., 2015; Wu et al., 2024). Lastly, one can see that P^* is anti-symmetric: $P^*(\boldsymbol{a}^1 \succ \boldsymbol{a}^2 \mid \boldsymbol{x}) + P^*(\boldsymbol{a}^2 \succ \boldsymbol{a}^1 \mid \boldsymbol{x}) = 1$.

Remark 2.2. *While recent works have introduced contextual bandit frameworks for general preference learning (Yang et al., 2025; Nayak et al., 2025; Wu et al., 2024), they predominantly rely on* item pair-wise *feature maps (linearizing payoffs as $\langle \varphi(\boldsymbol{a}^1, \boldsymbol{a}^2), \boldsymbol{\theta} \rangle$ for each pair of actions $\boldsymbol{a}^1, \boldsymbol{a}^2 \in \mathcal{A}$) or tabular structures (O'Donoghue et al., 2021); the former is conceptually similar to Wu et al. (2024), though the connection is not explicitly drawn in the literature. This contrasts with practical RLHF scenarios in which only* item-wise *features $\phi(\boldsymbol{a})$ are available, motivating our choice to consider **GBPM**; see Zhang et al. (2025b) and Lee et al. (2026, Appendix K) for more detailed discussions.*

2.2. Population Regularized Game and Nash Equilibrium

We first extend the action-level preference to evaluate policies. The ground-truth preference of π^1 (*max-player*) over π^2 (*min-player*) given a context $\boldsymbol{x}$ is defined by marginalizing over their action distributions, defined as $P^*(\pi^1 \succ \pi^2 \mid \boldsymbol{x}) := \mathbb{E}_{\boldsymbol{a}^1 \sim \pi^1(\cdot|\boldsymbol{x}),\, \boldsymbol{a}^2 \sim \pi^2(\cdot|\boldsymbol{x})}\left[P^*(\boldsymbol{a}^1 \succ \boldsymbol{a}^2 \mid \boldsymbol{x})\right]$.

To evaluate these policies globally, we take the expectation over the unknown context distribution $d_0 \in \Delta(\mathcal{X})$. For any parameter $\mathbf{\Theta} \in \mathrm{Skew}(d)$, we define the expected population game objective as:

$$J(\pi^1, \pi^2; \mathbf{\Theta}) := \mathbb{E}_{\boldsymbol{x} \sim d_0} \mathbb{E}_{\boldsymbol{a}^i \sim \pi^i(\cdot|\boldsymbol{x})} \left[\mu\left(\phi(\boldsymbol{x}, \boldsymbol{a}^1)^\top \mathbf{\Theta} \phi(\boldsymbol{x}, \boldsymbol{a}^2)\right)\right].$$

The ground-truth population preference is then precisely $\mathbb{E}_{\boldsymbol{x} \sim d_0}[P^*(\pi^1 \succ \pi^2 \mid \boldsymbol{x})] = J(\pi^1, \pi^2; \mathbf{\Theta}_\star)$. For notational convenience, we abbreviate this true objective as $J(\pi^1, \pi^2)$, and when analyzing realized features, we utilize the shorthand $J(\phi^1, \phi^2; \mathbf{\Theta}) := \mu\left((\phi^1)^\top \mathbf{\Theta} \phi^2\right)$. We now introduce its regularized counterpart. For $\eta \in (0, \infty]$ and a β^{-1}-strongly convex regularizer $\psi : \Delta(\mathcal{A}) \to \mathbb{R}_{\geq 0}$ w.r.t. $\|\cdot\|_1$, we define a *symmetric*, regularized game objective $J_\eta : \Pi \times \Pi \to \mathbb{R}$ as follows:

$$J_\eta(\pi, \pi'; \mathbf{\Theta}) := J(\pi, \pi'; \mathbf{\Theta}) - \eta^{-1} \mathbb{E}_{\boldsymbol{x} \sim d_0}[\psi(\pi(\cdot \mid \boldsymbol{x}))] + \eta^{-1} \mathbb{E}_{\boldsymbol{x} \sim d_0}[\psi(\pi'(\cdot \mid \boldsymbol{x}))].$$

Standard solution concepts (e.g., Condorcet winners) may not exist in general preference learning (Dudík et al., 2015; Bengs et al., 2021; Munos et al., 2024; Swamy et al., 2024). Thus, as in many recent literature in online RLHF (Munos et al., 2024), we consider *(regularized) Nash Equilibrium (NE)* (Nash, 1951; McKelvey & Palfrey, 1995):

Definition 2.3. *A pair $(\pi^1_\star, \pi^2_\star) \in \Pi \times \Pi$ is a **Nash equilibrium (NE)** if for all $\pi^1, \pi^2 \in \Pi$:*

$$J_\eta(\pi^1, \pi^2_\star) \leq J_\eta(\pi^1_\star, \pi^2_\star) \leq J_\eta(\pi^1_\star, \pi^2). \tag{2}$$

*If $\pi^1_\star = \pi^2_\star =: \pi_\star$, we refer to $\pi_\star$ as a **symmetric NE (SNE)**. By the minimax theorem (von Neumann, 1928; Sion, 1958), any **SNE** $\pi_\star$ is equivalently characterized as follows:*

$$\pi^\star \in \arg\max_{\pi^1 \in \Pi} \min_{\pi^2 \in \Pi} J_\eta(\pi^1, \pi^2). \tag{3}$$

2.3. Online Interaction Protocol and Regularized Max-Regret

The online contextual RLHF protocol proceeds as follows: At each $t = 1, \ldots, T$, a context $\boldsymbol{x}_t \sim d_0$ is revealed. The learner chooses policies $\hat{\pi}^1_t(\cdot|\boldsymbol{x}_t)$ and $\hat{\pi}^2_t(\cdot|\boldsymbol{x}_t)$, samples actions $\boldsymbol{a}^1_t$ and $\boldsymbol{a}^2_t$, and receives a bandit feedback $r_t \sim \mathrm{Ber}(P^*(\boldsymbol{a}^1_t \succ \boldsymbol{a}^2_t \mid \boldsymbol{x}_t))$. This constitutes a contextual symmetric two-player zero-sum game (Balduzzi et al., 2019) with bandit feedback. Note that this is a *self-play* framework, as the learner controls *both* players to learn by playing against itself to compute a NE.

We evaluate any resulting policy sequence $\{(\hat{\pi}^1_t, \hat{\pi}^2_t)\}_{t \in [T]}$ using ***regularized max-regret***:

$$\mathrm{MBR\text{-}Reg}_\eta(T) := \sum_{t=1}^{T} \mathrm{DGap}_\eta(\hat{\pi}^1_t), \tag{4}$$

$$\mathrm{DGap}_\eta(\hat{\pi}^1_t) := \frac{1}{2} - \min_{\pi^2 \in \Pi} J_\eta(\hat{\pi}^1_t, \pi^2) \tag{5}$$

where we refer to $\mathrm{DGap}_\eta(\hat{\pi}^1_t)$ as the **(symmetric) dual gap** of a policy $\hat{\pi}^1_t \in \Pi$, which quantifies how close $\mathrm{DGap}_\eta(\hat{\pi}^1_t)$ is to an ***SNE***.

Remark 2.4. *While there are multiple definitions of "regret" in a two-player zero-sum game, our work focuses on max-regret (specifically, the Max-Best-Response regret). As discussed in Section 1, this has been widely considered in self-play RL in two-player zero-sum games (Bai & Jin, 2020;*

Bai et al., 2020; Liu et al., 2021; Jin et al., 2022; Xiong et al., 2022) and theoretical analyses of online RLHF under general preferences (Ye et al., 2024; Wu et al., 2025). Notably, the max-regret can be converted to a sample for finding an NE via online-to-batch conversion (Freund & Schapire, 1999). We refer the reader to Appendix K for a detailed discussion of alternative regret definitions.

3. A New Analysis of Regularized Regret

We first present our main technical contribution: a novel bound on the instantaneous dual gap of any greedy NE policy. Denoting $\boldsymbol{\phi} \sim \pi$ as sampling a $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{a}) \in \mathcal{B}^d(1)$ from $\boldsymbol{a} \sim \pi(\cdot \mid \boldsymbol{x})$ and $\boldsymbol{x} \sim d_0$:

Theorem 3.1. *For any estimator $\widehat{\boldsymbol{\Theta}}_t \in \mathrm{Skew}(d)$ at time t, define the max-player's policy as*

$$\hat{\pi}_t \leftarrow \arg\max_{\pi^1} \min_{\pi^2} J_\eta(\pi^1, \pi^2; \widehat{\boldsymbol{\Theta}}_t). \tag{6}$$

Then, the instantaneous dual gap is bounded as follows: denoting $\boldsymbol{E}_t := \boldsymbol{\Theta}_\star - \widehat{\boldsymbol{\Theta}}_t$,

$$\mathrm{DGap}_\eta(\hat{\pi}_t) \le (2L_\mu^2 \eta\beta + L_\mu)\mathbb{E}_{\boldsymbol{\phi}\sim\hat{\pi}_t}\left[\|\boldsymbol{E}_t\boldsymbol{\phi}\|_2^2\right], \tag{7}$$

$$\mathrm{DGap}_\eta(\hat{\pi}_t) \le L_\mu\left(\sqrt{\mathbb{E}_{\boldsymbol{\phi}}\left[\|\boldsymbol{E}\boldsymbol{\phi}\|_2^2\right]} + \frac{1}{2}\mathbb{E}_{\boldsymbol{\phi}}\left[\|\boldsymbol{E}\boldsymbol{\phi}\|_2^2\right]\right). \tag{8}$$

Crucially, this bound holds for *any* choice of estimator and *any* β^{-1}-strongly convex regularizer $\psi(\cdot)$. As long as $\eta < \infty$ (i.e., the regularization by ψ exists), the instantaneous dual gap is bounded *quadratically* by the expected estimation error of $\boldsymbol{\Theta}_\star$ *along the features* of $\hat{\pi}_t$. As we will see in the proof, without strong convexity, one only recovers a linear dependence $\mathbb{E}_{\boldsymbol{\phi}}[\|\boldsymbol{E}_t\boldsymbol{\phi}\|_2]$.

3.1. Proof of Theorem 3.1

Regret Decomposition and Taylor Expansion. For simplicity let us omit dependencies on t. Let $\tilde{\pi} = \arg\min_{\pi\in\Pi} J_\eta(\hat{\pi}, \pi)$ be the min-player's best response to $\hat{\pi}$ with respect to the true objective. We denote the *dual gap* of $\hat{\pi}$ as $X := \frac{1}{2} - J_\eta(\hat{\pi}, \tilde{\pi})$. Decomposing the regret, we have that

$$X = J_\eta(\hat{\pi}, \tilde{\pi}; \widehat{\boldsymbol{\Theta}}) - J_\eta(\hat{\pi}, \tilde{\pi}) + \frac{1}{2} - J_\eta(\hat{\pi}, \tilde{\pi}; \widehat{\boldsymbol{\Theta}}) \tag{9}$$

$$\overset{(*)}{\le} J_\eta(\hat{\pi}, \tilde{\pi}; \widehat{\boldsymbol{\Theta}}) - J_\eta(\hat{\pi}, \tilde{\pi}) \tag{10}$$

$$= J(\hat{\pi}, \tilde{\pi}; \widehat{\boldsymbol{\Theta}}) - J(\hat{\pi}, \tilde{\pi}),$$

(regularization terms cancel out)

where the inequality $(*)$ holds due to the following reasoning. First, we establish the following important property of symmetric game:

Lemma 3.2. *For any $\widehat{\boldsymbol{\Theta}} \in \mathrm{Skew}(d)$, the value of the game, $\max_{\pi^1}\min_{\pi^2} J_\eta(\pi^1, \pi^2; \widehat{\boldsymbol{\Theta}})$, is $\frac{1}{2}$.*

Proof. For any $\widehat{\boldsymbol{\Theta}} \in \mathrm{Skew}(d)$, there always exists a ***SNE*** (Swamy et al., 2024, Lemma 2.1). As the game value of ***SNE*** is $\frac{1}{2}$, it must be so for any **NE** (von Neumann, 1928; Sion, 1958). □

With this, we have that $\frac{1}{2} = \min_{\pi\in\Pi} J_\eta(\hat{\pi}, \pi; \widehat{\boldsymbol{\Theta}})$ for any given $\widehat{\boldsymbol{\Theta}} \in \mathrm{Skew}(d)$. Then, it is easy to see that $\min_{\pi\in\Pi} J_\eta(\hat{\pi}, \pi; \widehat{\boldsymbol{\Theta}}) \le J_\eta(\hat{\pi}, \tilde{\pi}; \widehat{\boldsymbol{\Theta}})$.

Let us denote $\mathbb{E} := \mathbb{E}_{\boldsymbol{x}\sim d_0}\mathbb{E}_{\boldsymbol{\phi}\sim\hat{\pi}(\cdot|\boldsymbol{x}),\tilde{\boldsymbol{\phi}}\sim\tilde{\pi}(\cdot|\boldsymbol{x})}$ when clear from the context. Inspired by the regret analyses of logistic and generalized linear bandits (Abeille et al., 2021; Lee et al., 2024a;b), applying a ***Taylor expansion with integral remainder*** yields:

$$J(\hat{\pi}, \tilde{\pi}; \widehat{\boldsymbol{\Theta}}) - J(\hat{\pi}, \tilde{\pi}) = \underbrace{-\mathbb{E}\left[\dot{\mu}(\boldsymbol{\phi}^\top\boldsymbol{\Theta}_\star\tilde{\boldsymbol{\phi}})\boldsymbol{\phi}^\top\boldsymbol{E}\tilde{\boldsymbol{\phi}}\right]}_{(a)}$$

$$+ \underbrace{\mathbb{E}\left[\left[\int_0^1 (1-z)\ddot{\mu}\left(\boldsymbol{\phi}^\top\left(\boldsymbol{\Theta}_\star - z\boldsymbol{E}_t\right)\tilde{\boldsymbol{\phi}}\right)dz\right](\boldsymbol{\phi}^\top\boldsymbol{E}\tilde{\boldsymbol{\phi}})^2\right]}_{(b)}.$$

Bounding the First-Order Term (a). There are two key technical lemmas that are crucial in obtaining the self-bounding inequality later.

The first lemma relates the term (a) to $D := \sqrt{\mathbb{E}_{\boldsymbol{x}\sim d_0}\left[\|\hat{\pi}(\cdot \mid \boldsymbol{x}) - \tilde{\pi}(\cdot \mid \boldsymbol{x})\|_1^2\right]}$. Its proof, which combines skew-symmetry and the variational representation of the ℓ_1-norm (Müller, 1997), is presented at the end of this subsection:

Lemma 3.3. $\left|\mathbb{E}\left[\dot{\mu}(\boldsymbol{\phi}^\top\boldsymbol{\Theta}_\star\tilde{\boldsymbol{\phi}})\boldsymbol{\phi}^\top\boldsymbol{E}\tilde{\boldsymbol{\phi}}\right]\right| \le L_\mu D\sqrt{\mathbb{E}_{\boldsymbol{\phi}}\left[\|\boldsymbol{E}\boldsymbol{\phi}\|_2^2\right]}$.

The second lemma, whose proof is deferred to Appendix B, allows for us to bound D with the dual gap X, other than the naïve bound of $D \le 1$:

Lemma 3.4. $X = \frac{1}{2} - \mathbb{E}_{\boldsymbol{x}\sim d_0}[J_\eta(\hat{\pi}, \tilde{\pi} \mid \boldsymbol{x})] \ge (2\eta\beta)^{-1}D$.

We then chain everything, along with the naïve bound of

$D \le 1$, to obtain the following:

$$(a) \le L_\mu \sqrt{(1 \wedge 2\eta\beta X)\, \mathbb{E}_{\boldsymbol{\phi}} \left[\|\boldsymbol{E}\boldsymbol{\phi}\|_2^2 \right]}. \tag{11}$$

Bounding the Second-Order Term (b). We first bound (b) by noting that $\ddot{\mu}(\cdot) \le L_\mu$, giving $(b) \le L_\mu \mathbb{E}[(\boldsymbol{\phi}^\top \boldsymbol{E}\tilde{\boldsymbol{\phi}})^2] \int_0^1 (1-z)dz = \frac{L_\mu}{2}\mathbb{E}[(\boldsymbol{\phi}^\top \boldsymbol{E}\tilde{\boldsymbol{\phi}})^2]$. For clarity, let us distinguish between the independent expectations $\mathbb{E}_{\boldsymbol{\phi}}$ and $\mathbb{E}_{\tilde{\boldsymbol{\phi}}}$. We have:

$$\mathbb{E}[(\boldsymbol{\phi}^\top \boldsymbol{E}\tilde{\boldsymbol{\phi}})^2] = \mathbb{E}_{\boldsymbol{\phi}}\mathbb{E}_{\tilde{\boldsymbol{\phi}}} \left[\tilde{\boldsymbol{\phi}}^\top \boldsymbol{E}^\top \boldsymbol{\phi}\boldsymbol{\phi}^\top \boldsymbol{E}\tilde{\boldsymbol{\phi}} \right] \tag{12}$$

$$\le \mathbb{E}_{\boldsymbol{\phi}} \left[\max_{\tilde{\boldsymbol{\phi}} \in \mathcal{B}^d(1)} \tilde{\boldsymbol{\phi}}^\top (\boldsymbol{E}^\top \boldsymbol{\phi}\boldsymbol{\phi}^\top \boldsymbol{E})\tilde{\boldsymbol{\phi}} \right] = \mathbb{E}_{\boldsymbol{\phi}} \left[\|\boldsymbol{E}\boldsymbol{\phi}\|_2^2 \right], \tag{13}$$

where the last equality follows from the variational definition of the operator norm.[1]

Combining Everything. Combining these bounds for (a) and (b), we establish a ***self-bounding quadratic inequality*** in the dual gap X:

$$X \le L_\mu \sqrt{(1 \wedge 2\eta\beta X)\, \mathbb{E}_{\boldsymbol{\phi}} \left[\|\boldsymbol{E}\boldsymbol{\phi}\|_2^2 \right]} + \frac{L_\mu}{2} \mathbb{E}_{\boldsymbol{\phi}} \left[\|\boldsymbol{E}\boldsymbol{\phi}\|_2^2 \right]. \tag{14}$$

Solving for X *simultaneously* yields $X \le 2L_\mu^2 \eta\beta \mathbb{E}_{\boldsymbol{\phi}}[\|\boldsymbol{E}\boldsymbol{\phi}\|_2]^2 + L_\mu \mathbb{E}_{\boldsymbol{\phi}}[\|\boldsymbol{E}\boldsymbol{\phi}\|_2^2]$ *and* $X \le L_\mu \sqrt{\mathbb{E}_{\boldsymbol{\phi}} \left[\|\boldsymbol{E}\boldsymbol{\phi}\|_2^2 \right]} + \frac{L_\mu}{2} \mathbb{E}_{\boldsymbol{\phi}} \left[\|\boldsymbol{E}\boldsymbol{\phi}\|_2^2 \right]$. □

Proof of Lemma 3.3. First, using the symmetry of μ and the skew-symmetry of $\boldsymbol{E}$, we have:

$$Z \triangleq \mathbb{E}_{\boldsymbol{x} \sim d_0} \mathbb{E}_{\boldsymbol{\phi}, \tilde{\boldsymbol{\phi}} \sim \hat{\pi}(\cdot|\boldsymbol{x})} \left[\dot{\mu}(\boldsymbol{\phi}^\top \boldsymbol{\Theta}_\star \tilde{\boldsymbol{\phi}}) \boldsymbol{\phi}^\top \boldsymbol{E}\tilde{\boldsymbol{\phi}} \right] = 0, \tag{15}$$

where with a slight abuse of notation, $\boldsymbol{\phi} \sim \hat{\pi}(\cdot \mid \boldsymbol{x})$ denotes sampling $\boldsymbol{\phi}(\boldsymbol{x}, \boldsymbol{a})$ with $\boldsymbol{a} \sim \hat{\pi}(\cdot \mid \boldsymbol{x})$. Denote $f(\tilde{\boldsymbol{\phi}}; \boldsymbol{\phi}) := \dot{\mu}(\boldsymbol{\phi}^\top \boldsymbol{\Theta}_\star \tilde{\boldsymbol{\phi}}) \boldsymbol{\phi}^\top \boldsymbol{E}\tilde{\boldsymbol{\phi}}$, which satisfies $\max_{\tilde{\boldsymbol{\phi}} \in \mathcal{B}^d(1)} \left| f(\tilde{\boldsymbol{\phi}}; \boldsymbol{\phi}) \right| \le L_\mu \|\boldsymbol{E}\boldsymbol{\phi}\|_2$. Then,

$$\left| \mathbb{E} \left[\dot{\mu}(\boldsymbol{\phi}^\top \boldsymbol{\Theta}_\star \tilde{\boldsymbol{\phi}}) \boldsymbol{\phi}^\top \boldsymbol{E}\tilde{\boldsymbol{\phi}} \right] \right| \tag{16}$$

$$= \left| \mathbb{E}_{\boldsymbol{x} \sim d_0} \mathbb{E}_{\boldsymbol{\phi} \sim \hat{\pi}(\cdot|\boldsymbol{x}), \tilde{\boldsymbol{\phi}} \sim \tilde{\pi}(\cdot|\boldsymbol{x})} \left[f(\tilde{\boldsymbol{\phi}}; \boldsymbol{\phi}) \right] - Z \right| \tag{17}$$

$$\le \mathbb{E}_{\boldsymbol{x} \sim d_0} \mathbb{E}_{\boldsymbol{\phi} \sim \hat{\pi}(\cdot|\boldsymbol{x})} \tag{18}$$

$$\left[\left| \mathbb{E}_{\tilde{\boldsymbol{\phi}} \sim \hat{\pi}(\cdot|\boldsymbol{x})} \left[f(\tilde{\boldsymbol{\phi}}; \boldsymbol{\phi}) \right] - \mathbb{E}_{\tilde{\boldsymbol{\phi}} \sim \tilde{\pi}(\cdot|\boldsymbol{x})} \left[f(\tilde{\boldsymbol{\phi}}; \boldsymbol{\phi}) \right] \right| \right] \tag{19}$$

$$\overset{(*)}{\le} \mathbb{E}_{\boldsymbol{x} \sim d_0} \mathbb{E}_{\boldsymbol{\phi} \sim \hat{\pi}(\cdot|\boldsymbol{x})} \tag{20}$$

$$\left[L_\mu \|\boldsymbol{E}\boldsymbol{\phi}\|_2 \sup_{g \in \mathcal{G}_\infty(1)} \left| \int g(\tilde{\boldsymbol{\phi}}) d(\hat{\pi}(\cdot \mid \boldsymbol{x}) - \tilde{\pi}(\cdot \mid \boldsymbol{x}))(\tilde{\boldsymbol{\phi}}) \right| \right] \tag{21}$$

$$\overset{(**)}{=} L_\mu \mathbb{E}_{\boldsymbol{x} \sim d_0} \left[\mathbb{E}_{\boldsymbol{\phi} \sim \hat{\pi}(\cdot|\boldsymbol{x})} \left[\|\boldsymbol{E}\boldsymbol{\phi}\|_2 \right] \|\hat{\pi}(\cdot \mid \boldsymbol{x}) - \tilde{\pi}(\cdot \mid \boldsymbol{x})\|_1 \right], \tag{22}$$

where $(*)$ defines $\mathcal{G}_\infty(1) := \left\{ g : \mathcal{B}^d(1) \to [-1, 1] \mid g \text{ is measurable} \right\}$, and $(**)$ follows from the *integral probability metric representation (IPM) of the ℓ_1-norm* (Müller, 1997, Theorem 5.4).

We conclude by decoupling the $\boldsymbol{E}$ term and the ℓ_1-error term via Cauchy-Schwarz as follows:

$$\left| \mathbb{E} \left[\dot{\mu}(\boldsymbol{\phi}^\top \boldsymbol{\Theta}_\star \tilde{\boldsymbol{\phi}}) \boldsymbol{\phi}^\top \boldsymbol{E}\tilde{\boldsymbol{\phi}} \right] \right| \tag{23}$$

$$\le L_\mu \sqrt{\mathbb{E}_{\boldsymbol{x} \sim d_0} \left[\mathbb{E}_{\boldsymbol{\phi} \sim \hat{\pi}(\cdot|\boldsymbol{x})} \left[\|\boldsymbol{E}\boldsymbol{\phi}\|_2 \right]^2 \right]} \tag{24}$$

$$\times \sqrt{\mathbb{E}_{\boldsymbol{x} \sim d_0} \left[\|\hat{\pi}(\cdot \mid \boldsymbol{x}) - \tilde{\pi}(\cdot \mid \boldsymbol{x})\|_1^2 \right]} \tag{25}$$

$$\le L_\mu D \sqrt{\mathbb{E}_{\boldsymbol{\phi}} \left[\|\boldsymbol{E}\boldsymbol{\phi}\|_2^2 \right]}.$$

(Jensen's inequality w.r.t. $\mathbb{E}_{\boldsymbol{x} \sim d_0}[\cdot]$)

□

3.2. Discussions

Crucially, our proof relies strictly on the strong convexity of $\psi(\cdot)$ and entirely avoids any reliance on the specific algebraic properties of the KL divergence, departing from prior KL-centric analyses (Wu et al., 2025; Nayak et al., 2025; Ye et al., 2024). The central technical mechanism driving our proof is the ***self-bounding inequality*** presented in Eqn. (14). While our use of a Taylor expansion is inspired by the regret analyses of logistic and generalized linear bandits (Abeille et al., 2021; Lee et al., 2024a), our technical execution differs significantly. Prior works rely on self-concordance to control both terms and establish a similar self-bounding inequality; in contrast, our approach hinges on strong convexity, which is formalized through two key lemmas.

First, Lemma 3.3 ensures that the first-order term (a) is upper-bounded by the distance D between our policy $\hat{\pi}$ and the adversary's worst-possible policy $\tilde{\pi}$. To achieve this, the lemma leverages the skew-symmetry of the preference matrix and the symmetry of the link function. Notably, the proof establishes a surprising connection to the integral probability metric (IPM) representation of the ℓ_1-norm. This connection, combined with an appropriate application of the Cauchy-Schwarz inequality, cleanly decouples the estimation error D.

Second, Lemma 3.4 ensures that D is subsequently bounded

[1] $\|\boldsymbol{E}\|_{\mathrm{op}} = \sup_{\boldsymbol{x}, \boldsymbol{y} \in \mathcal{B}^d(1)} \boldsymbol{x}^\top \boldsymbol{E}\boldsymbol{y}$ for any $\boldsymbol{E} \in \mathbb{R}^{d \times d}$.

by the dual gap. By applying the zeroth-order characterization of strong convexity (Nesterov, 2018, Theorem 2.1.9), we effortlessly relate the policy divergence to the suboptimality gap without needing to compute functional derivatives (since our optimality is defined with respect to policies).

4. From Quadratic Error Bound to Fast Regrets

4.1. Feature Diversity and Computation Oracles

Feature Diversity. For the regret analyses, we consider the following assumption:

Assumption 4.1 (Feature Diversity)**.** *The learner has access to an exploration policy $\rho(\cdot \mid \boldsymbol{x})$ such that $\lambda_{\min}\left(\mathbb{E}_{\boldsymbol{x}\sim d_0}\mathbb{E}_{\boldsymbol{a}\sim\rho(\cdot|\boldsymbol{x})}\left[\phi(\boldsymbol{x},\boldsymbol{a})\phi(\boldsymbol{x},\boldsymbol{a})^\top\right]\right) \geq C_{\min}$ for some $C_{\min} > 0$.*

By considering this assumption, we can cleanly isolate the geometric impact of the strongly convex regularizer $\psi(\cdot)$ from the complexities of active exploration. By abstracting away the exploration mechanism, we can directly answer our core theoretical question: whether fast rates are achievable under generic regularization when given sufficient data coverage.

We highlight four important aspects regarding the above assumption:

- **Theoretical Necessity.** In standard bandit settings, achieving fast rates typically requires explicit algorithmic exploration (e.g., optimism or Thompson Sampling). When relying on greedy sampling strategies (which we consider in Section 4.2), assumptions regarding the diversity of the feature mapping are standard and necessary (Goldenshluger & Zeevi, 2013; Kannan et al., 2018; Wu et al., 2020; Bastani et al., 2021; Bogunovic et al., 2021; Kim & Oh, 2024). Furthermore, in the high-dimensional regime (which we consider in Section 4.3), sufficient feature space coverage is information-theoretically unavoidable for obtaining dimension-wise improved regret bounds (Hao et al., 2020; Li et al., 2022b; Zeng & Honorio, 2025).
- **Empirical Plausibility.** While motivated by theoretical tractability, this assumption has reasonable analogues in practical RLHF pipelines (Dong et al., 2024; Bai et al., 2022). Practitioners often use generative strategies that naturally induce diversity without explicit algorithmic exploration. For example, querying an ensemble of LLMs or sampling from LLMs with various temperatures yield a diverse set of responses across the feature space (Troshin et al., 2025; Minh et al., 2025). Under this interpretation, we argue that this assumption is a sensible theoretical simplification.
- **Statistical Tractability.** As detailed in Appendix C, obtaining an approximate exploration policy (e.g., a ρ satisfying the definition with $C_{\min}/2$ or $C_{\min}/d$) can be achieved in a computationally tractable manner with a T-independent statistical cost in regret. This relies on the minimal requirement that the underlying context distribution $d_0(\cdot)$ has adequate coverage over the feature space; without it, the learner would face "blind spots" in $\mathbb{R}^d$ and inherently fail to learn globally.
- **Scaling of $C_{\min}$.** The scaling of $C_{\min}$ is intrinsically tied to the geometry of the given feature map. Because $\|\phi(\boldsymbol{x},\boldsymbol{a})\|_2 \leq 1$, it strictly follows that $C_{\min}^{-1} \geq d$. Well-conditioned sets (e.g., unit-normalized hypercubes, standard basis) yield $C_{\min}^{-1} \asymp d$, while ill-conditioned sets yield worse.

Computation Oracles. We lastly describe the computation model. We assume the learner is *tractable* (not necessarily efficient), accessing $\mathcal{A}$ and $\mathrm{Skew}(d)$ only via:

Oracle 1. ***Sampling:*** *Given $\boldsymbol{x} \in \mathcal{X}$ and $\pi \in \Pi$, output a sample $\boldsymbol{a} \sim \pi(\cdot|\boldsymbol{x})$.*

Oracle 2. ***Regularized MLE:*** *Compute a regularized MLE over $\mathrm{Skew}(d)$.*[2]

Oracle 3. ***Population NE:*** *Given $\boldsymbol{\Theta}$, output population **SNE**: $\arg\max_{\pi^1}\min_{\pi^2} J_\eta(\pi^1,\pi^2;\boldsymbol{\Theta})$.*

The last oracle has been considered in prior online RLHF literature (Ye et al., 2024; Wu et al., 2025) and learning in regularized games under bandit feedback (Yang et al., 2025; Nayak et al., 2025).

4.2. Polylogarithmic Regret via Greedy Sampling: $\tilde{\mathcal{O}}(\eta(\log T)^2 \wedge \sqrt{T})$

In this section, we assume the link function is logistic, $\mu(z) = (1+e^{-z})^{-1}$, rendering the generalized linear model (GLM) well-specified as Bernoulli that admits a tight confidence sequence (Lee et al., 2024a, Theorem 3.2). We discuss extensions to generic link functions μ that preserve the dependencies on d and T in Remark 4.3. Additionally, we temporarily set aside the low-rank structure of $\boldsymbol{\Theta}_\star$ to focus purely on achieving polylogarithmic regret; we will revisit rank-exploitation in Section 4.3 to improve upon d dependencies at the cost of obtaining $\sqrt{T}$ regret.

We demonstrate that a surprisingly simple algorithm, `Greedy Sampling (GS)` (Algorithm 1 in Appendix A), is sufficient to obtain $\tilde{\mathcal{O}}(\eta(\log T)^2 \wedge \sqrt{T})$.

[2] In our case, as both the negative log-likelihood and nuclear norm penalty are convex, this is a convex optimization. This can be implemented by simply reparametrizing $\boldsymbol{\Theta} \in \mathrm{Skew}(d)$ as $\frac{\boldsymbol{\Theta}'-\boldsymbol{\Theta}'^\top}{2}$ for unconstrained $\boldsymbol{\Theta}' \in \mathbb{R}^{d\times d}$.

Under GS, the max-player perpetually plays the greedy NE policy with respect to the current MLE $\widehat{\boldsymbol{\Theta}}_t$, while the min-player explores using the coverage policy ρ (Assumption 4.1). We show that GS attains the following polylogarithmic regret:

Theorem 4.2. *Let $\delta \in (0,1)$ and suppose that $d^2 \log \frac{T}{d} \gtrsim \kappa^{-1} \log \frac{1}{\delta}$. Then, with probability at least $1-\delta$, GS simultaneously attains the following bounds:*

$$\text{MBR-Reg}_\eta(T) \lesssim \eta\beta\kappa^{-1} d^4 C_{\min}^{-1} \left(\log \frac{T}{d}\right)^2, \quad (26)$$

$$\text{MBR-Reg}_\eta(T) \lesssim \kappa^{-\frac{1}{2}} C_{\min}^{-\frac{1}{2}} d^2 \sqrt{T} \log \frac{T}{d}. \quad (27)$$

Proof Sketch. We provide a high-level sketch of the proof here; the full detail is deferred to Appendix D. The analysis proceeds in three main steps:

1. From Regret to Sum of Squared Errors. We begin with Theorem 3.1, which bounds the instantaneous dual gap by the *squared* estimation error: $\text{DGap}(\hat{\pi}_t) \lesssim \mathbb{E}_{\boldsymbol{\phi}\sim\hat{\pi}_t}[\|\boldsymbol{E}_t\boldsymbol{\phi}\|_2^2]$. Using a standard basis decomposition, Cauchy-Schwarz with respect to the regularized Hessian of the log-likelihood loss $\mathcal{L}_t(\cdot)$ at each time t (see Appendix D for the full definition), $\widehat{\boldsymbol{H}}_t \triangleq \boldsymbol{I}_{d^2} + \nabla^2 \mathcal{L}_t(\hat{\boldsymbol{\theta}}_t) \in \mathbb{R}^{d^2\times d^2}$, and the confidence sequence for the constrained MLE (Lee et al., 2024a, Theorem 3.2), we further bound this error by the sum of *expected* elliptical potentials. Here, one side is the standard basis $\boldsymbol{e}_j$ and the other is chosen by $\hat{\pi}_t$:

$$\text{DGap}(\hat{\pi}_t) \lesssim (d^2 \log T) \sum_{j=1}^{d} \mathbb{E}_{\boldsymbol{\phi}\sim\hat{\pi}_t}\left[\|\boldsymbol{\phi}\otimes\boldsymbol{e}_j\|^2_{\widehat{\boldsymbol{H}}_t^{-1}}\right]. \quad (28)$$

2. Towards Expected Elliptical Potentials. A discrepancy arises on the right-hand side: the upper bound involves a sum over fixed *basis* vectors on one side, whereas the empirical Hessian aggregates the *played* features on *both* sides. More specifically, $\widehat{\boldsymbol{H}}_t$ is the (weighted) sum of outer products of $\text{vec}(\boldsymbol{\phi}_t\tilde{\boldsymbol{\phi}}_t^\top)$, while the term inside the Mahalanobis norm is $\text{vec}(\boldsymbol{\phi}_t\boldsymbol{e}_j^\top)$. We resolve this via our ***Coverage Lemma*** (Lemma D.2), which leverages the feature diversity assumption (Assumption 4.1) to "transform" $\boldsymbol{e}_j$ to $\tilde{\boldsymbol{\phi}}_t$ at the cost of $C_{\min}^{-1}$. Then, $\sum_t \text{DGap}(\hat{\pi}_t)$ is bounded by:

$$\sum_{t=1}^{T} \text{DGap}(\hat{\pi}_t) \lesssim (d^2\log T)\kappa^{-1}C_{\min}^{-1} \quad (29)$$

$$\times \underbrace{\sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\phi}_t\sim\hat{\pi}_t,\tilde{\boldsymbol{\phi}}_t\sim\rho}\left[\left\|\text{vec}(\boldsymbol{\phi}_t\tilde{\boldsymbol{\phi}}_t^\top)\right\|^2_{\boldsymbol{V}_t^{-1}}\right]}_{\triangleq S_T}, \quad (30)$$

where $\boldsymbol{V}_t \triangleq \boldsymbol{I}_{d^2} + \sum_{s=1}^{t-1} \text{vec}(\boldsymbol{\phi}_s\tilde{\boldsymbol{\phi}}_s^\top)\text{vec}(\boldsymbol{\phi}_s\tilde{\boldsymbol{\phi}}_s^\top)^\top$.

3. Martingale Concentration for Realized Variance. The final challenge is to bound the sum of *expected* elliptical potentials S_T. The standard Elliptical Potential Lemma (Abbasi-Yadkori et al., 2011, Lemma 11) controls the sum of *realized* potentials, and thus cannot be applied directly. To bridge this gap, we decompose the term into the realized sum plus a *sum of martingale differences* as follows: denoting $\boldsymbol{v}_t \coloneqq \text{vec}(\boldsymbol{\phi}_s\tilde{\boldsymbol{\phi}}_s^\top)$,

$$S_T = \underbrace{\sum_{t=1}^{T} \|\boldsymbol{v}_t\|^2_{\boldsymbol{V}_t^{-1}}}_{(a)} + \underbrace{\sum_{t=1}^{T}\left\{\mathbb{E}_{t-1}[\|\boldsymbol{v}_t\|^2_{\boldsymbol{V}_t^{-1}}] - \|\boldsymbol{v}_t\|^2_{\boldsymbol{V}_t^{-1}}\right\}}_{(b)}. \quad (31)$$

As mentioned, (a) is bounded by the standard Elliptical Potential Lemma. Bounding (b) represents another technical novelty. We first apply an empirical Freedman's inequality (Freedman, 1975; Beygelzimer et al., 2011; Lee et al., 2024b), which bounds (b) as $(b) \lesssim AS_T$ for some constant $A \in (0,1)$. This leads to a ***linear self-bounding inequality*** of the form $S_T \le AS_T + \tilde{\mathcal{O}}(d^2\log T)$. Solving this recursion yields the final $\tilde{\mathcal{O}}(d^4(\log T)^2)$ regret.

Lastly, $\tilde{\mathcal{O}}(d^2\sqrt{T})$ is similarly obtained via the η-free (linear) error bound of Theorem 3.1. □

Discussions. While the quadratic error bound (Theorem 3.1) plays a critical role in achieving our results, we highlight two additional technical contributions stemming from our regret analysis.

First, we clarify the theoretical trade-offs compared to recent literature. Wu et al. (2025) established an $\tilde{\mathcal{O}}(e^{9\eta}(\log T)^2)$ regret bound for GS[3] *without* requiring a feature coverage assumption. However, their bound scales exponentially with η, is restricted exclusively to reverse KL regularization ($\psi(\cdot) = D_{\text{KL}}(\cdot,\pi_{\text{ref}})$), and fails to yield a meaningful guarantee as $\eta\to\infty$ (where it should technically recover the standard $\sqrt{T}$ unregularized regret; see Appendix H for a detailed discussion). In contrast, our analysis utilizes the coverage assumption to completely eliminate the $e^{\mathcal{O}(\eta)}$ penalty, trading it for the geometric, regularization-independent quantity $C_{\min}^{-1}$. Furthermore, echoing the spirit of Nayak et al. (2025, Theorem 2.1), our regret bound gracefully adapts to both regimes: it yields an $e^{\mathcal{O}(\eta)}$-free $(\log T)^2$ rate for small η, and seamlessly transitions to $\sqrt{T}$ as η increases. In Appendix L, we illustrate our theoretical regret bounds with preliminary numerical experiments.

Second, our martingale concentration technique offers a streamlined alternative to recent analyses. To handle a sim-

[3] Technically, their algorithm is intractable under ***GBPM***, as it requires minimizing the cross-entropy loss over the constrained space $\text{Skew}(d, 2r; S)$.

ilar expected potential sum in the KL-regularized multi-armed bandit scenario, Ji et al. (2026) employ a peeling argument that inherently incurs an additional doubly logarithmic terms. By utilizing Freedman's inequality to establish a ***linear self-bounding inequality***, our approach is strictly tighter and cleanly avoids any such logarithmic artifacts.

Relations to Prior Works. Prior online RLHF analyses have predominantly relied on the KL-specific properties of the exponential family (e.g., bounded log-density ratios) to enable oracle reductions to least squares regression (Cesa-Bianchi & Lugosi, 2006; Foster & Rakhlin, 2020; Zhang, 2022); we provide a rigorous instantiation of Wu et al. (2025) to ***GBPM*** in Appendix H to illustrate this reliance. While generic regularized formulations have appeared in recent literature (Tang et al., 2025), rigorous statistical guarantees for them have remained absent.

Crucially, our analysis unifies these diverse regularizers in the online setting. We demonstrate that the specific geometry of KL is not strictly necessary for fast rates; rather, it is the *strong convexity* of the regularizer that drives our results. This broadens the theoretical horizon to encompass a wide array of regularizers, including the sum of reverse KLs (Le et al., 2025; Aminian et al., 2025), Shannon entropy (McKelvey & Palfrey, 1995; Mertikopoulos & Sandholm, 2016; Cen et al., 2024), Tsallis entropy (Tsallis, 1988; Lee et al., 2018; Yang et al., 2019; Zimmert & Seldin, 2021), the χ^2-divergence (Huang et al., 2025b), and general f-divergences (Liese & Vajda, 2006; Go et al., 2023; Wang et al., 2024; Han et al., 2025).

Very recently, Zhang et al. (2026b) demonstrated that for KL-regularized zero-sum games, greedy sampling—computing a least squares estimate and performing equilibrium computation without any explicit pessimism—suffices to achieve fast rates. Their analysis also bounds the instantaneous duality gap by the *squared* ℓ_1-distance between policies using strong convexity, though it inherently relies on the explicit closed-form softmax structure of the KL-regularized best responses. An interesting future direction is whether our techniques can be combined with theirs to establish similar pessimism-free fast rates for *generally* regularized zero-sum games in the offline scenario.

Finally, we acknowledge recent works establishing the distinct statistical properties of different regularizers in the context of offline (reward-based) RL (Jiang & Xie, 2025; Huang et al., 2025b; Zhao et al., 2026). For instance, Huang et al. (2025b) demonstrated that the χ^2-divergence permits single-policy concentratability (unlike KL), and Zhao et al. (2026) showed that strongly convex f-divergences can achieve fast rates *without* single-policy concentratability. We leave the extension of these divergence-specific offline RLHF nuances within the ***GBPM*** framework to future work.

Remark 4.3 (Beyond Parametric Models)**.** *A bandit problem assuming a specific parametric distribution (e.g., GLM) for the reward is known as a parametric bandit (Filippi et al., 2010). For semi-parametric settings—where we only assume $r_t = \mu(\langle \boldsymbol{\theta}_\star, \boldsymbol{\phi}_t \rangle) + \varepsilon_t$ with bounded noise ε_t—one can adopt the maximum quasi-likelihood estimator approach of Li et al. (2017, Lemma 3), dating back to Chen et al. (1999). This strategy utilizes a T-independent warm-up phase (random sampling) to ensure the design matrix is sufficiently well-conditioned, preserving the dependencies on d and T.*

4.3. $\mathrm{poly}(d)$-free Regret via Explore-Then-Commit: $\widetilde{\mathcal{O}}(\sqrt{\eta r T} \wedge r^{1/3} T^{2/3})$

We now ask: what statistical gains can we achieve by maximally exploiting the *low-rank structure* of $\boldsymbol{\Theta}_\star$? This question is particularly relevant in the *high-dimensional* or *data-poor* regime, where T is not sufficiently large relative to d (specifically, $d^{c_1} \lesssim T \lesssim d^{c_2}$ for constants $0 < c_1 < c_2$). Such regimes are characteristic of modern applications involving high-dimensional features (Tucker et al., 2020; Li et al., 2024).

Here, it is imperative to avoid explicit $\mathrm{poly}(d)$ dependencies in the regret bound, isolating the complexity to the intrinsic rank r and unavoidable dependencies on the feature coverage $C_{\min}^{-1}$ (Assumption 4.1) (Zeng & Honorio, 2025, Table 1), which is known to be unavoidable. To achieve this, we leverage the low-rank structure of $\boldsymbol{\Theta}_\star$ using techniques standard in high-dimensional bandits (Carpentier & Munos, 2012; Hao et al., 2020; Kim & Paik, 2019; Oh et al., 2021; Li et al., 2022b; Lu et al., 2021; Kang et al., 2022; Jang et al., 2022; 2024). In this regime, sufficient initial exploration is requisite to identify and exploit the underlying low-rank subspace.

Following standard approaches in high-dimensional contextual bandits (Hao et al., 2020; Li et al., 2022b; Jang et al., 2024), we employ the `Explore-Then-Commit (ETC)` algorithm. The players explore for T_0 rounds using the coverage policy ρ, compute a symmetric Nash equilibrium (SNE) based on the resulting *nuclear-norm regularized MLE* (Fan et al., 2019; Lee et al., 2026), and symmetrically commit to this policy for the remaining rounds (see Algorithm 2 in Appendix A).

We now present the regret bound for `ETC`, with the full proof deferred to Appendix E, which also relies critically on Theorem 3.1:

Theorem 4.4. *Let $\delta \in (0,1)$, and suppose $T \gtrsim \kappa^{-2} C_{\min}^{-4} dr \log \frac{d}{\delta}$. By setting the regularization parameter $\lambda_{T_0} = \sqrt{\frac{32 L_\mu \log(4d/\delta)}{T_0}}$, `ETC` attains*

$\mathrm{MBR\text{-}Reg}_\eta(T) \lesssim T_0$ *with probability at least* $1-\delta$. *Specifically, depending on the choice of* T_0, `ETC` *achieves the following bounds: If* $T_0 \asymp \kappa^{-1} C_{\min}^{-2} \sqrt{T\eta\beta r \log \frac{d}{\delta}}$, *then*

$$\mathrm{MBR\text{-}Reg}_\eta(T) \lesssim \kappa^{-1} C_{\min}^{-2} \sqrt{T\eta\beta r \log \frac{d}{\delta}} \tag{32}$$

If $T_0 \asymp \left(\kappa^{-2} C_{\min}^{-4} r T^2 \log \frac{d}{\delta}\right)^{1/3}$, *then*

$$\mathrm{MBR\text{-}Reg}_\eta(T) \lesssim \left(\kappa^{-2} C_{\min}^{-4} r T^2 \log \frac{d}{\delta}\right)^{1/3}. \tag{33}$$

Discussions. This result highlights two critical theoretical insights for the high-dimensional regime. First, thanks to the quadratic error bound established in Theorem 3.1, `ETC` is able to achieve a fast $\tilde{\mathcal{O}}(\sqrt{\eta r T})$ rate. Interestingly, this surpasses the $\tilde{\mathcal{O}}(T^{2/3})$ rate typically associated with `ETC` algorithms (Lattimore & Szepesvári, 2020). The tightness of these bounds depends directly on the regularization coefficient η; specifically, focusing on η, d, and T, the $\tilde{\mathcal{O}}(\sqrt{\eta r T})$ bound is asymptotically tighter than the $T^{2/3}$ regret whenever $\eta \lesssim (T/\log d)^{1/3}$. Second, both regrets in Theorem 4.4 explicitly scale with r rather than d. This confirms that our framework can effectively exploit the low-rank structure of general preference games.

5. Conclusion

In this work, we investigated regularized max-regret minimization under the GBPM with bandit feedback, utilizing it as a theoretical abstraction to analyze the statistical impact of general regularizers beyond reverse KL. Under this framework, we demonstrated that "fast" regret rates are *not* an exclusive artifact of KL-geometry, but can be achieved for *any* strongly convex regularizer. Specifically, we first established a novel quadratic error bound on the dual gap that is central to our subsequent results. Its proof, which combines the strong convexity of the regularizer with the skew-symmetry of the preference matrix and the IPM representation of the ℓ_1-distance, stands as a key technical contribution that may be of independent interest. Armed with this crucial theorem, we showed that under a feature diversity assumption (Assumption 4.1), `Greedy Sampling` and `Explore-Then-Commit` yield polylogarithmic and $\mathrm{poly}(d)$-free (up to $C_{\min}^{-1}$) regret, respectively. We detail further future directions in Appendix M.

Impact Statement

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

Acknowledgments and Disclosure of Funding

This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2022-II220311 Development of Goal-Oriented Reinforcement Learning Techniques for Contact-Rich Robotic Manipulation of Everyday Objects, RS-2019-II190075 Artificial Intelligence Graduate School Support Program (KAIST), and RS-2019-II191906 Artificial Intelligence Graduate School Program (POSTECH)), and the National Science Foundation under grant CCF-2327013 and Meta Platforms, Inc.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. Improved Algorithms for Linear Stochastic Bandits. In *Advances in Neural Information Processing Systems*, volume 24, pp. 2312–2320. Curran Associates, Inc., 2011. URL https://sites.ualberta.ca/~szepesva/papers/linear-bandits-NeurIPS2011.pdf.

Abeille, M., Faury, L., and Calauzènes, C. Instance-Wise Minimax-Optimal Algorithms for Logistic Bandits. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 3691–3699. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/abeille21a.html.

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

Aminian, G., Asadi, A. R., Shenfeld, I., and Mroueh, Y. 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.

Auer, P., Cesa-Bianchi, N., Freund, Y., and Schapire, R. Gambling in a rigged casino: The adversarial multi-armed bandit problem. In *Proceedings of IEEE 36th Annual Foundations of Computer Science*, pp. 322–331, 1995. doi:10.1109/SFCS.1995.492488.

Bai, Y. and Jin, C. Provable Self-Play Algorithms for Competitive Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine*

Learning Research, pp. 551–560. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/bai20a.html.

Bai, Y., Jin, C., and Yu, T. Near-Optimal Reinforcement Learning with Self-Play. In *Advances in Neural Information Processing Systems*, volume 33, pp. 2159–2170. Curran Associates, Inc., 2020. URL https://arxiv.org/abs/2006.12007.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Balduzzi, D., Garnelo, M., Bachrach, Y., Czarnecki, W., Perolat, J., Jaderberg, M., and Graepel, T. Open-ended learning in symmetric zero-sum games. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 434–443. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/balduzzi19a.html.

Bastani, H., Bayati, M., and Khosravi, K. Mostly Exploration-Free Algorithms for Contextual Bandits. *Management Science*, 67(3):1329–1349, 2021. doi:10.1287/mnsc.2020.3605.

Bengs, V., Busa-Fekete, R., El Mesaoudi-Paul, A., and Hüllermeier, E. Preference-based Online Learning with Dueling Bandits: A Survey. *Journal of Machine Learning Research*, 22(7):1–108, 2021. URL http://jmlr.org/papers/v22/18-546.html.

Bengs, V., Saha, A., and Hüllermeier, E. Stochastic Contextual Dueling Bandits under Linear Stochastic Transitivity Models. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 1764–1786. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/bengs22a.html.

Besson, L. and Kaufmann, E. What Doubling Tricks Can and Can't Do for Multi-Armed Bandits. *arXiv preprint arXiv:1803.06971*, 2018. URL https://arxiv.org/abs/1803.06971.

Beygelzimer, A., Langford, J., Li, L., Reyzin, L., and Schapire, R. Contextual Bandit Algorithms with Supervised Learning Guarantees. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pp. 19–26, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/beygelzimer11a.html.

Bogunovic, I., Losalka, A., Krause, A., and Scarlett, J. Stochastic Linear Bandits Robust to Adversarial Attacks. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 991–999. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/bogunovic21a.html.

Bradley, R. A. and Terry, M. E. Rank Analysis of Incomplete Block Designs: The Method of Paired Comparisons. *Biometrika*, 39(3-4):324–345, 1952. doi:10.1093/biomet/39.3-4.324.

Candès, E. J. and Plan, Y. Tight Oracle Inequalities for Low-Rank Matrix Recovery From a Minimal Number of Noisy Random Measurements. *IEEE Transactions on Information Theory*, 57(4):2342–2359, 2011. doi:10.1109/TIT.2011.2111771.

Carpentier, A. and Munos, R. Bandit Theory meets Compressed Sensing for high dimensional Stochastic Linear Bandit. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pp. 190–198, La Palma, Canary Islands, 21–23 Apr 2012. PMLR. URL https://proceedings.mlr.press/v22/carpentier12.html.

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T. T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P. J., Damani, M., Slocum, S., Anwar, U., Siththaranjan, A., Nadeau, M., Michaud, E. J., Pfau, J., Krasheninnikov, D., Chen, X., Langosco, L., Hase, P., Biyik, E., Dragan, A., Krueger, D., Sadigh, D., and Hadfield-Menell, D. Open problems and fundamental limitations of reinforcement learning from human feedback. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=bx24KpJ4Eb. Survey Certification, Featured Certification.

Cen, S., Wei, Y., and Chi, Y. Fast Policy Extragradient Methods for Competitive Games with Entropy Regularization. *Journal of Machine Learning Research*, 25(4):1–48, 2024. URL http://jmlr.org/papers/v25/21-1205.html.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, Learning and Games*. Cambridge University Press, 2006.

Chen, K., Hu, I., and Ying, Z. Strong consistency of maximum quasi-likelihood estimators in generalized linear models with fixed and adaptive designs. *The Annals of Statistics*, 27(4):1155 – 1163, 1999. doi:10.1214/aos/1017938919.

Chow, Y., Nachum, O., and Ghavamzadeh, M. Path consistency learning in tsallis entropy regularized mdps. In *International conference on machine learning*, pp. 979–988. PMLR, 2018.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep Reinforcement Learning from Human Preferences. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4302–4310. Curran Associates, Inc., 2017. URL https://arxiv.org/abs/1706.03741.

Cui, G., Yuan, L., Ding, N., Yao, G., He, B., Zhu, W., Ni, Y., Xie, G., Xie, R., Lin, Y., Liu, Z., and Sun, M. ULTRAFEEDBACK: Boosting Language Models with Scaled AI Feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 9722–9744. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/cui24f.html.

Daskalakis, C., Deckelbaum, A., and Kim, A. Near-Optimal No-Regret Algorithms for Zero-Sum Games. In *Proceedings of the 2011 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pp. 235–254, 2011. doi:10.1137/1.9781611973082.21.

Daskalakis, C., Deckelbaum, A., and Kim, A. Near-optimal no-regret algorithms for zero-sum games. *Games and Economic Behavior*, 92:327–348, 2015. ISSN 0899-8256. doi:https://doi.org/10.1016/j.geb.2014.01.003.

Daskalakis, C., Ilyas, A., Syrgkanis, V., and Zeng, H. Training GANs with Optimism. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SJJySbbAZ.

Dong, H., Xiong, W., Pang, B., Wang, H., Zhao, H., Zhou, Y., Jiang, N., Sahoo, D., Xiong, C., and Zhang, T. Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863*, 2024.

Dudík, M., Hofmann, K., Schapire, R. E., Slivkins, A., and Zoghi, M. Contextual Dueling Bandits. In *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 563–587, Paris, France, 03–06 Jul 2015. PMLR. URL https://proceedings.mlr.press/v40/Dudik15.html.

Fan, J., Gong, W., and Zhu, Z. Generalized high-dimensional trace regression via nuclear norm regularization. *Journal of Econometrics*, 212(1):177–202, 2019. ISSN 0304-4076. doi:10.1016/j.jeconom.2019.04.026.

Faury, L., Abeille, M., Calauzènes, C., and Fercoq, O. Improved Optimistic Algorithms for Logistic Bandits. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3052–3060. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/faury20a.html.

Filippi, S., Cappe, O., Garivier, A., and Szepesvári, C. Parametric Bandits: The Generalized Linear Case. In *Advances in Neural Information Processing Systems*, volume 23, pp. 586–594. Curran Associates, Inc., 2010. URL https://sites.ualberta.ca/~szepesva/papers/GenLinBandits-NeurIPS2010.pdf.

Foster, D. and Rakhlin, A. Beyond UCB: Optimal and Efficient Contextual Bandits with Regression Oracles. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3199–3210. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/foster20a.html.

Foster, D., Rakhlin, A., Simchi-Levi, D., and Xu, Y. Instance-Dependent Complexity of Contextual Bandits and Reinforcement Learning: A Disagreement-Based Perspective. In *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 2059–2059. PMLR, 15–19 Aug 2021. URL https://arxiv.org/abs/2010.03104.

Freedman, D. A. On Tail Probabilities for Martingales. *The Annals of Probability*, 3(1):100 – 118, 1975. doi:10.1214/aop/1176996452.

Freund, Y. and Schapire, R. E. Adaptive Game Playing Using Multiplicative Weights. *Games and Economic Behavior*, 29(1):79–103, 1999. ISSN 0899-8256. doi:https://doi.org/10.1006/game.1999.0738.

Gajane, P., Urvoy, T., and Clérot, F. A Relative Exponential Weighing Algorithm for Adversarial Utility-based Dueling Bandits. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 218–227, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/gajane15.html.

Geist, M., Scherrer, B., and Pietquin, O. 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*, pp. 2160–2169. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/geist19a.html.

Go, D., Korbak, T., Kruszewski, G., Rozen, J., Ryu, N., and Dymetman, M. 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*, pp. 11546–11583. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/go23a.html.

Goldenshluger, A. and Zeevi, A. A linear response bandit problem. *Stochastic Systems*, 3(1):230 – 261, 2013. doi:10.1214/11-SSY032.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. Pmlr, 2018.

Han, J., Jiang, M., Song, Y., Ermon, S., and Xu, M. 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*, pp. 1144–1152. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/han25a.html.

Hao, B., Lattimore, T., and Wang, M. High-Dimensional Sparse Linear Bandits. In *Advances in Neural Information Processing Systems*, volume 33, pp. 10753–10763. Curran Associates, Inc., 2020. URL https://arxiv.org/abs/2011.04020.

Huang, A., Zhan, W., Xie, T., Lee, J. D., Sun, W., Krishnamurthy, A., and Foster, D. J. Correcting the mythos of KL-regularization: Direct alignment without overoptimization via chi-squared preference optimization. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=hXm0Wu2U9K.

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

Ito, S., Luo, H., Tsuchiya, T., and Wu, Y. Instance-Dependent Regret Bounds for Learning Two-Player Zero-Sum Games with Bandit Feedback. In *Proceedings of Thirty Eighth Conference on Learning Theory*, volume 291 of *Proceedings of Machine Learning Research*, pp. 2858–2892. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/ito25a.html.

Jang, K., Zhang, C., and Jun, K.-S. PopArt: Efficient Sparse Regression and Experimental Design for Optimal Sparse Linear Bandits. In *Advances in Neural Information Processing Systems*, volume 35, pp. 2102–2114. Curran Associates, Inc., 2022. URL https://openreview.net/forum?id=GWcdXz0M6a.

Jang, K., Zhang, C., and Jun, K.-S. Efficient Low-Rank Matrix Estimation, Experimental Design, and Arm-Set-Dependent Low-Rank Bandits. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21329–21372. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/jang24e.html.

Ji, K., Zhao, Q., Zhao, H., Di, Q., and Gu, Q. Near-Optimal Regret for KL-Regularized Multi-Armed Bandits. *arXiv preprint arXiv:2603.02155*, 2026. URL https://arxiv.org/abs/2603.02155.

Jiang, N. and Xie, T. Offline Reinforcement Learning in Large State Spaces: Algorithms and Guarantees. *Statistical Science*, 40(4):570 – 596, 2025. doi:10.1214/25-STS1000.

Jin, C., Liu, Q., and Yu, T. The Power of Exploiter: Provable Multi-Agent RL in Large State Spaces. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 10251–10279. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/jin22c.html.

Kang, Y., Hsieh, C.-J., and Lee, T. C. M. Efficient Frameworks for Generalized Low-Rank Matrix Bandit Problems. In *Advances in Neural Information Processing Systems*, volume 35, pp. 19971–19983. Curran Associates, Inc., 2022. URL https://arxiv.org/abs/2401.07298.

Kannan, S., Morgenstern, J. H., Roth, A., Waggoner, B., and Wu, Z. S. A Smoothed Analysis of the Greedy Algorithm for the Linear Contextual Bandit Problem. In *Advances in Neural Information Processing Systems*, volume 31, pp. 2231–2241. Curran Associates, Inc., 2018. URL https://arxiv.org/abs/1801.03423.

Kato, M. and Ito, S. 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*, pp. 3655–3663. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/kato25a.html.

Kim, G.-S. and Paik, M. C. Doubly-Robust Lasso Bandit. In *Advances in Neural Information Processing Systems*, volume 32, pp. 5877 – 5887. Curran Associates, Inc., 2019. URL https://arxiv.org/abs/1907.11362.

Kim, S.-J. and Oh, M.-h. Local Anti-Concentration Class: Logarithmic Regret for Greedy Linear Contextual Bandit. In *Advances in Neural Information Processing Systems*, volume 37, pp. 77525–77592. Curran Associates, Inc., 2024. doi:10.52202/079017-2465. URL https://openreview.net/forum?id=rblaF2euXQ.

Kuroki, Y., Rumi, A., Tsuchiya, T., Vitale, F., and Cesa-Bianchi, N. Best-of-Both-Worlds Algorithms for Linear Contextual Bandits. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1216–1224. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/kuroki24a.html.

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

Le, H., Tran, Q. H., Nguyen, D., Do, K., Mittal, S., Ogueji, K., and Venkatesh, S. 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.

Lee, J., Yun, S.-Y., and Jun, K.-S. A Unified Confidence Sequence for Generalized Linear Models, with Applications to Bandits. In *Advances in Neural Information Processing Systems*, volume 37, pp. 124640–124685. Curran Associates, Inc., 2024a. doi:10.52202/079017-3960. URL https://arxiv.org/abs/2407.13977.

Lee, J., Yun, S.-Y., and Jun, K.-S. Improved Regret Bounds of (Multinomial) Logistic Bandits via Regret-to-Confidence-Set Conversion. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4474–4482. PMLR, 02–04 May 2024b. URL https://proceedings.mlr.press/v238/lee24c.html.

Lee, J., Jang, K., Jun, K.-S., Vojnović, M., and Yun, S.-Y. GL-LowPopArt: A Nearly Instance-Wise Minimax-Optimal Estimator for Generalized Low-Rank Trace Regression. In *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research. PMLR, 02–05 May 2026. URL https://arxiv.org/abs/2506.03074.

Lee, K., Choi, S., and Oh, S. Sparse Markov Decision Processes With Causal Sparse Tsallis Entropy Regularization for Reinforcement Learning. *IEEE Robotics and Automation Letters*, 3(3):1466–1473, 2018. doi:10.1109/LRA.2018.2800085.

Li, G., Kamath, P., Foster, D. J., and Srebro, N. Understanding the Eluder Dimension. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23737–23750. Curran Associates, Inc., 2022a. URL https://openreview.net/forum?id=jHIn0U9U6RO.

Li, L., Lu, Y., and Zhou, D. Provably Optimal Algorithms for Generalized Linear Contextual Bandits. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2071–2080. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/li17c.html.

Li, W., Barik, A., and Honorio, J. A Simple Unified Framework for High Dimensional Bandit Problems. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12619–12655. PMLR, 17–23 Jul 2022b. URL https://proceedings.mlr.press/v162/li22a.html.

Li, Z., Ji, X., Chen, M., and Wang, M. Policy Evaluation for Reinforcement Learning from Human Feedback: A Sample Complexity Analysis. In *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2737–2745. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/li24l.html.

Liese, F. and Vajda, I. On Divergences and Informations in Statistics and Information Theory. *IEEE Transactions on Information Theory*, 52(10):4394–4412, 2006. doi:10.1109/TIT.2006.881731.

Liu, Q., Yu, T., Bai, Y., and Jin, C. A Sharp Analysis of Model-based Reinforcement Learning with Self-Play. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7001–7010. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/liu21z.html.

Llama Team. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Lu, Y., Meisami, A., and Tewari, A. Low-Rank Generalized Linear Bandit Problems. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 460–468. PMLR, 13–15 Apr

2021. URL https://proceedings.mlr.press/v130/lu21a.html.

May, K. O. Intransitivity, Utility, and the Aggregation of Preference Patterns. *Econometrica*, 22(1):1–13, 1954. doi:https://doi.org/10.2307/1909827.

McKelvey, R. D. and Palfrey, T. R. Quantal Response Equilibria for Normal Form Games. *Games and Economic Behavior*, 10(1):6–38, 1995. ISSN 0899-8256. doi:https://doi.org/10.1006/game.1995.1023.

Mertikopoulos, P. and Sandholm, W. H. Learning in Games via Reinforcement and Regularization. *Mathematics of Operations Research*, 41(4):1297–1324, 2016. doi:10.1287/moor.2016.0778.

Michaud, E. J., Gleave, A., and Russell, S. Understanding learned reward functions. *arXiv preprint arXiv:2012.05862*, 2020.

Minh, N. N., Baker, A., Neo, C., Roush, A. G., Kirsch, A., and Shwartz-Ziv, R. Turning up the heat: Min-p sampling for creative and coherent LLM outputs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=FBkpCyujtS.

Minka, T. P. Old and New Matrix Algebra Useful for Statistics, December 1997. URL https://tminka.github.io/papers/matrix/. MIT Media Lab note, 1997; revised 12/00.

Munos, R., Valko, M., Calandriello, D., Gheshlaghi Azar, M., Rowland, M., Guo, Z. D., Tang, Y., Geist, M., Mesnard, T., Fiegel, C., Michi, A., Selvi, M., Girgin, S., Momchev, N., Bachem, O., Mankowitz, D. J., Precup, D., and Piot, B. Nash Learning from Human Feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 36743–36768. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/munos24a.html.

Müller, A. Integral Probability Metrics and Their Generating Classes of Functions. *Advances in Applied Probability*, 29(2):429–443, 1997. doi:10.2307/1428011.

Nash, J. Non-Cooperative Games. *Annals of Mathematics*, 54(2):286–295, 1951. ISSN 0003486X, 19398980. doi:10.2307/1969529.

Nayak, A., Yang, T., Yagan, O., Joshi, G., and Chi, Y. Achieving Logarithmic Regret in KL-Regularized Zero-Sum Markov Games. *arXiv preprint arXiv:2510.13060*, 2025. URL https://arxiv.org/abs/2510.13060.

Nesterov, Y. *Lectures on Convex Optimization*. Springer Optimization and Its Applications. Springer Cham, 2 edition, 2018.

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

O'Donoghue, B., Lattimore, T., and Osband, I. Matrix Games with Bandit Feedback. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, volume 161 of *Proceedings of Machine Learning Research*, pp. 279–289. PMLR, 27–30 Jul 2021. URL https://proceedings.mlr.press/v161/o-donoghue21a.html.

Oh, M.-h., Iyengar, G., and Zeevi, A. Sparsity-Agnostic Lasso Bandit. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8271–8280. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/oh21a.html.

OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, 2024. URL https://arxiv.org/abs/2303.08774.

Osband, I. and Van Roy, B. Model-based Reinforcement Learning and the Eluder Dimension. In *Advances in Neural Information Processing Systems*, volume 27, pp. 1466–1474. Curran Associates, Inc., 2014. URL https://arxiv.org/abs/1406.1853.

Plackett, R. L. The Analysis of Permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 12 1975. ISSN 0035-9254. doi:10.2307/2346567.

Qwen Team. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Rakhlin, A. and Sridharan, K. Online Learning with Predictable Sequences. In *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pp. 993–1019, Princeton, NJ, USA, 12–14 Jun 2013a. PMLR. URL https://proceedings.mlr.press/v30/Rakhlin13.html.

Rakhlin, S. and Sridharan, K. Optimization, Learning, and Games with Predictable Sequences. In *Advances in Neural Information Processing Systems*, volume 26, pp. 3066–3074. Curran Associates, Inc., 2013b. URL https://arxiv.org/abs/1311.1869.

Rosset, C., Cheng, C.-A., Mitra, A., Santacroce, M., Awadallah, A., and Xie, T. Direct Nash Optimization: Teaching Language Models to Self-Improve with General Preferences. *arXiv preprint arXiv:2404.03715*, 2024. URL https://arxiv.org/abs/2404.03715.

Russac, Y., Faury, L., Cappé, O., and Garivier, A. Self-Concordant Analysis of Generalized Linear Bandits with Forgetting. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 658–666. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/russac21a.html.

Russo, D. and Van Roy, B. Eluder Dimension and the Sample Complexity of Optimistic Exploration. In *Advances in Neural Information Processing Systems*, volume 26, pp. 2256–2264. Curran Associates, Inc., 2013. URL https://web.stanford.edu/~bvr/pubs/Eluder.pdf.

Saha, A. Optimal Algorithms for Stochastic Contextual Preference Bandits. In *Advances in Neural Information Processing Systems*, volume 34, pp. 30050–30062. Curran Associates, Inc., 2021. URL https://openreview.net/forum?id=1lCZrXJBpM.

Saha, A. and Krishnamurthy, A. Efficient and Optimal Algorithms for Contextual Dueling Bandits under Realizability. In *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pp. 968–994. PMLR, 29 Mar–01 Apr 2022. URL https://proceedings.mlr.press/v167/saha22a.html.

Saha, A., Koren, T., and Mansour, Y. Adversarial Dueling Bandits. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 9235–9244. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/saha21a.html.

Saha, A., Pacchiano, A., and Lee, J. Dueling RL: Reinforcement Learning with Trajectory Preferences. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 6263–6289. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/saha23a.html.

Sion, M. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171 – 176, 1958. doi:10.2140/pjm.1958.8.171.

Swamy, G., Dann, C., Kidambi, R., Wu, S., and Agarwal, A. A Minimaximalist Approach to Reinforcement Learning from Human Feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 47345–47377. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/swamy24a.html.

Syrgkanis, V., Agarwal, A., Luo, H., and Schapire, R. E. Fast Convergence of Regularized Learning in Games. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2989–2997. Curran Associates, Inc., 2015. URL https://arxiv.org/abs/1507.00407.

Tang, X., Yoon, S., Son, S., Yuan, H., Gu, Q., and Bogunovic, I. RSPO: Regularized Self-Play Alignment of Large Language Models. *arXiv preprint arXiv:2503.00030*, 2025. URL https://arxiv.org/abs/2503.00030.

Tien, J., He, J. Z.-Y., Erickson, Z., Dragan, A., and Brown, D. S. Causal confusion and reward misidentification in preference-based reward learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=R0Xxvr_X3ZA.

Tropp, J. A. User-Friendly Tail Bounds for Sums of Random Matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. doi:10.1007/s10208-011-9099-z.

Troshin, S., Mohammed, W., Meng, Y., Monz, C., Fokkens, A., and Niculae, V. Control the temperature: Selective sampling for diverse and high-quality llm outputs. *arXiv preprint arXiv:2510.01218*, 2025.

Tsallis, C. Possible Generalization of Boltzmann-Gibbs Statistics. *Journal of Statistical Physics*, 52(1):479–487, 1988. doi:10.1007/BF01016429.

Tucker, M., Cheng, M., Novoseller, E., Cheng, R., Yue, Y., Burdick, J. W., and Ames, A. D. Human Preference-Based Learning for High-dimensional Optimization of Exoskeleton Walking Gaits. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3423–3430, 2020. doi:10.1109/IROS45743.2020.9341416.

Tversky, A. Intransitivity of preferences. *Psychological Review*, 76(1):31–48, 1969. doi:10.1037/h0026750.

von Neumann, J. Zur Theorie der Gesellschaftsspiele. *Mathematische Annalen*, 100(1):295–320, Dec 1928. ISSN 1432-1807. doi:10.1007/BF01448847.

Wainwright, M. J. and Jordan, M. I. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305, 2008. ISSN 1935-8237. doi:10.1561/2200000001.

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

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

Wu, W., Yang, J., and Shen, C. Stochastic Linear Contextual Bandits with Diverse Contexts. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2392–2401. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/wu20c.html.

Wu, Y., Jin, T., Di, Q., Lou, H., Farnoud, F., and Gu, Q. Borda Regret Minimization for Generalized Linear Dueling Bandits. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 53571–53596. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wu24m.html.

Xie, T., Foster, D. J., Bai, Y., Jiang, N., and Kakade, S. M. The Role of Coverage in Online Reinforcement Learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=LQIjzPdDt3q.

Xiong, W., Zhong, H., Shi, C., Shen, C., and Zhang, T. A Self-Play Posterior Sampling Algorithm for Zero-Sum Markov Games. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24496–24523. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/xiong22b.html.

Xiong, W., Dong, H., Ye, C., Wang, Z., Zhong, H., Ji, H., Jiang, N., and Zhang, T. 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*, pp. 54715–54754. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/xiong24a.html.

Yang, T., Dai, B., Xiao, L., and Chi, Y. Incentivize without Bonus: Provably Efficient Model-based Online Multi-agent RL for Markov Games. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=ciUHD7jcT9.

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

Ye, C., Xiong, W., Zhang, Y., Jiang, N., and Zhang, T. Online Iterative Reinforcement Learning from Human Feedback with General Preference Model. In *Advances in Neural Information Processing Systems*, volume 37, pp. 81773–81807. Curran Associates, Inc., 2024. URL https://openreview.net/forum?id=TwdX1W3M6S.

Zanette, A., Dong, K., Lee, J., and Brunskill, E. Design of Experiments for Stochastic Contextual Linear Bandits. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22720–22731. Curran Associates, Inc., 2021. URL https://openreview.net/forum?id=KsfuvGB3vco.

Zeng, G. and Honorio, J. A Novel General Framework for Sharp Lower Bounds in Succinct Stochastic Bandits. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=xvsPQuUHef.

Zhang, H., Li, B., Tian, W., and Sun, Q. Tail-Aware Information-Theoretic Generalization for RLHF and SGLD. *arXiv preprint arXiv:2604.10727*, 2026a. URL https://arxiv.org/abs/2604.10727.

Zhang, T. Feel-Good Thompson Sampling for Contextual Bandits and Reinforcement Learning. *SIAM Journal on Mathematics of Data Science*, 4(2):834–857, 2022. doi:10.1137/21M140924X.

Zhang, T. *Mathematical Analysis of Machine Learning Algorithms*. Cambridge University Press, 2023.

Zhang, Y., Yu, D., Ge, T., Song, L., Zeng, Z., Mi, H., Jiang, N., and Yu, D. Improving LLM General Preference Alignment via Optimistic Online Mirror Descent. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025a. URL https://openreview.net/forum?id=kZstGANG8D.

Zhang, Y., Zhang, G., Wu, Y., Xu, K., and Gu, Q. Beyond Bradley-Terry Models: A General Preference Model for Language Model Alignment. 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=UgKcpPE0ZX.

Zhang, Y., Chen, C., and Jiang, N. Beyond Pessimism: Offline Learning in KL-regularized Games. *arXiv preprint arXiv:2604.06738*, 2026b. URL https://arxiv.org/abs/2604.06738.

Zhang, Y.-J., Xu, S.-A., Zhao, P., and Sugiyama, M. Generalized Linear Bandits: Almost Optimal Regret with One-Pass Update. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025c. URL https://openreview.net/forum?id=H7vg3IgvHU.

Zhao, H., Ye, C., Gu, Q., and Zhang, T. 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.

Zhao, H., Ye, C., Xiong, W., Gu, Q., and Zhang, T. Logarithmic Regret for Online KL-Regularized Reinforcement Learning. In *Proceedings of the 41st International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 13–19 Jul 2025b. URL https://openreview.net/forum?id=6QH9IB53uy.

Zhao, Q., Ji, K., Zhao, H., Zhang, T., and Gu, Q. Towards a Sharp Analysis of Learning Offline f-Divergence-Regularized Contextual Bandits. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=ly6MB2Cfx2.

Zhu, B., Jordan, M., and Jiao, J. Principled Reinforcement Learning with Human Feedback from Pairwise or K-wise Comparisons. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 43037–43067. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/zhu23f.html.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., Dey, A. K., et al. Maximum entropy inverse reinforcement learning. In *Aaai*, volume 8, pp. 1433–1438. Chicago, IL, USA, 2008.

Zimmert, J. and Seldin, Y. 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

A. Pseudocodes for `Greedy Sampling` and `Explore-Then-Commit`

Algorithm 1: `Greedy Sampling`

Input: Exploration policy ρ;
Initialize $\hat{\pi}_1 \leftarrow \rho$;
for $t = 1, 2, \cdots, T$ **do**
  Observe $\boldsymbol{x}_t \sim d_0$;
  Sample $\boldsymbol{a}_t^1 \sim \hat{\pi}_t(\cdot|\boldsymbol{x}_t)$ and $\boldsymbol{a}_t^2 \sim \rho(\cdot|\boldsymbol{x}_t)$;
  Observe $r_t := \mathbb{1}[\boldsymbol{a}_t^1 \succ \boldsymbol{a}_t^2] \sim \mathrm{Ber}(\mu(\phi_t^{1\top} \boldsymbol{\Theta}_\star \phi_t^2) \mid \boldsymbol{x}_t)$, where $\phi_t^i := \phi(\boldsymbol{x}_t, \boldsymbol{a}_t^i)$;
  Compute an estimator $\widehat{\boldsymbol{\Theta}}_{t+1} \in \mathrm{Skew}(d)$;
  Compute a (symmetric) Nash equilibrium: $\hat{\pi}_{t+1} \leftarrow \arg\max_{\pi^1 \in \Pi} \min_{\pi^2 \in \Pi} J_\eta(\pi^1, \pi^2; \widehat{\boldsymbol{\Theta}}_{t+1})$.;

Algorithm 2: `Explore-Then-Commit`

Input: Exploration policy ρ and budget T_0;
for $t = 1, 2, \cdots, T_0$ **do**
  Observe $\boldsymbol{x}_t \sim d_0$;
  Sample $\boldsymbol{a}_t^1 \sim \rho(\cdot|\boldsymbol{x}_t)$ and $\boldsymbol{a}_t^2 \sim \rho(\cdot|\boldsymbol{x}_t)$;
  Observe $r_t := \mathbb{1}[\boldsymbol{a}_t^1 \succ \boldsymbol{a}_t^2] \sim \mathrm{Ber}(\mu(\phi_t^{1\top} \boldsymbol{\Theta}_\star \phi_t^2) \mid \boldsymbol{x}_t)$, where $\phi_t^i := \phi(\boldsymbol{x}_t, \boldsymbol{a}_t^i)$;
Compute an estimator $\widehat{\boldsymbol{\Theta}} \in \mathrm{Skew}(d)$;
Compute a (symmetric) Nash equilibrium: $\hat{\pi} \leftarrow \arg\max_{\pi^1 \in \Pi} \min_{\pi^2 \in \Pi} J_\eta(\pi^1, \pi^2; \widehat{\boldsymbol{\Theta}})$.;
for $t = T_0 + 1, \cdots, T$ **do**
  Symmetrically commit to $(\hat{\pi}, \hat{\pi})$;

B. Proof of Lemma 3.4

For each fixed context $\boldsymbol{x} \in \mathcal{X}$, $J(\hat{\pi}, \pi|\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{a}\sim\hat{\pi}(\cdot|\boldsymbol{x}),\boldsymbol{b}\sim\pi(\cdot|\boldsymbol{x})}[\mu(\boldsymbol{a}^\top\Theta\boldsymbol{b})]$ is linear in $\pi(\cdot|\boldsymbol{x})$. Because the sum of a linear function and a strongly convex function remains strongly convex, the regularized objective $J_\eta(\hat{\pi}, \pi|x)$ is $(\eta\beta)^{-1}$-strongly convex in ℓ_1. By the zeroth-order characterization of strong convexity (Nesterov, 2018, Theorem 2.1.9), for any $\pi, \tilde{\pi} \in \Pi$ and $\alpha \in [0, 1]$, the following holds point-wise for any $\boldsymbol{x}$:

$$J_\eta(\hat{\pi}, \alpha\pi + (1-\alpha)\tilde{\pi}|\boldsymbol{x}) \le \alpha J_\eta(\hat{\pi}, \pi|\boldsymbol{x}) + (1-\alpha)J_\eta(\hat{\pi}, \tilde{\pi}|\boldsymbol{x}) - \frac{\alpha(1-\alpha)}{2\eta\beta}\|\pi(\cdot|\boldsymbol{x}) - \tilde{\pi}(\cdot|\boldsymbol{x})\|_1^2 \tag{34}$$

Taking the expectation over $\boldsymbol{x} \sim d_0$ on both sides, choosing $\pi = \hat{\pi}$, and denoting the global objective as $F(\pi) := \mathbb{E}_{\boldsymbol{x}\sim d_0}[J_\eta(\hat{\pi}, \pi|\boldsymbol{x})]$ for simplicity, we obtain:

$$F(\alpha\pi + (1-\alpha)\tilde{\pi}) \le \alpha F(\pi) + (1-\alpha)F(\tilde{\pi}) - \frac{\alpha(1-\alpha)}{2\eta\beta}\mathbb{E}_{\boldsymbol{x}\sim d_0}\left[\|\pi(\cdot|\boldsymbol{x}) - \tilde{\pi}(\cdot|\boldsymbol{x})\|_1^2\right], \tag{35}$$

Choosing $\tilde{\pi} = \arg\min_{\pi\in\Pi} F(\pi)$, we have that $F(\tilde{\pi}) \le F(\alpha\hat{\pi} + (1-\alpha)\tilde{\pi})$, for *any* $\alpha \in [0, 1]$. Rearranging the inequality and dividing both sides by α, the following holds for any $\alpha \in (0, 1]$:

$$\underbrace{\mathbb{E}_{\boldsymbol{x}\sim d_0}[J_\eta(\hat{\pi}, \hat{\pi} \mid \boldsymbol{x})]}_{=\frac{1}{2}} - E_{\boldsymbol{x}\sim d_0}[J_\eta(\hat{\pi}, \tilde{\pi} \mid \boldsymbol{x})] = F(\hat{\pi}) - F(\tilde{\pi}) \ge \frac{1-\alpha}{2\eta\beta}\mathbb{E}_{\boldsymbol{x}\sim d_0}\left[\|\pi(\cdot|\boldsymbol{x}) - \tilde{\pi}(\cdot|\boldsymbol{x})\|_1^2\right]. \tag{36}$$

We then conclude by taking the limit $\alpha \to 0^+$. □

C. Statistical Cost of Obtaining ρ

Let us define the population E-optimal design:

$$\rho^\star \leftarrow \underset{\rho:\mathcal{X}\to\Delta(\mathcal{A})}{\arg\max} \left\{ L(\rho) \triangleq \lambda_{\min}\left(\mathbb{E}_{\boldsymbol{x}\sim d_0}\mathbb{E}_{\boldsymbol{a}\sim\rho(\cdot|\boldsymbol{x})}\left[\boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{a})\boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{a})^\top\right]\right)\right\}, \tag{37}$$

and let us denote $C_{\min} := L(\rho^\star)$. Recall that $\|\boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{a})\|_2 \le 1$, always.

Offline Scenario. Suppose that we have a $\{\boldsymbol{x}_i\}_{i\in[N]}$ with $\boldsymbol{x}_i \sim d_0$ i.i.d., and suppose that we have access to an exploratory policy class $\mathrm{P} \subset \{\rho : \mathcal{X} \to \Delta(\mathcal{A})\}$ that satisfies realizability, i.e., $\rho^\star \in \mathrm{P}$. For simplicity, suppose that $|\mathrm{P}| < \infty$, as if not, then one should be able to extend the arguments using standard covering and uniform convergence arguments, provided that P has a finite complexity measure.

We define the empirical E-optimal design:

$$\hat{\rho}_N \leftarrow \underset{\rho:\mathcal{X}\to\Delta(\mathcal{A})}{\arg\max} \left\{ \hat{L}_N(\rho) \triangleq \lambda_{\min}\left(\frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\boldsymbol{a}\sim\rho(\cdot|\boldsymbol{x}_i)}\left[\boldsymbol{\phi}(\boldsymbol{x}_i,\boldsymbol{a})\boldsymbol{\phi}(\boldsymbol{x}_i,\boldsymbol{a})^\top\right]\right)\right\} \tag{38}$$

Then we have the following guarantee:

Proposition C.1. $\mathbb{P}\left(L(\hat{\rho}_N) \ge C_{\min}/2\right) \ge 1-\delta$, *provided that* $N \ge 32 C_{\min}^{-2} \log \frac{2d|\Pi|}{\delta}$.

Proof. Note that each matrix $\boldsymbol{Z}_i(\rho) \triangleq \mathbb{E}_{\boldsymbol{x}\sim d_0}\mathbb{E}_{\boldsymbol{a}\sim\rho(\cdot|\boldsymbol{x})}\left[\boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{a})\boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{a})^\top\right] - \mathbb{E}_{\boldsymbol{a}\sim\rho(\cdot|\boldsymbol{x}_i)}\left[\boldsymbol{\phi}(\boldsymbol{x}_i,\boldsymbol{a})\boldsymbol{\phi}(\boldsymbol{x}_i,\boldsymbol{a})^\top\right]$ is i.i.d. that satisfies $\mathbb{E}[\boldsymbol{Z}_i(\rho)] = 0$ and $\boldsymbol{Z}_i(\rho)^2 \preceq \boldsymbol{I}_d$.[4] Thus, invoking matrix Hoeffding (Tropp, 2012, Theorem 1.3 & Remark 7.4), we have that *for each* $\rho \in \mathrm{P}$, with probability at least $2d\exp\left(-\frac{N\varepsilon^2}{2}\right)$,

$$\left\| \mathbb{E}_{\boldsymbol{x}\sim d_0}\mathbb{E}_{\boldsymbol{a}\sim\rho(\cdot|\boldsymbol{x})}\left[\boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{a})\boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{a})^\top\right] - \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\boldsymbol{a}\sim\rho(\cdot|\boldsymbol{x}_i)}\left[\boldsymbol{\phi}(\boldsymbol{x}_i,\boldsymbol{a})\boldsymbol{\phi}(\boldsymbol{x}_i,\boldsymbol{a})^\top\right] \right\|_{\mathrm{op}} \ge \varepsilon. \tag{39}$$

Union bound over $\rho \in \mathrm{P}$ and rearranging, we have that with probability at least $1-\delta$,

$$\mathcal{E}_N \triangleq \sup_{\rho\in\mathrm{P}} \left\| \frac{1}{N}\sum_{i=1}^N \boldsymbol{Z}_i(\rho) \right\|_{\mathrm{op}} \le \sqrt{\frac{2}{N}\log\frac{2d|\mathrm{P}|}{\delta}}. \tag{40}$$

Then, we have that with probability at least $1-\delta$,

$$L(\hat{\rho}_N) = L(\rho^\star) + \underbrace{\hat{L}_N(\rho^\star) - L(\rho^\star)}_{\ge -\mathcal{E}_N} + \underbrace{L(\hat{\rho}_N) - \hat{L}_N(\hat{\rho}_N)}_{\ge -\mathcal{E}_N} + \underbrace{\hat{L}_N(\hat{\rho}_N) - \hat{L}_N(\rho^\star)}_{\ge 0} \tag{41}$$

$$\ge C_{\min} - 2\mathcal{E}_N \qquad \text{(Weyl's inequality, } \hat{\rho}_N = \arg\max_\rho \hat{L}_N(\rho)\text{)}$$

$$\ge C_{\min} - \sqrt{\frac{8}{N}\log\frac{2d|\mathrm{P}|}{\delta}} \ge \frac{C_{\min}}{2}, \tag{42}$$

and thus, the statement follows. □

Online Scenario. This is largely inspired by the "design static or nonadaptive policies that can be used to gather data to identify optimal contextualized decision policies" as in Zanette et al. (2021).

Suppose that we interact with the environment for T_0 iterations via the following elliptical exploration algorithm:

1. **Initialize:** $\boldsymbol{V}_1 = \lambda \boldsymbol{I}_d$ for some regularization parameter $\lambda > 0$.

[4] This is because for $\boldsymbol{0} \preceq \boldsymbol{A}, \boldsymbol{B} \preceq \boldsymbol{I}_d$, $-\boldsymbol{I}_d \preceq \boldsymbol{A} - \boldsymbol{B} \preceq \boldsymbol{I}_d$, which then implies that $(\boldsymbol{A}-\boldsymbol{B})^2 \preceq \boldsymbol{I}_d$.

2. **For** $t = 1, \dots, T_0$:
   (a) Observe context $\boldsymbol{x}_t \sim d_0$.
   (b) Define the exploratory policy $\rho_t(\boldsymbol{x}_t)$ to *deterministically* select the action maximizing the elliptical bonus:

$$\rho_t(\boldsymbol{x}) := \arg\max_{\boldsymbol{a}\in\mathcal{A}} \|\boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{a})\|^2_{\boldsymbol{V}_{t-1}^{-1}}. \tag{43}$$

   (c) Choose action $\boldsymbol{a}_t = \rho_t(\boldsymbol{x}_t)$ and update the covariance matrix: $\boldsymbol{V}_{t+1} = \boldsymbol{V}_t + \boldsymbol{\phi}(\boldsymbol{x}_t,\boldsymbol{a}_t)\boldsymbol{\phi}(\boldsymbol{x}_t,\boldsymbol{a}_t)^\top$.
3. **Return:** The uniform mixture policy $\bar{\rho}_{T_0} \triangleq \frac{1}{T_0}\sum_{t=1}^{T_0}\rho_t$.

Then we have the following guarantee:

Proposition C.2. *Set* $T_0 \asymp \frac{d^2}{C_{\min}^2}\log\frac{d}{C_{\min}^2}\log\frac{d}{\delta}$ *and* $\lambda = 1$. *Then, the above elliptical exploration algorithm guarantees the following:*

$$\mathbb{P}\left(L(\bar{\rho}_{T_0}) \gtrsim \frac{C_{\min}}{d\log\frac{d}{C_{\min}^2}\log\frac{1}{\delta}}\right) \geq 1-\delta. \tag{44}$$

Proof. The proof proceeds similarly to our Freedman-based bounding of expected elliptical potentials.

Let $\Sigma(\rho) \triangleq \mathbb{E}_{\boldsymbol{x}\sim d_0}\mathbb{E}_{\boldsymbol{a}\sim\rho(\cdot|\boldsymbol{x})}[\boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{a})\boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{a})^\top]$, and M_t be the conditional expectation of the elliptical bonus, conditioned on the history $\mathcal{F}_{t-1} = \sigma(\boldsymbol{x}_1,\boldsymbol{a}_1,\cdots,\boldsymbol{x}_{t-1},\boldsymbol{a}_{t-1})$:

$$M_t \triangleq \mathbb{E}_{\boldsymbol{x}\sim d_0}\left[\max_{\boldsymbol{a}\in\mathcal{A}}\|\boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{a})\|^2_{\boldsymbol{V}_t^{-1}} \;\middle|\; \mathcal{F}_{t-1}\right] \leq 1. \tag{45}$$

Because $\max \geq \mathbb{E}$, it upper bounds the expectation over the optimal $\rho^\star$:

$$M_t \geq \mathbb{E}_{\boldsymbol{x}\sim d_0}\mathbb{E}_{\boldsymbol{a}\sim\rho^\star(\cdot|\boldsymbol{x})}\left[\|\boldsymbol{\phi}(\boldsymbol{x},\boldsymbol{a})\|^2_{\boldsymbol{V}_t^{-1}} \;\middle|\; \mathcal{F}_{t-1}\right] \overset{(*)}{=} \operatorname{tr}\left(\boldsymbol{V}_t^{-1}\Sigma(\rho^\star)\right) \geq C_{\min}\operatorname{tr}(\boldsymbol{V}_t^{-1}), \tag{46}$$

where $(*)$ follows from linearity of the expectation.

Because $\lambda = 1$ and $\|\boldsymbol{\phi}\|_2 \leq 1$, we have $y_t \triangleq \|\boldsymbol{\phi}(\boldsymbol{x}_t,\boldsymbol{a}_t)\|^2_{\boldsymbol{V}_t^{-1}} \in [0,1]$. Thus, by the elliptical potential lemma (Lemma J.3), we have that

$$\sum_{t=1}^{T_0} y_t \leq 2d\log\left(1+\frac{T_0}{d}\right), \tag{47}$$

where the last inequality follows from our choice of $\lambda = 1$.

By the same variant of the Freedman's inequality (Lemma D.3), for any $\xi \in (0,1]$, the following holds with probability at least $1-\frac{\delta}{2}$:

$$\sum_{t=1}^{T_0}(M_t - y_t) \leq (e-2)\xi\sum_{t=1}^{T_0}\underbrace{\mathbb{E}[(M_t-y_t)^2 \mid \mathcal{F}_{t-1}]}_{(*)} + \frac{1}{\xi}\log\frac{2}{\delta}. \tag{48}$$

We bound $(*)$ as follows: denoting $\mathbb{E}_{t-1} \triangleq \mathbb{E}[\cdot \mid \mathcal{F}_{t-1}]$,

$$\mathbb{E}_{t-1}[(M_t-y_t)^2] = \mathbb{E}_{t-1}[y_t^2] - M_t^2 \leq \mathbb{E}_{t-1}[y_t] - M_t^2 \leq M_t. \tag{49}$$

Choosing $\xi = \frac{1}{2(e-2)}$, we have the following linear self-bounding inequality:

$$\sum_{t=1}^{T_0} M_t = \sum_{t=1}^{T_0} y_t + \sum_{t=1}^{T_0}(M_t - y_t) \leq 2d\log\left(1+\frac{T_0}{d}\right) + \frac{1}{2}\sum_{t=1}^{T_0} M_t + 2(e-2)\log\frac{2}{\delta}. \tag{50}$$

Combining this with the lower bound of M_t, we have that

$$C_{\min} \sum_{t=1}^{T_0} \operatorname{tr}(\boldsymbol{V}_t^{-1}) \leq \sum_{t=1}^{T_0} M_t \leq 4d \log\left(1 + \frac{T_0}{d}\right) + 4(e-2)\log\frac{2}{\delta}. \tag{51}$$

As $\operatorname{tr}(\boldsymbol{V}_t^{-1}) \geq \frac{1}{\lambda_{\min}(\boldsymbol{V}_t)} \geq \frac{1}{\lambda_{\min}(\boldsymbol{V}_{T_0})}$, we have that

$$\frac{\lambda_{\min}(\boldsymbol{V}_{T_0})}{T_0} \geq \frac{C_{\min}}{4d\log\left(1+\frac{T_0}{d}\right) + 4(e-2)\log\frac{2}{\delta}}. \tag{52}$$

Let $\boldsymbol{Z}_t \triangleq \Sigma(\rho_t) - \boldsymbol{\phi}(\boldsymbol{x}_t, \boldsymbol{a}_t)\boldsymbol{\phi}(\boldsymbol{x}_t, \boldsymbol{a}_t)^\top$. Because ρ_t is fully determined (measurable) by $\mathcal{F}_{t-1}$, we have that $\mathbb{E}[\boldsymbol{\phi}(\boldsymbol{x}_t, \boldsymbol{a}_t)\boldsymbol{\phi}(\boldsymbol{x}_t, \boldsymbol{a}_t)^\top \mid \mathcal{F}_{t-1}] = \Sigma(\rho_t)$, i.e., $\{\boldsymbol{Z}_t\}_{t=1}^{T_0}$ is a matrix martingale difference sequence adapted to $\mathcal{F}_t$. Furthermore, because $\boldsymbol{0} \preceq \Sigma(\rho_t) \preceq \boldsymbol{I}$ and $\boldsymbol{0} \preceq \boldsymbol{\phi}\boldsymbol{\phi}^\top \preceq \boldsymbol{I}$, the differences are bounded: $\lambda_{\max}(\boldsymbol{Z}_t) \leq 1$ and $\lambda_{\min}(\boldsymbol{Z}_t) \geq -1$.

By the matrix Azuma inequality (Tropp, 2012, Theorem 7.1, Remark 7.8), with probability at least $1 - \frac{\delta}{2}$:

$$\lambda_{\min}\left(\sum_{t=1}^{T_0} \boldsymbol{Z}_t\right) \geq -\sqrt{2T_0 \log\frac{2d}{\delta}}. \tag{53}$$

Notice that the population covariance of the uniform mixture is $\Sigma(\bar{\rho}_{T_0}) = \frac{1}{T_0}\sum_{t=1}^{T_0}\Sigma(\rho_t)$. We can rewrite this sum as:

$$\frac{1}{T_0}\sum_{t=1}^{T_0}\Sigma(\rho_t) = \frac{1}{T_0}\left(\boldsymbol{V}_{T_0} - \lambda\boldsymbol{I} + \sum_{t=1}^{T_0}\boldsymbol{Z}_t\right). \tag{54}$$

Taking the minimum eigenvalue on both sides, we lower bound $L(\bar{\rho}_{T_0})$ as follows:

$$\begin{aligned}
L(\bar{\rho}_{T_0}) &= \frac{1}{T_0}\lambda_{\min}\left(\boldsymbol{V}_{T_0} - \boldsymbol{I} + \sum_{t=1}^{T_0}\boldsymbol{Z}_t\right) && (55)\\
&= \frac{1}{T_0}\lambda_{\min}\left(\boldsymbol{V}_{T_0} + \sum_{t=1}^{T_0}\boldsymbol{Z}_t\right) - \frac{\lambda}{T_0} && (56)\\
&\geq \frac{1}{T_0}\lambda_{\min}(\boldsymbol{V}_{T_0}) + \frac{1}{T_0}\lambda_{\min}\left(\sum_{t=1}^{T_0}\boldsymbol{Z}_t\right) - \frac{1}{T_0} && (\lambda_{\min}(\cdot) \text{ is concave})\\
&\geq \frac{C_{\min}}{4d\log\left(1+\frac{T_0}{d}\right) + 4(e-2)\log\frac{2}{\delta}} - \sqrt{\frac{2}{T_0}\log\frac{2d}{\delta}} - \frac{1}{T_0} && (57)\\
&\geq \frac{C_{\min}}{4d\log\left(1+\frac{T_0}{d}\right) + 4(e-2)\log\frac{2}{\delta}} - \sqrt{\frac{8}{T_0}\log\frac{2d}{\delta}}, && (58)
\end{aligned}$$

where the last inequality follows given that $T_0 \geq \left(8\log\frac{2d}{\delta}\right)^{-1}$.

We lastly solve for a condition on T_0 such that the first term dominates, i.e.,

$$\frac{C_{\min}}{d\log\left(1+\frac{T_0}{d}\right) + \log\frac{1}{\delta}} \gtrsim \sqrt{\frac{1}{T_0}\log\frac{d}{\delta}} \implies \sqrt{T_0} \gtrsim \frac{d\log\left(1+\frac{T_0}{d}\right) + \log\frac{1}{\delta}}{C_{\min}}\sqrt{\log\frac{d}{\delta}}. \tag{59}$$

Solving this condition yields the required exploration budget:

$$T_0 \asymp \frac{d^2}{C_{\min}^2}\log\frac{d}{C_{\min}^2}\log\frac{d}{\delta}. \tag{60}$$

Plugging this T_0 back into our lower bound, we conclude that with probability at least $1-\delta$ (via a union bound over the Freedman and Azuma events):

$$L(\bar{\rho}_{T_0}) \gtrsim \frac{C_{\min}}{d\log\frac{d}{C_{\min}^2}\log\frac{1}{\delta}}, \tag{61}$$

which completes the proof. □

D. Proof of Theorem 4.2: Regret Bound of `Greedy Sampling`

D.1. Main Proof

We will use several properties of the Kronecker product $\otimes$ throughout the proof; see Minka (1997) for a reference. Let us denote $\mathrm{vec} : \mathbb{R}^{d\times d} \to \mathbb{R}^{d^2}$ as the (column-wise) vectorization operator, and $\mathrm{mat} \triangleq \mathrm{vec}^{-1} : \mathbb{R}^{d^2} \to \mathbb{R}^{d\times d}$ to be the matrization operator.

Part I. $\tilde{\mathcal{O}}(\eta\beta(\log T)^2)$ Regret Bound. To mirror the proof sketch in the main text, we divide the proof into three parts.

1. From Regret to Sum of Squared Errors. With $\boldsymbol{v}_t := \mathrm{vec}(\boldsymbol{\phi}_t^2(\boldsymbol{\phi}_t^1)^\top) \in \mathbb{R}^{d^2}$, our MLE is defined as follows:

$$\widehat{\boldsymbol{\Theta}}_t \leftarrow \mathrm{mat}(\hat{\boldsymbol{\theta}}_t), \quad \hat{\boldsymbol{\theta}}_t \leftarrow \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\mathcal{K}_S} \mathcal{L}_t(\boldsymbol{\theta}), \quad \mathcal{L}_t(\boldsymbol{\theta}) := \sum_{s=1}^{t-1} \left\{ m(\langle \boldsymbol{\theta}, \boldsymbol{v}_t\rangle) - r_t\langle \boldsymbol{\theta}, \boldsymbol{v}_t\rangle \right\}, \tag{62}$$

where $m(\cdot)$ is the log-partition function (Wainwright & Jordan, 2008) such that $m' = \mu$, and

$$\mathcal{K}_S := \left\{ \boldsymbol{\theta} \in \mathbb{R}^{d^2} : \|\boldsymbol{\theta}\|_2 \le S \text{ and } \mathrm{mat}(\boldsymbol{\theta})^\top = -\mathrm{mat}(\boldsymbol{\theta}) \right\}. \tag{63}$$

As $\mu(z) = (1+e^{-z})^{-1}$ for our proof, this implies that $R_s = 1$ and $L_\mu = \frac{1}{4}$.

Let us denote the regularized Hessian of $\mathcal{L}_t(\cdot)$ at the MLE $\hat{\boldsymbol{\theta}}_t$ as

$$\widehat{\boldsymbol{H}}_t := \boldsymbol{I}_{d^2} + \sum_{s=1}^{t-1} \dot{\mu}(\langle \hat{\boldsymbol{\theta}}_s, \boldsymbol{v}_s\rangle)\boldsymbol{v}_s\boldsymbol{v}_s^\top. \tag{64}$$

We first recall the following elliptical confidence sequence:

Lemma D.1 (Theorem 3.2 of Lee et al. (2024a)). *For any adaptively collected*[a] $\{(\boldsymbol{v}_t, r_t)\}$ *and any* $\delta \in (0,1)$ *we have*

$$\mathbb{P}\left(\left\| \boldsymbol{\theta}_\star - \hat{\boldsymbol{\theta}}_t \right\|_{\widehat{\boldsymbol{H}}_t}^2 \lesssim \gamma_t(\delta), \ \ \forall t \ge 1 \right) \ge 1-\delta, \tag{65}$$

where $\gamma_t(\delta) \lesssim S^5 + S\log\frac{1}{\delta} + Sd^2\log\frac{St}{d}$.

[a] This can be formalized via the canonical bandit model as described in Lattimore & Szepesvári (2020, Chapter 4.6).

We note that this is used solely for the proof and does not affect the algorithm in any way.

From the η-dependent bound of Theorem 3.1, we have that

$$\mathrm{MBR\text{-}Reg}(T) \le (4L_\mu^2\eta\beta + L_\mu) \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\phi}_t\sim\hat{\pi}_t} \left[\|\boldsymbol{E}_t\boldsymbol{\phi}_t\|_2^2 \right]. \tag{66}$$

We decompose the squared Euclidean norm as follows: denoting the standard basis of $\mathbb{R}^d$ as $\{\boldsymbol{e}_j\}_{j\in[d]}$,

$$\begin{aligned}
\|\boldsymbol{E}_t\boldsymbol{\phi}_t\|_2^2 &= \sum_{j=1}^{d} \left(\boldsymbol{e}_j^\top \boldsymbol{E}_t\boldsymbol{\phi}_t\right)^2 && (67)\\
&= \sum_{j=1}^{d} \langle \mathrm{vec}(\boldsymbol{E}_t), \mathrm{vec}(\boldsymbol{e}_j\boldsymbol{\phi}_t^\top)\rangle^2 && (68)\\
&\le \sum_{j=1}^{d} \|\mathrm{vec}(\boldsymbol{E}_t)\|_{\widehat{\boldsymbol{H}}_t}^2 \left\|\mathrm{vec}(\boldsymbol{e}_j\boldsymbol{\phi}_t^\top)\right\|_{\widehat{\boldsymbol{H}}_t^{-1}}^2 && \text{(Cauchy-Schwarz)}
\end{aligned}$$

$$\lesssim \gamma_t(\delta)^2 \sum_{j=1}^{d} \|\boldsymbol{\phi}_t \otimes \boldsymbol{e}_j\|^2_{\widehat{\boldsymbol{H}}_t^{-1}}. \qquad \text{(Lemma D.1)}$$

2. Towards Expected Elliptical Potentials.

We now present our key technical lemma:

Lemma D.2 (Coverage Lemma). *Let $\tilde{\boldsymbol{\phi}} \sim \rho$ be such that $\mathbb{E}_{\tilde{\boldsymbol{\phi}}\sim\rho}\left[\tilde{\boldsymbol{\phi}}\tilde{\boldsymbol{\phi}}^\top\right] \succeq C_{\min}\boldsymbol{I}_d$. Then, for any positive semi-definite $\boldsymbol{M} \in \mathbb{R}^{d^2\times d^2}$ and any vector $\boldsymbol{\phi} \in \mathbb{R}^d$, the following holds:*

$$\sum_{j=1}^{d} (\boldsymbol{\phi}\otimes\boldsymbol{e}_j)^\top \boldsymbol{M}(\boldsymbol{\phi}\otimes\boldsymbol{e}_j) \le C_{\min}^{-1}\mathbb{E}_{\tilde{\boldsymbol{\phi}}\sim\rho}\left[\mathrm{vec}(\boldsymbol{\phi}\tilde{\boldsymbol{\phi}}^\top)^\top \boldsymbol{M}\mathrm{vec}(\boldsymbol{\phi}\tilde{\boldsymbol{\phi}}^\top)\right]. \tag{69}$$

From hereon, we denote $\mathbb{E}_{t-1}[\cdot] \triangleq \mathbb{E}_{\boldsymbol{\phi}_t\sim\hat{\pi}_t,\tilde{\boldsymbol{\phi}}_t\sim\rho}[\cdot]$, where $\mathbb{E}$ is to indicate that the expectation is conditional on the history, due to $\hat{\pi}_t$ being history-dependent.

Applying the above lemma with $\boldsymbol{M} = \widehat{\boldsymbol{H}}_t^{-1}$, we have:

$$\begin{aligned}\text{MBR-Reg}(T) &\lesssim \eta\beta d^2 C_{\min}^{-1}\log\frac{T}{d}\sum_{t=1}^{T}\mathbb{E}_{t-1}\left[\mathrm{vec}(\boldsymbol{\phi}_t\tilde{\boldsymbol{\phi}}_t^\top)^\top\widehat{\boldsymbol{H}}_t^{-1}\mathrm{vec}(\boldsymbol{\phi}_t\tilde{\boldsymbol{\phi}}_t^\top)\right]\\ &\le \eta\beta\kappa^{-1}d^2C_{\min}^{-1}\log\frac{T}{d}\sum_{t=1}^{T}\mathbb{E}_{t-1}\left[\mathrm{vec}(\boldsymbol{\phi}_t\tilde{\boldsymbol{\phi}}_t^\top)^\top\boldsymbol{V}_t^{-1}\mathrm{vec}(\boldsymbol{\phi}_t\tilde{\boldsymbol{\phi}}_t^\top)\right], \end{aligned}\tag{70}$$

where we have bounded $\gamma_t(\delta)^2 \le \gamma_T(\delta)^2 \lesssim d^2\log\frac{T}{d}$, and we denote $\boldsymbol{v}_t = \boldsymbol{\phi}_t\tilde{\boldsymbol{\phi}}_t^\top$ and $\boldsymbol{V}_t := \frac{1}{\kappa\wedge 1}\boldsymbol{I} + \sum_{s=1}^{t-1}\boldsymbol{v}_s\boldsymbol{v}_s^\top$.

3. Martingale Concentration for Realized Variance.

We now consider the elliptical-type quantity:

$$\begin{aligned}\underbrace{\sum_{t=1}^{T}\mathbb{E}_{t-1}\left[\boldsymbol{v}_t^\top\boldsymbol{V}_t^{-1}\boldsymbol{v}_t\right]}_{\triangleq S_T} &= \sum_{t=1}^{T}\boldsymbol{v}_t^\top\boldsymbol{V}_t^{-1}\boldsymbol{v}_t + \sum_{t=1}^{T}\underbrace{\left(\mathbb{E}_{t-1}\left[\boldsymbol{v}_t^\top\boldsymbol{V}_t^{-1}\boldsymbol{v}_t\right] - \boldsymbol{v}_t^\top\boldsymbol{V}_t^{-1}\boldsymbol{v}_t\right)}_{\triangleq M_t}\\ &= \underbrace{\sum_{t=1}^{T}\boldsymbol{v}_t^\top\boldsymbol{V}_t^{-1}\boldsymbol{v}_t}_{(a)} + \underbrace{\sum_{t=1}^{T}M_t}_{(b)}. \end{aligned}\tag{71}$$

We bound (a) via the usual elliptical potential lemma (Lemma J.3):

$$\sum_{t=1}^{T}\boldsymbol{v}_t^\top\boldsymbol{V}_t^{-1}\boldsymbol{v}_t \le 2d^2\log\left(1+\frac{\kappa T}{d^2}\right). \tag{72}$$

We now bound (b). The crucial observation is that M_t is a martingale difference sequence that satisfies $\max_{t\ge1}|M_t| \le \kappa$, which prompts us to use the following variant of Freedman's inequality (Freedman, 1975; Beygelzimer et al., 2011):

Lemma D.3 (Lemma 3 of Lee et al. (2024b)). *Let M_t be a martingale difference sequence that satisfies $\max_{t\ge1}|M_t| \le R$. Then for any $\delta \in (0,1)$ and $\xi \in (0, 1/R]$, we have:*

$$\mathbb{P}\left(\sum_{s=1}^{t}M_s \le (e-2)\xi\sum_{s=1}^{t}\mathbb{E}_{s-1}[M_s^2] + \frac{1}{\xi}\log\frac{2}{\delta},\quad \forall t\ge1\right) \ge 1-\frac{\delta}{2}. \tag{73}$$

We will now show, with high probability, that the variance term induces a linear self-bounding inequality. Denoting $Z_s := \boldsymbol{v}_s^\top \boldsymbol{V}_s^{-1} \boldsymbol{v}_s$, which satisfies $0 \le Z_s \le \kappa \wedge 1$,

$$\mathbb{E}_{s-1}[M_s^2] \le \mathbb{E}_{s-1}[{Z_s}^2] \le (\kappa \wedge 1)\mathbb{E}_{s-1}[Z_s], \tag{74}$$

where the first inequality follows from the fact that for any random variable X, $\mathrm{Var}[X] \le \mathbb{E}[X^2]$.

Choosing $\xi = \frac{\kappa \wedge 1}{2(e-2)} \le \kappa$ in Lemma D.3, the following holds with probability at least $1 - \frac{\delta}{2}$:

$$(b) = \sum_{t=1}^{T} M_t \le \frac{(\kappa \wedge 1)^2}{2} \underbrace{\sum_{t=1}^{T} \mathbb{E}_{t-1}[Z_t]}_{=S_T!} + \frac{2(e-2)}{\kappa \wedge 1} \log \frac{2}{\delta}. \tag{75}$$

Now bringing everything together for Eqn. (71), the following holds with probability at least $1 - \frac{\delta}{2}$:

$$\begin{aligned} S_T = (a) + (b) &\le \underbrace{2d^2 \log\left(1 + \frac{\kappa T}{d^2}\right)}_{(a)\le} + \underbrace{\frac{(\kappa \wedge 1)^2}{2} S_T + \frac{2(e-2)}{\kappa \wedge 1} \log \frac{2}{\delta}}_{(b)\le} \\ &\le \frac{1}{2} S_T + 2d^2 \log\left(1 + \frac{\kappa T}{d^2}\right) + \frac{2(e-2)}{\kappa \wedge 1} \log \frac{2}{\delta}, \end{aligned} \tag{76}$$

which is a ***linear, self-bounding inequality***!

Solving for S_T, we have that with probability at least $1 - \frac{\delta}{2}$,

$$S_T = \sum_{t=1}^{T} \mathbb{E}_{t-1}\left[\boldsymbol{v}_t^\top \boldsymbol{V}_t^{-1} \boldsymbol{v}_t\right] \le 4d^2 \log\left(1 + \frac{\kappa T}{d^2}\right) + \frac{4(e-2)}{\kappa \wedge 1} \log \frac{2}{\delta}. \tag{77}$$

Combining everything at Eqn. (70), we have that with probability at least $1 - \delta$ (after union bound with the confidence sequence),

$$\text{MBR-Reg}_\eta(T) \lesssim \eta\beta\kappa^{-1} d^2 C_{\min}^{-1} \left(\log \frac{T}{d}\right) \left(d^2 \log \frac{\kappa T}{d} + \kappa^{-1} \log \frac{1}{\delta}\right). \tag{78}$$

Part II. $\tilde{\mathcal{O}}(\sqrt{T})$ Regret Bound. By the η-independent bound of Theorem 3.1, we get

$$\begin{aligned} \text{MBR-Reg}_\eta(T) &\le L_\mu \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\phi}_t \sim \hat{\pi}_t} \left[\|\boldsymbol{E}_t \boldsymbol{\phi}_t\|_2 + \|\boldsymbol{E}_t \boldsymbol{\phi}_t\|_2^2\right] && (79) \\ &\le L_\mu \sqrt{T \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\phi}_t \sim \hat{\pi}_t} \left[\|\boldsymbol{E}_t \boldsymbol{\phi}_t\|_2^2\right]} + L_\mu \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\phi}_t \sim \hat{\pi}_t} \left[\|\boldsymbol{E}_t \boldsymbol{\phi}_t\|_2^2\right] && \text{(Cauchy-Schwarz \& Jensen)} \\ &\lesssim \sqrt{T\kappa^{-1} d^4 C_{\min}^{-1} \left(\log \frac{T}{d}\right)^2} + \kappa^{-1} d^4 C_{\min}^{-1} \left(\log \frac{T}{d}\right)^2 && \text{(Part I)} \\ &= \kappa^{-\frac{1}{2}} C_{\min}^{-\frac{1}{2}} d^2 \sqrt{T} \log \frac{T}{d} + \kappa^{-1} C_{\min}^{-1} d^4 \left(\log \frac{T}{d}\right)^2. && (80) \end{aligned}$$

□

D.2. Proof of Lemma D.2: Coverage Lemma

We begin with this simple yet effective matrix lemma that will be useful throughout (we provide its proof at Section D.3 for completeness):

Lemma D.4. *If $\boldsymbol{A} \succeq \mathbf{0}$ and $\boldsymbol{B} \succeq \boldsymbol{C} \succeq \mathbf{0}$ of compatible sizes, then $\boldsymbol{A} \otimes \boldsymbol{B} \succeq \boldsymbol{A} \otimes \boldsymbol{C}$ and $\operatorname{tr}(\boldsymbol{AB}) \geq \operatorname{tr}(\boldsymbol{AC})$.*

We begin from the expectation on the RHS and work our way to the LHS:

$$\begin{aligned}
\mathbb{E}_{\tilde{\phi}\sim\rho}\left[(\phi\otimes\tilde{\phi})^\top \boldsymbol{M}(\phi\otimes\tilde{\phi})\right] &= \mathbb{E}_{\tilde{\phi}\sim\rho}\left[\operatorname{tr}\left((\phi\otimes\tilde{\phi})^\top \boldsymbol{M}(\phi\otimes\tilde{\phi})\right)\right] && (81)\\
&= \mathbb{E}_{\tilde{\phi}\sim\rho}\left[\operatorname{tr}\left(\boldsymbol{M}(\phi\otimes\tilde{\phi})(\phi\otimes\tilde{\phi})^\top\right)\right] && \text{(Cyclic property of } \operatorname{tr}(\cdot))\\
&= \operatorname{tr}\left(\boldsymbol{M}\mathbb{E}_{\tilde{\phi}\sim\rho}\left[(\phi\otimes\tilde{\phi})(\phi\otimes\tilde{\phi})^\top\right]\right) && \text{(Linearity of expectation)}\\
&= \operatorname{tr}\left(\boldsymbol{M}\mathbb{E}_{\tilde{\phi}\sim\rho}\left[\phi\phi^\top\otimes\tilde{\phi}\tilde{\phi}^\top\right]\right). && \text{(Mixed-product property of } \otimes)
\end{aligned}$$

Now, let us examine $\mathbb{E}_{\tilde{\phi}\sim\rho}\left[\phi\phi^\top\otimes\tilde{\phi}\tilde{\phi}^\top\right]$. Due to the linearity of the expectation and Assumption 4.1, we have

$$\mathbb{E}_{\tilde{\phi}\sim\rho}\left[\phi\phi^\top\otimes\tilde{\phi}\tilde{\phi}^\top\right] = \phi\phi^\top\otimes\mathbb{E}_{\tilde{\phi}\sim\rho}\left[\tilde{\phi}\tilde{\phi}^\top\right] \succeq C_{\min}\phi\phi^\top\otimes\boldsymbol{I}_d, \tag{82}$$

where the last Lowener order follows from Lemma D.4. Then, again by Lemma D.4, we have that

$$\begin{aligned}
\mathbb{E}_{\tilde{\phi}\sim\rho}\left[(\phi\otimes\tilde{\phi})^\top \boldsymbol{M}(\phi\otimes\tilde{\phi})\right] &= \operatorname{tr}\left(\boldsymbol{M}\phi\phi^\top\otimes\mathbb{E}_{\tilde{\phi}\sim\rho}\left[\tilde{\phi}\tilde{\phi}^\top\right]\right) && (83)\\
&\geq C_{\min}\operatorname{tr}\left(\boldsymbol{M}\phi\phi^\top\otimes\boldsymbol{I}_d\right) && (84)\\
&= C_{\min}\operatorname{tr}\left(\boldsymbol{M}\phi\phi^\top\otimes\left(\sum_{j=1}^d \boldsymbol{e}_j\boldsymbol{e}_j^\top\right)\right) && (85)\\
&= C_{\min}\sum_{j=1}^d \operatorname{tr}\left(\boldsymbol{M}\phi\phi^\top\otimes\boldsymbol{e}_j\boldsymbol{e}_j^\top\right) && (86)\\
&= C_{\min}\sum_{j=1}^d \operatorname{tr}\left(\boldsymbol{M}(\phi\otimes\boldsymbol{e}_j)(\phi\otimes\boldsymbol{e}_j)^\top\right) && \text{(Mixed-product property of } \otimes)\\
&= C_{\min}\sum_{j=1}^d (\phi\otimes\boldsymbol{e}_j)^\top\boldsymbol{M}(\phi\otimes\boldsymbol{e}_j). && (87)
\end{aligned}$$

□

D.3. Proof of Lemma D.4: PSD Matrix Lemma

Let $\boldsymbol{D} := \boldsymbol{B} - \boldsymbol{C}$. Since $\boldsymbol{B} \succeq \boldsymbol{C}$, we have $\boldsymbol{D} \succeq \mathbf{0}$.

We first show that $\boldsymbol{A}\otimes\boldsymbol{B} \succeq \boldsymbol{A}\otimes\boldsymbol{C}$. By bilinearity of the Kronecker product,

$$\boldsymbol{A}\otimes\boldsymbol{B} - \boldsymbol{A}\otimes\boldsymbol{C} = \boldsymbol{A}\otimes(\boldsymbol{B}-\boldsymbol{C}) = \boldsymbol{A}\otimes\boldsymbol{D}. \tag{88}$$

It therefore suffices to prove that $\boldsymbol{A}\otimes\boldsymbol{D}$ is positive semidefinite.

Let $\{\lambda_i(\boldsymbol{A})\}_{i=1}^n$ and $\{\lambda_j(\boldsymbol{D})\}_{j=1}^m$ denote the eigenvalues of $\boldsymbol{A}$ and $\boldsymbol{D}$, respectively. A standard property of Kronecker products implies that the eigenvalues of $\boldsymbol{A}\otimes\boldsymbol{D}$ are

$$\left\{\lambda_i(\boldsymbol{A})\lambda_j(\boldsymbol{D})\right\}_{i\in[n],\, j\in[m]}.$$

Since $\boldsymbol{A}\succeq\mathbf{0}$ and $\boldsymbol{D}\succeq\mathbf{0}$, all eigenvalues $\lambda_i(\boldsymbol{A})$ and $\lambda_j(\boldsymbol{D})$ are nonnegative, hence so are all products $\lambda_i(\boldsymbol{A})\lambda_j(\boldsymbol{D})$. Therefore $\boldsymbol{A}\otimes\boldsymbol{D}\succeq\mathbf{0}$, which yields $\boldsymbol{A}\otimes\boldsymbol{B}\succeq\boldsymbol{A}\otimes\boldsymbol{C}$.

Now we prove $\operatorname{tr}(\boldsymbol{AB}) \geq \operatorname{tr}(\boldsymbol{AC})$. Since $\boldsymbol{A}\succeq 0$, there exists a symmetric square root $\boldsymbol{A}^{1/2}$ such that $\boldsymbol{A} = \boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2}$. Using cyclicity of trace,

$$\operatorname{tr}(\boldsymbol{AD}) = \operatorname{tr}(\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2}\boldsymbol{D}) = \operatorname{tr}(\boldsymbol{A}^{1/2}\boldsymbol{D}\boldsymbol{A}^{1/2}). \tag{89}$$

Moreover, $\boldsymbol{A}^{1/2}\boldsymbol{D}\boldsymbol{A}^{1/2}$ is positive semidefinite. For any $\boldsymbol{x} \in \mathbb{R}^n$, letting $\boldsymbol{y} := \boldsymbol{A}^{1/2}\boldsymbol{x}$, we have

$$\boldsymbol{x}^\top \boldsymbol{A}^{1/2}\boldsymbol{D}\boldsymbol{A}^{1/2}\boldsymbol{x} = \boldsymbol{y}^\top \boldsymbol{D}\boldsymbol{y} \geq 0, \tag{90}$$

where the inequality uses $\boldsymbol{D} \succeq \boldsymbol{0}$. Hence $\boldsymbol{A}^{1/2}\boldsymbol{D}\boldsymbol{A}^{1/2} \succeq \boldsymbol{0}$, and therefore $\mathrm{tr}(\boldsymbol{A}^{1/2}\boldsymbol{D}\boldsymbol{A}^{1/2}) \geq 0$. This implies $\mathrm{tr}(\boldsymbol{A}\boldsymbol{D}) \geq 0$, completing the proof. □

E. Proof of Theorem 4.4: Regret Bound of `Explore-Then-Commit`

Recall the nuclear-norm regularized MLE (Fan et al., 2019; Lee et al., 2026): denoting $\boldsymbol{X}_t := \phi(\boldsymbol{x}_t, \boldsymbol{a}_t^1)\phi(\boldsymbol{x}_t, \boldsymbol{a}_t^2)^\top$,

$$\widehat{\boldsymbol{\Theta}}_{T_0} \leftarrow \operatorname*{arg\,min}_{\boldsymbol{\Theta} \in \mathrm{Skew}(d)} \mathcal{L}_{T_0}(\boldsymbol{\Theta}) + \lambda_{T_0} \|\boldsymbol{\Theta}\|_{\mathrm{nuc}}, \tag{91}$$

$$\mathcal{L}_{T_0}(\boldsymbol{\Theta}) := \frac{1}{T_0} \sum_{t=1}^{T_0} \{m(\langle \boldsymbol{\Theta}, \boldsymbol{X}_t \rangle) - r_t \langle \boldsymbol{\Theta}, \boldsymbol{X}_t \rangle\}. \tag{92}$$

We also introduce the following assumption used commonly in logistic and generalized linear bandits (Abeille et al., 2021; Russac et al., 2021; Lee et al., 2024a):

Assumption E.1. *The link function μ is **self-concordant with constant** $R_s \geq 0$, i.e.,*

$$\left|\ddot{\mu}\left(\boldsymbol{\phi}^\top \boldsymbol{\Theta} \boldsymbol{\phi}'\right)\right| \leq R_s \dot{\mu}\left(\boldsymbol{\phi}^\top \boldsymbol{\Theta} \boldsymbol{\phi}'\right), \quad \forall \boldsymbol{\phi}, \boldsymbol{\phi}' \in \mathcal{B}^d(1), \forall \boldsymbol{\Theta} \in \mathrm{Skew}(d, 2r; S). \tag{93}$$

Lastly, we define the following instance-specific curvature quantity:

$$\kappa_\star := \min_{\boldsymbol{\phi}, \boldsymbol{\phi}' \in \mathcal{B}^d(1)} \dot{\mu}\left(\boldsymbol{\phi}^\top \boldsymbol{\Theta}_\star \boldsymbol{\phi}'\right). \tag{94}$$

This parameter has been identified as a fundamental instance-dependent factor in the analysis of logistic and generalized linear bandits (Abeille et al., 2021; Lee et al., 2024a). In contrast to the global curvature κ (see Definition 2.1), which involves an additional minimization over $\boldsymbol{\Theta}_\star \in \mathrm{Skew}(d, 2r; S)$, $\kappa_\star$ captures the local geometry around the true parameter $\boldsymbol{\Theta}_\star$. We note that in many practical regimes, the global bound may be overly pessimistic, such that $\kappa^{-1} \gg \kappa_\star^{-1}$. In this Appendix, we show that under the additional assumption that μ is self-concordant, we can obtain dependencies w.r.t. $\kappa_\star^{-1}$ instead of κ^{-1}.

The following lemma provides a Frobenius error guarantee for the nuclear-norm regularized MLE $\widehat{\boldsymbol{\Theta}}$:

Lemma E.2 (Theorem 3.2 of Lee et al. (2026)). *Let $\delta \in (0,1)$, and $\tilde{\kappa} \in \{\kappa, \kappa_\star\}$ is such that $\tilde{\kappa} = \kappa_\star$ if Assumption E.1 holds, and $\tilde{\kappa} = \kappa$ if otherwise. Suppose that $T_0 \gtrsim \tilde{\kappa}^{-2} C_{\min}^{-4} dr \log \frac{d}{\delta}$. Then, with $\lambda_{T_0} = \sqrt{\frac{32 L_\mu \log \frac{4d}{\delta}}{T_0}}$, we have the following:*

$$\mathbb{P}\left(\left\| \widehat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}_\star \right\|_F \lesssim \frac{1}{\tilde{\kappa} C_{\min}^2} \sqrt{\frac{L_\mu r \log \frac{d}{\delta}}{T_0}} \right) \geq 1 - \delta. \tag{95}$$

We first invoke the η-dependent bound of Theorem 3.1 and the above lemma to $t \in [[T_0 + 1, T]]$:

$$\mathrm{MBR\text{-}Reg}_\eta(T) \lesssim T_0 + \eta\beta \sum_{t=T_0+1}^{T} \mathbb{E}_{\boldsymbol{\phi}_t \sim \hat{\pi}_t}\left[\|\boldsymbol{E}_t \boldsymbol{\phi}_t\|_2^2 \right] \leq T_0 + T\eta\beta \|\boldsymbol{E}_{T_0}\|_{\mathrm{op}}^2 \lesssim T_0 + \frac{T\eta\beta}{\tilde{\kappa}^2 C_{\min}^4} \frac{r \log \frac{d}{\delta}}{T_0}. \tag{96}$$

This balances out when $T_0 \asymp \tilde{\kappa}^{-1} C_{\min}^{-2} \sqrt{T\eta\beta r \log \frac{d}{\delta}}$.

Now, we invoke the η-independent bound of Theorem 3.1, which yields

$$\mathrm{MBR\text{-}Reg}_\eta(T) \lesssim T_0 + \sum_{t=T_0+1}^{T} \mathbb{E}_{\boldsymbol{\phi}_t \sim \hat{\pi}_t}\left[\|\boldsymbol{E}_t \boldsymbol{\phi}_t\|_2 \right] + \sum_{t=T_0+1}^{T} \mathbb{E}_{\boldsymbol{\phi}_t \sim \hat{\pi}_t}\left[\|\boldsymbol{E}_t \boldsymbol{\phi}_t\|_2^2 \right] \tag{97}$$

$$\leq T_0 + T \|\boldsymbol{E}_{T_0}\|_{\mathrm{op}} + T \|\boldsymbol{E}_{T_0}\|_{\mathrm{op}}^2 \tag{98}$$

$$\lesssim T_0 + \frac{T}{\tilde{\kappa} C_{\min}^2} \sqrt{\frac{L_\mu r \log \frac{d}{\delta}}{T_0}} + \frac{T}{\tilde{\kappa}^2 C_{\min}^4} \frac{L_\mu r \log \frac{d}{\delta}}{T_0}. \tag{99}$$

This balances out when $T_0 \asymp \left(T^2 \tilde{\kappa}^{-2} C_{\min}^{-4} r \log \frac{d}{\delta}\right)^{1/3}$ and $T \gtrsim \tilde{\kappa} C_{\min}^2 r \log \frac{d}{\delta}$, the latter which we can assume to hold without loss of any generality. □

Remark E.3 (Unknown T). *We assume T is known to optimally tune T_0. If T is unknown, the standard doubling trick (Auer et al., 1995; Besson & Kaufmann, 2018) yields the same regret bound up to a constant factor.*

F. Eluder Dimension of GBPM

F.1. Upper Bounding the Eluder Dimension of Wu et al. (2025)

To instantiate the regret bound of Wu et al. (2025) in Appendix H, we first recall their specific notion of the eluder dimension:

Definition F.1 (Eluder dimension, General Preference Model (Wu et al., 2025)). *Under the general preference model, for any $\mathcal{D}_{t-1} = \{(\boldsymbol{x}_i, \boldsymbol{a}_i^1, \boldsymbol{a}_i^2)\}_{i=1}^{t-1}$, we define the uncertainty of $(\boldsymbol{x}, \boldsymbol{a}^1, \boldsymbol{a}^2)$ with respect to $\mathcal{P}$ as*

$$U_{GP}(\lambda, \boldsymbol{x}, \boldsymbol{a}^1, \boldsymbol{a}^2; \mathcal{P}, \mathcal{D}_{t-1}) = \sup_{P_1, P_2 \in \mathcal{P}} \frac{|P_1(\boldsymbol{a}^1 \succ \boldsymbol{a}^2 \mid \boldsymbol{x}) - P_2(\boldsymbol{a}^1 \succ \boldsymbol{a}^2 \mid \boldsymbol{x})|}{\sqrt{\lambda + \sum_{s=1}^{t-1} (P_1(\boldsymbol{a}_s^1 \succ \boldsymbol{a}_s^2 \mid \boldsymbol{x}_s) - P_2(\boldsymbol{a}_s^1 \succ \boldsymbol{a}_s^2 \mid \boldsymbol{x}_s))^2}}. \tag{100}$$

Then the eluder dimension of $\mathcal{P}$ is defined as

$$d(\mathcal{P}, \lambda, T) := \sup_{\boldsymbol{x}_{1:T}, \boldsymbol{a}_{1:T}^1, \boldsymbol{a}_{1:T}^2} \sum_{t \in [T]} \min \left\{ 1, \left[U_{GP}(\lambda, \boldsymbol{x}_t, \boldsymbol{a}_t^1, \boldsymbol{a}_t^2; \mathcal{P}, \mathcal{D}_{t-1}) \right]^2 \right\}. \tag{101}$$

Using the standard elliptical potential arguments, we now derive an upper bound for this complexity measure under the GBPM:

Proposition F.2. *For GBPM, the eluder dimension $d(\mathcal{P}, \lambda, T)$ is bounded as*

$$d(\mathcal{P}, \lambda, T) \leq \frac{2d^2 L_\mu^2}{\kappa^2} \log \left(1 + \frac{4\kappa^2 S^2 T}{d^2 \lambda} \right). \tag{102}$$

Proof. For the proof, let us arbitrarily fix a sequence of context-action-action pairs $(\boldsymbol{x}_{1:T}, \boldsymbol{a}_{1:T}^1, \boldsymbol{a}_{1:T}^2)$, and let us denote the induced sequence of features as $\boldsymbol{\phi}_{1:T}^1$ and $\boldsymbol{\phi}_{1:T}^2$, where $\boldsymbol{\phi}_t^i := \boldsymbol{\phi}(\boldsymbol{x}_t, \boldsymbol{a}_t^i)$ for $t \in [T]$ and $i \in \{1, 2\}$. Let us also denote $\boldsymbol{\Phi}_t := \boldsymbol{\phi}_t^1 (\boldsymbol{\phi}_t^2)^\top$. For any $\boldsymbol{\Theta} \in \Theta$, the induced preference model is defined as

$$P_{\boldsymbol{\Theta}}(\boldsymbol{\phi}_t^1, \boldsymbol{\phi}_t^2) := \mu\left((\boldsymbol{\phi}_t^1)^\top \boldsymbol{\Theta} \boldsymbol{\phi}_t^2 \right) = \mu\left(\langle \boldsymbol{\Theta}, \boldsymbol{\Phi}_t \rangle \right). \tag{103}$$

We bound the uncertainty via the elliptical potential lemma (Lemma J.3). Let us denote $P_i = P_{\boldsymbol{\Theta}_i}$ for some arbitrary $\boldsymbol{\Theta}_i \in \Theta$. We first upper bound the numerator as follows: denoting $\alpha(\boldsymbol{x}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2) := \int_0^1 \dot{\mu}\big(\boldsymbol{x}^\top \boldsymbol{\theta}_1 + z\, \boldsymbol{x}^\top (\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)\big)\, dz$ for $\boldsymbol{x}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^d$,

$$\begin{aligned}
|\mu(\langle \boldsymbol{\Theta}_1, \boldsymbol{\Phi}_t \rangle) - \mu(\langle \boldsymbol{\Theta}_2, \boldsymbol{\Phi}_t \rangle)| &= |\alpha(\mathrm{vec}(\boldsymbol{\Phi}_t), \mathrm{vec}(\boldsymbol{\Theta}_1), \mathrm{vec}(\boldsymbol{\Theta}_2))\, \langle \boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2, \boldsymbol{\Phi}_t \rangle| && \text{(Mean-value theorem)} \\
&= \left| \left\langle \boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2, \underbrace{\alpha(\mathrm{vec}(\boldsymbol{\Phi}_t), \mathrm{vec}(\boldsymbol{\Theta}_1), \mathrm{vec}(\boldsymbol{\Theta}_2))\, \boldsymbol{\Phi}_t}_{\triangleq \bar{\boldsymbol{\Phi}}_t(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)} \right\rangle \right| && (104) \\
&\leq \|\mathrm{vec}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\|_{\boldsymbol{G}_t(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)} \left\| \mathrm{vec}(\bar{\boldsymbol{\Phi}}_t(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)) \right\|_{\boldsymbol{G}_t(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)^{-1}} \\
& && \text{(Cauchy-Schwarz inequality)}
\end{aligned}$$

for some matrix $\boldsymbol{G}_t(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) \succ \boldsymbol{0}$ to be determined later.

For the denominator squared:

$$\lambda + \sum_{s=1}^{t-1} \left(\mu(\langle \boldsymbol{\Theta}_1, \boldsymbol{\Phi}_s \rangle) - \mu(\langle \boldsymbol{\Theta}_2, \boldsymbol{\Phi}_s \rangle) \right)^2 \tag{105}$$

$$= \lambda + \sum_{s=1}^{t-1} \left[\alpha(\mathrm{vec}(\boldsymbol{\Phi}_s), \mathrm{vec}(\boldsymbol{\Theta}_1), \mathrm{vec}(\boldsymbol{\Theta}_2))^2\, \langle \boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2, \boldsymbol{\Phi}_s \rangle^2 \right] \tag{106}$$

$$= \lambda + \mathrm{vec}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)^\top \left[\sum_{s=1}^{t-1} \left(\alpha(\mathrm{vec}(\boldsymbol{\Phi}_s), \mathrm{vec}(\boldsymbol{\Theta}_1), \mathrm{vec}(\boldsymbol{\Theta}_2))^2\, \mathrm{vec}(\boldsymbol{\Phi}_s)\, \mathrm{vec}(\boldsymbol{\Phi}_s)^\top \right) \right] \mathrm{vec}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2) \tag{107}$$

$$\geq \operatorname{vec}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)^\top \underbrace{\left[\frac{\lambda}{4S^2}\boldsymbol{I} + \sum_{s=1}^{t-1}\left[\alpha(\operatorname{vec}(\boldsymbol{\Phi}_s), \operatorname{vec}(\boldsymbol{\Theta}_1), \operatorname{vec}(\boldsymbol{\Theta}_2))^2 \operatorname{vec}(\boldsymbol{\Phi}_s)\operatorname{vec}(\boldsymbol{\Phi}_s)^\top\right]\right]}_{\triangleq \boldsymbol{G}_t(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)} \operatorname{vec}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2) \tag{108}$$

$$= \|\operatorname{vec}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\|^2_{\boldsymbol{G}_t(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)}. \tag{109}$$

Combining the above two inequalities, we have that

$$U_{\text{GP}}(\lambda, \boldsymbol{x}, \boldsymbol{a}^1, \boldsymbol{a}^2; \mathcal{P}, \mathcal{D}_{t-1}) \leq \sup_{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2 \in \Theta} \frac{\|\operatorname{vec}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\|_{\boldsymbol{G}_t} \left\|\operatorname{vec}(\bar{\boldsymbol{\Phi}}_t(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2))\right\|_{\boldsymbol{G}_t(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)^{-1}}}{\|\operatorname{vec}(\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2)\|_{\boldsymbol{G}_t(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)}} \tag{110}$$

$$= \sup_{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2 \in \Theta} \left\|\operatorname{vec}(\bar{\boldsymbol{\Phi}}_t(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2))\right\|_{\boldsymbol{G}_t(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)^{-1}} \tag{111}$$

We can lower-bound $\boldsymbol{G}_t(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2)$ in the Löwner sense using the fact that $\alpha(\cdot, \cdot, \cdot) \geq \kappa$,

$$\boldsymbol{G}_t(\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2) \succeq \frac{\lambda}{4S^2}\boldsymbol{I} + \sum_{s=1}^{t-1} \kappa^2 \operatorname{vec}(\boldsymbol{\Phi}_s)\operatorname{vec}(\boldsymbol{\Phi}_s)^\top = \frac{\lambda}{4S^2}\boldsymbol{I} + \sum_{s=1}^{t-1} \operatorname{vec}(\kappa\boldsymbol{\Phi}_s)\operatorname{vec}(\kappa\boldsymbol{\Phi}_s)^\top \triangleq \boldsymbol{V}_t. \tag{112}$$

Then, as $\alpha(\cdot, \cdot, \cdot) \leq L_\mu$, we have that

$$U_{\text{GP}}(\lambda, \boldsymbol{x}, \boldsymbol{a}^1, \boldsymbol{a}^2; \mathcal{P}, \mathcal{D}_{t-1}) \leq \frac{L_\mu}{\kappa} \|\operatorname{vec}(\kappa\boldsymbol{\Phi}_t)\|_{\boldsymbol{V}_t^{-1}}. \tag{113}$$

We then conclude the proof via the elliptical potential lemma (Lemma J.3):

$$d(\mathcal{P}, \lambda, T) \leq \sum_{t=1}^{T} \min\left\{1, \frac{L_\mu^2}{\kappa^2}\|\operatorname{vec}(\kappa\boldsymbol{\Phi}_t)\|^2_{\boldsymbol{V}_t^{-1}}\right\} \tag{114}$$

$$\leq \frac{L_\mu^2}{\kappa^2} \sum_{t=1}^{T} \min\left\{1, \|\operatorname{vec}(\kappa\boldsymbol{\Phi}_t)\|^2_{\boldsymbol{V}_t^{-1}}\right\} \qquad (\kappa \leq L_\mu)$$

$$\leq \frac{2d^2 L_\mu^2}{\kappa^2} \log\left(1 + \frac{4\kappa^2 S^2 T}{d^2\lambda}\right). \tag{115}$$

□

F.2. Connection to the Standard Eluder Dimensions

In this appendix, we will elucidate the connection between Definition F.1 and two other "standard" definitions of eluder-type complexities: *sequential extrapolation coefficient (SEC)* (Xie et al., 2023) and the original eluder dimension (Russo & Van Roy, 2013) For simplicity, we denote $\mathcal{Z} := \mathcal{X} \times \mathcal{A} \times \mathcal{A}$, and $P(\boldsymbol{z}) := P(\boldsymbol{a}^1 \succ \boldsymbol{a}^2 \mid \boldsymbol{x})$ for $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{a}^1, \boldsymbol{a}^2)$.

First, we recall the definition of SEC adopted for our setting:

Definition F.3 (λ-regularized SEC, Definition 7 of Xie et al. (2023))**.** *Let $\Pi \subseteq \Delta(\mathcal{Z})$ be a distribution class. Then, the SEC is defined as*

$$\operatorname{SEC}_\lambda(\mathcal{P}, \Pi, T) := \sup_{P^1_{1:T}, P^2_{1:T} \subseteq \mathcal{P}} \sup_{\nu_{1:T} \subseteq \Pi} \sum_{t=1}^{T} \frac{\left(\mathbb{E}_{\boldsymbol{z}_t \sim \nu_t}[P_t^1(\boldsymbol{z}_t) - P_t^2(\boldsymbol{z}_t)]\right)^2}{\lambda + \sum_{s=1}^{t-1} \mathbb{E}_{\boldsymbol{z}_s \sim \nu_s}\left[(P_t^1(\boldsymbol{z}_s) - P_t^2(\boldsymbol{z}_s))^2\right]}. \tag{116}$$

Note that when the distribution class is restricted to the set of Dirac measures $\boldsymbol{D} := \{\delta_{\boldsymbol{z}} : \boldsymbol{z} \in Z\}$, we have the following relationship between SEC and Definition F.1:

$$d(\mathcal{P}, \lambda, T) = \sup_{\boldsymbol{z}_{1:T}} \sum_{t \in [T]} \min\left\{1, \sup_{P_1, P_2 \in \mathcal{P}} \frac{(P_1(\boldsymbol{z}_t) - P_2(\boldsymbol{z}_t))^2}{\lambda + \sum_{s=1}^{t-1}(P_1(\boldsymbol{z}_s) - P_2(\boldsymbol{z}_s))^2}\right\} \tag{117}$$

$$\leq \sup_{\boldsymbol{z}_{1:T}} \sum_{t\in[T]} \sup_{P_1,P_2\in\mathcal{P}} \frac{(P_1(\boldsymbol{z}_t)-P_2(\boldsymbol{z}_t))^2}{\lambda+\sum_{s=1}^{t-1}(P_1(\boldsymbol{z}_s)-P_2(\boldsymbol{z}_s))^2} \tag{118}$$

$$= \sup_{P^1_{1:T},P^2_{1:T}} \sup_{\boldsymbol{z}_{1:T}} \sum_{t\in[T]} \frac{(P^1_t(\boldsymbol{z}_t)-P^2_t(\boldsymbol{z}_t))^2}{\lambda+\sum_{s=1}^{t-1}(P^1_s(\boldsymbol{z}_s)-P^2_s(\boldsymbol{z}_s))^2} \tag{119}$$

$$= \sup_{P^1_{1:T},P^2_{1:T}} \sup_{\nu_{1:T}\subseteq\boldsymbol{D}} \sum_{t\in[T]} \frac{(\mathbb{E}_{\boldsymbol{z}_t\sim\nu_t}[P^1_t(\boldsymbol{z}_t)-P^2_t(\boldsymbol{z}_t)])^2}{\lambda+\sum_{s=1}^{t-1}(\mathbb{E}_{\boldsymbol{z}_s\sim\nu_s}[P^1_t(\boldsymbol{z}_s)-P^2_t(\boldsymbol{z}_s)])^2} \tag{120}$$

$$= \mathrm{SEC}_\lambda(\mathcal{P},\boldsymbol{D},T). \tag{121}$$

Second, we recall the standard eluder dimension of Foster et al. (2021) and Li et al. (2022a):[5]

Definition F.4 (Definition 1 of Li et al. (2022a)). *For any fixed preference $P^*\in\mathcal{P}$, and scale $\varepsilon\geq 0$, the **exact eluder dimension** $\underline{\mathrm{Edim}}_{P^*}(\mathcal{P},\varepsilon)$ is the largest $m\in\mathbb{N}$ such that there exists a sequence $\{(\boldsymbol{z}_t,P_t)\}_{t\in[m]}\subset\mathcal{Z}\times\mathcal{P}$ such that the following holds: for all $t\in[m]$,*

$$|P_t(\boldsymbol{z}_t)-P^*(\boldsymbol{z}_t)|>\varepsilon, \quad \textit{and} \quad \sum_{s<t}(P_t(\boldsymbol{z}_s)-P^*(\boldsymbol{z}_s))^2<\varepsilon^2. \tag{122}$$

Then for all $\varepsilon>0$, we define:

- *The **eluder dimension** is $\mathrm{Edim}_{P^*}(\mathcal{P},\varepsilon):=\sup_{\varepsilon'\geq\varepsilon}\underline{\mathrm{Edim}}_{P^*}(\mathcal{P},\varepsilon')$.*
- *$\underline{\mathrm{Edim}}(\mathcal{P},\varepsilon):=\sup_{P^*\in\mathcal{P}}\underline{\mathrm{Edim}}_{P^*}(\mathcal{P},\varepsilon')$ and $\mathrm{Edim}(\mathcal{P},\varepsilon):=\sup_{P^*\in\mathcal{P}}\mathrm{Edim}_{P^*}(\mathcal{P},\varepsilon')$.*

We first prove that $\mathrm{Edim}(\mathcal{P},\varepsilon)$ and $\mathrm{SEC}_\lambda(\mathcal{P},\boldsymbol{D},T)$ are equivalent up to some constants and logarithmic factors:

Proposition F.5. *Suppose that $\mathcal{Z}\subseteq\mathcal{B}^d(1)$ and $\Theta\subseteq\mathrm{Skew}(d;2r,S)$ for some $S>0$. Let $\varepsilon>0$, $T\geq\mathrm{Edim}(\mathcal{P},\varepsilon)$, and $\lambda\geq 1$. Then,*

$$\frac{\varepsilon^2\mathrm{Edim}(\mathcal{P},\varepsilon)}{\lambda+\varepsilon^2}\leq\mathrm{SEC}_\lambda(\mathcal{P},\boldsymbol{D},T)\lesssim\mathrm{Edim}(\mathcal{P},T^{-1/2})\log T. \tag{123}$$

Proof. We prove each direction separately.

Upper Bound. Noting that for $\lambda\geq 1$, $\mathrm{SEC}_\lambda(\mathcal{P},\boldsymbol{D},T)\leq\mathrm{SEC}_1(\mathcal{P},\boldsymbol{D},T)$, this immediately follows from Xie et al. (2023, Proposition 7) with $\mathcal{D}=\boldsymbol{D}$.

Lower Bound. Consider the eluder witness, i.e., a sequence of $\{(\boldsymbol{z}_t,P_t)\}_{t\in d_e}$ and some fixed preference P^* that attains the eluder dimension $d_e:=\mathrm{Edim}(\mathcal{P},\varepsilon)$. Then, by definition,

$$\mathrm{SEC}_\lambda(\mathcal{P},\boldsymbol{D},T)=\sup_{P^1_{1:T},P^2_{1:T}\subseteq\mathcal{P}}\sup_{\boldsymbol{z}_{1:T}\subseteq\mathcal{Z}}\sum_{t=1}^{T}\frac{\left(P^1_t(\boldsymbol{z}_t)-P^2_t(\boldsymbol{z}_t)\right)^2}{\lambda+\sum_{s=1}^{t-1}(P^1_t(\boldsymbol{z}_s)-P^2_t(\boldsymbol{z}_s))^2} \tag{124}$$

$$\geq\sup_{P^1_{1:T}\subseteq\mathcal{P}}\sup_{\boldsymbol{z}_{1:T}\subseteq\mathcal{Z}}\sum_{t=1}^{T}\frac{\left(P^1_t(\boldsymbol{z}_t)-P^*(\boldsymbol{z}_t)\right)^2}{\lambda+\sum_{s=1}^{t-1}(P^1_t(\boldsymbol{z}_s)-P^*(\boldsymbol{z}_s))^2} \qquad (\text{Set } P^2_t=P^* \text{ for all } t\in[T])$$

$$\geq\sum_{t=1}^{d_e}\frac{(P_t(\boldsymbol{z}_t)-P^*(\boldsymbol{z}_t))^2}{\lambda+\sum_{s=1}^{t-1}(P_t(\boldsymbol{z}_s)-P^*(\boldsymbol{z}_s))^2} \qquad (\text{Set } \boldsymbol{z}_{1:T} \text{ and } P^1_{1:T} \text{ to be the eluder witness sequence})$$

$$>\sum_{t=1}^{d_e}\frac{\varepsilon^2}{\lambda+\varepsilon^2}=\frac{d_e\varepsilon^2}{\lambda+\varepsilon^2}. \tag{125}$$

□

[5]The original definition is due to Russo & Van Roy (2013) and slightly different, but as mentioned in Li et al. (2022a), the "new" definition is "never larger and is sufficient to analyze all the applications of eluder dimension in literature."

We conclude with a nearly-tight characterization of the eluder dimension of GBPM, whose proof is deferred to the next subsection:

Proposition F.6. *Let $\mathcal{X} \subseteq \mathcal{B}^d(1)$ and $\Theta \subseteq \mathrm{Skew}(d, 2r; S)$ for some $S > 0$. Define the function class as the **GBPM**:*

$$\mathcal{P} := \left\{(\boldsymbol{x}, \boldsymbol{y}) \mapsto \mu\left(\boldsymbol{x}^\top \boldsymbol{\Theta} \boldsymbol{y}\right) : \boldsymbol{\Theta} \in \Theta\right\}, \tag{126}$$

where we have the following properties: $\kappa \le \dot{\mu}(\boldsymbol{x}^\top \boldsymbol{\Theta} \boldsymbol{y}) \le L_\mu$ and $\left|\boldsymbol{x}^\top \boldsymbol{\Theta} \boldsymbol{y}\right| \le S$[a] for all $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\Theta}) \in \mathcal{X} \times \mathcal{X} \times \Theta$. Then, its eluder dimension is bounded as follows: for any $\varepsilon < SL_\mu$,

$$\mathrm{Edim}(\mathcal{P}, \varepsilon) \le \frac{3e}{e-1} \cdot d^2 \cdot \frac{L_\mu^2}{\kappa^2} \cdot \log \frac{24 S^2 L_\mu^2}{\varepsilon^2}. \tag{127}$$

Furthermore, if $\mu(z) = \frac{1}{2} + z$, then we have the following lower bound:

$$\mathrm{Edim}(\mathcal{P}, \varepsilon) \ge \binom{d}{2} \log_4 \frac{S}{\sqrt{3}\varepsilon}. \tag{128}$$

[a] This follows from matrix Hölder inequality: $\left|\boldsymbol{x}^\top \boldsymbol{\Theta} \boldsymbol{y}\right| = \left|\langle \boldsymbol{\Theta}, \boldsymbol{x}\boldsymbol{y}^\top \rangle\right| \le \|\boldsymbol{\Theta}\|_{\mathrm{nuc}} \left\|\boldsymbol{x}\boldsymbol{y}^\top\right\|_{\mathrm{op}} \le S$.

Note that the same $\widetilde{\Omega}(d^2)$ lower bound applies to the SEC due to Proposition F.5. This $\widetilde{\Omega}(d^2)$ scaling implies that eluder-dimension-based frameworks cannot efficiently exploit the low-rank structure of $\boldsymbol{\Theta}$. The high eluder dimension arises because $\mathrm{Skew}(d; 2r)$ contains rank-2 "coordinate spikes" of the form $(\boldsymbol{e}_i \boldsymbol{e}_j^\top - \boldsymbol{e}_j \boldsymbol{e}_i^\top)$. An adversary can query specific pairs to isolate these directions one-by-one. Since the eluder dimension measures worst-case separability rather than metric entropy (which scales as $\mathcal{O}(dr)$), it reflects the ambient basis size even when the parameter manifold is low-dimensional.

This limitation is best understood through the "global embedding" characterization. Specifically, a standard sufficient condition for bounding the eluder dimension by the μ-rank relies on constructing *global* maps $\phi : \mathcal{X} \times \mathcal{X} \to \mathcal{B}^{d_e}(1)$ and $w : \Theta \to \mathcal{B}^{d_e}(R)$ such that $\boldsymbol{x}^\top \boldsymbol{\Theta} \boldsymbol{y} = \langle \phi(\boldsymbol{x}, \boldsymbol{y}), w(\boldsymbol{\Theta}) \rangle$ (Li et al., 2022a, Proposition 4). While skew-symmetry admits a Schur decomposition $\boldsymbol{\Theta} = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^\top$ that allows the representation

$$\boldsymbol{x}^\top \boldsymbol{\Theta} \boldsymbol{y} = \left\langle \mathrm{vec}\big((\boldsymbol{Q}^\top \boldsymbol{x})(\boldsymbol{Q}^\top \boldsymbol{y})^\top\big),\ \mathrm{vec}(\boldsymbol{\Lambda}) \right\rangle,$$

this does not yield a valid low-dimensional witness. Crucially, the feature map depends on the basis $\boldsymbol{Q}$ (and thus on the specific parameter $\boldsymbol{\Theta}$), which violates the condition of having a single global feature map across the entire hypothesis class. Consequently, guarantees relying on such complexity measures (e.g., `GS` by Wu et al. (2025)) incur the full d^2 complexity, mirroring the statistical hardness of quadratic functions with full-rank Hessians (Osband & Van Roy, 2014, Proposition 3).

G. Proof of Proposition F.6

For the proof, we recall the notion of *generalized rank* and a useful proposition linking the above two concepts:

Definition G.1 (Definition 3 of Li et al. (2022a))**.** *For a given $\mu : \mathbb{R} \to \mathbb{R}$, the **$\mu$-rank** of $\mathcal{P}$ at scale $R > 0$, denoted as $\mu\text{-rk}(\mathcal{P}, R)$, is the smallest dimension $d \in \mathbb{N}$ for which there exist $R_\phi, R_w > 0$ with $R_\phi R_w = R$, and (global) mappings $\phi : \mathcal{X} \times \mathcal{X} \to \mathcal{B}^d(R_\phi)$ and $w : \mathcal{P} \to \mathcal{B}^d(R_w)$ such that*

$$P(\boldsymbol{x} \succ \boldsymbol{y}) = \mu\left(\langle \phi(\boldsymbol{x}, \boldsymbol{y}), w(P) \rangle\right), \quad \forall (\boldsymbol{x}, \boldsymbol{y}, P) \in \mathcal{X} \times \mathcal{X} \times \mathcal{P}, \tag{129}$$

or ∞ if no such d exists.

Proposition G.2 (Proposition 4(ii) of Li et al. (2022a))**.** *For all $\varepsilon < RL_\mu$,*

$$\underline{\mathrm{Edim}}(\mathcal{P}, \varepsilon) \le \frac{3e}{e-1} \cdot \mu\text{-rk}(\mathcal{P}) \cdot \frac{L_\mu^2}{\kappa^2} \cdot \log \frac{24 R^2 L_\mu^2}{\varepsilon^2}. \tag{130}$$

We prove the upper and lower bounds separately.

Upper Bound. This follows trivially from adapting Proposition G.2 to our setting by considering $\phi : (\boldsymbol{x}, \boldsymbol{y}) \mapsto \mathrm{vec}(\boldsymbol{x}\boldsymbol{y}^\top)$ and $w : \boldsymbol{\Theta} \mapsto \mathrm{vec}(\boldsymbol{\Theta})$.

Lower Bound. The construction is largely inspired by that of Li et al. (2022a, Proposition 5), which we adapt to our setting. We will construct a sequence $\{(\boldsymbol{x}_t, \boldsymbol{y}_t, \boldsymbol{\Theta}_t)\}_{t\in[m]}$ that witnesses the claimed lower bound with $\boldsymbol{\Theta}_\star = 0$. The key observation is that $\mathrm{Skew}(d)$ admits the following orthonormal basis: $\mathcal{B} \triangleq \left\{ \frac{1}{\sqrt{2}}(\boldsymbol{e}_i\boldsymbol{e}_j^\top - \boldsymbol{e}_j\boldsymbol{e}_i^\top) \right\}_{1\le i<j\le d}$.

For given ε, let $\alpha \in (\varepsilon, \sqrt{3}\varepsilon)$ and $k := \lfloor \log_4 \frac{S}{\alpha} \rfloor$. Then, we can first consider the following sequence of length $k+1$: for $t \in \{0\} \cup [k]$,

$$\boldsymbol{x}_t = 2^{t-k}\boldsymbol{e}_1,\ \boldsymbol{y}_t = 2^{t-k}\boldsymbol{e}_2, \quad \boldsymbol{\Theta}_t = \alpha \cdot 2^{2(k-t)}(\boldsymbol{e}_1\boldsymbol{e}_2^\top - \boldsymbol{e}_2\boldsymbol{e}_1^\top). \tag{131}$$

For each t, we have that

$$\mu(\boldsymbol{x}_t^\top \boldsymbol{\Theta}_t \boldsymbol{y}_t) - \mu(0) = \boldsymbol{x}_t^\top \boldsymbol{\Theta}_t \boldsymbol{y}_t = \alpha > \varepsilon, \tag{132}$$

and

$$\sum_{s<t} \left(\mu(\boldsymbol{x}_s^\top \boldsymbol{\Theta}_t \boldsymbol{y}_s) - \mu(0)\right)^2 = \sum_{s<t} \left(\boldsymbol{x}_s^\top \boldsymbol{\Theta}_t \boldsymbol{y}_s\right)^2 = \alpha^2 \sum_{s<t} 4^{2s-2t} < \frac{1}{15}\alpha^2 < \varepsilon^2. \tag{133}$$

As $\boldsymbol{x}_t, \boldsymbol{x}_t \in \mathcal{B}^d(1)$ and $\|\boldsymbol{\Theta}_t\|_{\mathrm{nuc}} \le \alpha 2^{2k} \le \alpha 4^{\log_4 \frac{S}{\alpha}} = S$, we have $\underline{\mathrm{Edim}}(\mathcal{P}, \varepsilon) \ge k+1 \ge \log_4 \frac{S}{\alpha} \ge \log_4 \frac{S}{\sqrt{3}\varepsilon}$.

Now we concatenate $\binom{d}{2} = \frac{d(d-1)}{2}$ times across the basis $\mathcal{B}$, i.e.,

$$\boldsymbol{x}_{t,i,j} = 2^{t-k}\boldsymbol{e}_i,\ \boldsymbol{y}_{t,i,j} = 2^{t-k}\boldsymbol{e}_j, \quad \boldsymbol{\Theta}_{t,i,j} = \alpha \cdot 2^{2(k-t)}(\boldsymbol{e}_i\boldsymbol{e}_j^\top - \boldsymbol{e}_j\boldsymbol{e}_i^\top) \tag{134}$$

for $1 \le i < j \le d$, and we are done. □

H. Instantiating Regret Bound of Wu et al. (2025) to GBPM

H.1. Regret Bound of `Greedy Sampling`

For this section, we will consider the reverse KL-regularization as in Wu et al. (2025), i.e., $\psi(\pi) = D_{\mathrm{KL}}(\pi, \pi_{\mathrm{ref}})$ for some fixed $\pi_{\mathrm{ref}} \in \Pi$. Recall that we defined the regularized and unregularized Max-Best-Response Regrets as

$$\text{MBR-Reg}_\eta(T) := \sum_{t=1}^{T} \max_{\pi \in \Pi} \left\{ \frac{1}{2} - J_\eta(\hat{\pi}_t^1, \pi) \right\}, \quad \text{MBR-Reg}(T) := \sum_{t=1}^{T} \max_{\pi \in \Pi} \left\{ \frac{1}{2} - J(\hat{\pi}_t^1, \pi) \right\}. \tag{135}$$

In this section, we derive the regularized and unregularized regret bound of `Greedy Sampling (GS)` of Wu et al. (2025) for GBPM, based on general function approximation. We first recall the regret bound from Wu et al. (2025):

Theorem H.1 (Theorem 1 of Wu et al. (2025))**.** *Suppose that the preference class $\mathcal{P}$ is finite with cardinality $N_{\mathcal{P}} = |\mathcal{P}| < \infty$. For any $\delta \in (0,1)$, with probability at least $1-\delta$, `GS` attains the following regret bound:*

$$\text{MBR-Reg}_\eta(T) = O\left(e^\eta \, d(\mathcal{P}, \lambda, T) \log(N_{\mathcal{P}} T/\delta)\right). \tag{136}$$

Instantiation for GBPM. We first instantiate the KL-regularized regret bound for GBPM:

Theorem H.2 (KL-regularized Regret Bound of `Greedy Sampling`)**.** *For any $\delta \in (0,1)$, with probability with at least $1-\delta$, `GS` (when applied to **GBPM**) attains the following regret bound:*

$$\text{MBR-Reg}_\eta(T) \lesssim e^\eta \cdot \frac{d^2 L_\mu^2}{\kappa^2} \cdot \log T \cdot \left(\log \frac{T}{\delta} + dr \log(LST) \right). \tag{137}$$

Proof. The proof consists of two parts: 1) Extending the preference model class size term for the infinite preference space, and 2) Bounding the eluder dimension of ***GBPM***.

We first extend the term $N_{\mathcal{P}} = |\mathcal{P}|$ for the infinite space case $\Theta \triangleq \mathrm{Skew}(d, 2r; S)$ by using covering number arguments. We denote $P_{\mathbf{\Theta}}$ as the preference probability given by ***GBPM*** for parameter $\mathbf{\Theta}$. For simplicity, we use the following notations introduced in the previous section. We denote $\mathcal{Z} := \mathcal{X} \times \mathcal{A} \times \mathcal{A}$ and $P(\boldsymbol{z}) := P(\boldsymbol{a}^1 \succ \boldsymbol{a}^2 \mid \boldsymbol{x})$ for $\boldsymbol{z} = (\boldsymbol{x}, \boldsymbol{a}^1, \boldsymbol{a}^2)$. Using these, we have the following lemma, whose proof is provided in Section H.2:

Lemma H.3. *Let $\{(\boldsymbol{z}_i, r_i)\}_{i \in [t]}$ be a potentially adaptively collected data with $r_i \sim \mathrm{Ber}(P(\boldsymbol{z}_i))$. Denote the (constrained) MLE as $\widehat{\mathbf{\Theta}}_t := \arg\max_{\mathbf{\Theta} \in \mathrm{Skew}(d,2r;S)} \sum_{i \in [t]} \ell_i(\mathbf{\Theta})$, where $\ell_i(\mathbf{\Theta}) := (r_i \log P_{\mathbf{\Theta}}(\boldsymbol{z}_i) + (1-r_i)\log(1 - P_{\mathbf{\Theta}}(\boldsymbol{z}_i)))$. Suppose that $\ell_i(\cdot)$ is L-Lipschitz w.r.t. the Frobenius norm. Then, for any $\delta \in (0,1)$ the following holds:*

$$\mathbb{P}\left(\sum_{i=1}^{t} (P_{\widehat{\mathbf{\Theta}}_t}(\boldsymbol{z}_i) - P_{\mathbf{\Theta}_\star}(\boldsymbol{z}_i))^2 \lesssim \log \frac{T}{\delta} + dr \log LST \right) \geq 1 - \delta, \quad \forall t \in [T]. \tag{138}$$

With this lemma and our eluder dimension bound (Proposition F.2), the derivation of the regret bound in Wu et al. (2025) follows through, with $\log\left(\frac{N_{\mathcal{P}} T}{\delta}\right)$ and λ both replaced with $\log \frac{T}{\delta} + dr \log LST$. □

Converting to Unregularized Regret Bound. We now convert the KL-regularized regret bound to its unregularized counterpart via the following lemma:

Lemma H.4. *Suppose $D_{\mathrm{ref}} := \max_{\pi \in \Pi} D_{\mathrm{KL}}(\pi, \pi_{\mathrm{ref}}) < \infty$. Then we have*

$$\text{MBR-Reg}(T) \leq \text{MBR-Reg}_\eta(T) + \eta^{-1} D_{\mathrm{ref}} T. \tag{139}$$

Proof. Recall that the KL-regularized objective is defined as

$$J_\eta(\pi, \pi') = J(\pi, \pi') - \eta^{-1} D_{\mathrm{KL}}(\pi, \pi_{\mathrm{ref}}) + \eta^{-1} D_{\mathrm{KL}}(\pi', \pi_{\mathrm{ref}}) \tag{140}$$

Then,

$$\text{MBR-Reg}(T) = \sum_{t=1}^{T} \max_{\pi \in \Pi} \left(\frac{1}{2} - J(\hat{\pi}_t^1, \pi) \right) \tag{141}$$

$$= \sum_{t=1}^{T} \max_{\pi \in \Pi} \left(\frac{1}{2} - \left(J(\hat{\pi}_t^1, \pi) - \eta^{-1} D_{\mathrm{KL}}(\hat{\pi}_t^1, \pi_{\mathrm{ref}}) + \eta^{-1} D_{\mathrm{KL}}(\pi, \pi_{\mathrm{ref}}) \right) - \eta^{-1} D_{\mathrm{KL}}(\hat{\pi}_t^1, \pi_{\mathrm{ref}}) + \eta^{-1} D_{\mathrm{KL}}(\pi, \pi_{\mathrm{ref}}) \right) \tag{142}$$

$$\leq \sum_{t=1}^{T} \max_{\pi \in \Pi} \left(\frac{1}{2} - J_\eta(\hat{\pi}_t^1, \pi) \right) + \sum_{t=1}^{T} \max_{\pi \in \Pi} \left(-\eta^{-1} D_{\mathrm{KL}}(\hat{\pi}_t^1, \pi_{\mathrm{ref}}) + \eta^{-1} D_{\mathrm{KL}}(\pi, \pi_{\mathrm{ref}}) \right) \tag{143}$$

$$\leq \text{MBR-Reg}_\eta(T) + \sum_{t=1}^{T} \max_{\pi \in \Pi} \eta^{-1} D_{\mathrm{KL}}(\pi, \pi_{\mathrm{ref}}) \tag{144}$$

$$= \text{MBR-Reg}_\eta(T) + \eta^{-1} D_{\mathrm{ref}} T. \tag{145}$$

□

Putting everything together, we have the following corollary of Theorem H.2 for the unregularized regret bound:

Corollary H.5 (Unregularized Regret Bound of GS)**.** *Suppose* $D_{\mathrm{ref}} := \max_{\pi \in \Pi} D_{\mathrm{KL}}(\pi, \pi_{\mathrm{ref}}) < \infty$. *Then with probability at least* $1 - \delta$*, the unregularized regret of GS under* ***GBPM*** *satisfies*

$$\text{MBR-Reg}(T) \lesssim e^{\eta} d^2 \kappa^{-2} L_\mu^2 \cdot \log T \cdot \left(\log \frac{T}{\delta} + dr \log(LST) \right) + \eta^{-1} D_{\mathrm{ref}} T. \tag{146}$$

Furthermore, there exists no η *(even dependent on* T*) such that the RHS is* $\mathcal{O}(T^{1-\gamma})$ *for any* $\gamma \in (0, 1]$.

Proof. The regret bound is a direct result from the KL-regularized regret bound in Theorem H.2, converted to unregularized regret via Lemma H.4.

For the second claim, suppose that this is true. Then, for the second term, we require $\eta^{-1} D_{\mathrm{ref}} T = \mathcal{O}(T^{1-\gamma})$ to hold, which implies $\eta = \Omega(T^{\gamma})$. Plugging this into the first term leads to an additive term of $\mathcal{O}(e^{T^{\gamma}})$, which is superpolynomial: a contradiction. This concludes the proof. □

H.2. Proof of Lemma H.3: MLE Estimator Bound

We define the probability mass function (pmf) of $r \mid \boldsymbol{z} \sim \mathrm{Ber}(P(\boldsymbol{z}))$ as

$$P(r \mid \boldsymbol{z}) = P(\boldsymbol{z})^r (1 - P(\boldsymbol{z}))^{1-r}, \quad r \in \{0, 1\}. \tag{147}$$

Then we have the following lemma, whose proof is deferred to Appendix I:

Lemma H.6. *For each* $\boldsymbol{\Theta} \in \Theta$ *and* $t \in [T]$*, the following holds:*

$$\mathbb{P}\left(\sum_{i=1}^{t} (P_{\boldsymbol{\Theta}}(\boldsymbol{z}_i) - P_{\boldsymbol{\Theta}_\star}(\boldsymbol{z}_i))^2 \leq \log \frac{1}{\delta} + \sum_{i=1}^{t} \log \frac{P_{\boldsymbol{\Theta}_\star}(r_i \mid \boldsymbol{z}_i)}{P_{\boldsymbol{\Theta}}(r_i \mid \boldsymbol{z}_i)} \right) \geq 1 - \delta \tag{148}$$

Let Θ_ε be an ε-net of Θ in terms of the Frobenius norm. Then by the union bound, we have:

$$\mathbb{P}\left(\sum_{i=1}^{t} (P_{\boldsymbol{\Theta}}(\boldsymbol{z}_i) - P_{\boldsymbol{\Theta}_\star}(\boldsymbol{z}_i))^2 \leq \log \frac{|\Theta_\varepsilon|}{\delta} + \sum_{i=1}^{t} \log \frac{P_{\boldsymbol{\Theta}_\star}(r_i \mid \boldsymbol{z}_i)}{P_{\boldsymbol{\Theta}}(r_i \mid \boldsymbol{z}_i)}, \quad \forall \boldsymbol{\Theta} \in \Theta_\varepsilon \right) \geq 1 - \delta. \tag{149}$$

Let $\widehat{\boldsymbol{\Theta}}_{\varepsilon,t}$ be the epsilon net element corresponding to $\widehat{\boldsymbol{\Theta}}_t$, i.e., $\|\widehat{\boldsymbol{\Theta}}_t - \widehat{\boldsymbol{\Theta}}_{\varepsilon,t}\|_F \le \varepsilon$. Then,

$$\mathbb{P}\left(\sum_{i=1}^{t}(P_{\widehat{\boldsymbol{\Theta}}_{\varepsilon,t}}(\boldsymbol{z}_i) - P_{\boldsymbol{\Theta}_\star}(\boldsymbol{z}_i))^2 \le \log\frac{|\Theta_\varepsilon|}{\delta} + \sum_{i=1}^{t}\log\frac{P_{\boldsymbol{\Theta}_\star}(r_i \mid \boldsymbol{z}_i)}{P_{\widehat{\boldsymbol{\Theta}}_{\varepsilon,t}}(r_i \mid \boldsymbol{z}_i)}\right) \ge 1-\delta. \tag{150}$$

Using the inequality $(a-b)^2 \ge \frac{1}{2}(a-c)^2 - (b-c)^2$ and the optimality of the MLE, with probability at least $1-\delta$ the following holds:

$$\frac{1}{2}\sum_{i=1}^{t}(P_{\widehat{\boldsymbol{\Theta}}_t}(\boldsymbol{z}_i) - P_{\boldsymbol{\Theta}_\star}(\boldsymbol{z}_i))^2 - \sum_{i=1}^{t}(P_{\widehat{\boldsymbol{\Theta}}_t}(\boldsymbol{z}_i) - P_{\widehat{\boldsymbol{\Theta}}_{\varepsilon,t}}(\boldsymbol{z}_i))^2 \tag{151}$$

$$= \log\frac{|\Theta_\varepsilon|}{\delta} + \sum_{i=1}^{t}\log\frac{P_{\boldsymbol{\Theta}_\star}(r_i \mid \boldsymbol{z}_i)}{P_{\widehat{\boldsymbol{\Theta}}_t}(r_i \mid \boldsymbol{z}_i)} + \sum_{i=1}^{t}\log\frac{P_{\widehat{\boldsymbol{\Theta}}_t}(r_i \mid \boldsymbol{z}_i)}{P_{\widehat{\boldsymbol{\Theta}}_{\varepsilon,t}}(r_i \mid \boldsymbol{z}_i)} \tag{152}$$

$$\le \log\frac{|\Theta_\varepsilon|}{\delta} + \sum_{i=1}^{t}\log\frac{P_{\widehat{\boldsymbol{\Theta}}_t}(r_i \mid \boldsymbol{z}_i)}{P_{\widehat{\boldsymbol{\Theta}}_{\varepsilon,t}}(r_i \mid \boldsymbol{z}_i)}. \tag{153}$$

Now, note that for any $\boldsymbol{\Theta}$,

$$\log P_{\boldsymbol{\Theta}}(r_i \mid \boldsymbol{z}_i) = r_i \log P_{\boldsymbol{\Theta}}(\boldsymbol{z}_i) + (1-r_i)\log(1 - P_{\boldsymbol{\Theta}}(\boldsymbol{z}_i)), \tag{154}$$

which is L-Lipschitz in $\boldsymbol{\Theta}$ by given. With this, we can bound the log sum on the right as

$$\sum_{i=1}^{t}\log\frac{P_{\widehat{\boldsymbol{\Theta}}_t}(r_i \mid \boldsymbol{z}_i)}{P_{\widehat{\boldsymbol{\Theta}}_{\varepsilon,t}}(r_i \mid \boldsymbol{z}_i)} \le Lt\left\|\widehat{\boldsymbol{\Theta}}_{\varepsilon,t} - \widehat{\boldsymbol{\Theta}}_t\right\|_F \le Lt\varepsilon. \tag{155}$$

Since we assumed that $\log P_{\boldsymbol{\Theta}}$ is L-Lipschitz, it follows that $P_{\boldsymbol{\Theta}}$ is also L-Lipschitz.[6] Therefore,

$$\sum_{i=1}^{t}(P_{\widehat{\boldsymbol{\Theta}}_t}(\boldsymbol{z}_i) - P_{\widehat{\boldsymbol{\Theta}}_{\varepsilon,t}}(\boldsymbol{z}_i))^2 \le Lt\varepsilon^2. \tag{156}$$

We now bound the cardinality of the ε-net $|\Theta_\varepsilon|$ by bounding the covering number of a slightly larger set:

Lemma H.7 (Lemma 3.1 of Candès & Plan (2011)). *Let* $\Theta(d, 2r; S) \coloneqq \{\boldsymbol{X} \in \mathbb{R}^{d\times d} \mid \|\mathbf{X}\|_F \le S, \mathrm{rank}(\boldsymbol{X}) \le 2r\} \supseteq \mathrm{Skew}(d, 2r; S)$. *For any* $\varepsilon > 0$, *there exists an* ε*-net* Θ_ε *of* $\Theta(d, 2r; S)$ *w.r.t.* $\|\cdot\|_F$ *with* $|\Theta_\varepsilon| \le \left(\frac{9S}{\varepsilon}\right)^{2(2d+1)r}$.

Putting the bounds together, we have:

$$\mathbb{P}\left(\sum_{i=1}^{t}(P_{\widehat{\boldsymbol{\Theta}}_t}(\boldsymbol{z}_i) - P_{\boldsymbol{\Theta}_\star}(\boldsymbol{z}_i))^2 \lesssim \log\frac{1}{\delta} + Lt\varepsilon + Lt\varepsilon^2 + dr\log\frac{S}{\varepsilon}\right) \ge 1-\delta. \tag{157}$$

Choosing $\varepsilon \approx \frac{1}{(Lt)^2}$,

$$\mathbb{P}\left(\sum_{i=1}^{t}(P_{\widehat{\boldsymbol{\Theta}}_t}(\boldsymbol{z}_i) - P_{\boldsymbol{\Theta}_\star}(\boldsymbol{z}_i))^2 \lesssim \log\frac{1}{\delta} + dr\log(LSt)\right) \ge 1-\delta. \tag{158}$$

Setting $\delta = \delta/T$ and taking the union bound over t, we have:

$$\mathbb{P}\left(\forall t \in [T], \sum_{i=1}^{t}(P_{\widehat{\boldsymbol{\Theta}}_t}(\boldsymbol{z}_i) - P_{\boldsymbol{\Theta}_\star}(\boldsymbol{z}_i))^2 \lesssim \log\frac{T}{\delta} + dr\log(LST)\right) \ge 1-\delta. \tag{159}$$

which concludes the proof. □

[6] As the Lipschitz constant is the maximum gradient norm by the Rademacher's theorem, $\|\nabla_{\boldsymbol{\Theta}} P_{\boldsymbol{\Theta}}\| = P_{\boldsymbol{\Theta}} \cdot \|\nabla_{\boldsymbol{\Theta}} \log P_{\boldsymbol{\Theta}}\| \le L$.

I. Proof of Lemma H.6

The proof closely follows that of Ye et al. (2024, Lemma 1) and Wu et al. (2025, Lemma 3).

For the function $P_{\boldsymbol{\Theta}}$ defined by the fixed $\boldsymbol{\Theta} \in \Theta$, we first upper bound its logarithmic moment generating function as

$$\log \mathbb{E} \exp\left(\sum_{i=1}^{t} \log \frac{P_{\boldsymbol{\Theta}}(r_i \mid \boldsymbol{z}_i)}{P_{\boldsymbol{\Theta}_\star}(r_i \mid \boldsymbol{z}_i)}\right) \tag{160}$$

$$= \log \mathbb{E} \exp\left(\sum_{i=1}^{t-1} \log \frac{P_{\boldsymbol{\Theta}}(r_i \mid \boldsymbol{z}_i)}{P_{\boldsymbol{\Theta}_\star}(r_i \mid \boldsymbol{z}_i)} + \log\left(2\, \mathbb{E}_{r_t \mid \boldsymbol{z}_t} \sqrt{\frac{P_{\boldsymbol{\Theta}}(r_t \mid \boldsymbol{z}_t)}{P_{\boldsymbol{\Theta}_\star}(r_t \mid \boldsymbol{z}_t)}}\right)\right) \tag{161}$$

$$= \log \mathbb{E} \exp\left(\sum_{i=1}^{t-1} \log \frac{P_{\boldsymbol{\Theta}}(r_i \mid \boldsymbol{z}_i)}{P_{\boldsymbol{\Theta}}(r_i \mid \boldsymbol{z}_i)} + \log\left(1 - H\left(P_{\boldsymbol{\Theta}}(r_t \mid \boldsymbol{z}_t) \,\|\, P_{\boldsymbol{\Theta}_\star}(r_t \mid \boldsymbol{z}_t)\right)^2\right)\right) \tag{162}$$

$$\leq \log \mathbb{E} \exp\left(\sum_{i=1}^{t-1} \log \frac{P_{\boldsymbol{\Theta}}(r_i \mid \boldsymbol{z}_i)}{P_{\boldsymbol{\Theta}_\star}(r_i \mid \boldsymbol{z}_i)} - H\left(P_{\boldsymbol{\Theta}}(r_t \mid \boldsymbol{z}_t) \,\|\, P_{\boldsymbol{\Theta}_\star}(r_t \mid \boldsymbol{z}_t)\right)^2\right) \tag{163}$$

$$\leq \cdots \leq -\sum_{i=1}^{t} H\left(P_{\boldsymbol{\Theta}}(r_i \mid \boldsymbol{z}_i) \,\|\, P_{\boldsymbol{\Theta}_\star}(r_i \mid \boldsymbol{z}_i)\right)^2, \tag{164}$$

where $H(P\|Q)^2$ is the squared Hellinger distance between probability measures P and Q on Ω, defined as

$$H(P\|Q)^2 := \int_{\Omega} \left(\sqrt{p(z)} - \sqrt{q(z)}\right)^2 d\mu(z), \tag{165}$$

with p and q denoting their respective densities with respect to a base measure μ.

We continue to lower-bound the Hellinger distance by

$$\sum_{i=1}^{t} \left(H\left(P_{\boldsymbol{\Theta}}(r_i \mid \boldsymbol{z}_i) \,\|\, P_{\boldsymbol{\Theta}_\star}(r_i \mid \boldsymbol{z}_i)\right)\right)^2 \geq \sum_{i=1}^{t} \left(\mathrm{TV}\left(P_{\boldsymbol{\Theta}}(r_i \mid \boldsymbol{z}_i) \,\|\, P_{\boldsymbol{\Theta}_\star}(r_i \mid \boldsymbol{z}_i)\right)\right)^2 \tag{166}$$

$$= \sum_{i=1}^{t} \left(P_{\boldsymbol{\Theta}}(\boldsymbol{z}_i) - P_{\boldsymbol{\Theta}_\star}(\boldsymbol{z}_i)\right)^2, \tag{167}$$

where the inequality uses the fact that for any distribution p, q, $H(p, q) \geq \mathrm{TV}(p, q)$ (Zhang, 2023, Theorem B.9).

Then, by Lemma J.1, we obtain for each $\boldsymbol{\Theta} \in \Theta$, with probability at least $1 - \delta$,

$$\sum_{i=1}^{t} \log \frac{P_{\boldsymbol{\Theta}}(r_i \mid \boldsymbol{z}_i)}{P_{\boldsymbol{\Theta}_\star}(r_i \mid \boldsymbol{z}_i)} \leq \log\left(\frac{1}{\delta}\right) + \log \mathbb{E} \exp\left(\sum_{i=1}^{t} \log \frac{P_{\boldsymbol{\Theta}}(r_i \mid \boldsymbol{z}_i)}{P_{\boldsymbol{\Theta}_\star}(r_i \mid \boldsymbol{z}_i)}\right) \tag{168}$$

$$\leq -\sum_{i=1}^{t} H\big(P_{\boldsymbol{\Theta}}(r_i \mid \boldsymbol{z}_i) \,\|\, P_{\boldsymbol{\Theta}_\star}(r_i \mid \boldsymbol{z}_i)\big)^2 + \log\left(\frac{1}{\delta}\right) \tag{169}$$

$$\leq -\sum_{i=1}^{t} \left(P_{\boldsymbol{\Theta}}(\boldsymbol{z}_i) - P_{\boldsymbol{\Theta}_\star}(\boldsymbol{z}_i)\right)^2 + \log\left(\frac{1}{\delta}\right). \tag{170}$$

J. Auxiliary Lemmas

Lemma J.1 (Martingale Exponential Inequalities; Theorem 13.2 of Zhang (2023)). *Consider a sequence of random functions $\xi_1(\mathcal{Z}_1), \ldots, \xi_t(\mathcal{Z}_t), \ldots$ with respect to filtration $\{\mathcal{F}_t\}$. We have for any $\delta \in (0,1)$ and $\lambda > 0$:*

$$\mathbb{P}\left(\exists n > 0 : -\sum_{i=1}^{n} \xi_i \geq \frac{\log(1/\delta)}{\lambda} + \frac{1}{\lambda}\sum_{i=1}^{n} \log \mathbb{E}_{Z_i^{(y)}} \exp(-\lambda \xi_i)\right) \leq \delta,$$

where $Z_t = (Z_t^{(x)}, Z_t^{(y)})$ and $\mathcal{Z}_t = (Z_1, \ldots, Z_t)$.

Lemma J.2 (Multiplicative Chernoff Bounds; Corollary 2.18 of Zhang (2023)). *Assume that $X \in [0,1]$ with $\mathbb{E}X = \mu$. Then for all $\epsilon > 0$,*

$$\mathbb{P}\left(\bar{X}_n \geq (1+\epsilon)\mu\right) \leq \exp\left[\frac{-2n\mu\epsilon^2}{2+\epsilon}\right] \tag{171}$$

$$\mathbb{P}\left(\bar{X}_n \leq (1-\epsilon)\mu\right) \leq \exp\left[\frac{-2n\mu\epsilon^2}{2}\right]. \tag{172}$$

Moreover, for $t > 0$, we have

$$\mathbb{P}\left(\bar{X}_n \geq \mu + \sqrt{\frac{2\mu t}{n}} + \frac{t}{3n}\right) \leq \exp(-t).$$

Lemma J.3 (Elliptical Potential Lemma; Lemma 11 of Abbasi-Yadkori et al. (2011)). *Let $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_T \in \mathcal{B}^d(X)$ be a sequence of vectors and $\boldsymbol{V}_t := \lambda \boldsymbol{I} + \sum_{s=1}^{t-1} \boldsymbol{x}_s \boldsymbol{x}_s^\intercal$. Then, we have*

$$\sum_{t=1}^{T} \min\left\{1, \|\boldsymbol{x}_t\|_{\boldsymbol{V}_t^{-1}}^2\right\} \leq 2d \log\left(1 + \frac{X^2 T}{d\lambda}\right). \tag{173}$$

K. Discussions on Regrets

K.1. Four Regret Definitions and Discussions

A standard measure of performance in online learning is *regret*. However, because the interaction is two-player and self-play, there are several ways to define regret, arising from different yet closely related communities: no-regret learning in games, dueling bandits/RL, and RL in two-player zero-sum games. In the main text, we only consider $\mathrm{MBR\text{-}Reg}_\eta$, and so in this Appendix, we provide the deferred discussions regarding four regrets:

Definition K.1. *The four regrets are defined as follows.*

(a) ***Average-Nash Regret****:*

$$\mathrm{AN\text{-}Reg}_\eta(T) := \max_{\pi^1,\pi^2\in\Pi} \sum_{t=1}^{T} \left\{ J_\eta(\pi^1, \hat{\pi}_t^2) - J_\eta(\hat{\pi}_t^1, \pi^2) \right\}. \tag{174}$$

(b) ***Average-Best-Response Regret****:*

$$\mathrm{ABR\text{-}Reg}_\eta(T) := \sum_{t=1}^{T} \max_{\pi^1,\pi^2\in\Pi} \left\{ J_\eta(\pi^1, \hat{\pi}_t^2) - J_\eta(\hat{\pi}_t^1, \pi^2) \right\}. \tag{175}$$

(c) ***Max-Nash Regret****:*

$$\mathrm{MN\text{-}Reg}_\eta(T) := \max_{\pi\in\Pi} \sum_{t=1}^{T} \left\{ \frac{1}{2} - J_\eta(\hat{\pi}_t^1, \pi) \right\}. \tag{176}$$

(d) ***Max-Best-Response Regret****:*

$$\mathrm{MBR\text{-}Reg}_\eta(T) := \sum_{t=1}^{T} \max_{\pi\in\Pi} \left\{ \frac{1}{2} - J_\eta(\hat{\pi}_t^1, \pi) \right\}. \tag{177}$$

The unregularized variants are defined similarly and denoted without the η subscript.

Categorization Criteria. There are two criteria that determine each regret definition. The first criterion is, at each time t, whether to consider the regrets of both players simultaneously, or to consider the regret of the max-player only. This distinguishes between *Average* or *Max*. The second criterion is whether to compare against a fixed comparator or to compare against the best response at each time t, which is usually time-varying. This distinguishes between *Nash* and *Best-Response*.

Intuitively, the *Average* regret definitions consider the "suboptimality" of both policies simultaneously. The difference between *Nash* and *Best-Response* is whether the regret is defined w.r.t. a fixed comparator (*Nash*) or a dynamically changing comparator (*Best-Response*). Thus, in classical literature, they are also known as *external* and *internal (swap)* regrets.

Average Regrets. $\mathrm{AN\text{-}Reg}(T)$ is the notion originally considered in the seminal work of Freund & Schapire (1999), followed by numerous works on no-regret learning dynamics in games (Daskalakis et al., 2011; 2015; 2018; Rakhlin & Sridharan, 2013a;b; Syrgkanis et al., 2015), recently adopted to game-theoretic LLM alignment (Zhang et al., 2025a). $\mathrm{AN\text{-}Reg}(T)$ is precisely the regret considered in contextual dueling bandits (Dudík et al., 2015, Eqn. (3)) and dueling RL (Saha et al., 2023, Eqn. (4)); indeed, for any $\pi \in \Pi$, utilizing the anti-symmetry of $J(\cdot,\cdot)$, we can rewrite $J(\pi, \hat{\pi}_t^2) - J(\hat{\pi}_t^1, \pi) = J(\pi, \hat{\pi}_t^1) + J(\pi, \hat{\pi}_t^2) - 1$. This also slightly resembles Borda regret in dueling bandits (Saha et al., 2021; Wu et al., 2024), and average regret in dueling bandits under linear stochastic transitivity (Saha, 2021; Bengs et al., 2021; 2022).

On the other hand, $\mathrm{ABR\text{-}Reg}(T)$ strongly resembles the notion of *best response regret* in adversarial dueling bandits (Saha & Krishnamurthy, 2022, Eqn. (1)), but there is a key difference. In Saha & Krishnamurthy (2022), the comparator at time t is the same for both players, whereas in our regret setting, it differs for each player. Basically, each player must compete with the worst-case (strongest) adversary from her perspective, who chooses the best response from his perspective.

Max Regrets. The notion of considering the regret of the *max* player dates back to the self-play framework for RL in two-player zero-sum games (Bai & Jin, 2020; Bai et al., 2020; Liu et al., 2021; Jin et al., 2022; Xiong et al., 2022); this idea has been recently applied to theoretical analyses of online RLHF under general preference (Ye et al., 2024; Wu et al., 2025). Basically, the intuition is that the learner only cares about obtaining the NE policy for the max-player, which is the policy that is actually deployed in practice.

Note that the min-player's policies $\hat{\pi}_t^2$ do not contribute to the regret at all, and thus, often, the min-player acts as an exploration agent whose sole role is to collect as much information as possible to facilitate the learning of the max-player (Bai & Jin, 2020; Bai et al., 2020; Liu et al., 2021; Jin et al., 2022; Xiong et al., 2022; 2024; Ye et al., 2024).

K.2. Online-to-Batch Conversion

A standard consequence of no-regret learning in repeated zero-sum games is an *online-to-batch conversion*: the time-averaged (mixed) policies form an approximate Nash equilibrium. We formalize this statement in the following proposition.

Proposition K.2 (Online-to-batch conversion)**.** *Let $\{(\hat{\pi}_t^1, \hat{\pi}_t^2)\}_{t=1}^T \subseteq \Pi \times \Pi$ be any policy sequence. Define the uniform mixture policies as $\bar{\pi}_T^i := \frac{1}{T}\sum_{t=1}^T \hat{\pi}_t^i$ for $i \in \{1,2\}$ and for simplicity, let us denote $\bar{\pi}_T := \bar{\pi}_T^1$.*

(a) Average regrets. *For average regrets, $(\bar{\pi}_T^1, \bar{\pi}_T^2)$ is a $\frac{\mathrm{Reg}(T)}{T}$-approximate symmetric NE:*

$$\max_{\pi^1,\pi^2\in\Pi}\Big\{J_\eta(\pi^1,\bar{\pi}_T^2) - J_\eta(\bar{\pi}_T^1,\pi^2)\Big\} \;\le\; \frac{\text{AN-Reg}_\eta(T)}{T} \;\le\; \frac{\text{ABR-Reg}_\eta(T)}{T}.$$

(b) Max regrets. *For max regrets, $\bar{\pi}_T$ is a $\frac{2\mathrm{Reg}(T)}{T}$-approximate symmetric NE:*

$$\max_{\pi\in\Pi}\Big\{J_\eta(\pi,\bar{\pi}_T) - J_\eta(\bar{\pi}_T,\pi)\Big\} \;\le\; \frac{\text{2MN-Reg}_\eta(T)}{T} \;\le\; \frac{\text{2MBR-Reg}_\eta(T)}{T}.$$

Proof. **(a)** Fix any $\pi^1, \pi^2 \in \Pi$. By the bilinearity of J and Jensen's inequality w.r.t. $\psi(\cdot)$,

$$J_\eta(\pi^1,\bar{\pi}_T^2) = J(\pi^1,\bar{\pi}_T^2) - \eta^{-1}\psi(\pi^1) + \eta^{-1}\psi(\bar{\pi}_T^2) \le \frac{1}{T}\sum_{t=1}^T J_\eta(\pi^1,\hat{\pi}_t^2),$$

and similarly,

$$J_\eta(\bar{\pi}_T^1,\pi^2) = J(\bar{\pi}_T^1,\pi^2) - \eta^{-1}\psi(\bar{\pi}_T^1) + \eta^{-1}\psi(\pi^2) \ge \frac{1}{T}\sum_{t=1}^T J_\eta(\hat{\pi}_t^1,\pi^2).$$

Subtracting the two inequalities and taking $\max_{\pi^1,\pi^2}$ yields

$$\max_{\pi^1,\pi^2}\Big\{J_\eta(\pi^1,\bar{\pi}_T^2) - J_\eta(\bar{\pi}_T^1,\pi^2)\Big\} \le \frac{1}{T}\max_{\pi^1,\pi^2}\sum_{t=1}^T\Big(J_\eta(\pi^1,\hat{\pi}_t^2) - J_\eta(\hat{\pi}_t^1,\pi^2)\Big) \tag{178}$$

$$\le \frac{1}{T}\sum_{t=1}^T\max_{\pi^1,\pi^2}\Big(J_\eta(\pi^1,\hat{\pi}_t^2) - J_\eta(\hat{\pi}_t^1,\pi^2)\Big) \tag{179}$$

(b) By the same averaging argument (bilinearity of J and Jensen's inequality), for every $\pi \in \Pi$,

$$J_\eta(\pi,\bar{\pi}_T) - J_\eta(\bar{\pi}_T,\pi) \le \frac{1}{T}\sum_{t=1}^T\Big(J_\eta(\pi,\hat{\pi}_t^1) - J_\eta(\hat{\pi}_t^1,\pi)\Big).$$

Since for each t, $J_\eta(\pi,\hat{\pi}_t^1) - J_\eta(\hat{\pi}_t^1,\pi) = 2\Big(\frac{1}{2} - J_\eta(\hat{\pi}_t^1,\pi)\Big)$, taking $\max_\pi$ and substituting yields

$$\max_\pi\big\{J_\eta(\pi,\bar{\pi}_T) - J_\eta(\bar{\pi}_T,\pi)\big\} \le \frac{2}{T}\max_\pi\sum_{t=1}^T\Big(\tfrac{1}{2} - J_\eta(\hat{\pi}_t^1,\pi)\Big) \le \frac{2}{T}\sum_{t=1}^T\max_\pi\Big(\tfrac{1}{2} - J_\eta(\hat{\pi}_t^1,\pi)\Big). \qquad \square$$

L. Synthetic Experiments

In this section, we present empirical results to numerically validate the theoretical regret bounds of `Greedy Sampling (GS)` established in Section 4.2.

L.1. Experiment Setup.

We consider a bilinear preference model with logistic link function $\mu(z) = 1/(1 + e^{-z})$ and reverse KL regularizer. We use K feature vectors in an uncontextualized setting, randomly sampled via a uniform distribution, and normalized to ℓ_2 norm ≤ 1.

We use the following hyperparameters for our experiments:

- $d = 5,\ K = 20,\ S = 5$
- $d = 10,\ K = 40,\ S = 10$
- $r = 1$
- $\eta \in \{10^{-2}, \cdots, 10^{4}\}$
- $T = 10000$

All reported metrics are averaged over 20 independent random seeds, with standard deviations shown as shaded regions or error bars.

L.2. Implementation Details and Reproducibility

To ensure reproducibility and bridge the gap between continuous bounds and discrete floating-point arithmetic, we detail our experimental setup and numerical stabilizations (source code provided in the supplement).

Instance Generation. To ensure that the true preference matrix $\mathbf{\Theta}_\star$ rigorously satisfies the theoretical assumptions of being low-rank, exactly skew-symmetric ($\mathbf{\Theta}_\star + \mathbf{\Theta}_\star^\top = 0$), and bounded in norm, we construct it using its real spectral decomposition. We first draw a Gaussian matrix $\mathbf{G} \in \mathbb{R}^{d\times 2r}$ and compute its QR decomposition to obtain an orthonormal basis matrix $\mathbf{Q} \in \mathbb{R}^{d\times 2r}$. We then construct a block-diagonal skew-symmetric core matrix $\mathbf{B} \in \mathbb{R}^{2r\times 2r}$ consisting of r independent 2×2 blocks of the form $\left[\begin{smallmatrix} 0 & s_i \\ -s_i & 0 \end{smallmatrix}\right]$, where the singular values s_i are sampled uniformly from $[0.1, 1.0]$. The unnormalized parameter matrix is assembled as $\mathbf{\Theta} = \mathbf{Q}\mathbf{B}\mathbf{Q}^\top$, guaranteeing an exact rank of $2r$. We apply a minor anti-symmetrization $\frac{1}{2}(\mathbf{\Theta} - \mathbf{\Theta}^\top)$ solely to correct floating-point inaccuracies, and scale the matrix such that its Frobenius norm is exactly $\|\mathbf{\Theta}_\star\|_F = S$.

Equiliibrium and Estimation Procedures. Both the base exploration policy π_0 and the reference policy π_{ref} are initialized as the uniform distribution over the K available items ($\pi_0 = \pi_{\text{ref}} = \mathbf{1}/K$). Since our experiments focus on the reverse-KL regularizer, the regularized game admits exact log-ratio coordinates, which allows us to compute equilibria utilizing SciPy's `root` function with the `hybr` method. Computing fixed points and log-likelihoods in extreme regimes ($\eta \gg 1$) requires specific numerical heuristics. Directly solving the fixed-point equation $p = \text{BR}_\eta(p)$—where BR_η denotes the regularized best response operator against an opponent policy—is highly unstable in this low-temperature limit due to exponential sensitivity. To address this, we employ an η-continuation (homotopy) method, sequentially solving the fixed point via adaptively damped Mann iterations.

Numerical Stability and Reproducibility. We use several numerical safeguards to preserve stability. For preference estimation, we utilize the Online Newton Step (ONS) algorithm to sequentially update the estimated parameter matrix, alongside offline Maximum Likelihood Estimation (MLE). During these updates, inner products are clipped to $[-50, 50]$ to prevent exponential overflow in the link function, and the ONS logistic variance term (which acts as the Hessian approximation) is strictly bounded to $[10^{-6}, 0.25]$ to prevent inverse covariance degeneracy.

All experiments were executed on standard consumer-grade CPU hardware without the need for hardware accelerators. The complete evaluation suite, including the 20 independent random seeds across all dimension and regularization configura-

tions, executes to completion within a few hours. Full code can be found in https://github.com/minju-hong/online_rlhf_gbpm.git.

L.3. Main Results

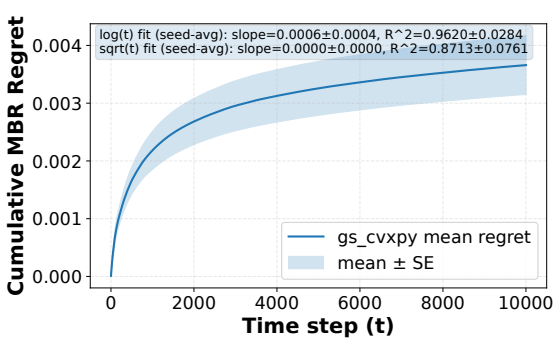
We first plot the cumulative MBR-Reg for varying regularization strengths η. As illustrated in Figure 1, the regret tightly fits a $\log T$ curve for small η and shifts to a $\sqrt{T}$ curve for large η, corroborating the bounds established in Theorem 4.2.

To quantify this phase transition, we fit the empirical regret to both logarithmic and square-root models and plot the goodness-of-fit (R^2) in Figure 2. Equating the two upper bounds in Theorem 4.2 gives the theoretical crossover

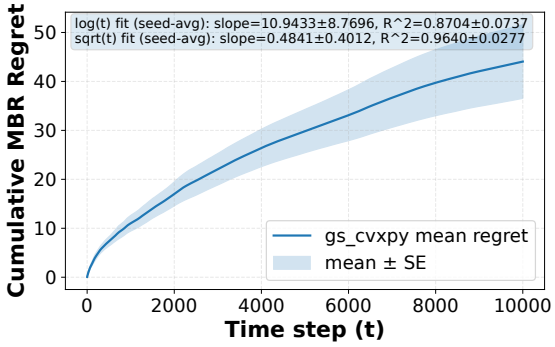
$$\eta_{\text{cross}}(d,T) = \frac{\kappa^{1/2} C_{\min}^{1/2}}{\beta} \frac{\sqrt{T}}{d^2 \log(T/d)}.$$

Thus, for $\eta \gtrsim \eta_{\text{cross}}(d,T)$, the $\sqrt{T}$ term is selected by the minimum, whereas for $\eta \lesssim \eta_{\text{cross}}(d,T)$, the logarithmic term is selected. Ignoring constants and the mild logarithmic dependence, this threshold scales as $d^{-2}\sqrt{T}$ when $C_{\min}$ is held fixed. Accordingly, increasing the dimension from $d = 5$ to $d = 10$ shifts the predicted crossover toward smaller η, up to the dependence of $C_{\min}$ on d.

Finally, we show the cumulative regret at $T = 10^4$ scaling with respect to η (Figure 3). The regret initially grows linearly but strictly plateaus in the large-η regime, perfectly matching the behavior of our unified $\min(\cdot,\cdot)$ bound.

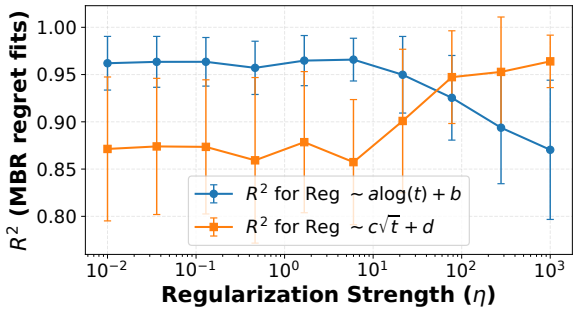


(a) Small $\eta = 0.01$: $\log T$ regime

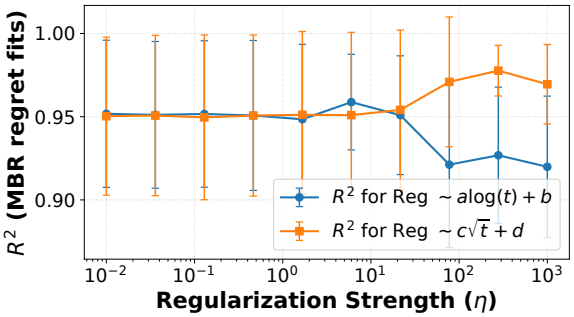


(b) Large $\eta = 1000$: $\sqrt{T}$ regime

Figure 1. MBR Regret Trajectories.



(a) $d = 5$



(b) $d = 10$

Figure 2. Crossover Point Analysis.

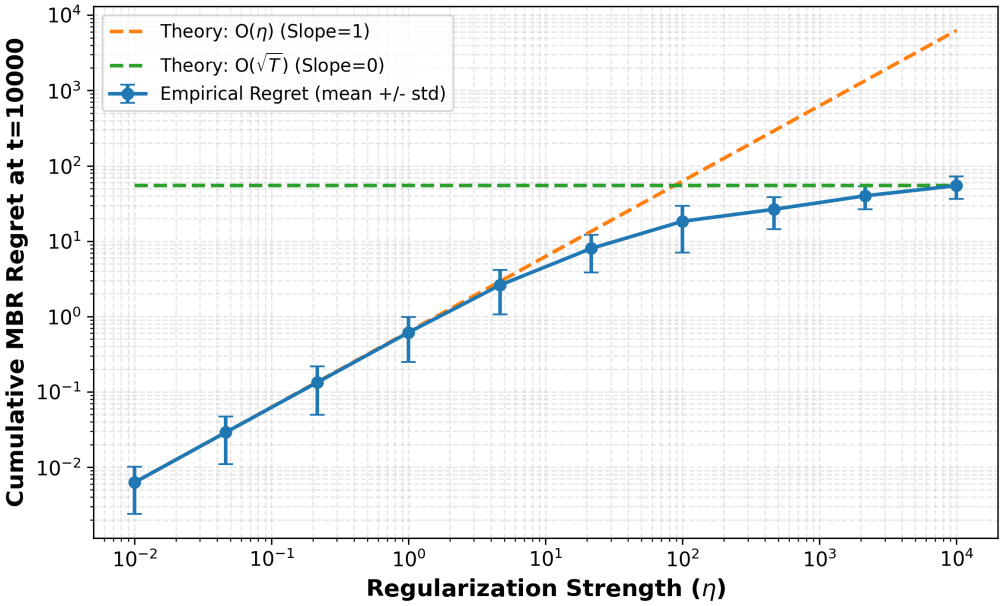


Figure 3. Final Regret Scaling.

M. Future Directions

Relaxing the Feature Diversity Assumption. Our current regret bounds rely on the feature diversity assumption (Assumption 4.1), characterized by the minimum eigenvalue $C_{\min}$. While this assumption is standard in the contextual bandits literature that involves either greedy sampling (e.g., algorithms without sophisticated exploration strategies) or high dimensions, it may still be restrictive for general RLHF applications. Recent works have investigated minimal assumptions required for greedy strategies, such as the local anti-concentration (LAC) property proposed by Kim & Oh (2024). Investigating the impact of such relaxed conditions on online RLHF with ***GBPM*** (e.g., whether we can still obtain polylogarithmic regret with `GS`) remains an important open question.

Instance-Specific Guarantees for Unregularized Regret. While our work establishes $\tilde{\mathcal{O}}(\sqrt{T})$ guarantees for *unregularized* regret (via Theorem 4.2), these bounds reflect worst-case hardness. For instance, Ito et al. (2025) demonstrated that in tabular games with bandit feedback, the Nash regret for the Tsallis-INF algorithm (Tsallis, 1988; Abernethy et al., 2015; Zimmert & Seldin, 2021) scales with the "sparsity" or "entropy" of the NE set, potentially achieving logarithmic regret $\mathcal{O}(\log T)$ when the NE is unique and deterministic (a pure strategy), and even rates of the form $\mathcal{O}(T^c)$ for some $c \in (0, 1)$, depending on the geometry of the set of Nash Equilibria. Adapting such instance-dependent guarantees to the contextual ***GBPM*** setting is non-trivial, even when the link function μ is linear. Recent advances in "Best-of-Both-Worlds" algorithms for linear contextual bandits (Kuroki et al., 2024; Kato & Ito, 2025) may provide a promising starting point.

Computationally Efficient Algorithms. Our current theoretical framework assumes access to a computational oracle for finding the NE (Oracle 3), which may be computationally expensive in practice. Developing efficient variants of our algorithms is a practical priority. Promising approaches include leveraging online estimation techniques such as Online Mirror Descent (OMD) (Zhang et al., 2025c), minimax optimization techniques such as optimistic OMD (Rakhlin & Sridharan, 2013a; Syrgkanis et al., 2015; Zhang et al., 2025a), or reductions to offline/online regression oracles (Foster & Rakhlin, 2020).