Provably Efficient Regularized Online RLHF with Generalized Bilinear Preferences

Junghyun Lee
KAIST AI
jh_lee00@kaist.ac.kr

Minju Hong
KAIST EE
minju23@kaist.ac.kr

Kwang-Sung Jun
POSTECH GSAI/CSE
kwangsungjun@postech.ac.kr

Chulhee Yun
KAIST AI
chulhee.yun@kaist.ac.kr

Se-Young Yun
KAIST AI
yunseyoung@kaist.ac.kr

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 regularization. We prove that for any preference-model estimator and strongly convex regularizer, the dual gap of its greedy Nash policy is bounded by the *squared* estimation error, using only strong convexity and preference anti-symmetry. We instantiate this general result under the *Generalized Bilinear Preference Model (GBPM)*, capturing intransitive preferences over d-dimensional item-wise features via a rank-$2r$ skew-symmetric matrix. Under suitable feature-coverage conditions with shared coverage constant $C_{\min} > 0$, `Greedy Sampling` achieves $\widetilde{\mathcal{O}}(\eta\beta\kappa^{-1}d^4C_{\min}^{-1} \wedge \kappa^{-1/2}d^2C_{\min}^{-1/2}\sqrt{T})$ regret, while `Explore-Then-Commit` achieves $\widetilde{\mathcal{O}}(\kappa^{-1}C_{\min}^{-2}\sqrt{\eta\beta rT} \wedge \kappa^{-2/3}C_{\min}^{-4/3}r^{1/3}T^{2/3})$ in admissible regimes, utilizing low-rank structure. These results demonstrate that "fast" regrets are *not* KL-specific, but rather a fundamental consequence of generic strongly convex geometry of the regularizer.

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; OpenAI, 2024; Qwen Team, 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 and 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 (McKelvey and Palfrey, 1995; Nash, 1951; von Neumann, 1928). Recently, this game-theoretic perspective has demonstrated notable empirical promise in LLM alignment (Cui et al., 2024; Munos et al., 2024; Rosset et al., 2024; Ye 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 (Casper et al., 2023; Michaud et al., 2020; Tien 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 and Palfrey (1995).

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

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; Nayak et al., 2026; Wu et al., 2025; Xiong et al., 2024; Yang et al., 2025; Ye et al., 2024). In particular, our analysis focuses on ***regularized max-regret***, which measures suboptimality of the *max* player only (see Section 2.2 for its definition). This notion originates from self-play frameworks for RL in two-player zero-sum games (Bai and Jin, 2020; Bai et al., 2020; Jin et al., 2022; Liu et al., 2021; Xiong et al., 2022), and has become the standard metric in theoretical analyses of online RLHF under general preferences (Wu et al., 2025; Ye et al., 2024). The intuition is that the learner ultimately only cares about obtaining the NE policy for the max-player, the policy actually deployed in practice.

The KL-Centric Theories. The current theoretical landscape for establishing polylogarithmic regularized best-response regret, often referred to as *fast rates* in the learning literature (Hazan et al., 2007; van Erven et al., 2015) 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 (Ji et al., 2026; Wu et al., 2025; Xiong et al., 2024; Zhao et al., 2025a,b). Consequently, the corresponding literature on regularized best-response regret has followed this same KL-specific trajectory (Nayak et al., 2026; Wu 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 (Haarnoja et al., 2018; Neu et al., 2017; Ziebart et al., 2008); mixtures of reverse KL divergences are used to yield more diverse and less biased alignment (Aminian et al., 2025; Le et al., 2025); the chi-squared divergence is known to provably mitigate overoptimization (Huang et al., 2025); and Tsallis entropy can encourage sparse policy selection (Chow et al., 2018; Lee et al., 2018). Broader classes of regularizers – such as α-Rényi entropies (Zhang et al., 2026a), Csiszár f-divergences (Go et al., 2023; Han et al., 2025; Wang et al., 2024), and expected or strongly convex regularizers (Geist et al., 2019; Yang 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 (Nayak et al., 2026; Wu et al., 2025; Ye et al., 2024), 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 and Terry, 1952; Plackett, 1975). To this end, we adopt the ***Generalized Bilinear Preference Model (GBPM)*** (Lee et al., 2026; Zhang et al., 2025a): 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 suitable feature coverage assumptions with a feature-coverage parameter $C_{\min} > 0$ **(Section 4.2)**. First, we prove that `Greedy Sampling (GS)` achieves *polylogarithmic* regrets of $\widetilde{O}\left(\eta\beta\kappa^{-1}d^4C_{\min}^{-1} \wedge \kappa^{-1/2}d^2C_{\min}^{-1/2}\sqrt{T}\right)$ **(Section 4.3)**. Second, to address the high-dimensional regime, we demonstrate that `Explore-Then-Commit (ETC)` with nuclear-norm regularized MLE

exploits low-rank structure to achieve $\widetilde{O}\left(\kappa^{-1}C_{\min}^{-2}\sqrt{\eta\beta rT} \wedge \kappa^{-2/3}C_{\min}^{-4/3}r^{1/3}T^{2/3}\right)$ in admissible regimes (**Section 4.4**).

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 $\boldsymbol{x}$. Let Π denote the set of all measurable policies $\pi : \mathcal{X} \to \Delta(\mathcal{A})$. 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 Preference, Regularized Game, and Nash Equilibrium

A *preference model* $P : \mathcal{A} \times \mathcal{A} \times \mathcal{X} \to [0, 1]$ is a function that satisfies $P(\boldsymbol{a}^1 \succ \boldsymbol{a}^2 \mid \boldsymbol{x}) + P(\boldsymbol{a}^2 \succ \boldsymbol{a}^1 \mid \boldsymbol{x}) = 1$; $P(\boldsymbol{a}^1 \succ \boldsymbol{a}^2 \mid \boldsymbol{x})$ is the probability that item $\boldsymbol{a}^1$ is preferred over $\boldsymbol{a}^2$ under context $\boldsymbol{x}$. We extend this action-level preference to evaluate policies as follows. The 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 $P(\pi^1 \succ \pi^2 \mid \boldsymbol{x}) := \mathbb{E}_{\boldsymbol{a}^1 \sim \pi^1(\cdot|\boldsymbol{x}),\, \boldsymbol{a}^2 \sim \pi^2(\cdot|\boldsymbol{x})}\left[P(\boldsymbol{a}^1 \succ \boldsymbol{a}^2 \mid \boldsymbol{x})\right]$. We refer to $J(\pi^1, \pi^2; P) := \mathbb{E}_{\boldsymbol{x} \sim d_0}[P(\pi^1 \succ \pi^2 \mid \boldsymbol{x})]$ as the population game objective corresponding to P.

We now introduce its regularized counterpart. For $\eta \in (0, \infty]$ and a β^{-1}-strongly convex regularizer $\psi : \Delta(\mathcal{A}) \to \mathbb{R}_{\geq 0}$ with respect to $\|\cdot\|_1$, we define a *symmetric*, regularized game objective $J_\eta : \Pi \times \Pi \to \mathbb{R}$ as follows:

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

Standard solution concepts (e.g., Condorcet winners) may not exist in general preference learning (Bengs et al., 2021; Dudík et al., 2015; 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)* (McKelvey and Palfrey, 1995; Nash, 1951):

Definition 2.1. *A pair $(\pi^1_\star, \pi^2_\star) \in \Pi \times \Pi$ is a **Nash equilibrium (NE)** if for all $\pi^1, \pi^2 \in \Pi$:*

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

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

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

We assume throughout that the population objectives are finite and that Nash equilibria and best responses exist for all preference models and $\eta \in (0, \infty]$ considered below. These conditions hold, for example, for finite $\mathcal{A}$, continuous ψ, and preference models measurable in $\boldsymbol{x}$. Game symmetry guarantees an SNE; for $0 < \eta < \infty$, strong concavity–convexity implies d_0-a.e. uniqueness.

2.2 Online Interaction Protocol and Regularized Max-Regret

Suppose that P^* is a fixed, ground-truth preference model. We omit the preference model argument in J and J_η when it is P^*. 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***:

$$\text{MBR-Reg}_\eta(T) := \sum_{t=1}^{T} \text{DGap}_\eta(\hat{\pi}^1_t), \quad \text{DGap}_\eta(\hat{\pi}^1_t) := \frac{1}{2} - \min_{\pi^2 \in \Pi} J_\eta(\hat{\pi}^1_t, \pi^2),$$

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

Remark 2.2. *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 and Jin, 2020; Bai et al., 2020; Jin et al., 2022; Liu et al., 2021; Xiong et al., 2022) and theoretical analyses of online RLHF under general preferences (Wu et al., 2025; Ye et al., 2024). Notably, the max-regret can be converted to a sample for finding an NE via online-to-batch conversion (Freund and Schapire, 1999). We refer the reader to Appendix J for a detailed discussion of alternative regret definitions.*

3 A New Analysis of Regularized Regret

We first present our main technical contribution: a quadratic bound on the instantaneous dual gap of any greedy NE policy under strongly convex regularization, beyond reverse KL. Below, $\mathbb{E}_{\boldsymbol{x},\boldsymbol{a}\sim\pi}$ denotes sampling $\boldsymbol{x} \sim d_0$ followed by $\boldsymbol{a} \sim \pi(\cdot \mid \boldsymbol{x})$.

Theorem 3.1. *For any preference model estimator $\widehat{P}_t$ at time t, define the max-player's policy as*

$$\hat{\pi}_t \leftarrow \arg\max_{\pi^1 \in \Pi} \min_{\pi^2 \in \Pi} J_\eta(\pi^1, \pi^2; \widehat{P}_t). \tag{1}$$

Then, the instantaneous dual gap is bounded as follows: denoting $e_t := P^* - \widehat{P}_t$,

$$\mathrm{DGap}_\eta(\hat{\pi}_t) \le 2\eta\beta \mathbb{E}_{\boldsymbol{x},\boldsymbol{a}\sim\hat{\pi}_t}\left[\sup_{\boldsymbol{b}\in\mathcal{A}} |e_t(\boldsymbol{a} \succ \boldsymbol{b} \mid \boldsymbol{x})|^2\right] \wedge \mathbb{E}_{\boldsymbol{x},\boldsymbol{a}\sim\hat{\pi}_t}\left[\sup_{\boldsymbol{b}\in\mathcal{A}} |e_t(\boldsymbol{a} \succ \boldsymbol{b} \mid \boldsymbol{x})|\right].$$

Crucially, this bound holds for *any* choice of preference model estimator and *any* β^{-1}-strongly convex regularizer $\psi(\cdot)$. For finite η, the instantaneous dual gap is bounded *quadratically* by the preference estimation error along the actions of $\hat{\pi}_t$, uniformly over the opposing action.

3.1 Proof of Theorem 3.1

Regret Decomposition. 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

$$\begin{aligned} X = J_\eta(\hat{\pi}, \tilde{\pi}; \widehat{P}) - J_\eta(\hat{\pi}, \tilde{\pi}) + \frac{1}{2} - J_\eta(\hat{\pi}, \tilde{\pi}; \widehat{P}) &\overset{(*)}{\le} J_\eta(\hat{\pi}, \tilde{\pi}; \widehat{P}) - J_\eta(\hat{\pi}, \tilde{\pi}) \\ &= J(\hat{\pi}, \tilde{\pi}; \widehat{P}) - J(\hat{\pi}, \tilde{\pi}), \\ &\qquad \text{(regularization terms cancel out)} \end{aligned}$$

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

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

Proof. For any $\widehat{P}$, there always exists a ***SNE*** as discussed in Section 2.1. As the game value of ***SNE*** is $\frac{1}{2}$, it must be so for any **NE** (Sion, 1958; von Neumann, 1928). □

We thus have $\frac{1}{2} = \min_{\pi\in\Pi} J_\eta(\hat{\pi}, \pi; \widehat{P})$ for any given $\widehat{P}$, i.e., $\min_{\pi\in\Pi} J_\eta(\hat{\pi}, \pi; \widehat{P}) \le J_\eta(\hat{\pi}, \tilde{\pi}; \widehat{P})$.

Bounding the Preference Error. Denote $D := \sqrt{\mathbb{E}_{\boldsymbol{x}\sim d_0}\left[\|\hat{\pi}(\cdot \mid \boldsymbol{x}) - \tilde{\pi}(\cdot \mid \boldsymbol{x})\|_1^2\right]}$. The first lemma relates the preference error to D; its proof, which combines anti-symmetry of $\widehat{P}$ and the variational representation of the ℓ_1-norm (Müller, 1997), is presented at the end of this subsection:

Lemma 3.3. $\left|J(\hat{\pi},\tilde{\pi};\widehat{P}) - J(\hat{\pi},\tilde{\pi})\right| \le D\sqrt{\mathbb{E}_{\boldsymbol{x},\boldsymbol{a}\sim\hat{\pi}}\left[\sup_{\boldsymbol{b}\in\mathcal{A}}|e(\boldsymbol{a}\succ\boldsymbol{b}\mid\boldsymbol{x})|^2\right]}$.

The second lemma, whose proof is deferred to Appendix C, bounds D^2 by the dual gap X:

Lemma 3.4. *For* $\eta < \infty$, $X = \frac{1}{2} - J_\eta(\hat{\pi},\tilde{\pi}) \ge (2\eta\beta)^{-1}D^2$.

Combining Everything. Combining the regret decomposition and the above two lemmas yields a ***self-bounding quadratic inequality*** in the dual gap X:

$$X \le J(\hat{\pi},\tilde{\pi};\widehat{P}) - J(\hat{\pi},\tilde{\pi}) \le D\sqrt{\mathbb{E}_{\boldsymbol{x},\boldsymbol{a}\sim\hat{\pi}}\left[\sup_{\boldsymbol{b}\in\mathcal{A}}|e(\boldsymbol{a}\succ\boldsymbol{b}\mid\boldsymbol{x})|^2\right]} \le \sqrt{2\eta\beta X\mathbb{E}_{\boldsymbol{x},\boldsymbol{a}\sim\hat{\pi}}\left[\sup_{\boldsymbol{b}\in\mathcal{A}}|e(\boldsymbol{a}\succ\boldsymbol{b}\mid\boldsymbol{x})|^2\right]} \tag{2}$$

Strong convexity makes this argument work by controlling the policy distance by $\sqrt{X}$. Squaring and dividing by X when $X > 0$ proves the quadratic bound; the case $X = 0$ is immediate. The linear bound follows directly from the regret decomposition by bounding the preference error uniformly over the opposing action, and also holds without regularization. □

Proof of Lemma 3.3. Denote $e(\boldsymbol{b};\boldsymbol{a},\boldsymbol{x}) := (\widehat{P} - P^*)(\boldsymbol{a}\succ\boldsymbol{b}\mid\boldsymbol{x})$. Since both preference models satisfy $P(\boldsymbol{a}\succ\boldsymbol{b}\mid\boldsymbol{x}) + P(\boldsymbol{b}\succ\boldsymbol{a}\mid\boldsymbol{x}) = 1$, we have $e(\boldsymbol{b};\boldsymbol{a},\boldsymbol{x}) = -e(\boldsymbol{a};\boldsymbol{b},\boldsymbol{x})$ and hence

$$\mathbb{E}_{\boldsymbol{x}\sim d_0}\mathbb{E}_{\boldsymbol{a},\boldsymbol{b}\sim\hat{\pi}(\cdot|\boldsymbol{x})}[e(\boldsymbol{b};\boldsymbol{a},\boldsymbol{x})] = 0.$$

Subtracting this zero term gives

$$\begin{aligned}\left|J(\hat{\pi},\tilde{\pi};\widehat{P}) - J(\hat{\pi},\tilde{\pi})\right| &= \left|\mathbb{E}_{\boldsymbol{x},\boldsymbol{a}\sim\hat{\pi}}\left[\mathbb{E}_{\boldsymbol{b}\sim\tilde{\pi}(\cdot|\boldsymbol{x})}[e(\boldsymbol{b};\boldsymbol{a},\boldsymbol{x})] - \mathbb{E}_{\boldsymbol{b}\sim\hat{\pi}(\cdot|\boldsymbol{x})}[e(\boldsymbol{b};\boldsymbol{a},\boldsymbol{x})]\right]\right| \\ &\le \mathbb{E}_{\boldsymbol{x}\sim d_0}\left[\|\hat{\pi}(\cdot\mid\boldsymbol{x}) - \tilde{\pi}(\cdot\mid\boldsymbol{x})\|_1\,\mathbb{E}_{\boldsymbol{a}\sim\hat{\pi}(\cdot|\boldsymbol{x})}\left[\sup_{\boldsymbol{b}\in\mathcal{A}}|e(\boldsymbol{b};\boldsymbol{a},\boldsymbol{x})|\right]\right] \\ &\le D\sqrt{\mathbb{E}_{\boldsymbol{x},\boldsymbol{a}\sim\hat{\pi}}\left[\sup_{\boldsymbol{b}\in\mathcal{A}}|e(\boldsymbol{b};\boldsymbol{a},\boldsymbol{x})|^2\right]},\end{aligned}$$

where the first inequality uses the variational representation of the ℓ_1-norm and the last uses Cauchy–Schwarz and Jensen's inequality. □

3.2 Discussions

Crucially, our proof relies on the strong convexity of $\psi(\cdot)$ and the anti-symmetry of the preference models, without relying on the specific algebraic properties of the KL divergence (Nayak et al., 2026; Wu et al., 2025; Ye et al., 2024). Lemma 3.3 introduces the policy distance D by cancelling the preference error under self-play, while Lemma 3.4 bounds D^2 by the dual gap X. Together, they yield the ***self-bounding quadratic inequality*** that leads the desired quadratic bound in Eqn. (2).

4 Fast Regrets under Generalized Bilinear Preferences

4.1 Generalized Bilinear Preference Model (GBPM)

As a concrete instantiation of our general theory, we consider a contextual linear variant of general preferences. Indeed, while recent works have introduced contextual bandit frameworks for general preference learning (Nayak et al., 2026; Wu et al., 2024; Yang et al., 2025), 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. This motivates our choice to consider ***generalized bilinear preference model (GBPM)*** (Lee et al., 2026; Zhang et al., 2025a). We denote $\mathrm{Skew}(d; 2r, S)$ 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***:

Definition 4.1 (***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:*

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

*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 twice-differentiable, monotone increasing, and symmetric link function ($\mu(z) + \mu(-z) = 1$).*

We briefly motivate the structure of ***GBPM***. Skew-symmetry, together with the symmetric link function, ensures reciprocity: $P(a^1 \succ a^2 \mid x) + P(a^2 \succ a^1 \mid x) = 1$. The low-rank structure provides an effective representation of pairwise interactions, as also motivated by prior work on low-rank preference models (Rajkumar and Agarwal, 2016). In particular, any skew-symmetric matrix of rank at most $2r$ can be decomposed into at most r two-dimensional antisymmetric interaction modes. Moreover, the model strictly generalizes linear Bradley–Terry–Luce (BTL) models, recovering them with $r = 1$, a logistic link, and a constant feature coordinate. Unlike BTL, even rank-two models can capture cyclic preferences, while higher ranks allow richer combinations of interaction modes. We refer interested readers to Zhang et al. (2025a) and Lee et al. (2026, Appendix K) for more detailed discussions.

We make the following assumptions on μ:

Assumption 4.2. *There exist $L_\mu \ge \kappa > 0$, with $L_\mu \ge \frac{1}{4}$, such that $\kappa \le \dot{\mu}(z) \le L_\mu$ and $|\ddot{\mu}(z)| \le L_\mu$ for all $|z| \le S + 1$.*[1]

The conditions for μ are standard in logistic and generalized linear (dueling) bandits (Abeille et al., 2021; Bengs et al., 2022; Faury et al., 2020; Lee et al., 2024a,b, 2026; Wu et al., 2024). The logistic link $\mu(z) = (1 + e^{-z})^{-1}$ satisfies the above with $L_\mu = \frac{1}{4}$. Affine link functions are also admissible on the bounded score interval, provided they take values in $[0, 1]$ (Gajane et al., 2015; Wu et al., 2024). Lastly, one can see that P^* is anti-symmetric: $P^*(\boldsymbol{a}^1 \succ \boldsymbol{a}^2 \mid \boldsymbol{x}) + P^*(\boldsymbol{a}^2 \succ \boldsymbol{a}^1 \mid \boldsymbol{x}) = 1$.

We will consider two simple algorithms, `Greedy Sampling` and `Explore-Then-Commit`, and show that under suitable feature coverage assumptions, they achieve polylogarithmic and dimension-wise improved regrets, respectively. We assume the learner is *tractable* (not necessarily efficient), accessing the action-set $\mathcal{A}$ 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 MLE (quasi-MLE for a generic link $\mu(\cdot)$) 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 (Wu et al., 2025; Ye et al., 2024) and learning in regularized games under bandit feedback (Nayak et al., 2026; Yang et al., 2025). We detail the computational complexities of implementing these oracles in Appendix B.

4.2 Feature Coverage Assumptions

Under ***GBPM***, let $\boldsymbol{E}_t := \boldsymbol{\Theta}_\star - \widehat{\boldsymbol{\Theta}}_t$. When the true and estimated scores lie in the interval specified in Assumption 4.2, skew-symmetry and $\|\phi(\boldsymbol{x}, \boldsymbol{b})\|_2 \le 1$ imply $\sup_{\boldsymbol{b}\in\mathcal{A}} |(P^* - \widehat{P}_t)(\boldsymbol{a} \succ \boldsymbol{b} \mid \boldsymbol{x})| \le L_\mu \|\boldsymbol{E}_t \phi(\boldsymbol{x}, \boldsymbol{a})\|_2$. Consequently, Theorem 3.1 yields:

$$\mathrm{DGap}_\eta(\hat{\pi}_t) \le 2\eta\beta L_\mu^2 \mathbb{E}_{\boldsymbol{x},\boldsymbol{a}\sim\hat{\pi}_t}\left[\|\boldsymbol{E}_t\phi(\boldsymbol{x}, \boldsymbol{a})\|_2^2\right] \wedge L_\mu \mathbb{E}_{\boldsymbol{x},\boldsymbol{a}\sim\hat{\pi}_t}\left[\|\boldsymbol{E}_t\phi(\boldsymbol{x}, \boldsymbol{a})\|_2\right]. \tag{3}$$

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

[2] In our case, as both the loss 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 $(\boldsymbol{\Theta}' - \boldsymbol{\Theta}'^\top)/2$ for unconstrained $\boldsymbol{\Theta}' \in \mathbb{R}^{d\times d}$.

If the squared estimation error in (3) decays at the usual $1/t$ rate, the cumulative max-regret scales as $\widetilde{\mathcal{O}}(\eta\beta \log T \wedge \sqrt{T})$, up to problem-dependent factors. Since (3) applies to greedy Nash policies (1), we focus on algorithms that exploit such policies without optimism-based exploration. To ensure sufficient estimation accuracy, we consider the following feature-diversity assumptions:

Assumption 4.3 (Feature Coverage)**.** *The learner has access to an exploration policy $\rho(\cdot \mid \boldsymbol{x})$ satisfying the following conditions for some $C_{\min} > 0$:*

(i) ***Marginal Feature Coverage.*** $\lambda_{\min}\left(\mathbb{E}_{\boldsymbol{x}\sim d_0}\mathbb{E}_{\boldsymbol{a}\sim\rho(\cdot|\boldsymbol{x})}\left[\phi(\boldsymbol{x},\boldsymbol{a})\phi(\boldsymbol{x},\boldsymbol{a})^\top\right]\right) \geq C_{\min}$.

(ii) ***Comparison Coverage.*** *For every $\pi \in \Pi$ and $\boldsymbol{E} \in \mathrm{Skew}(d)$,*

$$\mathbb{E}_{\boldsymbol{x}\sim d_0}\mathbb{E}_{\boldsymbol{a}\sim\pi(\cdot|\boldsymbol{x}),\boldsymbol{b}\sim\rho(\cdot|\boldsymbol{x})}\left[\left(\phi(\boldsymbol{x},\boldsymbol{a})^\top \boldsymbol{E}\phi(\boldsymbol{x},\boldsymbol{b})\right)^2\right] \geq C_{\min}\mathbb{E}_{\boldsymbol{x}\sim d_0}\mathbb{E}_{\boldsymbol{a}\sim\pi(\cdot|x)}\left[\|\boldsymbol{E}\phi(\boldsymbol{x},\boldsymbol{a})\|_2^2\right].$$

In part (ii), the two responses are sampled independently conditional on the same context $\boldsymbol{x}$.

Part (i) ensures aggregate feature coverage, while part (ii) controls the estimation error in (3) through same-context comparisons. The former does not necessarily imply the latter in contextual settings, although it does when $|\mathcal{X}| = 1$. `GS` requires only (ii), whereas `ETC` uses both to achieve a $1/\sqrt{T_0}$ parameter estimation rate.

Feasibility and Scaling. For finite action sets, if some policy satisfies both coverage conditions with constant $C_{\min}$, a contextwise optimal design achieves $C_{\min}/d$ using only known features, without preference observations or knowledge of d_0(Appendix D). Marginal feature coverage alone does not guarantee that both conditions are feasible. Moreover, the scaling of $C_{\min}$ is intrinsically tied to the geometry of the given feature map. Since $\|\phi(\boldsymbol{x},\boldsymbol{a})\|_2 \leq 1$, (i) implies $C_{\min}^{-1} \geq d$. Well-conditioned, context-independent feature sets (e.g., normalized hypercubes, standard bases) achieve $C_{\min}^{-1} \asymp d$; ill-conditioned sets may scale worse.

Role of Coverage. Our assumptions isolate the benefit of strong convexity from the challenges of active exploration. Marginal feature coverage is standard in analyses of greedy contextual bandits (Bastani et al., 2021; Bogunovic et al., 2021; Goldenshluger and Zeevi, 2013; Kannan et al., 2018; Kim and Oh, 2024; Wu et al., 2020) and high-dimensional bandits (Hao et al., 2020; Li et al., 2022b; Zeng and Honorio, 2025). Unlike KL-regularized `GS`, where the Gibbs form provides implicit coverage at the cost of $\exp(\mathrm{poly}(\eta))$ factors (Wu et al., 2025), generic strongly convex regularizers may yield zero-probability actions, motivating our explicit coverage assumption. In practice, sampling multiple LLMs or temperatures can encourage diversity (Bai et al., 2022; Dong et al., 2024; Minh et al., 2025; Troshin et al., 2025). Whether active exploration can eliminate this assumption under generic strongly convex regularization remains open.

Coverage Alone is Insufficient. Our fast-rate guarantees rely on both feature coverage and strong convexity. In our setting, the min-player can collect information without directly incurring max-regret, reminiscent of decoupled exploration and exploitation (Avner et al., 2012). Nevertheless, even maximal feature coverage and "free" exploration do not guarantee fast regret without regularization:

Proposition 4.4 (Unregularized Lower Bound)**.** *For any $d \geq 2$, $T \geq d^2$, and any learner choosing both policies adaptively, there exists a noncontextual* ***GBPM*** *instance with logistic link, $|\mathcal{A}| = d$, $r = 1$, and $S = 1$, for which uniform ρ satisfies both parts of Assumption 4.3 with the maximal constant $C_{\min} = 1/d$, yet $\mathbb{E}\left[\text{MBR-Reg}_\infty(T)\right] \geq \frac{\sqrt{dT}}{64}$.*

The proof is deferred to Appendix D.2. The lower bound holds even when the learner adaptively controls both players. In contrast, Theorem 3.1 shows that strong convexity yields a quadratic dependence of the dual gap on estimation error. Thus, while coverage enables accurate estimation, strong convexity drives the improved error dependence underlying our fast-rate guarantees.

4.3 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 4.6. 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 4.4 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 $\tilde{\mathcal{O}}(\eta(\log T)^2 \wedge \sqrt{T})$. Under `GS`, after initializing $\hat{\pi}_1 = \rho$, the max-player 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 4.3). We show that `GS` attains the following polylogarithmic regret:

Theorem 4.5. *Let $\delta \in (0,1)$ and suppose that $d^2 \log(e + T/d) \gtrsim \kappa \log \frac{1}{\delta}$. Then, with probability at least $1-\delta$, `GS` simultaneously attains the following bounds:*

$$\mathrm{MBR\text{-}Reg}_\eta(T) \lesssim \mathrm{DGap}_\eta(\rho) + \min\left\{ \begin{matrix} \eta\beta\kappa^{-1} d^4 C_{\min}^{-1} \big(\log(e + T/d)\big)^2, \\ \kappa^{-\frac{1}{2}} C_{\min}^{-\frac{1}{2}} d^2 \sqrt{T} \log(e + T/d) \end{matrix} \right\}.$$

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

1. From Regret to Sum of Squared Errors. We begin with Theorem 3.1. For $t \geq 2$, the logistic link is L_μ-Lipschitz and $\|\phi(\boldsymbol{x}, \boldsymbol{b})\|_2 \leq 1$, so the instantaneous dual gap is bounded by the *squared* estimation error:

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

2. Towards Expected Elliptical Potentials. The comparison condition in Assumption 4.3(ii) converts this error into the squared prediction error on the *played* response pairs. Our ***Coverage Lemma*** (Lemma E.3) combines this condition with Cauchy–Schwarz with respect to the regularized Hessian $\widehat{\boldsymbol{H}}_t = \boldsymbol{I}_{d^2} + \nabla^2 \mathcal{L}_t(\hat{\boldsymbol{\theta}}_t)$. Applying the confidence sequence for the constrained MLE (Lee et al., 2024a, Theorem 3.2), we obtain:

$$\sum_{t=2}^{T} \mathrm{DGap}_\eta(\hat{\pi}_t) \lesssim \eta\beta L_\mu^2 \kappa^{-1} C_{\min}^{-1} \gamma_T(\delta/2) \underbrace{\sum_{t=1}^{T} \mathbb{E}_{t-1} \left[\|\boldsymbol{v}_t\|_{\boldsymbol{V}_t^{-1}}^2 \right]}_{\triangleq S_T},$$

where $\boldsymbol{v}_t = \mathrm{vec}(\boldsymbol{\phi}_t^1 (\boldsymbol{\phi}_t^2)^\top)$, $\boldsymbol{V}_t = \kappa^{-1} \boldsymbol{I}_{d^2} + \sum_{s=1}^{t-1} \boldsymbol{v}_s \boldsymbol{v}_s^\top$, and $\gamma_T(\delta/2) \lesssim d^2 \log(e + T/d)$ is the squared confidence radius. The expectation is conditional on the history and includes the shared context and both sampled responses.

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*:

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

(a) is deterministically bounded via the standard Elliptical Potential Lemma (Abbasi-Yadkori et al., 2011). For (b), we apply Freedman's inequality (Beygelzimer et al., 2011; Freedman, 1975; Lee et al., 2024b) which yields a ***contractive linear recursion*** of the form $S_T \leq \frac{1}{2} S_T + \tilde{\mathcal{O}}(d^2 \log T)$. The coefficient is strictly below one, so the expected potential sum on the right can be absorbed into the left. 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 $\tilde{\mathcal{O}}(\eta e^{3\eta} d^2 (\log T)^2)$ regret bound for GS[3] *without* assuming feature coverage. However, their bound scales exponentially with η, is restricted exclusively to reverse KL regularization ($\psi(\cdot) = D_{\mathrm{KL}}(\cdot, \pi_{\mathrm{ref}})$), 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. (2026, Theorem 2.1), our regret bound gracefully adapts to both regimes: it yields a $(\log T)^2$ rate for small η, and seamlessly transitions to $\sqrt{T}$ as η increases. In Appendix K, 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 similar 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 ***contractive linear recursion***, 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 and Lugosi, 2006; Foster and 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., 2026), rigorous statistical guarantees for them have remained absent.

Crucially, our analysis unifies these diverse regularizers in the online setting (for greedy policies). 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 (Aminian et al., 2025; Le et al., 2025), Shannon entropy (Cen et al., 2024; McKelvey and Palfrey, 1995; Mertikopoulos and Sandholm, 2016), Tsallis entropy (Lee et al., 2018; Tsallis, 1988; Yang et al., 2019; Zimmert and Seldin, 2021), the χ^2-divergence (Huang et al., 2025), and general f-divergences (Go et al., 2023; Han et al., 2025; Liese and Vajda, 2006; Wang et al., 2024).

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 (Huang et al., 2025; Jiang and Xie, 2025; Zhao et al., 2026). For instance, Huang et al. (2025) 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 4.6 (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.*

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

4.4 Dimension-wise Improved 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 fully exploiting the *low-rank structure* of $\mathbf{\Theta}_\star$? Our goal is to isolate the explicit statistical complexity to the intrinsic rank r, while retaining dependence on the feature coverage $C_{\min}^{-1}$ (Assumption 4.3). In particular, although $C_{\min}^{-1} \geq d$, exploiting low rank allows us to improve the dimension-dependent factors in our regret guarantees.

To this end, we leverage techniques for structured estimation in high-dimensional bandits (Carpentier and Munos, 2012; Hao et al., 2020; Jang et al., 2022, 2024; Kang et al., 2022; Kim and Paik, 2019; Li et al., 2022b; Lu et al., 2021; Oh et al., 2021). Following standard algorithmic approaches in this literature (Hao et al., 2020; Jang et al., 2024; Li et al., 2022b), we employ the `Explore-Then-Commit (ETC)` algorithm. The players first explore for T_0 rounds using the coverage policy ρ, then 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`, which builds on Theorem 3.1, with the full proof deferred to Appendix F:

Theorem 4.7. *Let $\delta \in (0,1)$, and suppose $\kappa^{-2}C_{\min}^{-4}L_\mu r \log\frac{4d}{\delta} \lesssim T_0 \leq T$. By setting $\lambda_{T_0} = \sqrt{32L_\mu \log(4d/\delta)/T_0}$, `ETC` satisfies, with probability at least $1-\delta$,*

$$\mathrm{MBR\text{-}Reg}_\eta(T) \leq T_0 \mathrm{DGap}_\eta(\rho) + \mathcal{O}\left(\frac{T\eta\beta r \log\frac{d}{\delta}}{\kappa^2 C_{\min}^4 T_0} \wedge \frac{T}{\kappa C_{\min}^2}\sqrt{\frac{r\log\frac{d}{\delta}}{T_0}}\right).$$

In particular, appropriately choosing T_0 gives the following regret bounds:

$$\mathrm{MBR\text{-}Reg}_\eta(T) \lesssim \begin{cases} \kappa^{-1}C_{\min}^{-2}\sqrt{T\eta\beta r\log\frac{d}{\delta}} & \text{if } T_0 \asymp \kappa^{-1}C_{\min}^{-2}\sqrt{T\eta\beta r\log\frac{d}{\delta}}, \\ \left(\kappa^{-2}C_{\min}^{-4} r T^2 \log\frac{d}{\delta}\right)^{1/3} & \text{if } T_0 \asymp \left(\kappa^{-2}C_{\min}^{-4} r T^2 \log\frac{d}{\delta}\right)^{1/3} \end{cases}.$$

Discussions. Our result highlights two benefits of exploiting regularization and low-rank structure. First, the quadratic error bound in Theorem 3.1 enables `ETC` to achieve a $\tilde{\mathcal{O}}(\sqrt{\eta r T})$ rate when $\mathrm{DGap}_\eta(\rho) = \mathcal{O}(1)$ and the exploration budget satisfies the theorem's sample-size condition, improving upon the $\tilde{\mathcal{O}}(T^{2/3})$ rate typically associated with `ETC` algorithms (Lattimore and Szepesvári, 2020). For sufficiently small η, the quadratic-error bound improves upon the linear-error bound, highlighting the benefit of stronger regularization. Second, the statistical error terms in Theorem 4.7 scale with the intrinsic rank r, retaining dimension dependence through $C_{\min}$ and logarithmic factors. Compared with `GS`, `ETC` offers improved dimension dependence at the expense of a worse dependence on T. In particular, when $C_{\min}^{-1} \asymp d$, $r \ll d$, and $T = \widetilde{\Theta}(d^5)$ with other problem parameters fixed, the quadratic-error `ETC` upper bound is smaller than that of `GS` by a factor of $\widetilde{O}(\sqrt{r/d})$, provided the exploration budget condition in Theorem 4.7 is satisfied.

5 Conclusion

In this work, we investigated regularized max-regret minimization in online RLHF under general preferences and bandit feedback. We established a novel quadratic error bound on the dual gap of greedy Nash policies for any strongly convex regularizer, using only strong convexity and preference anti-symmetry. Instantiating this result under ***GBPM***, we showed that, under feature coverage(Assumption 4.3), `Greedy Sampling` achieves polylogarithmic regret of $\widetilde{\mathcal{O}}(\eta(\log T)^2 \wedge \sqrt{T})$, while `Explore-Then-Commit` exploits low-rank structure to obtain rank-dependent regret of $\widetilde{\mathcal{O}}\left(C_{\min}^{-2}\sqrt{\eta r T} \wedge C_{\min}^{-4/3} r^{1/3} T^{2/3}\right)$. We further established that maximal feature coverage alone does not eliminate $\Omega(\sqrt{dT})$ regret without regularization. We detail further future directions in Appendix L.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for helpful comments and suggestions that helped us improve the manuscript. This work was partly supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2019-II190075 Artificial Intelligence Graduate School Support Program (KAIST), 5%; No. RS-2019-II191906 Artificial Intelligence Graduate School Program (POSTECH), 5%; No. RS-2024-00457882, AI Research Hub Project, 5%; No. RS-2022-II220311 Development of Goal-Oriented Reinforcement Learning Techniques for Contact-Rich Robotic Manipulation of Everyday Objects, 40%; No. RS-2022-II220184, Development and Study of AI Technologies to Inexpensively Conform to Evolving Policy on Ethics, 45%), and the National Science Foundation under grant CCF-2327013 and Meta Platforms, Inc.

References

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

Marc Abeille, Louis Faury, and Clément Calauzènes. 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*, pages 3691–3699. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/abeille21a.html.

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

Alekh Agarwal, Sahand Negahban, and Martin J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *The Annals of Statistics*, 40(5):2452 – 2482, 2012. doi:10.1214/12-AOS1032.

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

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

Orly Avner, Shie Mannor, and Ohad Shamir. Decoupling Exploration and Exploitation in Multi-Armed Bandits. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1107–1114, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851. URL https://arxiv.org/abs/1205.2874.

Yu Bai and Chi Jin. 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*, pages 551–560. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/bai20a.html.

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

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

David Balduzzi, Marta Garnelo, Yoram Bachrach, Wojciech Czarnecki, Julien Perolat, Max Jaderberg, and Thore Graepel. 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*, pages 434–443. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/balduzzi19a.html.

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

Viktor Bengs, Ròbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. 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.

Viktor Bengs, Aadirupa Saha, and Eyke Hüllermeier. 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*, pages 1764–1786. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/bengs22a.html.

Lilian Besson and Emilie Kaufmann. 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.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. 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*, pages 19–26, Fort Lauderdale, FL, USA, 11–13 Apr 2011. PMLR. URL https://proceedings.mlr.press/v15/beygelzimer11a.html.

Ilija Bogunovic, Arpan Losalka, Andreas Krause, and Jonathan Scarlett. 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*, pages 991–999. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/bogunovic21a.html.

Ralph Allan Bradley and Milton E. Terry. 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.

Emmanuel J. Candès and Yaniv Plan. 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.

Alexandra Carpentier and Rémi Munos. 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*, pages 190–198, La Palma, Canary Islands, 21–23 Apr 2012. PMLR. URL https://proceedings.mlr.press/v22/carpentier12.html.

Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomek Korbak, David Lindner, Pedro Freire, Tony Tong Wang, Samuel Marks, Charbel-Raphael Segerie, Micah Carroll, Andi Peng, Phillip J.K. Christoffersen, Mehul Damani, Stewart Slocum, Usman Anwar, Anand Siththaranjan, Max Nadeau, Eric J Michaud, Jacob Pfau, Dmitrii Krasheninnikov, Xin Chen, Lauro Langosco, Peter Hase, Erdem Biyik, Anca Dragan, David Krueger, Dorsa Sadigh, and Dylan Hadfield-Menell. Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=bx24KpJ4Eb. Survey Certification, Featured Certification.

Shicong Cen, Yuting Wei, and Yuejie Chi. 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.

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

Kani Chen, Inchi Hu, and Zhiliang Ying. 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.

Yinlam Chow, Ofir Nachum, and Mohammad Ghavamzadeh. Path Consistency Learning in Tsallis Entropy Regularized MDPs. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 979–988. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/chow18a.html.

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

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 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*, pages 9722–9744. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/cui24f.html.

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

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

Constantinos Daskalakis, Andrew Ilyas, Vasilis Syrgkanis, and Haoyang Zeng. Training GANs with Optimism. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=SJJySbbAZ.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. RLHF Workflow: From Reward Modeling to Online RLHF. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=a13aYUU9eU.

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

Jianqing Fan, Wenyan Gong, and Ziwei Zhu. 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.

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

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

Dylan Foster and Alexander Rakhlin. 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*, pages 3199–3210. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/foster20a.html.

Dylan Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. 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*, pages 2059–2059. PMLR, 15–19 Aug 2021. URL https://arxiv.org/abs/2010.03104.

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

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

Pratik Gajane, Tanguy Urvoy, and Fabrice Clérot. 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*, pages 218–227, Lille, France, 07–09 Jul 2015. PMLR. URL https://proceedings.mlr.press/v37/gajane15.html.

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

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

Alexander Goldenshluger and Assaf Zeevi. A Linear Response Bandit Problem. *Stochastic Systems*, 3(1):230 – 261, 2013. doi:10.1214/11-SSY032.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1861–1870. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/haarnoja18b.html.

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

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

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

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

Shinji Ito, Haipeng Luo, Taira Tsuchiya, and Yue Wu. 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*, pages 2858–2892. PMLR, 30 Jun–04 Jul 2025. URL https://proceedings.mlr.press/v291/ito25a.html.

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

Kyoungseok Jang, Chicheng Zhang, and Kwang-Sung Jun. 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*, pages 21329–21372. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/jang24e.html.

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

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

Chi Jin, Qinghua Liu, and Tiancheng Yu. 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*, pages 10251–10279. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/jin22c.html.

Yue Kang, Cho-Jui Hsieh, and Thomas Chun Man Lee. Efficient Frameworks for Generalized Low-Rank Matrix Bandit Problems. In *Advances in Neural Information Processing Systems*, volume 35, pages 19971–19983. Curran Associates, Inc., 2022. URL https://arxiv.org/abs/2401.07298.

Sampath Kannan, Jamie H Morgenstern, Aaron Roth, Bo Waggoner, and Zhiwei Steven Wu. A Smoothed Analysis of the Greedy Algorithm for the Linear Contextual Bandit Problem. In *Advances in Neural Information Processing Systems*, volume 31, pages 2231–2241. Curran Associates, Inc., 2018. URL https://arxiv.org/abs/1801.03423.

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

J. Kiefer and J. Wolfowitz. The Equivalence of Two Extremum Problems. *Canadian Journal of Mathematics*, 12:363–366, 1960. doi:10.4153/CJM-1960-030-4.

Gi-Soo Kim and Myunghee Cho Paik. Doubly-Robust Lasso Bandit. In *Advances in Neural Information Processing Systems*, volume 32, pages 5877 – 5887. Curran Associates, Inc., 2019. URL https://arxiv.org/abs/1907.11362.

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

Georgios Kotsalis, Guanghui Lan, and Tianjiao Li. Simple and Optimal Methods for Stochastic Variational Inequalities, I: Operator Extrapolation. *SIAM Journal on Optimization*, 32(3):2041–2073, 2022. doi:10.1137/20M1381678.

Yuko Kuroki, Alberto Rumi, Taira Tsuchiya, Fabio Vitale, and Nicolò Cesa-Bianchi. 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*, pages 1216–1224. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/kuroki24a.html.

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

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

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

Junghyun Lee, Se-Young Yun, and Kwang-Sung Jun. 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*, pages 4474–4482. PMLR, 02–04 May 2024b. URL https://proceedings.mlr.press/v238/lee24c.html.

Junghyun Lee, Kyoungseok Jang, Kwang-Sung Jun, Milan Vojnović, and Se-Young Yun. 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.

Kyungjae Lee, Sungjoon Choi, and Songhwai Oh. 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.

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

Lihong Li, Yu Lu, and Dengyong Zhou. 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*, pages 2071–2080. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/li17c.html.

Long-Fei Li, Yu-Yang Qian, Peng Zhao, and Zhi-Hua Zhou. Provably Efficient Online RLHF with One-Pass Reward Modeling. In *Advances in Neural Information Processing Systems*, volume 38, pages 185115–185144. Curran Associates, Inc., 2025. URL https://openreview.net/forum?id=41mBrhy6U4.

Wenjie Li, Adarsh Barik, and Jean Honorio. 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*, pages 12619–12655. PMLR, 17–23 Jul 2022b. URL https://proceedings.mlr.press/v162/li22a.html.

F. Liese and I. Vajda. 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.

Qinghua Liu, Tiancheng Yu, Yu Bai, and Chi Jin. 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*, pages 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.

Yangyi Lu, Amirhossein Meisami, and Ambuj Tewari. 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*, pages 460–468. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/lu21a.html.

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

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

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

Eric J Michaud, Adam Gleave, and Stuart Russell. Understanding Learned Reward Functions. *arXiv preprint arXiv:2012.05862*, 2020. URL https://arxiv.org/abs/2012.05862.

Nguyen Nhat Minh, Andrew Baker, Clement Neo, Allen G Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. 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.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J Mankowitz, Doina Precup, and Bilal Piot. Nash Learning from Human Feedback. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 36743–36768. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/munos24a.html.

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

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

Anupam Nayak, Tong Yang, Osman Yagan, Gauri Joshi, and Yuejie Chi. Achieving Logarithmic Regret in KL-Regularized Zero-Sum Markov Games. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 92044–92096. PMLR, 06–11 Jul 2026. URL https://openreview.net/forum?id=GcCBoFM8u9.

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

Brendan O'Donoghue, Tor Lattimore, and Ian Osband. 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*, pages 279–289. PMLR, 27–30 Jul 2021. URL https://proceedings.mlr.press/v161/o-donoghue21a.html.

Min-hwan Oh, Garud Iyengar, and Assaf Zeevi. Sparsity-Agnostic Lasso Bandit. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 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.

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

R. L. Plackett. 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.

Arun Rajkumar and Shivani Agarwal. When can we rank well from comparisons of $O(n \log(n))$ non-actively chosen pairs? In *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pages 1376–1401, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL https://proceedings.mlr.press/v49/rajkumar16.html.

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

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

Corby Rosset, Ching-An Cheng, Arindam Mitra, Michael Santacroce, Ahmed Awadallah, and Tengyang Xie. 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.

Yoan Russac, Louis Faury, Olivier Cappé, and Aurélien Garivier. 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*, pages 658–666. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/russac21a.html.

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

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

Aadirupa Saha and Akshay Krishnamurthy. 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*, pages 968–994. PMLR, 29 Mar–01 Apr 2022. URL https://proceedings.mlr.press/v167/saha22a.html.

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

Aadirupa Saha, Aldo Pacchiano, and Jonathan Lee. 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*, pages 6263–6289. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/saha23a.html.

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

Gokul Swamy, Christoph Dann, Rahul Kidambi, Steven Wu, and Alekh Agarwal. 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*, pages 47345–47377. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/swamy24a.html.

Vasilis Syrgkanis, Alekh Agarwal, Haipeng Luo, and Robert E Schapire. Fast Convergence of Regularized Learning in Games. In *Advances in Neural Information Processing Systems*, volume 28, pages 2989–2997. Curran Associates, Inc., 2015. URL https://arxiv.org/abs/1507.00407.

Xiaohang Tang, Sangwoong Yoon, Seongho Son, Huizhuo Yuan, Quanquan Gu, and Ilija Bogunovic. RSPO: Regularized Self-Play Alignment of Large Language Models. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 119328–119350. PMLR, 06–11 Jul 2026. URL https://openreview.net/forum?id=PysPgClt8e.

Jeremy Tien, Jerry Zhi-Yang He, Zackory Erickson, Anca Dragan, and Daniel S. Brown. 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.

Joel A. Tropp. 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.

Sergey Troshin, Wafaa Mohammed, Yan Meng, Christof Monz, Antske Fokkens, and Vlad Niculae. Control the Temperature: Selective Sampling for Diverse and High-Quality LLM Outputs. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=IyOC5GCzv4.

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

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

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

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

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Martin J. Wainwright and Michael I. Jordan. 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.

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

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

Weiqiang Wu, Jing Yang, and Cong Shen. 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*, pages 2392–2401. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/wu20c.html.

Yue Wu, Tao Jin, Qiwei Di, Hao Lou, Farzad Farnoud, and Quanquan Gu. 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*, pages 53571–53596. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wu24m.html.

Tengyang Xie, Dylan J Foster, Yu Bai, Nan Jiang, and Sham M. Kakade. The Role of Coverage in Online Reinforcement Learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=LQIjzPdDt3q.

Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, and Tong Zhang. 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*, pages 24496–24523. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/xiong22b.html.

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

Tong Yang, Bo Dai, Lin Xiao, and Yuejie Chi. 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.

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

Chenlu Ye, Wei Xiong, Yuheng Zhang, Nan Jiang, and Tong Zhang. Online Iterative Reinforcement Learning from Human Feedback with General Preference Model. In *Advances in Neural Information Processing Systems*, volume 37, pages 81773–81807. Curran Associates, Inc., 2024. URL https://openreview.net/forum?id=TwdX1W3M6S.

Guo Zeng and Jean Honorio. 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.

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

Tong Zhang. 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.

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

Yifan Zhang, Ge Zhang, Yue Wu, Kangping Xu, and Quanquan Gu. 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 2025a. URL https://openreview.net/forum?id=UgKcpPE0ZX.

Yu-Jie Zhang, Sheng-An Xu, Peng Zhao, and Masashi Sugiyama. Generalized Linear Bandits: Almost Optimal Regret with One-Pass Update. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025b. URL https://openreview.net/forum?id=H7vg3IgvHU.

Yuheng Zhang, Dian Yu, Tao Ge, Linfeng Song, Zhichen Zeng, Haitao Mi, Nan Jiang, and Dong Yu. Improving LLM General Preference Alignment via Optimistic Online Mirror Descent. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025c. URL https://openreview.net/forum?id=kZstGANG8D.

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

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

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

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

Banghua Zhu, Michael Jordan, and Jiantao Jiao. 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*, pages 43037–43067. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/zhu23f.html.

Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum Entropy Inverse Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 3:1433–1438, 2008. URL https://aaai.org/papers/01433-aaai08-227-maximum-entropy-inverse-reinforcement-learning/.

Julian Zimmert and Yevgeny Seldin. Tsallis-INF: An Optimal Algorithm for Stochastic and Adversarial Bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021. URL http://jmlr.org/papers/v22/19-753.html.

Contents

1 Introduction **1**

2 Problem Setting **3**

2.1 Preference, Regularized Game, and Nash Equilibrium 3

2.2 Online Interaction Protocol and Regularized Max-Regret 3

3 A New Analysis of Regularized Regret **4**

3.1 Proof of Theorem 3.1 4

3.2 Discussions 5

4 Fast Regrets under Generalized Bilinear Preferences **5**

4.1 Generalized Bilinear Preference Model (GBPM) 5

4.2 Feature Coverage Assumptions 6

4.3 Polylogarithmic Regret via Greedy Sampling: $\tilde{\mathcal{O}}(\eta(\log T)^2 \wedge \sqrt{T})$ 7

4.4 Dimension-wise Improved Regret via Explore-Then-Commit: $\widetilde{\mathcal{O}}(\sqrt{\eta r T} \wedge r^{1/3}T^{2/3})$ 10

5 Conclusion **10**

A Pseudocodes for `Greedy Sampling` and `Explore-Then-Commit` **24**

B Computational Complexities of Implementing the Oracles **25**

C Proof of Lemma 3.4 **26**

D Feature Diversity **27**

D.1 Constructing the Exploration Policy ρ 27

D.2 Proof of Proposition 4.4: Unregularized Lower Bound 27

E Proof of Theorem 4.5: Regret Bound of `Greedy Sampling` **29**

E.1 Main Proof 29

E.2 Proof of Lemma E.3: Coverage Lemma 31

F Proof of Theorem 4.7: Regret Bound of `Explore-Then-Commit` **32**

G Eluder Dimension of GBPM **35**

G.1 Upper Bounding the Eluder Dimension of Wu et al. (2025) 35

G.2 Connection to the Standard Eluder Dimensions 36

G.3 Proof of Proposition G.6 38

H Instantiating Regret Bound of Wu et al. (2025) to GBPM **39**

H.1 Regret Bound of `Greedy Sampling` 39

H.2 Proof of Lemma H.3: MLE Estimator Bound 40

H.3 Proof of Lemma H.6 42

I Comparison with Prior Work **43**

I.1 Existing Regret Guarantees 43

I.2 Discussions 43

J Discussions on Regrets **44**

J.1 Four Regret Definitions and Discussions 44

J.2 Online-to-Batch Conversion 45

K Synthetic Experiments **47**

K.1 Experiment Setup. 47

K.2 Implementation Details and Reproducibility 47

K.3 Main Results 48

L Future Directions **50**

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 Computational Complexities of Implementing the Oracles

We discuss the computational complexity of the oracles used in `Greedy Sampling` (GS) and `Explore-Then-Commit` (ETC). Throughout, let t denote the current round, and suppress dimension-dependent factors for clarity. Let $\epsilon_{2,t}$ be the target suboptimality of the averaged empirical (log-likelihood) loss at iteration t, and $\epsilon_{3,t}$ be the dual-gap optimization error at iteration t.

Sampling Oracle. Oracle 1 is a standard sampling oracle that, given a context $\boldsymbol{x} \in \mathcal{X}$ and a policy $\pi \in \Pi$, draws an action $\boldsymbol{a} \sim \pi(\cdot \mid \boldsymbol{x})$ (e.g., by sampling from an LLM).

Regularized MLE Oracle. Oracle 2 computes a regularized MLE over the space of skew-symmetric matrices $\mathrm{Skew}(d)$. Under suitable restricted strong convexity and smoothness conditions, projected or composite gradient methods achieve geometric convergence up to the statistical precision of the estimator (Agarwal et al., 2012). Each gradient evaluation requires $\mathcal{O}(t)$ operations over the accumulated observations, and $\mathcal{O}(\log \epsilon_{2,t}^{-1})$ iterations suffice to reach accuracy $\epsilon_{2,t}$, provided the target lies within this convergence regime. For $\epsilon_{2,t} = \mathcal{O}(\log t)/t)$, the per-round iteration complexity is $\mathcal{O}(t \log t)$, and hence, and a cumulative cost of $\mathcal{O}(T^2 \log T)$ for `GS` and $\mathcal{O}(T_0 \log T_0)$ for `ETC`.

Population SNE Oracle. Given an estimated parameter $\boldsymbol{\Theta}$, Oracle 3 computes the population saddle point

$$\arg\max_{\pi^1 \in \Pi} \min_{\pi^2 \in \Pi} J_\eta(\pi^1, \pi^2; \boldsymbol{\Theta}).$$

For finite $\eta > 0$, strong convexity of the regularizer makes the objective strongly concave in π^1 and strongly convex in π^2. Under suitable smoothness and stochastic-oracle conditions, the resulting saddle-point problem can be solved using stochastic variational inequality methods, such as Stochastic Operator Extrapolation (Kotsalis et al., 2022). With Monte-Carlo evaluations based on the estimated model $\widehat{\boldsymbol{\Theta}}_t$ rather than $\boldsymbol{\Theta}_\star$, the method requires $\mathcal{O}(\eta\beta\, \epsilon_{3,t}^{-1} \log \epsilon_{3,t}^{-1})$ iterations to reach accuracy $\epsilon_{3,t}$, provided the required oracle accuracy is controlled by the solver's error guarantee. Taking $\epsilon_{3,t} = \mathcal{O}(1/t)$ to preserve our regret bounds yields a cumulative cost of $\mathcal{O}(\eta\beta T^2 \log T)$ for GS.

Computational Considerations. Unlike `GS`, `ETC` invokes the MLE and SNE oracles only once after its exploration phase, avoiding repeated optimization over the growing dataset. Further improvements may be possible by replacing repeated MLE computations with online updates, as in recent one-pass reward-modeling methods based on online mirror descent (Li et al., 2025). Extending such approaches to skew-symmetric matrix estimation, particularly with nuclear-norm regularization in ETC, remains an interesting direction for future work.

C Proof of Lemma 3.4

For each fixed context $\boldsymbol{x} \in \mathcal{X}$, $J(\hat{\pi}, \pi|\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{a}\sim\hat{\pi}(\cdot|\boldsymbol{x}),\boldsymbol{b}\sim\pi(\cdot|\boldsymbol{x})}[P(\boldsymbol{a} \succ \boldsymbol{b} \mid \boldsymbol{x})]$ 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 $\pi(\cdot|\boldsymbol{x}) \mapsto J_\eta(\hat{\pi}, \pi|\boldsymbol{x})$ is $(\eta\beta)^{-1}$-strongly convex in ℓ_1. By strong convexity, for any $\pi, \tilde{\pi} \in \Pi$ and $\alpha \in [0, 1]$, the following holds point-wise for any $\boldsymbol{x}$:

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

Taking the expectation over $\boldsymbol{x} \sim d_0$ on both sides 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:

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

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

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

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

D Feature Diversity

D.1 Constructing the Exploration Policy ρ

We show how to obtain a policy satisfying both parts of Assumption 4.3, whenever such a policy exists. The construction uses the known feature map and does not require preference observations or knowledge of d_0.

Proposition D.1. *Suppose that $\mathcal{A}$ is finite and some policy $\rho^\star$ satisfies both parts of Assumption 4.3 with constant $C_{\min}$. For each x, let $\rho(\cdot \mid x)$ be a G-optimal design of $\{\phi(x,a) : a \in \mathcal{A}\}$ on its linear span. Then ρ satisfies both parts of Assumption 4.3 with constant $C_{\min}/d$.*

Proof. The standard properties of the G-optimal design (Kiefer and Wolfowitz, 1960) (Lattimore and Szepesvári, 2020, Theorem 21.1) and Cauchy-Schwarz give:

$$\left(\boldsymbol{u}^\top \phi(\boldsymbol{x},\boldsymbol{b})\right)^2 \le d\,\mathbb{E}_{\boldsymbol{a}\sim\rho(\cdot|\boldsymbol{x})}\left[\left(\boldsymbol{u}^\top\phi(\boldsymbol{x},\boldsymbol{a})\right)^2\right], \quad \forall \boldsymbol{x}\in\mathcal{X}, \boldsymbol{b}\in\mathcal{A}, \boldsymbol{u}\in\mathbb{R}^d.$$

Averaging over $\boldsymbol{b}\sim\rho^\star(\cdot\mid\boldsymbol{x})$ yields

$$\mathbb{E}_{\boldsymbol{a}\sim\rho(\cdot|\boldsymbol{x})}\left[\left(\boldsymbol{u}^\top\phi(\boldsymbol{x},\boldsymbol{a})\right)^2\right] \ge \frac{1}{d}\mathbb{E}_{\boldsymbol{b}\sim\rho^\star(\cdot|\boldsymbol{x})}\left[\left(\boldsymbol{u}^\top\phi(\boldsymbol{x},\boldsymbol{b})\right)^2\right]. \tag{4}$$

Averaging over $\boldsymbol{x}$ and applying part (i) for $\rho^\star$ proves marginal feature diversity with constant $C_{\min}/d$.

For comparison coverage, fix $\pi\in\Pi$ and $\boldsymbol{E}\in\mathrm{Skew}(d)$. Applying Eqn. (4) with $\boldsymbol{u}=\boldsymbol{E}^\top\phi(\boldsymbol{x},\boldsymbol{a})$ and then averaging over $\boldsymbol{x},\boldsymbol{a}\sim\pi$, we obtain

$$\begin{aligned}&\mathbb{E}_{\boldsymbol{x},\boldsymbol{a}\sim\pi,\boldsymbol{b}\sim\rho}\left[\left(\phi(\boldsymbol{x},\boldsymbol{a})^\top\boldsymbol{E}\phi(\boldsymbol{x},\boldsymbol{b})\right)^2\right]\\ &\quad\ge\frac{1}{d}\mathbb{E}_{\boldsymbol{x},\boldsymbol{a}\sim\pi,\boldsymbol{b}\sim\rho^\star}\left[\left(\phi(\boldsymbol{x},\boldsymbol{a})^\top\boldsymbol{E}\phi(\boldsymbol{x},\boldsymbol{b})\right)^2\right]\ge\frac{C_{\min}}{d}\mathbb{E}_{\boldsymbol{x},\boldsymbol{a}\sim\pi}\left[\|\boldsymbol{E}\phi(\boldsymbol{x},\boldsymbol{a})\|_2^2\right],\end{aligned}$$

where the last inequality uses part (ii) for $\rho^\star$. □

This construction allows singular feature covariances at individual contexts. For finite $\mathcal{A}$, computing the design is a convex optimization over action probabilities, although it may be expensive for large action sets. This optimization is additional to the three oracles in Section 4.1. The proposition establishes constructibility when both coverage conditions are feasible; marginal diversity alone does not imply their joint feasibility.

D.2 Proof of Proposition 4.4: Unregularized Lower Bound

We use the standard change-of-measure argument (Lattimore and Szepesvári, 2020, Chapter 15), accounting for information from both responses.

1. Hard Instances and Coverage. Let $|\mathcal{X}|=1$, $\mathcal{A}=\{\boldsymbol{a}_1,\dots,\boldsymbol{a}_d\}$, $\phi(\boldsymbol{a}_i)=\boldsymbol{e}_i$, and ρ be uniform. Set $\mu(z)=(1+e^{-z})^{-1}$, $S=1$, and

$$\Delta=\frac{1}{2}\sqrt{\frac{d}{T}},\qquad \boldsymbol{\Theta}_0=\boldsymbol{0},\qquad \boldsymbol{\Theta}_i=\Delta(\boldsymbol{e}_i\boldsymbol{1}^\top-\boldsymbol{1}\boldsymbol{e}_i^\top),\quad i\in[d].$$

Each $\boldsymbol{\Theta}_i$, $i\ge 1$, has rank two and two nonzero singular values equal to $\Delta\sqrt{d-1}$. Hence $\|\boldsymbol{\Theta}_i\|_{\mathrm{nuc}}=2\Delta\sqrt{d-1}\le 1$ when $T\ge d^2$. The logistic link satisfies Assumption 4.2. Moreover, $\mathbb{E}_{\boldsymbol{a}\sim\rho}[\phi(\boldsymbol{a})\phi(\boldsymbol{a})^\top]=\boldsymbol{I}_d/d$, and for every $\pi\in\Pi$ and $\boldsymbol{E}\in\mathrm{Skew}(d)$,

$$\mathbb{E}_{\boldsymbol{a}\sim\pi,\boldsymbol{b}\sim\rho}\left[(\phi(\boldsymbol{a})^\top\boldsymbol{E}\phi(\boldsymbol{b}))^2\right]=\frac{1}{d}\mathbb{E}_{\boldsymbol{a}\sim\pi}\left[\|\boldsymbol{E}\phi(\boldsymbol{a})\|_2^2\right].$$

Thus, both coverage conditions hold with $C_{\min}=1/d$, which is maximal since the marginal covariance has trace one.

2. Unregularized Dual Gap. Under $\mathbf{\Theta}_i$, the preference probability for (a_j, a_k) is

$$P_i(a_j \succ a_k) = \frac{1}{2} + \left(\mu(\Delta) - \frac{1}{2}\right)(\mathbf{1}\{j = i\} - \mathbf{1}\{k = i\}).$$

Thus a_i is a best response for the min-player against every π, and

$$\mathrm{DGap}_\infty(\pi) = \left(\mu(\Delta) - \frac{1}{2}\right)(1 - \pi(a_i)).$$

3. Change of Measure. Write $\mathbb{P}_i$ and $\mathbb{E}_i$ for the law and expectation of the complete interaction history under $\mathbf{\Theta}_i$, including the learner's randomization. Let

$$N_i := \sum_{t=1}^{T} \left(\mathbb{1}\{\boldsymbol{a}_t^1 = \boldsymbol{a}_i\} + \mathbb{1}\{\boldsymbol{a}_t^2 = \boldsymbol{a}_i\}\right), \qquad \sum_{i=1}^{d} N_i = 2T.$$

The conditional observation laws differ from the null instance only when the compared pair contains $\boldsymbol{a}_i$. Since

$$D_{\mathrm{KL}}\left(\mathrm{Ber}(1/2), \mathrm{Ber}(\mu(\pm\Delta))\right) = \log\cosh(\Delta/2) \le \Delta^2/8,$$

the chain rule gives $D_{\mathrm{KL}}(\mathbb{P}_0, \mathbb{P}_i) \le (\Delta^2/8)\mathbb{E}_0[N_i]$. This bound allows arbitrary adaptive choices of both policies; self-comparisons simply contribute zero information.

Using $\sum_{i=1}^{d} \hat{\pi}_t^1(\boldsymbol{a}_i) = 1$ and $\sum_{i=1}^{d} N_i = 2T$, we obtain

$$\begin{aligned}
\frac{1}{d}\sum_{i=1}^{d} \mathbb{E}_i\left[\sum_{t=1}^{T} \hat{\pi}_t^1(\boldsymbol{a}_i)\right] &\le \frac{T}{d} + \frac{T}{d}\sum_{i=1}^{d} \mathrm{TV}(\mathbb{P}_0, \mathbb{P}_i) \\
&\le \frac{T}{d} + \frac{T}{d}\sum_{i=1}^{d} \sqrt{\frac{1}{2} D_{\mathrm{KL}}(\mathbb{P}_0, \mathbb{P}_i)} && \text{(Pinsker's inequality)} \\
&\le \frac{T}{d} + \frac{T\Delta}{4d}\sum_{i=1}^{d} \sqrt{\mathbb{E}_0[N_i]} \\
&\le \frac{T}{d} + \frac{T\Delta}{4}\sqrt{\frac{1}{d}\sum_{i=1}^{d} \mathbb{E}_0[N_i]} && \text{(Jensen's inequality)} \\
&= \frac{T}{d} + T\Delta\sqrt{\frac{T}{8d}} \\
&= \frac{T}{d} + \frac{T}{4\sqrt{2}} \le \frac{3T}{4},
\end{aligned}$$

where the first inequality uses $\sum_{t=1}^{T} \hat{\pi}_t^1(\boldsymbol{a}_i) \in [0, T]$, the third uses $D_{\mathrm{KL}}(\mathbb{P}_0, \mathbb{P}_i) \le (\Delta^2/8)\mathbb{E}_0[N_i]$, and the last line uses $\Delta = \frac{1}{2}\sqrt{d/T}$ and $d \ge 2$.

Finally, $\Delta \le 1$ and $\dot{\mu}(z) \ge 1/8$ for $z \in [0, 1]$ imply $\mu(\Delta) - 1/2 \ge \Delta/8$. Therefore,

$$\frac{1}{d}\sum_{i=1}^{d} \mathbb{E}_i\left[\text{MBR-Reg}_\infty(T)\right] \ge \left(\mu(\Delta) - \frac{1}{2}\right)\frac{T}{4} \ge \frac{\Delta T}{32} = \frac{\sqrt{dT}}{64}.$$

At least one instance attains this lower bound, proving the proposition. □

E Proof of Theorem 4.5: Regret Bound of `Greedy Sampling`

E.1 Main Proof

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 $\boldsymbol{v}_t := \mathrm{vec}(\boldsymbol{\phi}_t^1(\boldsymbol{\phi}_t^2)^\top) \in \mathbb{R}^{d^2}$, our MLE is defined as follows:

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

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

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

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

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

We first recall the following elliptical confidence sequence:

Lemma E.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*

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

where $\gamma_t(\delta) \asymp (1+S)^5 + (1+S)\log\frac{1}{\delta} + (1+S)d^2\log\left(e + \frac{S(2+S)t}{d^2}\right)$.

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

The confidence bound is used only in the analysis and need not be computed by the algorithm.

Remark E.2. *The confidence bound of Lee et al. (2024a) assumes a full-dimensional parameter set. Although our feasible set is lower-dimensional in* $\mathbb{R}^{d^2}$*, the result applies after reparameterizing its span. Let* $\boldsymbol{U} \in \mathbb{R}^{d^2\times p}$ *have orthonormal columns spanning* $\mathrm{vec}(\mathrm{Skew}(d))$*, where* $p = d(d-1)/2$*. Writing* $\boldsymbol{\theta} = \boldsymbol{U}\boldsymbol{u}$*, the feasible set becomes* $\{\boldsymbol{u} \in \mathbb{R}^p : \|\boldsymbol{u}\|_2 \le S\}$*, and the covariates become* $\boldsymbol{U}^\top\boldsymbol{v}_s$*, whose norms are at most one. The true parameter is feasible since* $\|\boldsymbol{\Theta}_\star\|_F \le \|\boldsymbol{\Theta}_\star\|_{\mathrm{nuc}} \le S$*. Applying the theorem with Hessian regularization parameter* $\lambda = 1$ *gives the stated bound, since* $\|\boldsymbol{\theta}_\star - \hat{\boldsymbol{\theta}}_t\|_{\widehat{\boldsymbol{H}}_t}^2 = \|\boldsymbol{u}_\star - \hat{\boldsymbol{u}}_t\|_{\boldsymbol{U}^\top\widehat{\boldsymbol{H}}_t\boldsymbol{U}}^2$*, where* $\boldsymbol{u}_\star = \boldsymbol{U}^\top\boldsymbol{\theta}_\star$ *and* $\hat{\boldsymbol{u}}_t = \boldsymbol{U}^\top\hat{\boldsymbol{\theta}}_t$*, and* $p \le d^2$*.*

Recall that $\boldsymbol{E}_t = \boldsymbol{\Theta}_\star - \widehat{\boldsymbol{\Theta}}_t \in \mathrm{Skew}(d)$. Since μ is L_μ-Lipschitz and $\|\phi(\boldsymbol{x},\boldsymbol{b})\|_2 \le 1$, skew-symmetry gives

$$\sup_{\boldsymbol{b}\in\mathcal{A}} \left| (P^* - \widehat{P}_t)(\boldsymbol{a} \succ \boldsymbol{b} \mid \boldsymbol{x}) \right| \le L_\mu \left\| \boldsymbol{E}_t\phi(\boldsymbol{x},\boldsymbol{a}) \right\|_2.$$

In `GS`, the policy $\hat{\pi}_t$ is a greedy NE w.r.t. $\widehat{\boldsymbol{\Theta}}_t$ for $t \ge 2$. Thus, the η-dependent bound of Theorem 3.1 yields

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

2. Towards Expected Elliptical Potentials. We now present our key technical lemma:

Lemma E.3 (Coverage Lemma)**.** *Suppose that Assumption 4.3(ii) holds. Then, for any $\pi \in \Pi$, $\boldsymbol{E} \in \mathrm{Skew}(d)$, and positive definite $\boldsymbol{M} \in \mathbb{R}^{d^2 \times d^2}$, the following holds:*

$$\mathbb{E}_{\boldsymbol{\phi}\sim\pi}\left[\|\boldsymbol{E}\boldsymbol{\phi}\|_2^2\right] \le C_{\min}^{-1}\|\mathrm{vec}(\boldsymbol{E})\|_{\boldsymbol{M}}^2 \mathbb{E}_{\boldsymbol{\phi}\sim\pi,\tilde{\boldsymbol{\phi}}\sim\rho}\left[\|\mathrm{vec}(\boldsymbol{\phi}\tilde{\boldsymbol{\phi}}^\top)\|_{\boldsymbol{M}^{-1}}^2\right],$$

where both responses are sampled independently conditional on the same context $\boldsymbol{x} \sim d_0$.

From hereon, $\mathbb{E}_{t-1}[\cdot]$ denotes expectation conditional on the history before round t, averaging over the fresh context $\boldsymbol{x}_t \sim d_0$ and responses $\boldsymbol{a}_t^1 \sim \hat{\pi}_t(\cdot \mid \boldsymbol{x}_t)$ and $\boldsymbol{a}_t^2 \sim \rho(\cdot \mid \boldsymbol{x}_t)$. Set $\boldsymbol{V}_t := \kappa^{-1}\boldsymbol{I}_{d^2} + \sum_{s=1}^{t-1}\boldsymbol{v}_s\boldsymbol{v}_s^\top$. Since $\kappa \le 1/4$ and the link derivative is at least κ throughout the estimation domain, $\widehat{\boldsymbol{H}}_t \succeq \kappa \boldsymbol{V}_t$.

Conditional on the history, $\hat{\pi}_t$ and $\boldsymbol{E}_t$ are fixed. Applying the above lemma with $\boldsymbol{M} = \widehat{\boldsymbol{H}}_t$ and Lemma E.1 at confidence level $\delta/2$, we have, simultaneously for all t,

$$\mathbb{E}_{\boldsymbol{\phi}_t\sim\hat{\pi}_t}\left[\|\boldsymbol{E}_t\boldsymbol{\phi}_t\|_2^2\right] \lesssim C_{\min}^{-1}\gamma_t(\delta/2)\mathbb{E}_{t-1}\left[\|\boldsymbol{v}_t\|_{\widehat{\boldsymbol{H}}_t^{-1}}^2\right] \le \kappa^{-1}C_{\min}^{-1}\gamma_t(\delta/2)\mathbb{E}_{t-1}\left[\|\boldsymbol{v}_t\|_{\boldsymbol{V}_t^{-1}}^2\right]. \quad (5)$$

Consequently,

$$\text{MBR-Reg}_\eta(T) - \text{DGap}_\eta(\rho) \lesssim \eta\beta L_\mu^2\kappa^{-1}C_{\min}^{-1}\gamma_T(\delta/2)\sum_{t=1}^T \mathbb{E}_{t-1}\left[\|\boldsymbol{v}_t\|_{\boldsymbol{V}_t^{-1}}^2\right]. \quad (6)$$

Under the theorem's condition, $\gamma_T(\delta/2) \lesssim d^2\log(e+T/d)$, with implicit constants depending on S.

3. Martingale Concentration for Realized Variance.

We now consider the elliptical-type quantity:

$$\underbrace{\sum_{t=1}^T \mathbb{E}_{t-1}\left[\boldsymbol{v}_t^\top\boldsymbol{V}_t^{-1}\boldsymbol{v}_t\right]}_{\triangleq S_T} = \sum_{t=1}^T \boldsymbol{v}_t^\top\boldsymbol{V}_t^{-1}\boldsymbol{v}_t + \sum_{t=1}^T \underbrace{\left(\mathbb{E}_{t-1}\left[\boldsymbol{v}_t^\top\boldsymbol{V}_t^{-1}\boldsymbol{v}_t\right] - \boldsymbol{v}_t^\top\boldsymbol{V}_t^{-1}\boldsymbol{v}_t\right)}_{\triangleq M_t}$$

$$= \underbrace{\sum_{t=1}^T \boldsymbol{v}_t^\top\boldsymbol{V}_t^{-1}\boldsymbol{v}_t}_{(a)} + \underbrace{\sum_{t=1}^T M_t}_{(b)}. \quad (7)$$

We bound (a) via the usual elliptical potential lemma (Lemma E.4), which we recall here:

Lemma E.4 (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*

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

We thus have that

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

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 (Beygelzimer et al., 2011; Freedman, 1975):

Lemma E.5 (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:*

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

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

$$\mathbb{E}_{s-1}[M_s^2] \le \mathbb{E}_{s-1}[{Z_s}^2] \le \kappa \mathbb{E}_{s-1}[Z_s],$$

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

Choosing $\xi = \frac{1}{2\kappa(e-2)} \le \frac{1}{\kappa}$ in Lemma E.5, the following holds with probability at least $1 - \frac{\delta}{2}$:

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

Now bringing everything together for Eqn. (7), the following holds with probability at least $1 - \frac{\delta}{2}$:

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

which is a ***contractive linear recursion***. The coefficient of S_T is strictly below one, allowing us to absorb this term into the left-hand side.

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

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

Combining everything at Eqn. (6), we have that with probability at least $1 - \delta$ (after union bound with the confidence sequence),

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

The theorem's condition gives the claimed quadratic-error regret bound.

Part II. $\tilde{\mathcal{O}}(\sqrt{T})$ Regret Bound. On the same event as in **Part I**, the η-independent bound of Theorem 3.1 yields

$$\begin{aligned} \text{MBR-Reg}_\eta(T) - \text{DGap}_\eta(\rho) &\le L_\mu \sum_{t=2}^{T} \mathbb{E}_{\phi_t \sim \hat{\pi}_t}\left[\|\boldsymbol{E}_t \phi_t\|_2\right] \\ &\le L_\mu \sqrt{T \sum_{t=2}^{T} \mathbb{E}_{\phi_t \sim \hat{\pi}_t}\left[\|\boldsymbol{E}_t \phi_t\|_2^2\right]} \quad \text{(Cauchy–Schwarz \& Jensen)} \\ &\lesssim L_\mu \sqrt{T \kappa^{-1} C_{\min}^{-1} \gamma_T(\delta/2) S_T} \quad \text{((5))} \\ &\lesssim \kappa^{-\frac{1}{2}} C_{\min}^{-\frac{1}{2}} d^2 \sqrt{T} \log(e + T/d). \end{aligned}$$

□

E.2 Proof of Lemma E.3: Coverage Lemma

By Assumption 4.3(ii),

$$\begin{aligned} \mathbb{E}_{\phi \sim \pi}\left[\|\boldsymbol{E}\phi\|_2^2\right] &\le C_{\min}^{-1} \mathbb{E}_{\phi \sim \pi, \tilde{\phi} \sim \rho}\left[(\phi^\top \boldsymbol{E} \tilde{\phi})^2\right] \\ &= C_{\min}^{-1} \mathbb{E}_{\phi \sim \pi, \tilde{\phi} \sim \rho}\left[\langle \mathrm{vec}(\boldsymbol{E}), \mathrm{vec}(\phi \tilde{\phi}^\top) \rangle^2\right] \\ &\le C_{\min}^{-1} \|\mathrm{vec}(\boldsymbol{E})\|_{\boldsymbol{M}}^2 \mathbb{E}_{\phi \sim \pi, \tilde{\phi} \sim \rho}\left[\|\mathrm{vec}(\phi \tilde{\phi}^\top)\|_{\boldsymbol{M}^{-1}}^2\right], \end{aligned}$$

where the last inequality is Cauchy–Schwarz. □

F Proof of Theorem 4.7: Regret Bound of `Explore-Then-Commit`

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

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

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

Here, $m' = \mu$; this is the Bernoulli negative log-likelihood for the logistic link and a convex quasi-likelihood for a general link. For a general link, set $\widehat{\boldsymbol{\Theta}}_{T_0} = \mathbf{0}$ if the minimum is not attained; the event used below guarantees attainment.

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

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

$$|\ddot{\mu}(z)| \leq R_s \dot{\mu}(z), \qquad \forall z \in [-S-1, S+1].$$

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

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

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, which controls the entire score interval, $\kappa_\star$ 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}, with constants depending on R_s.

The following lemma provides a Frobenius error guarantee for the nuclear-norm regularized MLE:

Lemma F.2. *Let $\delta \in (0,1)$, and let $\tilde{\kappa} = \kappa$ or, under Assumption F.1, $\tilde{\kappa} = \kappa_\star$. Suppose that*

$$T_0 \gtrsim C_{\min}^{-2} \log\frac{4d}{\delta} + \tilde{\kappa}^{-2} C_{\min}^{-4} L_\mu r \log\frac{4d}{\delta}.$$

With $\lambda_{T_0} = \sqrt{32 L_\mu \log(4d/\delta)/T_0}$, we have

$$\mathbb{P}\left(\left\|\widehat{\boldsymbol{\Theta}}_{T_0} - \boldsymbol{\Theta}_\star\right\|_F \lesssim \frac{1}{\tilde{\kappa} C_{\min}^2}\sqrt{\frac{L_\mu r \log(4d/\delta)}{T_0}}\right) \geq 1-\delta.$$

In the self-concordant case, the implicit constants may depend on R_s.

Proof. We follow the standard nuclear-norm regularized estimation argument (Fan et al., 2019; Wainwright, 2019), verifying the design and curvature conditions for our same-context comparisons.

1. Population and Empirical Coverage. For any $\Delta \in \mathrm{Skew}(d)$, Item (ii) with $\pi = \rho$, followed by Item (i), gives

$$\begin{aligned}\mathbb{E}\left[\langle \Delta, \boldsymbol{X}_t\rangle^2\right] &\geq C_{\min}\mathbb{E}_{\boldsymbol{x}\sim d_0, \boldsymbol{a}\sim\rho(\cdot|\boldsymbol{x})}\left[\|\Delta\phi(\boldsymbol{x},\boldsymbol{a})\|_2^2\right] \\ &= C_{\min}\,\mathrm{tr}\left(\Delta^\top \Delta\, \mathbb{E}_{\boldsymbol{x},\boldsymbol{a}\sim\rho}\left[\phi(\boldsymbol{x},\boldsymbol{a})\phi(\boldsymbol{x},\boldsymbol{a})^\top\right]\right) \geq C_{\min}^2 \|\Delta\|_F^2.\end{aligned}$$

All expectations here use one shared context for the two actions. In fact, part (i) can be weakened for `ETC` to

$$\mathbb{E}_{\boldsymbol{x},\boldsymbol{a}\sim\rho}\left[\|\Delta\phi(\boldsymbol{x},\boldsymbol{a})\|_2^2\right] \geq C_{\min}\|\Delta\|_F^2, \qquad \forall \Delta \in \mathrm{Skew}(d).$$

The proof only uses this inequality; the covariance formulation in the main text is a sufficient, familiar condition. The projection of $\boldsymbol{X}_t$ onto $\mathrm{Skew}(d)$ is $(\boldsymbol{X}_t - \boldsymbol{X}_t^\top)/2$, whose squared Frobenius norm is at most $1/2$. Thus, matrix Chernoff (Tropp, 2012) on the $d(d-1)/2$-dimensional space $\mathrm{Skew}(d)$ gives, with probability at least $1 - \delta/2$,

$$\frac{1}{T_0}\sum_{t=1}^{T_0}\langle \Delta, \boldsymbol{X}_t\rangle^2 \geq \frac{C_{\min}^2}{2}\,\|\Delta\|_F^2\,, \qquad \forall \Delta \in \mathrm{Skew}(d), \tag{8}$$

provided $T_0 \gtrsim C_{\min}^{-2}\log(4d/\delta)$.

2. Score Bound and Local Curvature. Since r_t is Bernoulli and $\|\boldsymbol{X}_t\|_{\mathrm{op}} \leq 1$, matrix Bernstein (Tropp, 2012), together with $L_\mu \geq 1/4$, gives

$$\|\nabla \mathcal{L}_{T_0}(\boldsymbol{\Theta}_\star)\|_{\mathrm{op}} \leq \lambda_{T_0}/2 \tag{9}$$

with probability at least $1 - \delta/2$, provided $T_0 \gtrsim L_\mu^{-1}\log(4d/\delta)$. We condition on the intersection of these two events. The score bound also makes the penalized objective coercive, since its nuclear-norm penalty has coefficient λ_{T_0} while the linear lower bound on the loss has coefficient at most $\lambda_{T_0}/2$; hence a minimizer exists on this event. For $\Delta \in \mathrm{Skew}(d)$ with $\|\Delta\|_F \leq 1$, Taylor's formula gives

$$\begin{aligned}&\mathcal{L}_{T_0}(\boldsymbol{\Theta}_\star + \Delta) - \mathcal{L}_{T_0}(\boldsymbol{\Theta}_\star) - \langle \nabla \mathcal{L}_{T_0}(\boldsymbol{\Theta}_\star), \Delta\rangle \\ &\quad = \frac{1}{T_0}\sum_{t=1}^{T_0}\langle \Delta, \boldsymbol{X}_t\rangle^2 \int_0^1 (1-z)\dot{\mu}\big(\langle \boldsymbol{\Theta}_\star + z\Delta, \boldsymbol{X}_t\rangle\big)\,dz.\end{aligned}$$

The scores in this integral lie in $[-S-1, S+1]$, since $\|\boldsymbol{\Theta}_\star\|_{\mathrm{op}} \leq S$ and $|\langle \Delta, \boldsymbol{X}_t\rangle| \leq 1$. Without self-concordance, the integral is at least $\kappa/2$. Under Assumption F.1, integrating the bound on the logarithmic derivative of $\dot{\mu}$ gives

$$\int_0^1 (1-z)\dot{\mu}\big(\langle \boldsymbol{\Theta}_\star + z\Delta, \boldsymbol{X}_t\rangle\big)\,dz \geq \kappa_\star \int_0^1 (1-z)e^{-R_s z}\,dz = \kappa_\star \tilde{R}_s,$$

where $\tilde{R}_s = 1/2$ if $R_s = 0$ and $\tilde{R}_s = (R_s - 1 + e^{-R_s})/R_s^2$ otherwise. Set $\tilde{R}_s = 1/2$ when using the global curvature $\tilde{\kappa} = \kappa$. Combining with (8), we have

$$\begin{aligned}&\mathcal{L}_{T_0}(\boldsymbol{\Theta}_\star + \Delta) - \mathcal{L}_{T_0}(\boldsymbol{\Theta}_\star) - \langle \nabla \mathcal{L}_{T_0}(\boldsymbol{\Theta}_\star), \Delta\rangle \\ &\quad \geq \frac{\tilde{R}_s \tilde{\kappa} C_{\min}^2}{2}\,\|\Delta\|_F^2\,, \qquad \|\Delta\|_F \leq 1.\end{aligned} \tag{10}$$

3. Nuclear-Norm Estimation Error. For any $\boldsymbol{\Theta}_\star + \Delta$ whose penalized objective is no larger than that at $\boldsymbol{\Theta}_\star$, the basic inequality, (9), and decomposability of the nuclear norm give

$$\begin{aligned}&\mathcal{L}_{T_0}(\boldsymbol{\Theta}_\star + \Delta) - \mathcal{L}_{T_0}(\boldsymbol{\Theta}_\star) - \langle \nabla \mathcal{L}_{T_0}(\boldsymbol{\Theta}_\star), \Delta\rangle \\ &\quad \leq 3\lambda_{T_0}\sqrt{r}\,\|\Delta\|_F\,.\end{aligned} \tag{11}$$

Indeed, the tangent component of Δ at $\boldsymbol{\Theta}_\star$ has rank at most $2\mathrm{rank}(\boldsymbol{\Theta}_\star) \leq 4r$; its nuclear norm is at most $2\sqrt{r}\,\|\Delta\|_F$, while the orthogonal component has a nonpositive coefficient in the basic inequality. For $\|\Delta\|_F \leq 1$, (10) and (11) imply

$$\|\Delta\|_F \leq \frac{6\lambda_{T_0}\sqrt{r}}{\tilde{R}_s \tilde{\kappa} C_{\min}^2}. \tag{12}$$

The assumed sample size makes the right-hand side strictly smaller than 1. If the unconstrained estimator were outside this unit ball around $\boldsymbol{\Theta}_\star$, convexity of the penalized objective would give a point on the connecting segment with Frobenius distance 1 and objective no larger than that at $\boldsymbol{\Theta}_\star$, contradicting (12). Hence (12) applies to $\Delta = \widehat{\boldsymbol{\Theta}}_{T_0} - \boldsymbol{\Theta}_\star$ and proves the result. □

We now prove Theorem 4.7, suppressing factors depending only on L_μ (and on R_s for the self-concordant refinement). The exploration contribution is exactly $T_0\mathrm{DGap}_\eta(\rho)$. We first invoke the η-dependent bound of Theorem 3.1 and the above lemma for $t \in [[T_0 + 1, T]]$:

$$\text{MBR-Reg}_\eta(T) - T_0\mathrm{DGap}_\eta(\rho) \leq 2\eta\beta L_\mu^2 (T - T_0)\,\|\boldsymbol{E}_{T_0}\|_F^2$$

$$\lesssim \frac{T\eta\beta}{\tilde{\kappa}^2 C_{\min}^4} \frac{r \log \frac{d}{\delta}}{T_0}.$$

The error bound is local, so the estimated scores also lie in $[-S-1, S+1]$, where μ is L_μ-Lipschitz. For $T_0 \asymp \tilde{\kappa}^{-1} C_{\min}^{-2} \sqrt{T\eta\beta r \log(d/\delta)}$, the commitment contribution is $\mathcal{O}(T_0)$, provided this budget meets the estimator's sample-size condition and does not exceed T.

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

$$\begin{aligned} \text{MBR-Reg}_\eta(T) - T_0 \text{DGap}_\eta(\rho) &\leq L_\mu (T - T_0) \, \|\boldsymbol{E}_{T_0}\|_F \\ &\lesssim \frac{T}{\tilde{\kappa} C_{\min}^2} \sqrt{\frac{r \log \frac{d}{\delta}}{T_0}}. \end{aligned}$$

For $T_0 \asymp (\tilde{\kappa}^{-2} C_{\min}^{-4} r T^2 \log(d/\delta))^{1/3}$ satisfying the same budget requirements, this contribution is also $\mathcal{O}(T_0)$. Combining the two bounds and taking $\tilde{\kappa} = \kappa$ proves the theorem. The version with $\kappa_\star$ follows under Assumption F.1. □

Remark F.3 (Unknown T)**.** *We assume T is known to tune T_0. If T is unknown, the standard doubling trick (Auer et al., 1995; Besson and Kaufmann, 2018) applies once the epoch budgets meet the sample-size condition; the regret from earlier epochs is retained as an additive initial cost.*

G Eluder Dimension of GBPM

G.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 G.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*

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

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

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

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

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

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

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

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

We bound the uncertainty via the elliptical potential lemma (Lemma E.4). 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 $\boldsymbol{x}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^{d^2}$,

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

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

For the denominator squared:

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

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

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

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

Combining the above two inequalities, we have that

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

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

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

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

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

We then conclude the proof via the elliptical potential lemma (Lemma E.4):

$$\begin{aligned} d(\mathcal{P}, \lambda, T) &\leq \sum_{t=1}^{T} \min\left\{1, \frac{L_\mu^2}{\kappa^2} \|\operatorname{vec}(\kappa\boldsymbol{\Phi}_t)\|^2_{\boldsymbol{V}_t^{-1}}\right\} \\ &\leq \frac{L_\mu^2}{\kappa^2} \sum_{t=1}^{T} \min\left\{1, \|\operatorname{vec}(\kappa\boldsymbol{\Phi}_t)\|^2_{\boldsymbol{V}_t^{-1}}\right\} && (\kappa \leq L_\mu) \\ &\leq \frac{2d^2 L_\mu^2}{\kappa^2} \log\left(1 + \frac{4\kappa^2 S^2 T}{d^2 \lambda}\right). \end{aligned}$$

□

G.2 Connection to the Standard Eluder Dimensions

In this appendix, we will elucidate the connection between Definition G.1 and two other "standard" definitions of eluder-type complexities: *sequential extrapolation coefficient (SEC)* (Xie et al., 2023) and the original eluder dimension (Russo and 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 G.3 (λ-regularized SEC, Adapted from Definition 7 of Xie et al. (2023))**.** *Let $\Pi \subseteq \Delta(\mathcal{Z})$ be a distribution class. Then, the SEC is defined as*

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

Let $\mathcal{D} := \{\delta_z : z \in \mathcal{Z}\}$ be the class of Dirac measures. Since every $P \in \mathcal{P}$ takes values in $[0, 1]$, the clipping in Definition G.1 is inactive when $\lambda \geq 1$. Consequently,

$$\begin{aligned} d(\mathcal{P}, \lambda, T) &= \sup_{z_{1:T} \in \mathcal{Z}^T} \sup_{P_{1:T}^1, P_{1:T}^2 \in \mathcal{P}^T} \sum_{t=1}^{T} \frac{\left(P_t^1(z_t) - P_t^2(z_t)\right)^2}{\lambda + \sum_{s<t}\left(P_t^1(z_s) - P_t^2(z_s)\right)^2} \\ &= \mathrm{SEC}_\lambda(\mathcal{P}, \mathcal{D}, T). \end{aligned}$$

Second, we recall the standard eluder dimension of Foster et al. (2021) and Li et al. (2022a):[4]

Definition G.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 $t \in [m]$,*

$$|P_t(\boldsymbol{z}_t) - P^*(\boldsymbol{z}_t)| > \varepsilon, \quad \text{and} \quad \sum_{s<t} \left(P_t(\boldsymbol{z}_s) - P^*(\boldsymbol{z}_s)\right)^2 \leq \varepsilon^2.$$

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 $\mathrm{Edim}(\mathcal{P}, \varepsilon) := \sup_{P^* \in \mathcal{P}} \mathrm{Edim}_{P^*}(\mathcal{P}, \varepsilon)$.*

The following proposition transfers an exact eluder-dimension lower bound to the sequential complexities above.

Proposition G.5. *Let $\mathcal{P} \subseteq (\mathcal{Z} \to [0, 1])$, $\varepsilon > 0$, and $\lambda \geq 1$. If $T \geq \underline{\mathrm{Edim}}(\mathcal{P}, \varepsilon)$, then*

$$\frac{\varepsilon^2}{\lambda + \varepsilon^2} \underline{\mathrm{Edim}}(\mathcal{P}, \varepsilon) \leq d(\mathcal{P}, \lambda, T) = \mathrm{SEC}_\lambda(\mathcal{P}, \mathcal{D}, T).$$

Proof. Let $m := \underline{\mathrm{Edim}}(\mathcal{P}, \varepsilon)$. The case $m = 0$ is immediate. Otherwise, take a witness $\{(z_t, P_t)\}_{t=1}^m$ with reference model $P^\star \in \mathcal{P}$. Extend the sequence arbitrarily to length T, and use the pairs $(P_t, P^\star)$ for the first m terms. By the preceding identity,

$$d(\mathcal{P}, \lambda, T) \geq \sum_{t=1}^{m} \frac{(P_t(z_t) - P^\star(z_t))^2}{\lambda + \sum_{s<t}(P_t(z_s) - P^\star(z_s))^2} \geq \frac{m\varepsilon^2}{\lambda + \varepsilon^2}.$$

□

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

Proposition G.6. *Let $\mathcal{X} \subseteq \mathcal{B}^d(1)$, $\Theta \subseteq \mathrm{Skew}(d, 2r; S)$ for some $S > 0$, and $\kappa \leq \dot{\mu}(z) \leq L_\mu$ for all $|z| \leq S$. Define the function class as the **GBPM**:*

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

Then, its eluder dimension is bounded as follows: for any $\varepsilon < SL_\mu$,

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

For the lower bound, suppose additionally that $\mathcal{X} = B^d(1)$, $\Theta = \mathrm{Skew}(d; 2r, S)$, $r \geq 1$, and $0 < S < 1/2$. If $\mu(z) = 1/2 + z$ on $[-S, S]$, then, for $0 < \varepsilon < S/(2\sqrt{3})$,

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

For fixed S, $\lambda \geq 1$, and a sufficiently small constant $\varepsilon > 0$, Proposition G.5 transfers this worst-case $\Omega(d^2)$ dependence to both $d(\mathcal{P}, \lambda, T)$ and $\mathrm{SEC}_\lambda(\mathcal{P}, \mathcal{D}, T)$, provided T is sufficiently large. Thus, low rank alone does not guarantee a reduction in these worst-case sequential complexities. The construction uses rank-two coordinate directions $\boldsymbol{e}_i \boldsymbol{e}_j^\top - \boldsymbol{e}_j \boldsymbol{e}_i^\top$, which can be isolated by successive comparisons, even though the covering-number term can scale with dr.

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

[4] The original definition is due to Russo and 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."

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

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

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, upper bound 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 and Van Roy, 2014, Proposition 3).

G.3 Proof of Proposition G.6

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

Definition G.7 (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*

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

or ∞ if no such d exists.

Proposition G.8 (Proposition 4(ii) of Li et al. (2022a)). *Suppose that $0 < \kappa \le \dot{\mu}(z) \le L_\mu$ for all $|z| \le R$. Then, for every $0 < \varepsilon < R L_\mu$,*

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

We prove the upper and lower bounds separately.

Upper Bound. This follows trivially from adapting Proposition G.8 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 $k := \lfloor \log_4 \frac{S}{2\alpha} \rfloor$. Then, we can first consider the following sequence of length $k+1$: for $t \in \{0\} \cup [k]$,

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

For each t, we have that

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

and

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

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.,

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

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

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

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:*

$$\text{MBR-Reg}_\eta(T) = \mathcal{O}\left(\eta e^{3\eta}(\eta^2 e^{6\eta} + 1)\, d(\mathcal{P}, \lambda, T) \log(N_{\mathcal{P}} T/\delta)\right).$$

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:*

$$\text{MBR-Reg}_\eta(T) \lesssim \eta e^{3\eta}(\eta^2 e^{6\eta} + 1) \cdot \frac{d^2 L_\mu^2}{\kappa^2} \cdot \log T \cdot \left(\log \frac{T}{\delta} + dr \log(LST) \right).$$

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 $P_{\boldsymbol{\Theta}}$ as the preference probability given by ***GBPM*** for parameter $\boldsymbol{\Theta}$. 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 Appendix 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 $\widehat{\boldsymbol{\Theta}}_t := \arg\max_{\boldsymbol{\Theta} \in \mathrm{Skew}(d,2r;S)} \sum_{i \in [t]} \ell_i(\boldsymbol{\Theta})$, where $\ell_i(\boldsymbol{\Theta}) := (r_i \log P_{\boldsymbol{\Theta}}(\boldsymbol{z}_i) + (1-r_i)\log(1 - P_{\boldsymbol{\Theta}}(\boldsymbol{z}_i)))$. Suppose that $\ell_i(\cdot)$ is L-Lipschitz w.r.t. the Frobenius norm. Then, for any $\delta \in (0,1)$ the following holds:*

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

With this lemma and our eluder dimension bound (Proposition G.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*

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

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

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

Then,

$$\begin{aligned}
&\text{MBR-Reg}(T) = \sum_{t=1}^{T} \max_{\pi \in \Pi} \left(\frac{1}{2} - J(\hat{\pi}_t^1, \pi) \right) \\
&= \sum_{t=1}^{T} \max_{\pi \in \Pi} \left(\frac{1}{2} - \left(J(\hat{\pi}_t^1, \pi) - \eta^{-1} D_{\mathrm{KL}}(\hat{\pi}_t^1, \pi_{\mathrm{ref}}) + \eta^{-1} D_{\mathrm{KL}}(\pi, \pi_{\mathrm{ref}}) \right) - \eta^{-1} D_{\mathrm{KL}}(\hat{\pi}_t^1, \pi_{\mathrm{ref}}) + \eta^{-1} D_{\mathrm{KL}}(\pi, \pi_{\mathrm{ref}}) \right) \\
&\le \sum_{t=1}^{T} \max_{\pi \in \Pi} \left(\frac{1}{2} - J_\eta(\hat{\pi}_t^1, \pi) \right) + \sum_{t=1}^{T} \max_{\pi \in \Pi} \left(-\eta^{-1} D_{\mathrm{KL}}(\hat{\pi}_t^1, \pi_{\mathrm{ref}}) + \eta^{-1} D_{\mathrm{KL}}(\pi, \pi_{\mathrm{ref}}) \right) \\
&\le \text{MBR-Reg}_\eta(T) + \sum_{t=1}^{T} \max_{\pi \in \Pi} \eta^{-1} D_{\mathrm{KL}}(\pi, \pi_{\mathrm{ref}}) \\
&= \text{MBR-Reg}_\eta(T) + \eta^{-1} D_{\mathrm{ref}} T.
\end{aligned}$$

□

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*

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

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

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

Then we have the following lemma, whose proof is deferred to Appendix H.3:

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

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

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

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

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,

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

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:

$$\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{t}(P_{\widehat{\boldsymbol{\Theta}}_t}(\boldsymbol{z}_i) - P_{\boldsymbol{\Theta}_\star}(\boldsymbol{z}_i))^2 - \sum_{i=1}^{t}(P_{\widehat{\boldsymbol{\Theta}}_t}(\boldsymbol{z}_i) - P_{\widehat{\boldsymbol{\Theta}}_{\varepsilon,t}}(\boldsymbol{z}_i))^2 \\
&= \log\frac{|\Theta_\varepsilon|}{\delta} + \sum_{i=1}^{t}\log\frac{P_{\boldsymbol{\Theta}_\star}(r_i \mid \boldsymbol{z}_i)}{P_{\widehat{\boldsymbol{\Theta}}_t}(r_i \mid \boldsymbol{z}_i)} + \sum_{i=1}^{t}\log\frac{P_{\widehat{\boldsymbol{\Theta}}_t}(r_i \mid \boldsymbol{z}_i)}{P_{\widehat{\boldsymbol{\Theta}}_{\varepsilon,t}}(r_i \mid \boldsymbol{z}_i)} \\
&\le \log\frac{|\Theta_\varepsilon|}{\delta} + \sum_{i=1}^{t}\log\frac{P_{\widehat{\boldsymbol{\Theta}}_t}(r_i \mid \boldsymbol{z}_i)}{P_{\widehat{\boldsymbol{\Theta}}_{\varepsilon,t}}(r_i \mid \boldsymbol{z}_i)}.
\end{aligned}$$

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

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

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

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

Since we assumed that $\log P_{\boldsymbol{\Theta}}$ is L-Lipschitz, it follows that $P_{\boldsymbol{\Theta}}$ is also L-Lipschitz.[5] Therefore,

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

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 and Plan (2011)). *Let* $\Theta(d, 2r; S) \coloneqq \{\boldsymbol{X} \in \mathbb{R}^{d\times d} \mid \|\mathbf{X}\|_F \le S, \operatorname{rank}(\boldsymbol{X}) \le 2r\} \supseteq \operatorname{Skew}(d, 2r; S)$. *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:

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

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

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

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

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

which concludes the proof. □

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

H.3 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

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

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

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

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

We continue to lower-bound the Hellinger distance by

$$\begin{aligned}
\sum_{i=1}^{t}\left(H\left(P_{\boldsymbol{\Theta}}(r_i \mid \boldsymbol{z}_i) \,\|\, P_{\boldsymbol{\Theta}_\star}(r_i \mid \boldsymbol{z}_i)\right)\right)^2 &\ge \sum_{i=1}^{t}\left(\mathrm{TV}\left(P_{\boldsymbol{\Theta}}(r_i \mid \boldsymbol{z}_i) \,\|\, P_{\boldsymbol{\Theta}_\star}(r_i \mid \boldsymbol{z}_i)\right)\right)^2 \\
&= \sum_{i=1}^{t}\left(P_{\boldsymbol{\Theta}}(\boldsymbol{z}_i) - P_{\boldsymbol{\Theta}_\star}(\boldsymbol{z}_i)\right)^2,
\end{aligned}$$

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

Then, by the martingale exponential inequality (Lemma H.8, recalled after the proof), we obtain for each $\boldsymbol{\Theta} \in \Theta$, with probability at least $1 - \delta$,

$$\begin{aligned}
\sum_{i=1}^{t} \log \frac{P_{\boldsymbol{\Theta}}(r_i \mid \boldsymbol{z}_i)}{P_{\boldsymbol{\Theta}_\star}(r_i \mid \boldsymbol{z}_i)} &\le \log\left(\frac{1}{\delta}\right) + \log \mathbb{E} \exp\left(\sum_{i=1}^{t} \log \frac{P_{\boldsymbol{\Theta}}(r_i \mid \boldsymbol{z}_i)}{P_{\boldsymbol{\Theta}_\star}(r_i \mid \boldsymbol{z}_i)}\right) \\
&\le -\sum_{i=1}^{t} H\big(P_{\boldsymbol{\Theta}}(r_i \mid \boldsymbol{z}_i) \,\|\, P_{\boldsymbol{\Theta}_\star}(r_i \mid \boldsymbol{z}_i)\big)^2 + \log\left(\frac{1}{\delta}\right) \\
&\le -\sum_{i=1}^{t}\left(P_{\boldsymbol{\Theta}}(\boldsymbol{z}_i) - P_{\boldsymbol{\Theta}_\star}(\boldsymbol{z}_i)\right)^2 + \log\left(\frac{1}{\delta}\right).
\end{aligned}$$

□

Lemma H.8 (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 $\lambda > 0$:*

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

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

I Comparison with Prior Work

I.1 Existing Regret Guarantees

Here we compare existing regularized max-regret guarantees with our results under the ***GBPM*** framework. For the bound of Wu et al. (2025), instantiated to ***GBPM*** in Appendix H, we define $f(\eta) := \eta e^{3\eta}(1 + \eta^2 e^{6\eta})$.

Table 1: Comparison of regularized max-regret guarantees with prior work. All bounds are expressed in $\widetilde{O}$ notation, suppressing logarithmic factors in T, d, and $1/\delta$, while retaining polynomial dependence on model parameters.

Work	Preference model	Regularizer	Coverage	Max-regret bound
Zhao et al. (2025b)	Reward-based (BTL)	Reverse KL	Not required	$\eta\kappa^{-2}d^2$
Wu et al. (2025), GS	GBPM (instantiated)	Reverse KL	Not required	$f(\eta)\kappa^{-2}L_\mu^2 d^2(1+dr)$
Ours (GS)	GBPM	Any strongly convex	Comparison coverage	$\eta\beta\kappa^{-1}d^4C_{\min}^{-1} \wedge \kappa^{-1/2}d^2C_{\min}^{-1/2}\sqrt{T}$
Ours (ETC)	GBPM, rank $2r$	Any strongly convex	Both coverage conditions	$\kappa^{-1}C_{\min}^{-2}\sqrt{\eta\beta rT} \wedge (\kappa^{-2}C_{\min}^{-4}rT^2)^{1/3}$

I.2 Discussions

The results in the above table highlight several trade-offs compared with prior work. First, Zhao et al. (2025b) study reward-based preferences and are included for context rather than direct comparison. Among results for general preferences, Wu et al. (2025) achieve better dimension dependence than our `GS` bound without requiring an explicit feature-coverage assumption. However, their analysis is restricted to reverse KL regularization and incurs exponential dependence on η, whereas our bounds apply to any β^{-1}-strongly convex regularizer. Under comparison coverage, our `GS` bound has only linear dependence on η in its polylogarithmic branch and additionally admits an η-independent $\widetilde{O}(\kappa^{-1/2}d^2C_{\min}^{-1/2}\sqrt{T})$ bound, which remains meaningful as $\eta \to \infty$. Moreover, its explicit dependence on κ improves from κ^{-2} in our ***GBPM*** instantiation of Wu et al. (2025) to κ^{-1}, despite under stronger coverage assumptions.

Finally, `ETC` exploits low-rank structure to obtain rank-dependent regret guarantees, at the cost of requiring both coverage conditions and an admissible exploration budget. The simplified bounds assume $\mathrm{DGap}_\eta(\rho) = O(1)$, with the quadratic-error branches requiring finite η. Overall, our results establish the generality of fast regret under strongly convex regularization, rather than uniformly improving prior bounds across all parameters.

J Discussions on Regrets

J.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 MBR-Reg_η, and so in this Appendix, we provide the deferred discussions regarding four regrets:

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

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

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

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

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

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

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

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

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

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. 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.

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 with respect to 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. $\text{AN-Reg}(T)$ is the notion originally considered in the seminal work of Freund and Schapire (1999), followed by numerous works on no-regret learning dynamics in games (Daskalakis et al., 2011, 2015, 2018; Rakhlin and Sridharan, 2013a,b; Syrgkanis et al., 2015), recently adopted to game-theoretic LLM alignment (Zhang et al., 2025c). This is closely related to the regret used in contextual dueling bandits (Dudík et al., 2015) and dueling RL (Saha et al., 2023), where a common comparator is evaluated against both selected responses. Indeed, for any $\pi \in \Pi$, antisymmetry gives

$$J(\pi, \hat{\pi}_t^2) - J(\hat{\pi}_t^1, \pi) = J(\pi, \hat{\pi}_t^1) + J(\pi, \hat{\pi}_t^2) - 1.$$

Unlike this common-comparator formulation, AN-Reg allows the two comparators to be chosen independently and therefore upper-bounds the corresponding common-comparator quantity. 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 (Bengs et al., 2021, 2022; Saha, 2021).

On the other hand, $\text{ABR-Reg}(T)$ strongly resembles the notion of *best response regret* in adversarial dueling bandits (Saha and Krishnamurthy, 2022, Eqn. (1)), but there is a key difference. In Saha and 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 and Jin, 2020; Bai et al., 2020; Jin et al., 2022; Liu et al., 2021; Xiong et al., 2022); this idea has been recently applied to theoretical analyses of online RLHF under general preference (Wu et al., 2025; Ye et al., 2024). 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 and Jin, 2020; Bai et al., 2020; Jin et al., 2022; Liu et al., 2021; Xiong et al., 2022, 2024; Ye et al., 2024).

J.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 J.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 $\bar{\pi}_T^i := \frac{1}{T}\sum_{t=1}^T \hat{\pi}_t^i$ for $i \in \{1,2\}$ and for simplicity, 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 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. For brevity, write $\psi(\pi) := \mathbb{E}_{x\sim d_0}[\psi(\pi(\cdot \mid x))]$. The approximation guarantees below are expressed in terms of the displayed dual gaps.

(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, $J_\eta(\bar{\pi}_T^1,\pi^2) \ge \frac{1}{T}\sum_{t=1}^T J_\eta(\hat{\pi}_t^1,\pi^2)$. Subtracting the two inequalities and taking $\max_{\pi^1,\pi^2}$ yields

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

(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

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

□

K Synthetic Experiments

In this section, we present empirical results to numerically validate the theoretical regret bounds of `Greedy Sampling (GS)` established in Section 4.3.

K.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.

K.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 $\|\mathbf{\Theta}_\star\|_F = S$.

Equilibrium 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 configurations, executes to completion within a few hours. Full code can be found in `https://github.com/minju-hong/online-rl-alignment.git`.

K.3 Main Results

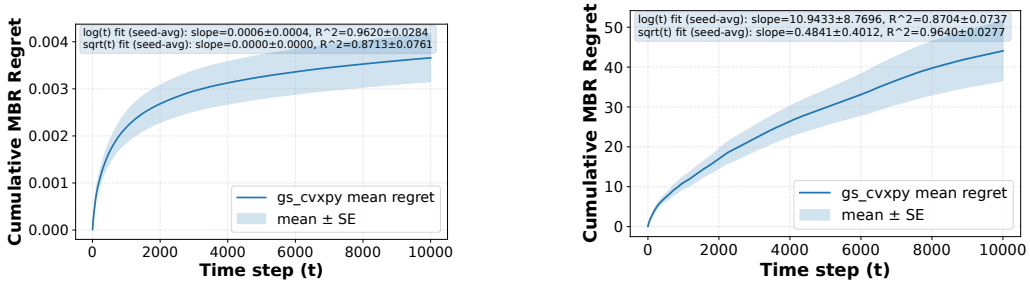
We first plot the cumulative MBR-Reg for varying regularization strengths η. As illustrated in Figure 1, the regret tightly fits a $\log T$ curve for small η and shifts to a $\sqrt{T}$ curve for large η, corroborating the bounds established in Theorem 4.5.

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 4.5 gives the theoretical crossover

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

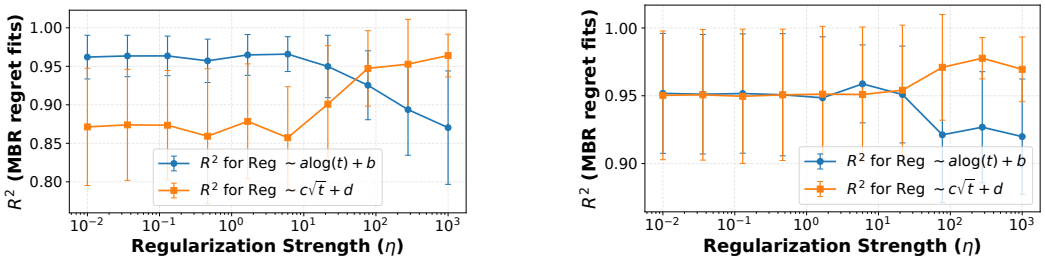
Thus, for $\eta \gtrsim \eta_{\text{cross}}(d, T)$, the $\sqrt{T}$ term is selected by the minimum, whereas for $\eta \lesssim \eta_{\text{cross}}(d, T)$, the logarithmic term is selected. Ignoring constants and the mild logarithmic dependence, this threshold scales as $d^{-2}\sqrt{T}$ when $C_{\min}$ 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.

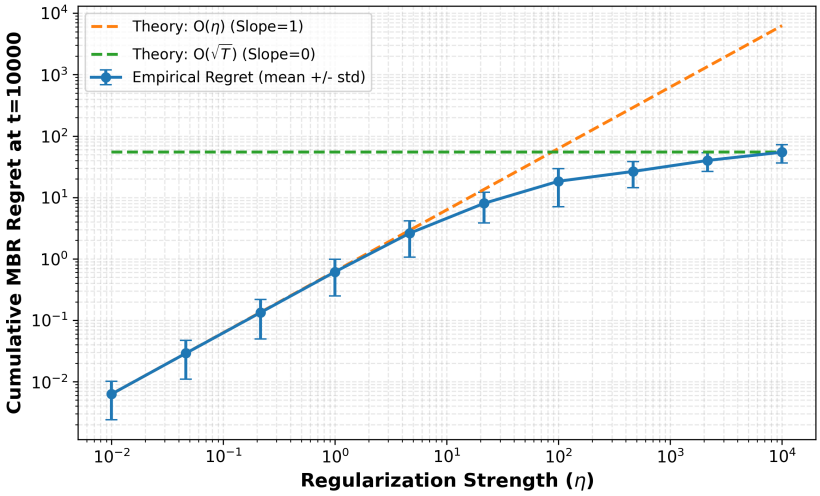


Figure 3: Final Regret Scaling.

L Future Directions

Relaxing the Feature Coverage Conditions. Our regret bounds rely on the feature-coverage conditions in Assumption 4.3, parameterized by $C_{\min}$. 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 and 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 4.5), 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 (Abernethy et al., 2015; Tsallis, 1988; Zimmert and 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 (Kato and Ito, 2025; Kuroki et al., 2024) 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., 2025b), minimax optimization techniques such as optimistic OMD (Rakhlin and Sridharan, 2013a; Syrgkanis et al., 2015; Zhang et al., 2025c), or reductions to offline/online regression oracles (Foster and Rakhlin, 2020).

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We present the quadratic bound on regret in Section 3, and the algorithm specific regret bounds in Section 4.3 and Section 4.4.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Limitations of the work and following future research directions are detailed in Appendix L.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: We state and explain the assumptions in Section 4.2, and provide the full proofs in Section 3, Appendix E and F.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We detail the experiment setup and implementation in Appendix K.1 and K.2.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

Answer: [Yes]

Justification: We provide an anonymous repository containing the full code and data generation scripts, alongside comprehensive instructions necessary to reproduce the main experimental results of the paper.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: We detail the model settings and hyperparameter choices in Appendix K.1, and implementation details in Appendix K.2.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: All reported empirical regret trajectories and metrics are averaged across 20 independent random seeds. We explicitly plot shaded regions and error bars denoting one standard deviation to illustrate the statistical variance and robustness of the algorithms across runs. See Appendix K.3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

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

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We detail the computational requirements in Appendix K.2. Because our empirical evaluations are synthetic and rely on convex optimization solvers, all experiments execute within a few hours on standard consumer-grade CPU hardware.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The research conducted in this paper conforms to the NeurIPS Code of Ethics in all respects.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

Justification: This paper presents foundational, theoretical research on online RLHF. As the contributions are purely mathematical and methodological, there are no direct or foreseeable societal impacts to discuss.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A]

Justification: The contributions of this work are primarily theoretical in nature, and therefore does not possess such risks.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: This paper does not utilize existing external datasets, pre-trained models, or proprietary assets. All data used in our empirical evaluations consists of purely synthetic datasets created specifically for this work. Standard open-source mathematical and machine learning libraries used for these generations and function implementations are cited in the manuscript.

Guidelines:

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

13. **New assets**

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

    Answer: [Yes]

    Justification: We release an anonymous repository containing the code and synthetic data generation scripts used in this paper. The repository includes a README with documentation detailing setup, dependencies, and instructions to reproduce the experimental results.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

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

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

    Answer: [N/A]

    Justification: This work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

Answer: [N/A]

Justification: This work does not involve crowdsourcing nor research with human subjects.

Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: While LLMs were utilized as productivity tools for drafting text, writing experimental code, and assisting with mathematical derivations (as declared in the submission portal), they do not form an important, original, or non-standard component of the core mathematical methodology proposed in this work.

Guidelines:

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