Towards a Sharp Analysis of Offline Policy Learning for f-Divergence-Regularized Contextual Bandits

Qingyue Zhao [*1] **Kaixuan Ji** [*1] **Heyang Zhao** [*1] **Tong Zhang** [2] **Quanquan Gu** [1]

Abstract

Although many popular reinforcement learning algorithms are underpinned by f-divergence regularization, their sample complexity with respect to the *regularized objective* still lacks a tight characterization. In this paper, we analyze f-divergence-regularized offline policy learning. For reverse Kullback–Leibler (KL) divergence, arguably the most commonly used one, we give the first $\widetilde{O}(\epsilon^{-1})$ sample complexity under single-policy concentrability for contextual bandits, surpassing existing $\widetilde{O}(\epsilon^{-1})$ bound under all-policy concentrability and $\widetilde{O}(\epsilon^{-2})$ bound under single-policy concentrability. Our analysis for general function approximation leverages the principle of pessimism in the face of uncertainty to refine a mean-value-type argument to its extreme. This in turn leads to a novel moment-based technique, effectively bypassing the need for uniform control over the discrepancy between any two functions in the function class. We further propose a lower bound, demonstrating that a multiplicative dependency on single-policy concentrability is necessary to maximally exploit the strong convexity of reverse KL. In addition, for f-divergences with strongly convex f, to which reverse KL *does not* belong, we show that the sharp sample complexity $\widetilde{\Theta}(\epsilon^{-1})$ is achievable even without single-policy concentrability. In this case, the algorithm design can get rid of pessimistic estimators. We further extend our analysis to dueling bandits, and we believe these results take a significant step toward a comprehensive understanding of f-divergence-regularized policy learning.

*Equal contribution [1]Department of Computer Science, University of California, Los Angeles, CA 90095, USA [2]University of Illinois Urbana-Champaign, Urbana, IL 61801, USA. Correspondence to: Quanquan Gu <qgu@cs.ucla.edu>.

ICML 2025 Workshop on Models of Human Feedback for AI Alignment, Vancouver, Canada. PMLR 267, 2025.

1 Introduction

Due to the data-hungry and instable nature of reinforcement learning (RL), divergences that are straightforward to estimate via Monte Carlo or amenable to constrained optimization stand out from numerous candidates (Rényi, 1961; Csiszár, 1967; Müller, 1997; Basseville, 2013) as regularizers; the former family is typically *f-divergence* (Rényi, 1961) because any of them is an expectation, for which empirical average is a good proxy (Levine, 2018; Levine et al., 2020); and the latter class subsumes those with nice positive curvatures (e.g., Bregman divergence (Bregman, 1967) induced by strongly convex functions). In particular, *Kullback-Leibler (KL) divergence* is the only one at the intersection of f-divergence and Bregman divergence (Jiao et al., 2014, Theorem 5), indicating its theoretical advantage among common choices from both computational and statistical aspects. Also, the *KL-regularized RL objective* is arguably the most popular one in practice:

$$J(\pi) = \mathbb{E}_\pi[r] - \eta^{-1}\mathsf{KL}(\pi\|\pi^{\mathsf{ref}}), \tag{1.1}$$

where r is the reward, π^{ref} is a reference policy, $\mathsf{KL}(\pi\|\pi^{\mathsf{ref}})$ is the reverse KL divergence, and $\eta > 0$ is the inverse temperature. When π^{ref} is uniform, (1.1) reduces to the entropy-regularized objective that encourages diverse actions and enhances robustness (Williams, 1992; Ziebart et al., 2008; Levine & Koltun, 2013; Levine et al., 2016; Haarnoja et al., 2018; Richemond et al., 2024; Liu et al., 2024). KL regularization has also been widely used in the RL fine-tuning of large language models (Ouyang et al., 2022; Rafailov et al., 2023), where π^{ref} is the base model. Given its widespread use, there has been a surge of interest in understanding the role of KL regularization in RL by both empirical studies (Ahmed et al., 2019; Liu et al., 2019) and theoretical analysis (Geist et al., 2019; Vieillard et al., 2020; Kozuno et al., 2022). There are also lines of research on KL regularization in online learning (Cai et al., 2020; He et al., 2022; Ji et al., 2023) and convex optimization (Neu et al., 2017). However, the performance metric in most of these studies is still the unregularized reward maximization objective, under which the sample complexity is at least $\Omega(\epsilon^{-2})$. (See Appendix A.1 for detailed discussions)

Several recent papers (Xiong et al., 2024; Xie et al., 2024; Zhao et al., 2024; Foster et al., 2025) switched the focus

Table 1. Comparison of sample complexity bounds for finding ϵ-optimal policy for offline contextual bandits with KL- and (strongly convex) f-divergence regularization. Constants and polylog factors are omitted here except the metric entropy $\log \mathcal{N}$. "Reverse-KL" stands for KL-regularized contextual bandits and "f-divergence w/ s.c., f" for the counterpart with an α-strongly convex f. The two existing upper bounds are adapted from the implicit form in Xiong et al. (2024, Theorem 3.1) and Zhao et al. (2024, Theorem 3.3 and Theorem 4.4.), of which the detailed adaptions are deferred to Appendix A. Also see Remark 2.8 for discussion on the relationship between $D^2_{\pi^*}$ and C^{π^*}

Regularizer		Xiong et al. (2024)	Zhao et al. (2024)	This work
Reverse KL	Upper	$d\epsilon^{-2}$	$\eta D^2 \epsilon^{-1} \log \mathcal{N}$	$\eta D^2_{\pi^*} \epsilon^{-1} \log \mathcal{N}$
	Lower	-	$\eta \epsilon^{-1} \log \mathcal{N}$	$\eta C^{\pi^*} \epsilon^{-1} \log \mathcal{N}$
f-divergence w/ s.c. f	Upper	-	-	$\alpha^{-1} \eta \epsilon^{-1} \log \mathcal{N}$
	Lower	-	-	$\alpha^{-1} \eta \epsilon^{-1} \log \mathcal{N}$

to analyzing the suboptimality defined via the regularized objective (1.1), under which an $\Omega(\epsilon^{-1})$ sample complexity is possible (Zhao et al., 2024, Theorem 3.6). However, even restricted to the pure i.i.d. setting, existing analyses in this vein either result in still $\widetilde{O}(\epsilon^{-2})$ bounds (Xiong et al., 2024; Xie et al., 2024) or has a stringent all-policy concentrabiity dependency in their upper bounds (Zhao et al., 2024) (See Section 2.1 for details on coverage conditions). Thus, there are by far no perfectly matching bounds down to both the dependency of ϵ^{-1} and concrete coverage conditions for KL-regularized offline decision making. Moreover, these theoretical analyses set KL as the right target by default; but reverse KL is the f-divergence with $f(x) = x \log x$, which is merely convex. Therefore, it is also unknown whether those choices with even nicer (e.g., strongly convex) f, whose performance against the reward maximization objective can be provably promising (Zhan et al., 2022; Gabbianelli et al., 2024; Huang et al., 2024), can enjoy a better coverage dependency in the sample complexity under the corresponding regularized objectives. These perspectives motivate an open problem:

What is the weakest coverage condition for offline *policy learning to be minimax optimal with respect to* f-divergence-regularized objectives?

We are the first to thoroughly investigate this question by proposing sharp sample complexity analyses for two representative ad subclasses of f-divergence. In particular, for contextual bandits with KL regularization, we present sample complexity guarantee with linearly dependence on *single*-policy coverage ratio. Our lower bound further indicates that a multiplicative dependency on *single*-policy concentrability is necessary. Surprisingly, for f-divergence with α-strongly-convex f, we prove nearly matching sample complexity bounds of $\widetilde{\Theta}(\alpha^{-1}\eta\epsilon^{-1})$ for the first time, settling the dependence on coverage completely. For the ease of comparison, we adapt previous sample complexity bounds to our setting and notation, and summarize them in Table 1.

1.1 Contributions

- For arguably the most commonly used KL regularization, we are the first to employ the principle of pessimism (Jin et al., 2021) to achieve a nearly tight sample complexity under *single*-policy concentrability. Our analysis roadmap leverages the strong convexity of KL and pessimism of the reward estimator, to refine a mean-value-type risk upper bound to its extreme which in turn leads to a novel moment-based analysis, effectively bypassing the need for uniform control over the discrepancy between any two functions in the function class.
- Under KL-regularization, we prove a lower bound growing linearly with the single-policy coverage coefficient, which also strictly improved upon previous works (Zhao et al., 2024; Foster et al., 2025) in offline setting. This lower bound manifests the necessity of a multiplicative dependency on single-policy concentrability and firstly closes the gap between upper and lower bounds under a technically nontrivial case.
- For any f-divergence with strongly convex f, we are able to design a truly lightweight algorithm free of pessimism-based gadgets and still obtain the $\widetilde{\Theta}(\epsilon^{-1})$ sample complexity certified by a matching lower bound without coverage conditions.
- We extend all algorithmic ideas and hard instance constructions above to f-divergence-regularized contextual dueling bandits, achieving $\widetilde{\Theta}(\epsilon^{-1})$ sample complexity bounds. Moreover, all algorithms work for general reward function classes with finite metric entropy.

1.2 Additional Related Work

We review two additional lines of theoretical progress that are closely related to our algorithm design and analysis.

Pessimism in offline RL. The principle of pessimism has been underpinning offline RL for both the tabular (Rashidinejad et al., 2021) and function approximation (Jin et al., 2021) settings under the name of lower confidence bound (LCB). For contextual bandits, it is behind the adaptively optimal sample complexity analysis (Li et al., 2022). Shi et al. (2022) proposed a LCB-based model-free algorithm for tabular RL with near-optimal guarantee. Jin et al. (2021); Xiong et al. (2022); Di et al. (2023) utilized LCB in conjunction with the classic least-square value iteration paradigm to derive $\widetilde{O}(\epsilon^{-2})$ sample complexity results for model-free RL with function approximation. The line of work from Rashidinejad et al. (2021); Xie et al. (2021b) to Li et al. (2024) settled the sample complexity of tabular model-based RL via pessimistic estimators exploiting the variance information. It is also possible to leverage the idea of pessimism to design model-based algorithms under general function approximation that are at least statistically efficient (Xie et al., 2021a; Uehara & Sun, 2021; Wang et al., 2024).

However, in terms of risk decomposition, to the best of our knowledge, none of these pessimism-based analyses really goes beyond the performance difference lemma (Foster & Rakhlin, 2023, Lemma 13) or simulation lemma (Foster & Rakhlin, 2023, Lemma 23); both of which are not able to capture the strong concavity of KL-regularized objectives even in the bandit setting. The algorithmic idea of using pessimistic least-square estimators under general function approximation in Jin et al. (2021); Di et al. (2023) is similar to ours, but their suboptimality gap is bounded by the sum of bonuses, which cannot lead to the desired sample complexity of our objective by direct adaption.

Offline contextual dueling bandits. CDBs (Dudík et al., 2015) is the contextual extension of dueling bandits in classic literature of online learning from pairwise comparisons (Yue et al., 2012; Zoghi et al., 2014). Since the empirical breakthrough of preference-based RL fine-tuning of LLMs (Ouyang et al., 2022), the theory of offline CDBs has received more attention under linear (Zhu et al., 2023; Xiong et al., 2024) and general (Zhan et al., 2022; Zhao et al., 2024; Song et al., 2024; Huang et al., 2024) function approximation. Preference models without stochastic transitivity (Munos et al., 2023; Ye et al., 2024; Wu et al., 2024; Zhang et al., 2024) are beyond the scope of this work, viz., our preference labels are assumed to follow the Bradley-Terry Model (Bradley & Terry, 1952).

Notation. The sets $\mathcal{S}$ and $\mathcal{A}$ are assumed to be countable throughout the paper. For nonnegative sequences $\{x_n\}$ and $\{y_n\}$, we write $x_n = O(y_n)$ if $\limsup_{n\to\infty} x_n/y_n < \infty$, $y_n = \Omega(x_n)$ if $x_n = O(y_n)$, and $y_n = \Theta(x_n)$ if $x_n = O(y_n)$ and $x_n = \Omega(y_n)$. We further employ $\widetilde{O}(\cdot), \widetilde{\Omega}(\cdot)$, and $\widetilde{\Theta}$ to hide polylog factors. For countable $\mathcal{X}$ and $\mathcal{Y}$, we denote the family of probability kernels from $\mathcal{X}$ to $\mathcal{Y}$ by $\Delta(\mathcal{Y}|\mathcal{X})$. For $g : \mathcal{X} \to \mathbb{R}$, its infinity norm is denoted by $\|g\|_\infty := \sup_{x\in\mathcal{X}} |g(x)|$. For a pair of probability measures $P \ll Q$ on the same space and function $f : \mathbb{R}_+ \to \mathbb{R}$, their f-divergence is $D_f(P\|Q) := \int f(\mathrm{d}P/\mathrm{d}Q)\,\mathrm{d}Q$. Specifically, when $f(x) = x\log x$, f-divergence becomes KL divergence denoted as $\mathsf{KL}(P\|Q) := \int \log(\mathrm{d}P/\mathrm{d}Q)\,\mathrm{d}P$, and when $f(x) = |x-1|/2$, it becomes the total variation (TV) distance, which is denoted as $\mathsf{TV}(P\|Q) := 0.5\int |\mathrm{d}P - \mathrm{d}Q|$.

2 KL-regularized Contextual Bandits

In this section, we propose a pessimism-based algorithm, dubbed PCB-KL, for offline KL-regularized contextual bandits. In the following subsections, we also showcase our key technical novelty which couples the pessimism of our reward estimator to the non-trivial curvature property of KL regularization.

2.1 Problem Setup

We consider contextual bandit, which is denoted by a tuple $(\mathcal{S}, \mathcal{A}, r, \pi^{\mathsf{ref}})$. Specifically, $\mathcal{S}$ is the context space, $\mathcal{A}$ is the action space and $r : \mathcal{S}\times\mathcal{A} \to [0,1]$ is the reward function. In the offline setting, the agent only has access to an i.i.d. dataset $\mathcal{D} = \{(s_i, a_i, r_i)\}_{i=1}^n$. Here $s_i's$ are states sampled from $\rho \in \Delta(\mathcal{S})$, $a_i \in \mathcal{A}$ is the action taken from a *behavior policy*, and r_i is the observed reward given by $r_i = r(s_i, a_i) + \varepsilon_i$, where ε_t is 1-sub-Gaussian (Lattimore & Szepesvári, 2020, Definition 5.2). In this work, we consider the KL-regularized objective

$$J(\pi) := \mathbb{E}_{(s,a)\sim\rho\times\pi}\left[r(s,a) - \eta^{-1}\log\frac{\pi(a|s)}{\pi^{\mathsf{ref}}(a|s)}\right], \quad (2.1)$$

where π^{ref} is a known reference policy and the "inverse temperature" η controls the intensity of regularization. For simplicity, we assume that π^{ref} is also the behavior policy that generates the dataset $\mathcal{D}$. This type of "behavior regularization" has been also been studied in Zhan et al. (2022). The unique optimal policy $\pi^* := \mathrm{argmax}_{\pi\in\Delta(\mathcal{A}|\mathcal{S})} J(\pi)$ is given by (See, e.g., Zhang 2023, Proposition 7.16)

$$\pi^*(\cdot|s) \propto \pi^{\mathsf{ref}}(\cdot|s)\exp\big(\eta\cdot r(s,\cdot)\big), \forall s \in \mathcal{S}. \quad (2.2)$$

A policy π is said to be ϵ-optimal if $J(\pi^*) - J(\pi) \le \epsilon$ and the goal of the agent is to find one such policy based on $\mathcal{D}$. To ensure ϵ-optimality is achievable, we assume that r lies in a known function class $\mathcal{G} \subset (\mathcal{S}\times\mathcal{A} \to [0,1])$, from which the agent obtains the estimator $\widehat{r}$. More specifically, we work with general function approximation under realizability, which is as follows.

Assumption 2.1. For this known function class $\mathcal{G} \subset (\mathcal{S}\times\mathcal{A} \to [0,1])$, $\exists g^* \in \mathcal{G}$ with $g^* = r$.

We also need a standard condition to control the complexity of $\mathcal{G}$ through the notion of covering number (Wainwright, 2019, Definition 5.1).

Definition 2.2 (ϵ-net and covering number)**.** Given a function class $\mathcal{G} \subset (\mathcal{S} \times \mathcal{A} \to \mathbb{R})$, a finite set $\mathcal{G}(\epsilon) \subset \mathcal{G}$ is an ϵ-net of $\mathcal{G}$ w.r.t. $\|\cdot\|_\infty$, if for any $g \in \mathcal{G}$, there exists $g' \in \mathcal{G}(\epsilon)$ such that $\|g - g'\|_\infty \le \epsilon$. The ϵ-covering number is the smallest cardinality $\mathcal{N}_\mathcal{G}(\epsilon)$ of such $\mathcal{G}(\epsilon)$.

Assumption 2.3. For any $\epsilon_c > 0$, the ϵ_c-covering number $\mathcal{N}_\mathcal{G}(\epsilon_c)$ of $\mathcal{G}$ is $\text{poly}(\epsilon_c^{-1})$.

When $\mathcal{G}$ is the class of linear functions of dimension d and radius R, the covering number of $\mathcal{G}$ is given by $\mathcal{N}_\mathcal{G} = O((1 + R\epsilon^{-1})^d)$ (Jin et al., 2020, Lemma D.6), indicating that the Assumption 2.3 on $\mathcal{N}_\mathcal{G}$ is mild.

Concentrability. The data quality of $\mathcal{D}$ collected by π^{ref} is typically characterized by *concentrability* in offline RL (Farahmand et al., 2010; Chen & Jiang, 2019; Jiang & Xie, 2024), which quantifies the ability of the behavioral policy to generate diverse actions. We first define the density-ratio-based concentrability as follows.

Definition 2.4 (*Density-ratio-based* concentrability)**.** For policy class Π, reference policy π^{ref}, the density-ratio-based all-policy concentrability C^Π is $C^\Pi := \sup_{\pi \in \Pi, s \in \mathcal{S}, a \in \mathcal{A}} \pi(a|s)/\pi^{\text{ref}}(a|s)$, whose single-policy counterpart under the optimal policy π^* is $C^{\pi^*} := \sup_{s \in \mathcal{S}, a \in \mathcal{A}} \pi^*(a|s)/\pi^{\text{ref}}(a|s)$.

In the definition above, small all-policy concentrability intuitively corresponds to supp(π^{ref}) covering all possible inputs. On the other hand, small single-policy concentrability means that supp(π^{ref}) only subsumes supp(π^*). In this paper, in addition to density-ratio-based concentrability, we also adopt the following D^2-based concentrabilites to better capturing the nature of function class $\mathcal{G}$. In detail, we first introduce D^2-divergence as follows.

Definition 2.5. Given a function class $\mathcal{G} \subset (\mathcal{S} \times \mathcal{A} \to \mathbb{R})$ and a fixed policy π, define the D^2-divergence $D^2_\mathcal{G}((s,a);\pi)$ as

$$\sup_{g,h \in \mathcal{G}} \frac{\big(g(s,a) - h(s,a)\big)^2}{\mathbb{E}_{(s',a') \sim \rho \times \pi}[(g(s',a') - h(s',a'))^2]}.$$

The "eluder dimension"-type Definition 2.5 is directly inspired by Di et al. (2023); Zhao et al. (2024), the intuition behind which is that given $(s,a) \in \mathcal{S} \times \mathcal{A}$, a small D^2-divergence indicates that for two functions g and h, if they are close under the behavior policy π, then they will also be close on such pair (s,a). Therefore, the D^2-divergence quantifies how well the estimation on dataset collected by the behavior policy π can be generalized to a specific state-action pair. We are now ready to define the two notions of concentrability conditions.

Assumption 2.6 (All-policy concentrability)**.** Given a reference policy π^{ref}, there exists $D < \infty$ such that $D^2 = \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} D^2_\mathcal{G}((s,a);\pi^{\text{ref}})$.

Assumption 2.6 indicates that the errors on any state-action pairs can be bounded by the error on the samples from $\rho \times \pi$ up to a factor D, whose relaxed counterpart under the same π^{ref} is as follows.

Assumption 2.7 (Single-policy concentrability)**.** $D^2_{\pi^*} := \mathbb{E}_{(s,a) \sim \rho \times \pi^*} D^2_\mathcal{G}((s,a);\pi^{\text{ref}}) < \infty$.

Assumption 2.7 indicates that the errors on the distributions of state-action pairs $\rho \times \pi^*$ can be bounded by the error on the samples from $\rho \times \pi^{\text{ref}}$ up to some constant. For both types, the single-policy concentrability assumption is strictly weaker than the all-policy concentrability assumption. However, in general, the two quantities characterizing single-policy concentrability C^{π^*} and $D^2_{\pi^*}$ cannot be bounded by each other up to constant factors.

Remark 2.8. In particular, we have $D^2_{\pi^*} \le |\mathcal{S}||\mathcal{A}|C^{\pi^*}$, indicating that C^{π^*} subsumes $D^2_{\pi^*}$ when $|\mathcal{S}|$ and $|\mathcal{A}|$ can be seen as constant.

2.2 Algorithm

In this subsection, we present an offline bandit algorithm, KL-PCB, for KL-regularized contextual bandits in Algorithm 1. KL-PCB first leverages least-square estimator to find a function $\bar{g} \in \mathcal{G}$ that minimizes its risk on the offline dataset. In Zhao et al. (2024), such $\bar{g}$ is directly applied to construct the estimated policy. In contrast, we construct a pessimistic estimator of g^* following the well-known pessimism principle in offline RL (Jin et al., 2021). Specifically, we define the bonus term Γ_n through the confidence radius $\beta = \sqrt{128 \log\big(2\mathcal{N}_\mathcal{G}(\epsilon_c)/\delta\big)/3n + 18\epsilon_c}$ as

$$\Gamma_n(s,a) = \beta D_\mathcal{G}\big((s,a), \pi^{\text{ref}}\big), \forall (s,a) \in \mathcal{S} \times \mathcal{A}. \quad (2.3)$$

We then obtain our pessimistic estimation $\widehat{g}$ by setting $\widehat{g} = \bar{g} - \Gamma_n$, which is less than g^* with high probability. After obtaining the pessimistic estimation, KL-PCB output the policy $\widehat{\pi}$, which maximizes the estimated objective

$$\widehat{J}(\pi) = \mathbb{E}_{(s,a) \sim \rho \times \pi}\left[\widehat{g}(s,a) - \eta^{-1} \log \frac{\pi(a|s)}{\pi^{\text{ref}}(a|s)}\right],$$

the maximizer of which is the counterpart of (2.2), i.e.,

$$\widehat{\pi}(a|s) \propto \pi^{\text{ref}}(a|s) \exp\big(\eta \cdot \widehat{g}(s,a)\big).$$

2.3 Theoretical Results

The sample complexity for KL-regularized contextual bandits is settled in this subsection. We first give the upper bound of KL-PCB.

Theorem 2.9. Under Assumption 2.7, for sufficiently small $\epsilon \in (0,1)$, if we set Γ_n as in (2.3), then $n = \widetilde{O}\big(\eta D^2_{\pi^*} \epsilon^{-1} \log \mathcal{N}_\mathcal{G}(\epsilon)\big)$ suffices to guarantee the output policy $\widehat{\pi}$ of Algorithm 1 to be ϵ-optimal with probability at least $1 - \delta$.

Algorithm 1 Offline KL-Regularized Pessimistic Contextual Bandits (KL-PCB)

Require: regularization η, reference policy π^{ref}, offline dataset $\mathcal{D}$, function class $\mathcal{G}$

1: Compute the least square estimation of reward function $\bar{g} \in \mathrm{argmin}_{g\in\mathcal{G}} \sum_{(s_i,a_i,r_i)\in\mathcal{D}} \big(g(s_i,a_i) - r_i\big)^2$

$$\bar{g} \in \operatorname*{argmin}_{g\in\mathcal{G}} \sum_{(s_i,a_i,r_i)\in\mathcal{D}} \big(g(s_i,a_i) - r_i\big)^2$$

2: Let $\widehat{g} \leftarrow \bar{g} - \Gamma_n$, where Γ_n is the bonus term in (2.3)

Ensure: $\widehat{\pi}(a|s) \propto \pi^{\mathsf{ref}}(a|s) \exp\big(\eta \cdot \widehat{g}(s,a)\big)$

Previously, Zhao et al. (2024) achieved an $\widetilde{O}(\epsilon^{-1})$ sample complexity under Assumption 2.6. As a comparison, KL-PCB achieves the same $\widetilde{O}(\epsilon^{-1})$ sample complexity but only requiring Assumption 2.7, which is weaker than Assumption 2.6. We also provide the sample complexity lower bound of KL-regularized contextual bandits in the following theorem.

Theorem 2.10. For $\forall S \geq 1$, $\eta > 4\log 2$, $C^* \in (2, \exp(\eta/4)]$, and any algorithm Alg, there is a KL-regularized contextual bandit with $C^{\pi^*} \leq C^*$ such that Alg requires $\Omega\big(\min\{\eta C^* \epsilon^{-1}, C^* \epsilon^{-2}\} \log \mathcal{N}_{\mathcal{G}}(\epsilon)\big)$ samples to find an ϵ-optimal policy for sufficiently small ϵ.

Previously, Zhao et al. (2024) provided a sample complexity lower bound of $\Omega(\eta \log \mathcal{N}_{\mathcal{G}}(\epsilon)/\epsilon)$ under KL-regularized objective. Foster et al. (2025) also provided a lower bound of $\Omega(C^{\pi^*})$ for KL-regularized objective to show the necessity of coverage. Compared to their results, our result shows that the *multiplicative* dependency on C^{π^*} is necessary for the first time.

Remark 2.11. Theorem 2.10 shows that when ϵ is sufficiently small, any algorithm for offline KL-regularized contextual bandits requires at least $\Omega(\eta C^{\pi^*})\epsilon^{-1} \log \mathcal{N}_{\mathcal{G}}(\epsilon))$ samples to output an ϵ-optimal policy. The presence of $\exp(\mathrm{poly}(\eta))$ in the range of C^* is inevitable, since we always have $C^{\pi^*} \leq \exp(\eta)$ in reverse KL regularized bandits with bounded rewards. As discussed in Remark 2.8, when $|\mathcal{S}|$ and $|\mathcal{A}|$ are constant, $D^2_{\pi^*} \lesssim C^{\pi^*}$, indicating that KL-PCB is near-optimal under this special case.

2.4 Proof of Theorem 2.9

In this section, we finish the proof of Theorem 2.9. At a high level, if we consider the regularized objective (1.1) multi-arm bandits, then $P \mapsto \mathsf{KL}(P\|Q)$ is 1-strongly convex w.r.t. $\mathsf{TV}(\cdot\|\cdot)$ (Polyanskiy & Wu, 2025, Exercise I.37), and thus $J(\pi)$ is strongly concave. Therefore, $J(\pi^*) - J(\widehat{\pi})$ is possible to be of the order $[\mathsf{TV}(\pi^*\|\widehat{\pi})]^2 \approx \widetilde{O}(n^{-1})$, pretending that π^* is the unconstrained maximizer. This intuition guides our analysis for contextual bandits.

We begin the proof with the definition of the event $\mathcal{E}(\delta)$ given $\delta > 0$ as

$$\mathcal{E}(\delta) := \Big\{ \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Big[\big|\bar{g} - g^*\big| - \Gamma_n\Big](s,a) \leq 0 \Big\}, \tag{2.4}$$

where Γ_n is defined in (2.3). Event $\mathcal{E}(\delta)$ holds indicates that the least square estimation $\bar{g}$ obtained in Line 1 of Algorithm 1 does not deviate too much from the true function g^*. More specifically, we have the following lemma, whose proof are deferred to Appendix B.1.

Lemma 2.12. For all $\delta > 0$, $\mathcal{E}(\delta)$ holds with probability at least $1 - \delta$.

The following covariance-type observation is the first pivot.

Lemma 2.13. If a bounded random variable $X \leq 0$ almost surely, then $\mathbb{E}[X^3] - \mathbb{E}[X^2]\mathbb{E}[X] \leq 0$.

Remark 2.14. While this lemma is elementary, to the best of our knowledge, we are the first to isolate this structure from our non-standard analysis of offline RL, from which the sharp upper bound is derived. The intuition behind Lemma 2.13 is very natural: X and X^2 cannot be positively correlated.

We further define the following quantities. For all $\gamma \in [0,1]$, we define $g_\gamma := \gamma\widehat{g} + (1-\gamma)g^*$ and further denote

$$\pi_\gamma(\cdot|s) \propto \pi^{\mathsf{ref}}(\cdot|s) \exp\big(\eta g_\gamma(s,\cdot)\big), \forall s \in \mathcal{S};$$
$$G(\gamma) := \mathbb{E}_{\rho\times\pi_\gamma}\Big[\big(\widehat{g} - g^*\big)^2(s,a)\Big].$$

The key to our analysis is the monotonicity of the function $G(\gamma)$ in γ, which is formally stated in the following lemma.

Lemma 2.15. On event $\mathcal{E}$, $0 \in \mathrm{argmax}_{\gamma\in[0,1]} G(\gamma)$.

Proof. For simplicity, we use $\triangle(s,a)$ to denote $\big(\widehat{g} - g^*\big)(s,a)$ in *this* proof. Then we know that $\triangle(s,a) \leq 0$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ on event $\mathcal{E}$. The most direct way to prove is to take derivative of G with respect to γ, which corresponds to the policy gradient (Sutton et al., 1999) of π_γ and thus implying a favorable structure. A direct calculation yields that

$$
\begin{aligned}
&= \mathbb{E}_{\rho\times\pi_\gamma}\big[\nabla_\gamma \log \pi_\gamma(a|s)\triangle(s,a)^2\big] \\
&= \eta\mathbb{E}_\rho\mathbb{E}_{a\sim\pi_\gamma}\big[\triangle^2(s,a)\big(\triangle(s,a)-\mathbb{E}_{a'\sim\pi_\gamma}[\triangle(s,a')]\big)\big] \\
&= \eta\mathbb{E}_\rho\Big[\mathbb{E}_{\pi_\gamma}\big[\triangle^3(s,a)\big]-\mathbb{E}_{\pi_\gamma}\big[\triangle^2(s,a)\big]\mathbb{E}_{\pi_\gamma}\big[\triangle(s,a)\big]\Big] \\
&\le 0,
\end{aligned}
$$

where $\mathbb{E}_\rho$ is the shorthand of $\mathbb{E}_{s\sim\rho}$, $\mathbb{E}_{\pi_\gamma}$ is the shorthand of $\mathbb{E}_{a\sim\pi_\gamma}$, the first equation is derived from standard policy gradient and the inequality holds conditioned on the event $\mathcal{E}(\delta)$ due to Lemma 2.13. □

Finally, we need the following lemma to bound the performance difference between two policy w.r.t. the KL-regularized objective J. For any given $g:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$, we define $\pi_g(\cdot|s)\propto\exp\big(\eta\cdot g(s,\cdot)\big), \forall s\in\mathcal{S}$ to facilitate presentation; then $\pi^*=\pi_{g^*}$ by definition. Please refer to Appendix B.3 for the proof of the lemma.

Lemma 2.16. Given $g:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$, $\exists\gamma\in[0,1]$ such that for $g_\gamma=\gamma g+(1-\gamma)g^*$ and $\pi_\gamma:=\pi_{g_\gamma}$,

$$J(\pi^*)-J(\pi_g)\le\eta\mathbb{E}_{(s,a)\sim\rho\times\pi_\gamma}\big[\big(g^*-g\big)^2(s,a)\big].$$

Now we are ready to prove Theorem 2.9.

Proof of Theorem 2.9. Following the proof of Zhao et al. (2024, Theorem 3.3), we know that there exists $\bar\gamma\in[0,1]$ such that

$$J(\pi^*)-J(\widehat\pi)\le\eta G(\bar\gamma)\le\eta G(0), \tag{2.5}$$

where the first inequality holds due to Lemma 2.16 and the second inequality holds due to the event $\mathcal{E}$ and Lemma 2.15. The term $G(0)$ can be further bounded with the D^2-based concentrability as follows

$$
\begin{aligned}
G(0) &= \eta\mathbb{E}_{(s,a)\sim\rho\times\pi^*}\Big[\big(\widehat g-g^*\big)^2(s,a)\Big] \\
&\le 4\eta\mathbb{E}_{(s,a)\sim\rho\times\pi^*}[\Gamma_n^2(s,a)] \\
&= 4\eta\beta^2\mathbb{E}_{(s,a)\sim\rho\times\pi^*}\big[D^2_{\mathcal{F}}((s,a);\pi^{\mathsf{ref}})\big] \\
&= \widetilde O(\eta D^2_{\pi^*}n^{-1}\log_{\mathcal{G}}(\epsilon_c)),
\end{aligned} \tag{2.6}
$$

where the second inequality holds conditioned on $\mathcal{E}(\delta)$ because of Lemma 2.15, and the last inequality follows from the definition of $\mathcal{E}(\delta)$ together with Line 1. By Lemma 2.12, we know that event $\mathcal{E}$ holds with probability at least $1-\delta$, which finishes the proof. □

3 f-divergence-regularized Contextual Bandits

As we discussed in Section 2, the fast rate is primarily achieved due to the strong convexity of $\pi\mapsto\mathsf{KL}(\cdot\|\pi^{\mathsf{ref}})$. However, KL is just an instance of f-divergence with $f(x)=x\log x$, which is not strongly convex, viz., only locally strongly convex. Motivated by this observation, we examine the use of f-divergence regularization with strongly convex f, which, in principle, may introduce a more favorable curvature in the performance metric of offline learning.

3.1 Problem Setup

We study the same contextual bandit problems which follows the definition in Section 2.1. In this section, we consider the following f-divergence regularized objective

$$J(\pi):=\mathbb{E}_{(s,a)\sim\rho\times\pi}[r(s,a)]-\eta^{-1}\mathbb{E}_{s\sim\rho}\big[D_f(\pi||\pi^{\mathsf{ref}})\big], \tag{3.1}$$

where η is the inverse temperature regularization intensity and $D_f(\pi||\pi^{\mathsf{ref}})$ is given by

$$D_f(\pi||\pi^{\mathsf{ref}})=\mathbb{E}_{a\sim\pi^{\mathsf{ref}}}\left[f\left(\frac{\pi(a|s)}{\pi^{\mathsf{ref}}(a|s)}\right)\right].$$

We consider those functions $f:(0,+\infty)\to\mathbb{R}$ with nice positive curvature as follows.

Assumption 3.1. f is α-strongly convex, twice continuously differentiable, and $f(1)=0$.

Many elementary functions like quadratic polynomials naturally satisfy Assumption 3.1. For instance, the 1-strongly convex $f(x)=(x-1)^2/2$ yields $D_f(P\|Q)=\chi^2(P\|Q)$, which is the popular χ^2-divergence recently considered in RL literature (see e.g., Zhan et al. (2022); Huang et al. (2024); Amortila et al. (2024)) and exhibits a promising theoretical potential on relaxing the data coverage requirement for efficient offline policy learning, at least against the unregularized objective.

3.2 Algorithm and Main Results

In this subsection, we present an offline learning algorithm for f-divergence regularized bandit, f-CB, in Algorithm 2. Algorithm 2 first leverages least-square estimator to find a function $\bar g\in\mathcal{G}$ that minimizes its risk on the offline dataset. The algorithm then uses the least squares estimation $\bar g$ to construct the output policy $\widehat\pi$. Compared to Algorithm 1, f-CB does not require any procedure to construct pessimistic reward estimation, whose sample complexity upper bound is given as follows.

Theorem 3.2. Under Assumption 3.1, for sufficiently small $\epsilon\in(0,1)$, with probability at least $1-\delta$, $n=\widetilde O(\alpha^{-1}\eta\epsilon^{-1}\log\mathcal{N}_{\mathcal{G}}(\epsilon))$ is sufficient to guarantee the output policy $\widehat\pi$ of f-CB to be ϵ-optimal.

Remark 3.3. Compared to the $D^2_{\pi^*}$ dependency in Theorem 2.9, Theorem 3.2 shows that the sample complexity of

Algorithm 2 Offline f-divergence Regularized Contextual Bandits (f-CB)

Require: regularization η, reference policy π^{ref}, function class $\mathcal{G}$, offline dataset $\mathcal{D}$

1: Compute the least square estimation of reward function

$$\bar{g} \in \operatorname*{argmin}_{g\in\mathcal{G}} \sum_{(s_i,a_i,r_i)\in\mathcal{D}} \big(g(s_i,a_i)-r_i\big)^2$$

2: Compute the optimal policy under the least-square reward estimator $\bar{g}$ for $s \in \mathcal{S}$ as

$$\widehat{\pi}(\cdot|s) \leftarrow \operatorname*{argmax}_{\pi(\cdot|s)\in\Delta(\mathcal{A})} \langle \pi(\cdot|s), \bar{g}(s,\cdot)\rangle + \eta^{-1} D_f\big(\pi(\cdot|s)\|\pi^{\mathsf{ref}}(\cdot|s)\big)$$

Ensure: $\widehat{\pi}$

Algorithm 2 gets rid of the dependency on any data coverage conditions when f is strongly convex. Intuitively, this is because the f-divergence regularization in this case is much stronger, so that both π^* and $\widehat{\pi}$ are close enough to π^{ref}.

The following hardness result justify the near-optimality of Theorem 3.2 for f-divergence-regularized bandits.

Theorem 3.4. For any $\epsilon \in (0,1), \alpha > 0, \eta > 0$, and algorithm Alg, there is an α-strongly-convex function f and an f-divergence-regularized contextual bandit instance such that Alg requires at least $\Omega\big(\alpha^{-1}\eta\epsilon^{-1}\log \mathcal{N}_{\mathcal{G}}(\epsilon)\big)$ samples to return an ϵ-optimal policy.

3.3 Insights Behind the Proof of Theorem 3.2

We provide an overview of key analysis techniques for proving Theorem 3.2. For the simplicity of presentation, we consider multi-arm bandits here and omit the subscript for context s.

We consider the function $H(\pi) = \eta^{-1}D_f(\pi\|\pi^{\mathsf{ref}})$, which is the regularizer in the objective. Then its convex conjugate is given by $H^*(r) = \sup_{\pi\in\Delta^d}\{\langle \pi, r\rangle - H_s(\pi)\}$, which is exactly the expected reward obtained by the optimal policy given reward function r. One observation is that when f is strongly convex, the induced f-divergence, and therefore the function H is also strongly convex. Therefore, let $\pi_r = \operatorname{argmax}_\pi\{\langle \pi, r\rangle - H_s(\pi)\}$ given some reward function r, the strong convexity of $H(\pi)$ gives that $\nabla H^*(r) = \pi_r$. This leads to the following regret decomposition, which is one of our key observations.

$$\begin{aligned} J(\pi^*) - J(\widehat{\pi}) &= \mathbb{E}_{a\sim\pi^*}[g^*(a)] - \mathbb{E}_{a\sim\widehat{\pi}}[g^*(a)] \\ &\quad - \eta^{-1}\big[D_f(\pi^*\|\pi^{\mathsf{ref}}) - D_f(\widehat{\pi}\|\pi^{\mathsf{ref}})\big] \\ &= H^*(g^*) - H^*(\bar{g}) - \langle \widehat{\pi}, g^* - \bar{g}\rangle \\ &= H^*(g^*) - H^*(\bar{g}) - \langle \nabla H^*(\bar{g}), g^* - \bar{g}\rangle, \end{aligned}$$

which is the Bregman divergence of the dual function H^* and therefore can be bounded by $(g^* - \bar{g})^\top H^*(\widetilde{g})(g^* - \bar{g})$ for some $\widetilde{g}$. By Proposition 3.2 in Penot (1994), we can bound $H^*(\widetilde{g})$ as follows

$$\begin{aligned} H^*(\widetilde{g}) &\preceq \big(\nabla^2 H(\pi_{\widetilde{g}})\big)^{-1} \\ &\preceq \alpha^{-1}\eta\,\mathrm{diag}\big(\pi^{\mathsf{ref}}(a_1), \cdots, \pi^{\mathsf{ref}}(a_{|\mathcal{A}|})\big), \end{aligned}$$

which enables us to bound $(g^* - \bar{g})^\top H^*(\widetilde{g})(g^* - \bar{g})$ by $\alpha^{-1}\eta\mathbb{E}_{\pi^{\mathsf{ref}}}[(g^* - \widehat{g})^2]$. Since $\mathbb{E}_{\pi^{\mathsf{ref}}}[(g^* - \widehat{g})^2]$ is not related to the optimal π^*, our resulting upper bound does not depend on any notion of concentrability.

4 Generalization to Contextual Dueling Bandits

In this section, we extend our algorithm to the problems of regularized contextual dueling bandits, where the learner receives preference comparison instead of absolute signals. Our data model follows Zhu et al. (2023); Zhan et al. (2023) and the notion of suboptimality follows Xiong et al. (2024); Zhao et al. (2024).

4.1 Problem Setup

We still consider contextual bandits $(\mathcal{S}, \mathcal{A}, r, \pi^{\mathsf{ref}})$ where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space and $r : \mathcal{S}\times\mathcal{A} \to [0,1]$ is the reward function.[1] But only relative preference feedback is available, viz., we have an i.i.d. offline dataset $\mathcal{D} = \{(s_i, a_i^1, a_i^2, y_i)\}_{i=1}^n$, where $s_i \in \mathcal{S}$ is generated from distribution ρ and $a_i^1, a_i^2 \sim \pi^{\mathsf{ref}}$. The binary preference label $y_i = 1$ indicates a_i^1 is preferred over a_i^2 (denoted by $a^1 \succ a^2$) and 0 for $a^2 \succ a^1$ given context s. In this work we consider the Bradley-Terry Model, where $\mathbb{P}[y = 1|s, a^1, a^2] = \sigma(r(s_i, a_i^1) - r(s_i, a_i^2))$, where $\sigma(x) = (1 + e^{-x})^{-1}$ is the link function. The objective here identical to (2.1) for KL-regularization and (3.1) for f-divergence regularization. Our goal is still to find an ϵ-optimal policy. To control the complexity of the function class $\mathcal{G}$, we assume that Assumption 2.1 still holds here.

[1] We overload some notations in Section 2 by their dueling counterparts for notational simplicity.

Concentrability. Analogous to Section 2, we need our estimation from offline dataset generalizable to the state-action pairs visited by our obtained policy. While density-ratio-based concentrability can be directly adapted to dueling bandit, we need a slightly different notion of D^2-divergence. This is because in dueling bandit, we cannot observe the absolute reward and best estimation g we can achieve is that for any state s and actions a^1, a^2, our estimated $g(s,a^1) - g(s,a^2) \approx r(s,a^1) - r(s,a^2)$. This implies that there exists some mapping $b : \mathcal{S} \to [-1,1]$ such that $g(s,a) - b(s) \approx r(s,a)$ on the offline data, which leads to the following definition.

Definition 4.1. Given a class of functions $\mathcal{G} \subset (\mathcal{S} \times \mathcal{A} \to \mathbb{R})$ and some policy π, let $\mathcal{B} = (\mathcal{S} \to [-1,1])$ be the function class, define the D^2-divergence $D^2_{\mathcal{G}}((s,a);\pi)$ as

$$\sup_{g,h\in\mathcal{G}} \inf_{b\in\mathcal{B}} \frac{\big(g(s,a) - h(s,a) - b(s)\big)^2}{\mathbb{E}_{s\sim\rho} \operatorname{Var}_{a'\sim\pi(\cdot|s')}[g(s',a') - h(s',a')]}.$$

A similar definition has been introduced in Zhao et al. (2024, Definition 2.6), which underpins the following two assumptions that characterize the coverage ability of π^{ref} similarly as in Section 2.

Given a reference policy π^{ref}, we define two coverage notions for contextual dueling bandits.

Assumption 4.2 (All-policy concentrability)**.** $D^2 := \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} D^2_{\mathcal{G}}((s,a);\pi^{\mathsf{ref}}) < \infty$.

Assumption 4.3 (Single-policy concentrability)**.** $D^2_{\pi^*} := \mathbb{E}_{(s,a)\sim\rho\times\pi^*}[D^2_{\mathcal{G}}((s,a);\pi^{\mathsf{ref}})] < \infty$.

Similar single-policy concentrability assumptions have appeared in previous work in offline contextual dueling bandits (Huang et al., 2024; Song et al., 2024) and similar notions has also appeared in the analysis of model-based RL (Uehara & Sun, 2021; Wang et al., 2024). Although Assumption 4.3 is strictly weaker than Assumption 4.2, $D^2_{\pi^*} \le |\mathcal{S}||\mathcal{A}|C^{\pi^*}$ still holds and, in general, C^{π^*} and $D^2_{\pi^*}$ cannot be bounded by each other up to some constant factors.

4.2 Main Results

In this subsection, we provide our main results for contextual dueling bandits with regularization. We refer the reader to Appendix D for a more detailed discussion. We first present the following results for KL-regularized contextual dueling bandit.

Theorem 4.4 (Informal)**.** There exists an algorithm generating ϵ-optimal policy with $n = \widetilde{O}(\eta D^2_{\pi^*}\epsilon^{-1}\log\mathcal{N}_{\mathcal{G}}(\epsilon))$ with high probability. Moreover, for any sufficiently small $\epsilon > 0$, and $\eta > 0$, and any algorithm Alg, there is a KL-regularized contextual dueling bandit such that Alg requires at least $\Omega\big(\min\{\eta C^{\pi^*}\epsilon^{-1}\log\mathcal{N}_{\mathcal{G}}(\epsilon), (C^{\pi^*})^2\epsilon^{-2}\log\mathcal{N}_{\mathcal{G}}(\epsilon)\}\big)$ samples to return an ϵ-optimal policy.

The following theorem shows a similar guarantee for f-divergence regularized contextual dueling bandits where f is α-strongly convex.

Theorem 4.5 (Informal)**.** Under Assumption 3.1, there exists an algorithm generating ϵ-optimal policy with $n = \widetilde{O}(\alpha^{-1}\eta\epsilon^{-1}\log\mathcal{N}_{\mathcal{G}}(\epsilon))$ with high probability. Moreover, for any sufficiently small $\epsilon > 0$, $\eta, \alpha > 0$, and any offline RL algorithm Alg, there is an α-strongly convex function f such that there exists an f-divergence regularized contextual dueling bandit such that Alg requires at least $\Omega\big(\alpha^{-1}\eta\epsilon^{-1}\log\mathcal{N}_{\mathcal{G}}(\epsilon)\big)$ samples to return an ϵ-optimal policy.

Remark 4.6. Theorems 4.4 and 4.5 are extension of Theorems 2.9 and 3.2 to offline contextual dueling bandits. These theorems indicate that, for both KL-regularization and f-divergence regularization, our techniques can be adapted to achieve algorithms and achieve similar sample complexities.

5 Conclusion and Limitation

In this work, we take the first step towards fully understanding the statistical efficiency *with respect to f-divergence-regularized objectives* of offline policy learning by sharp analyses for two empirically relevant subclasses. (1) We are the first to show that single-policy concentrability is nearly the right coverage condition for reverse KL to achieve the fast $\widetilde{\Theta}(\epsilon^{-1})$ sample complexity. The novel techniques in algorithm analysis leverages the curvature of KL-regularized objectives and integrates pessimism with a newly identified moment-based observation, enabling a neat refinement of a mean-value-type argument to the extreme; which are decoupled from tricky algorithmic tweaks, and thus might be of independent interest. (2) If strong convexity is further imposed on f, our fast $\widetilde{\Theta}(\epsilon^{-1})$ sample complexity is provably free of any coverage dependency. Unlike those for KL, the upper bound arguments for strongly convex f do not rely on the specific closed-form of the regularized objective minimizer.

All techniques in this work can be generalized beyond vanilla absolute reward feedback, as certified by contextual dueling bandits. However, there is still a gap between the upper bound and lower bound under reverse-KL regularization. Also, for general f-divergence other than reverse-KL, our results are limited to twice continuously differentiable and strongly convex f. The fully closing of the gap under reverse-KL regularization and the investigation of general f-divergence are left as future directions.

Impact Statement

This paper presents work whose goal is to advance the field of AI alignment from a theoretical perspective. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ahmed, Z., Le Roux, N., Norouzi, M., and Schuurmans, D. Understanding the impact of entropy on policy optimization. In *International conference on machine learning*, pp. 151–160. PMLR, 2019.

Amortila, P., Foster, D. J., and Krishnamurthy, A. Scalable online exploration via coverability. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 1392–1455, 2024.

Basseville, M. Divergence measures for statistical data processing—an annotated bibliography. *Signal Processing*, 93(4):621–633, 2013.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Bregman, L. M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3): 200–217, 1967.

Cai, Q., Yang, Z., Jin, C., and Wang, Z. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pp. 1283–1294. PMLR, 2020.

Chen, F., Foster, D. J., Han, Y., Qian, J., Rakhlin, A., and Xu, Y. Assouad, fano, and le cam with interaction: A unifying lower bound framework and characterization for bandit learnability. *Advances in Neural Information Processing Systems*, 37:75585–75641, 2024.

Chen, J. and Jiang, N. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pp. 1042–1051. PMLR, 2019.

Csiszár, I. On information-type measure of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.*, 2:299–318, 1967.

Di, Q., Zhao, H., He, J., and Gu, Q. Pessimistic nonlinear least-squares value iteration for offline reinforcement learning. *arXiv preprint arXiv:2310.01380*, 2023.

Dudík, M., Hofmann, K., Schapire, R. E., Slivkins, A., and Zoghi, M. Contextual dueling bandits. In *Conference on Learning Theory*, pp. 563–587. PMLR, 2015.

Farahmand, A.-m., Szepesvári, C., and Munos, R. Error propagation for approximate policy and value iteration. *Advances in neural information processing systems*, 23, 2010.

Foster, D. J. and Rakhlin, A. Foundations of reinforcement learning and interactive decision making. *arXiv preprint arXiv:2312.16730*, 2023.

Foster, D. J., Mhammedi, Z., and Rohatgi, D. Is a good foundation necessary for efficient reinforcement learning? the computational role of the base model in exploration. *arXiv preprint arXiv:2503.07453*, 2025.

Gabbianelli, G., Neu, G., and Papini, M. Importance-weighted offline learning done right. In *International Conference on Algorithmic Learning Theory*, pp. 614–634. PMLR, 2024.

Geist, M., Scherrer, B., and Pietquin, O. A theory of regularized markov decision processes. In *International Conference on Machine Learning*, pp. 2160–2169. PMLR, 2019.

Gilbert, E. N. A comparison of signalling alphabets. *The Bell system technical journal*, 31(3):504–522, 1952.

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

He, J., Zhou, D., and Gu, Q. Near-optimal policy optimization algorithms for learning adversarial linear mixture mdps. In *International Conference on Artificial Intelligence and Statistics*, pp. 4259–4280. PMLR, 2022.

Huang, A., Zhan, W., Xie, T., Lee, J. D., Sun, W., Krishnamurthy, A., and Foster, D. J. Correcting the mythos of kl-regularization: Direct alignment without overoptimization via chi-squared preference optimization. *arXiv preprint arXiv:2407.13399*, 2024.

Huang, A., Block, A., Liu, Q., Jiang, N., Foster, D. J., and Krishnamurthy, A. Is best-of-n the best of them? coverage, scaling, and optimality in inference-time alignment. *arXiv preprint arXiv:2503.21878*, 2025.

Ji, K., Zhao, Q., He, J., Zhang, W., and Gu, Q. Horizon-free reinforcement learning in adversarial linear mixture mdps. *arXiv preprint arXiv:2305.08359*, 2023.

Jiang, N. and Xie, T. Offline reinforcement learning in large state spaces: Algorithms and guarantees. *Statistical Science*, 2024.

Jiao, J., Courtade, T. A., No, A., Venkat, K., and Weissman, T. Information measures: the curious case of the binary alphabet. *IEEE Transactions on Information Theory*, 60 (12):7616–7626, 2014.

Jin, C., Yang, Z., Wang, Z., and Jordan, M. I. Provably efficient reinforcement learning with linear function approximation. In *Conference on learning theory*, pp. 2137–2143. PMLR, 2020.

Jin, Y., Yang, Z., and Wang, Z. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pp. 5084–5096. PMLR, 2021.

Kozuno, T., Yang, W., Vieillard, N., Kitamura, T., Tang, Y., Mei, J., Ménard, P., Azar, M. G., Valko, M., Munos, R., et al. Kl-entropy-regularized rl with a generative model is minimax optimal. *arXiv preprint arXiv:2205.14211*, 2022.

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

Le Cam, L. Convergence of estimates under dimensionality restrictions. *The Annals of Statistics*, pp. 38–53, 1973.

Levine, S. Reinforcement learning and control as probabilistic inference: Tutorial and review. *arXiv preprint arXiv:1805.00909*, 2018.

Levine, S. and Koltun, V. Guided policy search. In *International conference on machine learning*, pp. 1–9. PMLR, 2013.

Levine, S., Finn, C., Darrell, T., and Abbeel, P. End-to-end training of deep visuomotor policies. *Journal of Machine Learning Research*, 17(39):1–40, 2016.

Levine, S., Kumar, A., Tucker, G., and Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.

Li, G., Ma, C., and Srebro, N. Pessimism for offline linear contextual bandits using ℓ_p confidence sets. *Advances in Neural Information Processing Systems*, 35:20974–20987, 2022.

Li, G., Shi, L., Chen, Y., Chi, Y., and Wei, Y. Settling the sample complexity of model-based offline reinforcement learning. *The Annals of Statistics*, 52(1):233–260, 2024.

Liu, G., Ji, K., Zheng, R., Wu, Z., Dun, C., Gu, Q., and Yan, L. Enhancing multi-step reasoning abilities of language models through direct q-function optimization. *arXiv preprint arXiv:2410.09302*, 2024.

Liu, Z., Li, X., Kang, B., and Darrell, T. Regularization matters in policy optimization. *arXiv preprint arXiv:1910.09191*, 2019.

Müller, A. Integral probability metrics and their generating classes of functions. *Advances in applied probability*, 29 (2):429–443, 1997.

Munos, R., Valko, M., Calandriello, D., Azar, M. G., Rowland, M., Guo, Z. D., Tang, Y., Geist, M., Mesnard, T., Michi, A., et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

Neu, G., Jonsson, A., and Gómez, V. A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Penot, J.-P. Sub-hessians, super-hessians and conjugation. *Nonlinear Analysis: Theory, Methods & Applications*, 23 (6):689–702, 1994.

Polyanskiy, Y. and Wu, Y. *Information theory: From coding to learning*. Cambridge university press, 2025.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2023.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

Rényi, A. On measures of entropy and information. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, pp. 547–562. University of California Press, 1961.

Richemond, P. H., Tang, Y., Guo, D., Calandriello, D., Azar, M. G., Rafailov, R., Pires, B. A., Tarassov, E., Spangher, L., Ellsworth, W., et al. Offline regularised reinforcement learning for large language models alignment. *arXiv preprint arXiv:2405.19107*, 2024.

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International conference on machine learning*, pp. 19967–20025. PMLR, 2022.

Song, Y., Swamy, G., Singh, A., Bagnell, D., and Sun, W. The importance of online data: Understanding preference fine-tuning via coverage. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Sutton, R. S., McAllester, D., Singh, S., and Mansour, Y. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

Uehara, M. and Sun, W. Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*, 2021.

Varshamov, R. R. Estimate of the number of signals in error correcting codes. *Docklady Akad. Nauk, SSSR*, 117: 739–741, 1957.

Vieillard, N., Kozuno, T., Scherrer, B., Pietquin, O., Munos, R., and Geist, M. Leverage the average: an analysis of kl regularization in reinforcement learning. *Advances in Neural Information Processing Systems*, 33:12163–12174, 2020.

Wainwright, M. J. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

Wang, Z., Zhou, D., Lui, J., and Sun, W. Model-based rl as a minimalist approach to horizon-free and second-order bounds. *arXiv preprint arXiv:2408.08994*, 2024.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8:229–256, 1992.

Wu, Y., Sun, Z., Yuan, H., Ji, K., Yang, Y., and Gu, Q. Self-play preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675*, 2024.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021a.

Xie, T., Jiang, N., Wang, H., Xiong, C., and Bai, Y. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021b.

Xie, T., Foster, D. J., Krishnamurthy, A., Rosset, C., Awadallah, A., and Rakhlin, A. Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf. *arXiv preprint arXiv:2405.21046*, 2024.

Xiong, W., Zhong, H., Shi, C., Shen, C., Wang, L., and Zhang, T. Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and markov game. *arXiv preprint arXiv:2205.15512*, 2022.

Xiong, W., Dong, H., Ye, C., Wang, Z., Zhong, H., Ji, H., Jiang, N., and Zhang, T. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. In *Forty-first International Conference on Machine Learning*, 2024.

Ye, C., Xiong, W., Zhang, Y., Jiang, N., and Zhang, T. A theoretical analysis of nash learning from human feedback under general kl-regularized preference. *arXiv preprint arXiv:2402.07314*, 2024.

Yu, B. Assouad, fano, and le cam. In *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pp. 423–435. Springer, 1997.

Yue, Y., Broder, J., Kleinberg, R., and Joachims, T. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

Zhan, W., Huang, B., Huang, A., Jiang, N., and Lee, J. Offline reinforcement learning with realizability and single-policy concentrability. In *Conference on Learning Theory*, pp. 2730–2775. PMLR, 2022.

Zhan, W., Uehara, M., Kallus, N., Lee, J. D., and Sun, W. Provable offline preference-based reinforcement learning. *arXiv preprint arXiv:2305.14816*, 2023.

Zhang, T. *Mathematical Analysis of Machine Learning Algorithms*. Cambridge University Press, 2023. doi: 10.1017/9781009093057.

Zhang, Y., Zhang, G., Wu, Y., Xu, K., and Gu, Q. General preference modeling with preference representations for aligning language models. *arXiv preprint arXiv:2410.02197*, 2024.

Zhao, H., Ye, C., Gu, Q., and Zhang, T. Sharp analysis for kl-regularized contextual bandits and rlhf. *arXiv preprint arXiv:2411.04625*, 2024.

Zhao, H., Ye, C., Xiong, W., Gu, Q., and Zhang, T. Logarithmic regret for online kl-regularized reinforcement learning. *arXiv preprint arXiv:2502.07460*, 2025.

Zhou, X. On the fenchel duality between strong convexity and lipschitz continuous gradient. *arXiv preprint arXiv:1803.06573*, 2018.

Zhu, B., Jordan, M., and Jiao, J. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pp. 43037–43067. PMLR, 2023.

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

Zoghi, M., Whiteson, S., Munos, R., and Rijke, M. Relative upper confidence bound for the k-armed dueling bandit problem. In *International conference on machine learning*, pp. 10–18. PMLR, 2014.

A Review of Existing Results

Additional Notations. Besides the notation introduced in Section 1, we will use the following notations in Appendix. We denote $[N] := \{1, \cdots, N\}$ for any positive integer N. Boldfaced lower case (resp. upper case) letters are reserved for vectors (resp. matrices). Given a positive definite $\mathbf{\Sigma} \in \mathbb{R}^{d\times d}$ and $\mathbf{x} \in \mathbb{R}^d$, we denote the vector's Euclidean norm by $\|\mathbf{x}\|_2$ and define $\|\mathbf{x}\|_{\mathbf{\Sigma}} = \sqrt{\mathbf{x}^\top \mathbf{\Sigma} \mathbf{x}}$. We use $\mathsf{Bern}(p)$ to denote Bernoulli distribution with expectation p and $\mathsf{Unif}(\mathcal{X})$ for the uniform distribution on finite set $\mathcal{X}$. For $x \in \mathbb{R}^{\mathcal{A}}$, we denote $\|x\|_1 = \sum_{a\in\mathcal{A}} |x_a|$. We also denote $x_n = \Omega(y_n)$ by $x_n \gtrsim y_n$ in Appendix.

A.1 Previous Attempts on Understanding KL-regularized RL

There has been a surge of interest in understanding the principle behind KL-regularized RL. Ahmed et al. (2019); Liu et al. (2019) studied by ablation the effect of entropy regularization on the stability of policy improvement in policy optimization, the regret of which has been rigorously settled under the classic online mirror descent framework (Cai et al., 2020; He et al., 2022; Ji et al., 2023). Neu et al. (2017) unified popular KL-regularized policy optimization algorithms under a convex optimization framework, but the interplay with the data was left untouched. A series of work (Geist et al., 2019; Vieillard et al., 2020; Kozuno et al., 2022) then analyzed the sample complexity of algorithms using KL/entropy-type proximal terms with respect to the previous iteration or/and entropy regularizer with improved dependence on the effective horizon in discounted Markov decision processes. However, the performance metric in these studies is still the unregularized reward maximization objective, under which the sample complexity for finding an ϵ-optimal policy is at least equal to the statistical limit $\Omega(\epsilon^{-2})$.

Convergence under regularized objectives. Several recent studies (Xie et al., 2024; Xiong et al., 2024; Zhao et al., 2024; 2025; Foster et al., 2025) switched the focus to analyzing the suboptimality guarantee with respect to the regularized objective (1.1). In particular, Xie et al. (2024) studied token-level Markov decision processes (MDPs) and proposed a KL-regularized RL algorithm named XPO, which achieves $\widetilde{O}(\epsilon^{-2})$ sample complexity under their notion of all-policy concentrability. Xiong et al. (2024) proposed an Offline GSHF algorithm via the principle of *pessimism in the face of uncertainty*, and proved $\widetilde{O}(\epsilon^{-2})$ sample complexity under single-policy concentrability (See Section 2.1 for detailed definitions of concentrability). On the other hand, the sharp analysis in Zhao et al. (2024) yields the optimal sample complexity $\widetilde{O}(\epsilon^{-1})$, but requires all-policy concentrability (Zhao et al., 2024, Definition 2.6), i.e., the behavior policy π^{ref} is required to cover the entire function class for all possible policies. Zhao et al. (2025) considered the online episodic MDP setting, which inherently does not need any notion of data coverage and thus their results are not directly adaptable to our offline setting. Foster et al. (2025) considered an interesting hybrid setting in which the n state-action pairs are still from the offline dataset but $\Omega(n)$ online reward queries and policy switches are allowed; in contrast, in our setting, all reward signals are obtained in a purely offline fashion.

Here, we briefly discuss the direct adaptation of previous sample complexity results (with respect to KL-regularized objectives) to our setting and demonstrate the reason why theirs cannot imply an $\widetilde{O}(\epsilon^{-1})$ sample complexity without all-policy concentrability. Specifically, Xiong et al. (2024) obtained a performance gap upper bound under linear function approximation

$$J(\pi^*) - J(\pi) \le \big\|\mathbb{E}_{\rho\times\pi^*}[\boldsymbol{\phi}(s,a)] - \boldsymbol{\nu}\big\|_{\mathbf{\Sigma}_{\text{off}}^{-1}} =: \text{RHS},$$

where $\boldsymbol{\nu}$ is the reference vector, $\boldsymbol{\phi}(s,a) \in \mathbb{R}^d$ is the feature map, and $\Sigma_{\text{off}} = \sum_{i=1}^n \boldsymbol{\phi}(s_i,a_i)\boldsymbol{\phi}(s_i,a_i)^\top$ is the sample covariance matrix. However, we can show that RHS can be bounded from *below* by

$$\begin{aligned}
\big\|\mathbb{E}_{(s,a)\sim\rho\times\pi^*}[\boldsymbol{\phi}(s,a)] - \boldsymbol{\nu}\big\|\sqrt{\lambda_{\min}(\mathbf{\Sigma}_{\text{off}}^{-1})} &= \big\|\mathbb{E}_{(s,a)\sim\rho\times\pi^*}[\boldsymbol{\phi}(s,a)] - \boldsymbol{\nu}\big\|\lambda_{\max}(\mathbf{\Sigma}_{\text{off}})^{-1/2} \\
&\ge \big\|\mathbb{E}_{(s,a)\sim\rho\times\pi^*}[\boldsymbol{\phi}(s,a)] - \boldsymbol{\nu}\big\|\text{tr}(\mathbf{\Sigma}_{\text{off}})^{-1/2} \\
&= \big\|\mathbb{E}_{(s,a)\sim\rho\times\pi^*}[\boldsymbol{\phi}(s,a)] - \boldsymbol{\nu}\big\|\bigg(\sum_{i=1}^n \|\boldsymbol{\phi}(s_i,a_i)\|_2^2\bigg)^{-1/2} \\
&= \Omega(n^{-1/2}),
\end{aligned}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ is the minimum and maximum eigenvalue of a matrix, the first inequality holds due to the fact that $\mathbf{x}^\top\mathbf{\Sigma}\mathbf{x} \ge \|\mathbf{x}\|_2^2\lambda_{\min}(\mathbf{\Sigma})$ and the second inequality holds due to $\lambda_{\max}(\mathbf{\Sigma}) \le \text{tr}(\mathbf{\Sigma})$. Zhao et al. (2024) proposed a two-stage learning algorithm and obtained an $\widetilde{O}(\epsilon^{-1})$ sample complexity for online KL-regularized bandits. The algorithm can be adopted to offline learning by removing the second stage[2] and treat the samples from first stage as the offline dataset. An analogous analysis gives a sample complexity of $\widetilde{O}(D^2\epsilon^{-1})$, where D^2 is the all-policy concentrability.

B Missing Proofs from Section 2

B.1 Proof of Lemma 2.12

We first provide the following lemmas of concentration.

Lemma B.1 (Zhao et al. 2024, Lemma C.1)**.** For any policy π and state-action pairs $\{(s_i, a_i)\}_{i=1}^m$ generated i.i.d. from $\rho \times \pi$, and $\epsilon_c < 1$, with probability at least $1-\delta$, for any g_1 and g_2 we have

$$\mathbb{E}_{\rho\times\pi}\big[\big(g_1(s,a) - g_2(s,a)\big)^2\big] \le \frac{2}{n}\sum_{i=1}^n \big(g_1(s_i,a_i) - g_2(s_i,a_i)\big)^2 + \frac{32}{3n}\log(2\mathcal{N}_{\mathcal{G}}(\epsilon_c)/\delta) + 10\epsilon_c,$$

where $\mathcal{N}_{\mathcal{G}}(\epsilon_c)$ is the ϵ_c-covering number of $\mathcal{G}$.

Lemma B.2 (Zhao et al. 2024, Lemma C.2)**.** For arbitrary policy π and dataset $\{(s_i, a_i, r_i)\}_{i=1}^m$ generated i.i.d., from the product of π, ρ and the Bradley-Terry Model; let $\bar{g}$ be the least square estimator of g^*, then for any $0 < \epsilon_c < 1$ and $\delta > 0$, with probability at least $1-\delta$ we have

$$\sum_{i=1}^n \big(\bar{g}(s_i,a_i) - g^*(s_i,a_i)\big)^2 \le 16\log(a\mathcal{N}_{\mathcal{G}}(\epsilon_c)/\delta) + 4n\epsilon_c.$$

Proof of Lemma 2.12. We have the following inequality

$$\begin{aligned}
\big(\bar{g}(s,a) - g^*(s,a)\big)^2 &= \frac{\big(\bar{g}(s,a) - g^*(s,a)\big)^2}{\mathbb{E}_{\pi^{\mathsf{ref}}}\big[\big(\bar{g}(s,a) - g^*(s,a)\big)^2\big]}\mathbb{E}_{\pi^{\mathsf{ref}}}\big[\big(\bar{g}(s,a) - g^*(s,a)\big)^2\big] \\
&\le \sup_{g_1,g_2\in\mathcal{G}} \frac{\big(g_1(s,a) - g_2(s,a)\big)^2}{\mathbb{E}_{\pi^{\mathsf{ref}}}\big[\big(g_1(s,a) - g_2(s,a)\big)^2\big]}\mathbb{E}_{\pi^{\mathsf{ref}}}\big[\big(\bar{g}(s,a) - g^*(s,a)\big)^2\big] \\
&= D^2_{\mathcal{G}}((s,a),\pi^{\mathsf{ref}})\mathbb{E}_{\pi^{\mathsf{ref}}}\big[\big(\bar{g}(s,a) - g^*(s,a)\big)^2\big], \qquad \text{(B.1)}
\end{aligned}$$

where the inequality holds by taking supremum to $g_1, g_2 \in \mathcal{G}$. Now we have

$$\begin{aligned}
\mathbb{E}_{\pi^{\mathsf{ref}}}\big[\big(\bar{g}(s,a) - g^*(s,a)\big)^2\big] &\le \frac{2}{n}\sum_{i=1}^n \big(\bar{g}(s_i,a_i) - g^*(s_i,a_i)\big)^2 + \frac{32}{3n}\log(2\mathcal{N}_{\mathcal{G}}(\epsilon_c)/\delta) + 10\epsilon_c \\
&\le \frac{2}{n}\big[16\log(\mathcal{N}_{\mathcal{G}}(\epsilon_c)/\delta) + 4n\epsilon_c\big] + \frac{32}{3n}\log(2\mathcal{N}_{\mathcal{G}}(\epsilon_c)/\delta) + 10\epsilon_c \\
&= \frac{128}{3n}\log(2\mathcal{N}_{\mathcal{G}}(\epsilon_c)/\delta) + 18\epsilon_c, \qquad \text{(B.2)}
\end{aligned}$$

where the first inequality holds due to Lemma B.1 and second holds due to Lemma B.2. Plugging (B.2) into (B.1) and setting $\epsilon_c = O(n^{-1})$ complete the proof. □

B.2 Proof of Lemma 2.13

Proof of Lemma 2.13. We define $Y = -X$. Then it suffices to show that the covariance between Y and Y^2 is

$$\begin{aligned}
\mathrm{Cov}(Y, Y^2) &= \mathbb{E}[Y^3] - \mathbb{E}[Y^2]\mathbb{E}[Y] \\
&\ge \big(\mathbb{E}[Y^2]\big)^{3/2} - \mathbb{E}[Y^2]\mathbb{E}[Y] \\
&= \big(\mathbb{E}[Y^2]\big)\big(\sqrt{\mathbb{E}[Y^2]} - \mathbb{E}[Y]\big) \\
&\ge 0,
\end{aligned}$$

where both inequalities follow from Jensen's inequality. □

[2]This can be done by setting the n in their paper to 0.

B.3 Proof of Lemma 2.16

This proof is extracted from the proof of Zhao et al. (2024, Theorem 3.3) and we present it here for completeness. By definition of our objective in (2.1), we have

$$
\begin{aligned}
&J(\pi^*) - J(\pi_g) \\
&= \mathbb{E}_{(s,a)\sim\rho\times\pi^*}\left[g^*(s,a) - \eta^{-1}\log\frac{\pi^*(a|s)}{\pi^{\mathsf{ref}}(a|s)}\right] - \mathbb{E}_{(s,a)\sim\rho\times\pi_g}\left[g^*(s,a) - \frac{1}{\eta}\log\frac{\pi_g(a|s)}{\pi^{\mathsf{ref}}(a|s)}\right] \\
&= \frac{1}{\eta}\mathbb{E}_{(s,a)\sim\rho\times\pi^*}\left[\log\frac{\pi^{\mathsf{ref}}(a|s)\cdot\exp\big(\eta g^*(s,a)\big)}{\pi^*(a|s)}\right] - \frac{1}{\eta}\mathbb{E}_{(s,a)\sim\rho\times\pi_g}\left[\log\frac{\pi^{\mathsf{ref}}(a|s)\cdot\exp\big(\eta g^*(s,a)\big)}{\pi_g(a|s)}\right] \\
&= \frac{1}{\eta}\mathbb{E}_{s\sim\rho}\big[\log Z_{g^*}(s)\big] - \frac{1}{\eta}\mathbb{E}_{s\sim\rho}\big[\log Z_g(s)\big] - \mathbb{E}_{s\sim\rho}\left[\sum_{a\in\mathcal{A}}\pi_g(a|s)\cdot\big(g^*(s,a) - f(s,a)\big)\right],
\end{aligned}
$$

where for all $g \in \mathcal{G}$ we define $Z_g(\cdot)$ as follows,

$$
Z_g(\cdot) \coloneqq \sum_{a\in\mathcal{A}}\pi^{\mathsf{ref}}(a|\cdot)\exp\big(\eta g(\cdot,a)\big).
$$

We further denote $\Delta(s,a) = g(s,a) - g^*(s,a)$ and $H_s(g) = \log Z_g(s) - \eta\sum_{a\in\mathcal{A}}\pi_g(a|s)\cdot\Delta(s,a)$. It worth noticing that $\eta^{-1}\mathbb{E}_{s\sim\rho}[H_s(g^*) - H_s(g)] = J(\pi^*) - J(\pi_g)$. Now we take derivative of H with respect to $\Delta(s,a)$,

$$
\begin{aligned}
\frac{\partial H_s(g)}{\partial\Delta(s,a)} &= \frac{\partial}{\partial\Delta(s,a)}\left[\log Z_g(s) - \eta\sum_{a\in\mathcal{A}}\pi_g(a|s)\cdot\Delta(s,a)\right] \\
&= \frac{1}{Z_g(s)}\cdot\pi^{\mathsf{ref}}(a|s)\exp\big(\eta\cdot g(s,a)\big)\cdot\eta - \eta\cdot\pi_g(a|s) \\
&\quad - \eta^2\cdot\Delta(s,a)\cdot\frac{\pi^{\mathsf{ref}}(a|s)\cdot\exp\big(\eta\cdot g(s,a)\big)}{Z_g(s)} + \eta^2\cdot\Delta(s,a)\cdot\frac{\big[\pi^{\mathsf{ref}}(a|s)\cdot\exp\big(\eta\cdot g(s,a)\big)\big]^2}{[Z_g(s)]^2} \\
&\quad + \eta\sum_{a'\in\mathcal{A}\setminus\{a\}}\frac{\pi^{\mathsf{ref}}(a'|x)\cdot\exp\big(\eta\cdot g(s,a')\big)}{Z_g(s)}\cdot\eta\cdot\Delta(s,a')\cdot\frac{\pi^{\mathsf{ref}}(a|s)\cdot\exp\big(\eta\cdot g(s,a)\big)}{Z_g(s)} \\
&= -\eta^2\pi_g(a|s)\Delta(s,a) + \eta^2[\pi_g(a|s)]^2\cdot\Delta(s,a) + \eta^2\sum_{a'\in\mathcal{A}\setminus\{a\}}\pi_g(a'|x)\pi_g(a|s)\Delta(s,a').
\end{aligned}
$$

Therefore, by mean value theorem, there exists $\gamma \in [0,1]$ and $g_\gamma = \gamma g + (1-\gamma)g^*$ such that

$$
\begin{aligned}
H_s(g) - H_s(g^*) &= -\eta^2\gamma\sum_{a\in\mathcal{A}}\pi_{g_\gamma}(a|s)\Delta(s,a)^2 + \gamma\eta^2\sum_{a_1\in\mathcal{A}}\sum_{a_2\in\mathcal{A}}\pi_{g_\gamma}(a_1|x)\pi_{g_\gamma}(a_2|x)\Delta(s,a_1)\Delta(s,a_2) \\
&= -\eta^2\gamma\mathbb{E}_{a\sim\pi_{g_\gamma}}\left[\big(g^*(s,a) - g(s,a)\big)^2\right] + \gamma\eta^2\Big(\mathbb{E}_{a\sim\pi_{g_\gamma}}\left[\big(g^*(s,a) - g(s,a)\big)\right]\Big)^2 \\
&\geq -\eta^2\mathbb{E}_{a\sim\pi_{g_\gamma}}\left[\big(g^*(s,a) - g(s,a)\big)^2\right],
\end{aligned}
$$

where the inequality holds by omitting the second term and $\gamma \leq 1$. Now taking expectation over ρ, we have

$$
\begin{aligned}
J(\pi^*) - J(\pi_g) &= \eta^{-1}\mathbb{E}_{s\sim\rho}[H_s(g^*) - H_s(g)] \\
&\leq \eta\mathbb{E}_{(s,a)\sim\rho_{g_\gamma}}\left[\big(g^*(s,a) - g(s,a)\big)^2\right],
\end{aligned}
$$

which concludes the proof.

B.4 Proof of Theorem 2.10

Proof of Theorem 2.10. We consider the family of contextual bandits with $S \coloneqq |\mathcal{S}|, A \coloneqq |\mathcal{A}| < \infty$ and reward function in some function class $\mathcal{G}$ composed of function $\mathcal{S}\times\mathcal{A}\to[0,1]$ as follows.

$$
\mathrm{CB}_{\mathcal{G}} \coloneqq \{(\mathcal{S},\mathcal{A},\rho,r,\pi^{\mathsf{ref}},\eta) : r\in\mathcal{G}, \rho\in\Delta(\mathcal{S}), \pi^{\mathsf{ref}}\in\Delta(\mathcal{A}|\mathcal{S})\}. \tag{B.3}
$$

Our goal is to prove the following statement. Fixing any $S \geq 1, \eta > 4\log 2$ and $C^* \in (2, \exp(\eta/4)]$, then for any estimator $\mathcal{D} \mapsto \widehat{\pi} \in \Delta(\mathcal{A}|\mathcal{S})$, for any $n \geq 16SC^*$, there exist some function class $\mathcal{G}$, such that $\exists\, \mathrm{inst} = (\mathcal{S}, \mathcal{A}, \rho, r, \pi^{\mathsf{ref}}, \eta) \in \mathrm{CB}_{\mathcal{G}}$ with single-policy concentrability $C^{\pi^*} \leq C^*$, regularization coefficient η, $|\mathcal{S}| = S = \Theta(\log|\mathcal{G}|)$, and

$$\mathrm{SubOpt}_{\mathrm{RKL}}(\widehat{\pi}; \mathrm{inst}) \gtrsim \min\{\eta SC^* n^{-1}, (SC^*)^{1/2} n^{-1/2}\}. \tag{B.4}$$

Since $\log|\mathcal{G}| \geq \log \mathcal{N}_{\mathcal{G}}(\epsilon)$ for any $\epsilon \in (0,1)$, equation (B.4) yields the desired bound.
We set $\mathcal{S} = [S]$, $\mathcal{A} = \{\pm 1\}$, $\rho = \mathsf{Unif}(\mathcal{S})$, and the reference policy to be

$$\forall s \in \mathcal{S}, \pi^{\mathsf{ref}}(-1|s) = C^{-1}, \pi^{\mathsf{ref}}(+1|s) = 1 - C^{-1};$$

where $C \geq 1$ is a parameter to be specified later. We construct 2^S Bernoulli reward functions, in particular, $\forall \tau \in \{\pm 1\}^S$, the mean function r_τ of the reward (indexed by τ) is defined as

$$r_\tau(s, -1) = 0.5 + \tau_s \delta, r_\tau(s, +1) = 0.5 - \alpha$$

for any state $s \in \mathcal{S}$, where $\alpha \in (0, 1/2)$ and $\delta \in (0, 1/4]$ will be specified later. We omit the RKL subscript in the following argument when it is clear in context. By (2.2), the optimal policy π_τ^* under r_τ is

$$\forall s \in \mathcal{S}, \pi_\tau^*(-1|s) = \frac{\exp\big(\eta(\alpha + \tau_s\delta)\big)}{\exp\big(\eta(\alpha + \tau_s\delta)\big) + C - 1}, \pi_\tau^*(+1|s) = \frac{C-1}{\exp\big(\eta(\alpha + \tau_s\delta)\big) + C - 1}. \tag{B.5}$$

Since $C^* \leq \exp(\eta/4)$, we assign $C = C^*$ and $\alpha = \eta^{-1}\log(C-1) \Leftrightarrow C - 1 = \exp(\eta\alpha)$, which gives

$$\forall s \in \mathcal{S}, \frac{\pi_\tau^*(-1|s)}{\pi^{\mathsf{ref}}(-1|s)} \leq C\frac{\exp(\eta(\alpha + \tau_s\delta))}{C - 1 + \exp(\eta(\alpha + \tau_s\delta))} = C\frac{\exp(\eta\tau_s\delta)}{1 + \exp(\eta\tau_s\delta)} \leq C = C^*;$$
$$\forall s \in \mathcal{S}, \frac{\pi_\tau^*(+1|s)}{\pi^{\mathsf{ref}}(+1|s)} = \frac{C}{C-1} \cdot \frac{1}{\exp(\eta\tau_s\delta) + 1} \leq C = C^*;$$

where the last inequality is due to the assumption $C^* \geq 2$. Therefore, we obtain

$$\max_{\tau \in \{\pm 1\}^S} C^{\pi_\tau^*} \leq C^*. \tag{B.6}$$

We will abuse the notation $\mathrm{SubOpt}(\widehat{\pi}; \tau) \coloneqq \mathrm{SubOpt}(\widehat{\pi}; r_\tau)$. Since $\rho = \mathsf{Unif}(\mathcal{S})$,

$$\mathrm{SubOpt}(\widehat{\pi}; \tau) = \frac{1}{S}\sum_{s=1}^{S} \mathrm{SubOpt}_s(\widehat{\pi}; \tau), \tag{B.7}$$

where

$$\begin{aligned}
\mathrm{SubOpt}_s(\widehat{\pi}; \tau) &= \langle \pi_\tau^*(\cdot|s), r_\tau(s, \cdot) - \eta^{-1}\log\frac{\pi_\tau^*(\cdot|s)}{\pi^{\mathsf{ref}}(\cdot|s)}\rangle - \langle \widehat{\pi}(\cdot|s), r_\tau(s, \cdot) - \eta^{-1}\log\frac{\widehat{\pi}(\cdot|s)}{\pi^{\mathsf{ref}}(\cdot|s)}\rangle \\
&= \frac{1}{\eta}\mathbb{E}_{a \sim \pi_\tau^*(\cdot|s)}\left[\log\frac{\pi^{\mathsf{ref}}(a|s) \cdot \exp\big(\eta r_\tau(s,a)\big)}{\pi_\tau^*(a|s)}\right] \\
&\quad - \frac{1}{\eta}\mathbb{E}_{a \sim \widehat{\pi}(\cdot|s)}\left[\log\frac{\pi^{\mathsf{ref}}(a|s) \cdot \exp\big(\eta r_\tau(s,a)\big)}{\widehat{\pi}(a|s)}\right] \\
&= \frac{1}{\eta}\mathbb{E}_{a \sim \pi_\tau^*(\cdot|s)}\left[\log\Big(\sum_{b \in \mathcal{A}} \pi^{\mathsf{ref}}(b|s) \cdot \exp\big(\eta r_\tau(s,b)\big)\Big)\right] \\
&\quad - \frac{1}{\eta}\mathbb{E}_{a \sim \widehat{\pi}(\cdot|s)}\left[\log\frac{\pi^{\mathsf{ref}}(a|s) \cdot \exp\big(\eta r_\tau(s,a)\big)}{\widehat{\pi}(a|s)}\right] \\
&= \frac{1}{\eta}\mathbb{E}_{a \sim \widehat{\pi}(\cdot|s)}\left[\log\frac{\pi^{\mathsf{ref}}(a|s) \cdot \exp\big(\eta r_\tau(s,a)\big)}{\pi_\tau^*(a|s)} - \log\frac{\pi^{\mathsf{ref}}(a|s) \cdot \exp\big(\eta r_\tau(s,a)\big)}{\widehat{\pi}(a|s)}\right] \\
&= \eta^{-1}\mathsf{KL}\left(\widehat{\pi}\|\pi_\tau^*\right).
\end{aligned} \tag{B.8}$$

We write $\tau \sim_s \tau'$ if $\tau, \tau' \in \{\pm 1\}^{\mathcal{S}}$ differ in only the s-th coordinate and $\tau \sim \tau'$ if $\exists s \in \mathcal{S}, \tau \sim_s \tau'$. By (B.8), $\forall s \in \mathcal{S}$, $\forall \tau, \tau' \in \{\pm 1\}^{\mathcal{S}}$ with $\tau \sim_s \tau'$,

$$\begin{aligned}
&\mathrm{SubOpt}_s(\widehat{\pi}; \tau) + \mathrm{SubOpt}_s(\widehat{\pi}; \tau') \\
&= \eta^{-1} \mathsf{KL}\left(\widehat{\pi} \| \pi^*_\tau\right) + \eta^{-1} \mathsf{KL}\left(\widehat{\pi} \| \pi^*_{\tau'}\right) \\
&= 2\eta^{-1} \sum_{a \in \mathcal{A}} \widehat{\pi}(a|s) \log \frac{\widehat{\pi}(a|s)}{\sqrt{\pi^*_\tau(a|s)\pi^*_{\tau'}(a|s)}} \\
&= 2\eta^{-1} \mathsf{KL}\left(\widehat{\pi}(\cdot|s) \| \bar{\pi}_{\tau,\tau'}(\cdot|s)\right) - 2\eta^{-1} \mathbb{E}_{a \sim \widehat{\pi}(\cdot|s)} \log \Big(\sum_{b \in \mathcal{A}} \sqrt{\pi^*_\tau(b|s)\pi^*_{\tau'}(b|s)} \Big) \\
&\geq -2\eta^{-1} \log \Big(\sum_{b \in \mathcal{A}} \sqrt{\pi^*_\tau(b|s)\pi^*_{\tau'}(b|s)} \Big) \\
&= \frac{1}{\eta} \log \frac{(\exp(\eta\delta) + 1)(\exp(-\eta\delta) + 1)}{4},
\end{aligned} \tag{B.9}$$

where $\bar{\pi}(\cdot|s) = \sqrt{\pi^*_\tau(\cdot|s)\pi^*_{\tau'}(\cdot|s)} / \sum_{b \in \mathcal{A}} \sqrt{\pi^*_\tau(b|s)\pi^*_{\tau'}(b|s)}$ for every $s \in \mathcal{S}$, the inequality is due to the non-negativity of KL divergence, and the last equality follows from (B.5) together with the design choice $C - 1 = \exp(\eta\alpha)$.

Case $\eta\delta \leq 2$. Recall that $\forall x \in \mathbb{R}, (\mathrm{e}^x + \mathrm{e}^{-x})/2 - 1 = x^2 \sum_{k=0}^{\infty} \frac{x^{2k}}{(2k+2)!} \geq x^2/2$, which implies

$$(\text{B.9}) = \frac{1}{\eta} \log \left(1 + \frac{1}{2}\Big(\frac{\mathrm{e}^{\eta\delta} + \mathrm{e}^{-\eta\delta}}{2} - 1\Big)\right) \geq \frac{1}{\eta} \log \left(1 + \frac{\eta^2\delta^2}{4}\right) \geq \frac{1}{\eta} \cdot \frac{\eta^2\delta^2/4}{2} = \eta\delta^2/8. \tag{B.10}$$

Here, the last inequality is due to $\eta^2\delta^2/4 \leq 1$ and $\forall x \in [0, 1], \log(1 + x) \geq x/2$.

Case $\eta\delta > 2$. We have $-\eta^{-1} 2 \log 2 \geq -\delta \log 2$, which implies the following bound.

$$(\text{B.9}) \geq \frac{1}{\eta} \log \frac{\exp(\eta\delta) + 1}{4} \geq \frac{\eta\delta - 2\log 2}{\eta} = \delta - \eta^{-1} 2 \log 2 \geq (1 - \log 2)\delta \geq 3\delta/10. \tag{B.11}$$

In summary, (B.10) and (B.11) imply that $\forall s \in \mathcal{S}$, $\forall \tau, \tau' \in \{\pm 1\}^{\mathcal{S}}$ with $\tau \sim_s \tau'$,

$$\mathrm{SubOpt}_s(\widehat{\pi}; \tau) + \mathrm{SubOpt}_s(\widehat{\pi}; \tau') \geq \frac{\eta\delta^2}{8} \wedge \frac{3\delta}{10}. \tag{B.12}$$

Let P_τ be the distribution of (s, a, y) where $s \sim \rho, a \sim \pi^{\mathsf{ref}}(\cdot|s)$, and $y \sim \mathsf{Bern}(r_\tau(s, a))$. Then $\forall x \in \mathcal{S}$ $\forall \tau, \tau' \in \{\pm 1\}^{\mathcal{S}}$ with $\tau \sim_x \tau'$,

$$\begin{aligned}
\mathsf{KL}\left(P_\tau \| P_{\tau'}\right) &= \frac{1}{S} \sum_{s,a} \pi^{\mathsf{ref}}(a|s) \mathsf{KL}\left(\mathsf{Bern}(r_\tau(s, a)) \| \mathsf{Bern}(r_{\tau'}(s, a))\right) \\
&= \frac{1}{S} \cdot C^{-1} \mathsf{KL}\left(\mathsf{Bern}(r_\tau(x, -1)) \| \mathsf{Bern}(r_{\tau'}(x, -1))\right) \\
&\leq \frac{4\delta^2}{SC(0.25 - \delta^2)} \leq \frac{16\delta^2}{3SC},
\end{aligned} \tag{B.13}$$

where we use the requirement $\delta \leq 1/4$ and $\mathsf{KL}\left(\mathsf{Bern}(p) \| \mathsf{Bern}(q)\right) \leq (p - q)^2 / \big(q(1 - q)\big)$. Then let $P_{\mathcal{D}_\tau}$ be the distribution of $\mathcal{D}$ given the mean reward function r_τ, we employ (B.13) to get

$$\mathsf{KL}\left(P_{\mathcal{D}_\tau} \| P_{\mathcal{D}_{\tau'}}\right) = n \mathsf{KL}\left(P_\tau \| P_{\tau'}\right) \leq \frac{16n\delta^2}{3SC}. \tag{B.14}$$

Since $n \geq 16SC^* = 16SC$ by design, we can set $\delta = \sqrt{SC/n}$ (which ensures $\delta \leq 1/4$) to obtain

$$\begin{aligned}
\sup_{\text{inst}} \mathrm{SubOpt}(\widehat{\pi}; \text{inst}) &\geq \sup_{\tau \in \{\pm 1\}^{\mathcal{S}}} \mathrm{SubOpt}(\widehat{\pi}; \tau) \\
&\geq \frac{1}{S} \cdot S \cdot \frac{1}{4} \cdot \Big(\frac{\eta\delta^2}{8} \wedge \frac{3\delta}{10}\Big) \min_{\tau \sim \tau'} \exp\Big(-\mathsf{KL}\left(P_{\mathcal{D}_\tau} \| P_{\mathcal{D}_{\tau'}}\right)\Big) \\
&\geq \Big(\frac{\eta SC^*}{32n} \wedge \frac{3\sqrt{SC^*}}{40\sqrt{n}}\Big) \exp(-16/3) \gtrsim \frac{\eta SC^*}{n} \wedge \sqrt{\frac{SC^*}{n}}.
\end{aligned}$$

where the S^{-1} in the second inequality comes from (B.7), the second inequality is by substituting (B.12) into Assouad's Lemma (Lemma F.3), and the last inequality is due to (B.14). □

C Missing Proof from Section 3

C.1 Proof of Theorem 3.2

Before coming to the proof, we first introduce some useful properties. The following properties characterize the convexity of f-divergence when f is (strongly) convex.

The strong-convexity of f implies that the corresponding f-divergence, $D_f(\cdot||\pi^{\mathsf{ref}})$ is also strongly convex with respect to all $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ supported by π^{ref}.

Proposition C.1. Given context s, $D_f(\pi(\cdot|s)||\pi^{\mathsf{ref}}(\cdot|s))$ is strict convex with respect to π if f is strictly convex.

Proposition C.2. Given context s, $\pi(\cdot|s) \mapsto D_f(\pi(\cdot|s)\|\pi^{\mathsf{ref}}(\cdot|s))$ is 4α-strong convex with respect to the metric TV if f is α-strongly convex.

Proof of Proposition C.2. We first show the gradient of D_f with respect to π.

$$\frac{\partial D_f(\pi||\pi^{\mathsf{ref}})}{\pi(a)} = \frac{\partial}{\partial \pi(a)} \sum_{b\in\mathcal{A}} \pi^{\mathsf{ref}}(b) f\bigg(\frac{\pi(b)}{\pi^{\mathsf{ref}}(b)}\bigg) = f'\bigg(\frac{\pi(a)}{\pi^{\mathsf{ref}}(a)}\bigg).$$

Now consider $\pi_1, \pi_2 \in \Delta(\mathcal{A})$ supported by π^{ref}.

$$\begin{aligned}
&D_f(\pi_1||\pi^{\mathsf{ref}}) - D_f(\pi_2||\pi^{\mathsf{ref}}) - \big\langle \pi_1 - \pi_2, \nabla D_f(\pi_2||\pi^{\mathsf{ref}})\big\rangle \\
&\quad= \sum_{a\in\mathcal{A}} \pi^{\mathsf{ref}}(a)\bigg(f\bigg(\frac{\pi_1(a)}{\pi^{\mathsf{ref}}(a)}\bigg) - f\bigg(\frac{\pi_2(a)}{\pi^{\mathsf{ref}}(a)}\bigg)\bigg) - \sum_{a\in\mathcal{A}} \big(\pi_1(a) - \pi_2(a)\big) f'\bigg(\frac{\pi_2(a)}{\pi^{\mathsf{ref}}(a)}\bigg) \\
&\quad= \sum_{a\in\mathcal{A}} \pi^{\mathsf{ref}}(a)\bigg(f\bigg(\frac{\pi_1(a)}{\pi^{\mathsf{ref}}(a)}\bigg) - f\bigg(\frac{\pi_2(a)}{\pi^{\mathsf{ref}}(a)}\bigg) - \bigg(\frac{\pi_1(a)}{\pi^{\mathsf{ref}}(a)} - \frac{\pi_2(a)}{\pi^{\mathsf{ref}}(a)}\bigg) f'\bigg(\frac{\pi_2(a)}{\pi^{\mathsf{ref}}(a)}\bigg)\bigg) \\
&\quad\geq \frac{\alpha}{2} \sum_{a\in\mathcal{A}} \pi^{\mathsf{ref}}(a)\bigg(\frac{\pi_1(a)}{\pi^{\mathsf{ref}}(a)} - \frac{\pi_2(a)}{\pi^{\mathsf{ref}}(a)}\bigg)^2 \\
&\quad= \frac{\alpha}{2} \sum_{a\in\mathcal{A}} \frac{1}{\pi^{\mathsf{ref}}(a)} \big(\pi_1(a) - \pi_2(a)\big)^2 \\
&\quad\geq \frac{\alpha}{2} \bigg(\sum_{a\in\mathcal{A}} \big|\pi_1(a) - \pi_2(a)\big|\bigg)^2,
\end{aligned}$$

where the first inequality holds due to f's strong convexity and the second holds due to Cauchy–Schwarz. The proof finishes since $\|\pi_1 - \pi_2\|_1 = 2\mathsf{TV}(\pi_1\|\pi_2)$. □

We first introduce some notation and important properties concerning the convex conjugate of functions. Given some context s, we denote the regularization term as $H_s(\pi) = \eta^{-1} D_f(\pi(\cdot|s)\|\pi^{\mathsf{ref}}(\cdot|s))$. We use $H_s^*(r)$ to denote the convex conjugate of H_s, which is defined as $H_s^*(r) = \sup_{\pi\in\mathcal{S}\to\Delta^{|\mathcal{A}|}}\{\langle \pi(\cdot|s), r(s,\cdot)\rangle - H_s(\pi)\}$. We have the following properties for the convex conjugate. The first property gives the gradient of convex conjugate (see, e.g., Zhou 2018, Lemma 5).

Proposition C.3. Given context s, and convex f, let $\pi_r \in \mathrm{argmax}_\pi\{\langle \pi(\cdot|s), r(s,\cdot)\rangle - H_s(\pi)\}$ for some r, then the gradient of H_s^* is given by $\nabla H_s^*(r) = \pi_r(\cdot|s)$.

We also need some properties of $\nabla^2 H_s^*$, the Hessian matrix of the convex conjugate function. We first give the Hessian matrix of the original function H_s as follows.

$$\nabla^2 H_s(\pi) = \eta^{-1}\mathrm{diag}\bigg(\pi^{\mathsf{ref}}(a_1|s)^{-1} f''\bigg(\frac{\pi(a_1|s)}{\pi^{\mathsf{ref}}(a_1|s)}\bigg), \cdots, \pi^{\mathsf{ref}}(a_{|\mathcal{A}|}|s)^{-1} f''\bigg(\frac{\pi(a_{|\mathcal{A}|}|s)}{\pi^{\mathsf{ref}}(a_{|\mathcal{A}|}|s)}\bigg)\bigg). \tag{C.1}$$

Furthermore, when f is α-strongly convex, we have

$$\nabla^2 H_s(\pi) \succeq \alpha\eta^{-1}\mathrm{diag}\Big(\pi^{\mathsf{ref}}(a_1|s)^{-1}, \cdots, \pi^{\mathsf{ref}}(a_{|\mathcal{A}|}|s)^{-1}\Big).$$

The following lemma, which gives an estimate of $\nabla^2 H_s^*$, is the pivot of the proof.

Lemma C.4. For any reward $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$, we have

$$\nabla^2 H_s^*(r) \preceq \alpha^{-1}\eta\mathrm{diag}\big(\pi^{\mathsf{ref}}(a_1|s), \cdots, \pi^{\mathsf{ref}}(a_{|\mathcal{A}|}|s)\big).$$

Proof of Lemma C.4. Given reward function $r : \mathcal{S} \times \mathcal{A} \to [0, 1]$, we consider

$$\pi_r \in \underset{\pi\in\mathcal{S}\to\Delta^{|\mathcal{A}|}}{\mathrm{argmax}} \,\{\langle \pi(\cdot|s), r(\cdot|s)\rangle - H_s(\pi)\}.$$

From equation (C.1) we know that $\nabla^2 H_s(\pi_r)$ is invertible. Therefore, by Penot 1994, Proposition 3.2, we have $\nabla^2 H_s^*(r) \preceq (\nabla^2 H_s(\pi_r))^{-1}$. Since f is α-strongly convex, we have

$$\nabla^2 H_s^*(r) \preceq \alpha^{-1}\eta\mathrm{diag}\big(\pi^{\mathsf{ref}}(a_1|s), \cdots, \pi^{\mathsf{ref}}(a_{|\mathcal{A}|}|s)\big),$$

which finishes the proof. □

Now we are ready to prove Theorem 3.2.

Proof of Theorem 3.2. Consider our estimation $\bar{g}$ which approximates the ground truth reward function g^*, we know that

$$\widehat{\pi} = \underset{\pi\in\mathcal{S}\to\Delta(\mathcal{A})}{\mathrm{argmax}} \Big\{\mathbb{E}_{(s,a)\sim\rho\times\pi}[\bar{g}(s, a)] - \eta^{-1}\mathbb{E}_{s\sim\rho}\big[D_f(\pi||\pi^{\mathsf{ref}})\big]\Big\}.$$

We have the following sub-optimality decomposition

$$\begin{aligned}
J(\pi^*) - J(\widehat{\pi}) &= \mathbb{E}_{s\sim\rho}\Big[\mathbb{E}_{a\sim\pi^*}[g^*(s, a)] - \mathbb{E}_{a\sim\widehat{\pi}}[g^*(s, a)] - \eta^{-1}\big[D_f(\pi^*\|\pi^{\mathsf{ref}}) - D_f(\widehat{\pi}\|\pi^{\mathsf{ref}})\big]\Big] \\
&= \mathbb{E}_{s\sim\rho}\big[H_s^*(g^*) - H_s^*(\bar{g}) - \langle \widehat{\pi}, g^* - \bar{g}\rangle\,\big] \\
&= \mathbb{E}_{s\sim\rho}\big[H_s^*(g^*) - H_s^*(\bar{g}) - \langle \nabla H_s^*(\bar{g}), g^* - \bar{g}\rangle\,\big] \\
&= \mathbb{E}_{s\sim\rho}[(g^* - \bar{g})^\top \nabla^2 H_s^*(\widetilde{g})(g^* - \bar{g})],
\end{aligned}$$

where $\widetilde{g} = \gamma g^* + (1 - \gamma)\bar{g}$ and $\gamma \in [0, 1]$ and the last equation holds due to Taylor's expansion. Now, for any $\delta \in (0, 1)$ and $\epsilon_c > 0$, with probability at least $1 - \delta$

$$\begin{aligned}
J(\pi^*) - J(\widehat{\pi}) &= \mathbb{E}_{s\sim\rho}[(g^* - \bar{g})^\top \nabla^2 H_s^*(\widetilde{g})(g^* - \bar{g})] \\
&\le \alpha^{-1}\eta\mathbb{E}_{s\sim\rho}\Big[(g^* - \bar{g})^\top \mathrm{diag}\big(\pi^{\mathsf{ref}}(a_1|s), \cdots, \pi^{\mathsf{ref}}(a_{|\mathcal{A}|}|s)\big)(g^* - \bar{g})\Big] \\
&= \alpha^{-1}\eta\mathbb{E}_{(s,a)\sim\rho\times\pi^{\mathsf{ref}}}\big[\big(g^*(s, a) - \bar{g}(s, a)\big)^2\big] \\
&\le \alpha^{-1}\eta\bigg(\frac{128}{3n}\log(2\mathcal{N}_{\mathcal{G}}(\epsilon_c)/\delta) + 18\epsilon_c\bigg),
\end{aligned}$$

where the first inequality holds due to Lemma C.4 and last inequality holds due to equation (B.2). Setting $\epsilon_c = O(n^{-1})$ completes the proof. □

C.2 Proof of Theorem 3.4

We first provide the following lemma that gives the close form of optimal policy under χ^2-divergence regularization.

Lemma C.5 (Huang et al. (2025, Lemma G.2))**.** Let π^* be the optimal policy of χ^2-divergence regularized objective with reward function r, then π^* has the closed form

$$\pi^*(\cdot) = \pi^{\mathsf{ref}}(\cdot) \max\big\{0, \eta(r(\cdot) - \lambda)\big\}, \text{ where } \sum_{a\in\mathcal{A}} \pi^*_{f\mathrm{div}}(a) = 1.$$

By Proposition C.2, $\pi^*_{f\text{div}} = \text{argmax}_{\pi\in\Delta(\mathcal{A})} J_{f\text{div}}(\pi)$ is unique. The sub-optimality gap for f-divergence is consequently defined as

$$\text{SubOpt}_{f\text{div}}(\cdot) \coloneqq \text{SubOpt}_{f\text{div}}(\cdot;\mathcal{A},r,\pi^{\mathsf{ref}}) = J_{f\text{div}}(\pi^*_{f\text{div}}) - J_{f\text{div}}(\cdot). \tag{C.2}$$

Now we are ready to prove Theorem 3.4.

Proof of Theorem 3.4. We still consider the family of contextual bandits $\text{CB}_{\mathcal{G}}$ given by (B.3). We, still, aim to prove the following statement. Fixing any $S \geq 32\log 2$, $\eta > 4\log 2$ and α, we set $f(x) \coloneqq \alpha(x-1)^2/2$, then for any estimator $\mathcal{D} \mapsto \widehat{\pi} \in \Delta(\mathcal{A}|\mathcal{S})$, for any n sufficiently large, there exist some function class $\mathcal{G}$, such that $\exists\ \text{inst} = (\mathcal{S},\mathcal{A},\rho,r,\pi^{\mathsf{ref}},\eta) \in \text{CB}_{\mathcal{G}}$ with $|\mathcal{S}| = S = \Theta(\log|\mathcal{G}|)$, and

$$\text{SubOpt}_f(\widehat{\pi};\text{inst}) \gtrsim \alpha^{-1}\eta S n^{-1}. \tag{C.3}$$

Since $\log|\mathcal{G}| \geq \log\mathcal{N}_{\mathcal{G}}(\epsilon)$ for any $\epsilon \in (0,1)$, equation (C.3) yields the desired bound.
We again omit subscripts $f$div when it is clear in context. We set $\mathcal{S} = [S]$, $\mathcal{A} = \{-1,+1\}$, and $\rho = \mathsf{Unif}(\mathcal{S})$. For all $s \in \mathcal{S}$, $\pi^{\mathsf{ref}} = \mathsf{Unif}(\mathcal{A})$. We further consider the following reward function class. We leverage Lemma F.4 and obtain a set $\mathcal{V} \in \{-1,+1\}^S$ such that (1) $|\mathcal{V}| \geq \exp(S/8)$ and (2) for any $v,v' \in \mathcal{V}, v \neq v'$, one has $\|v-v'\|_1 \geq S/2$. We construct the following reward function class where the reward follows Bernoulli distribution and the mean functions are given by the function class

$$\mathcal{G} = \{r_v(s,-1) = 1/2 + v_s\delta, r_{v'}(s,+1) = 1/2 + v_s\delta,\ \forall s \in \mathcal{S}|v \in \mathcal{V}\},$$

where $\delta \in (0,\eta^{-1}\alpha]$ is to be specified later. Fix some context s and $v_1 \neq v_2$ different at entry s and corresponding reward r_1 and r_2. Without loss of generality, we assume $r_1(s,\cdot) = (1/2+\delta, 1/2-\delta)$ and $r_2(s,\cdot) = (1/2-\delta,1/2+\delta)$. Then direct calculation implies that

$$\pi_1^*(\cdot|s) = \frac{1}{2}\max\{0,\eta\alpha^{-1}(r_1(s,\cdot)-\lambda)\} = 0.5\eta\alpha^{-1}(r_1(s,\cdot)-\lambda),$$
$$\pi_2^*(\cdot|s) = \frac{1}{2}\max\{0,\eta\alpha^{-1}(r_2(s,\cdot)-\lambda)\} = 0.5\eta\alpha^{-1}(r_2(s,\cdot)-\lambda),$$

where $\lambda = 0.5 - \eta^{-1}\alpha$. Note that $2\chi^2(\mu\|\nu) + 1 = \sum_{a\in\mathcal{A}}[\mu(a)]^2/\nu(a)$ and $\chi^2 = D_f$, we obtain that $\forall\ \widehat{\pi}$,

$$\text{SubOpt}_s(\widehat{\pi}(\cdot|s);r_1) + \text{SubOpt}_s(\widehat{\pi}(\cdot|s);r_2) \tag{C.4}$$
$$\begin{aligned}
&= \langle r_1(s,\cdot),\pi_1^*(\cdot|s)\rangle + \langle r_2(s,\cdot),\pi_2^*(\cdot|s)\rangle - \overbrace{\langle r_1(s,\cdot)+r_2(s,\cdot),\widehat{\pi}(\cdot|s)\rangle}^{=1} + \overbrace{2\eta^{-1}\alpha\chi^2(\widehat{\pi}(\cdot|s)\|\pi^{\mathsf{ref}}(\cdot|s))}^{\geq 0} \\
&\quad - \eta^{-1}\alpha\cdot\chi^2(\pi_1^*(\cdot|s)\|\pi^{\mathsf{ref}}(\cdot|s)) - \eta^{-1}\alpha\cdot\chi^2(\pi_2^*(\cdot|s)\|\pi^{\mathsf{ref}}(\cdot|s)) \\
&\geq 2\langle r_1(s,\cdot),\pi_1^*(\cdot|s)\rangle - 1 - 2\eta^{-1}\alpha\cdot\chi^2(\pi_1^*(\cdot|s)\|\pi^{\mathsf{ref}}(\cdot|s)) \\
&= 1 + \frac{2\eta\delta^2}{\alpha} - 1 - \frac{\eta\delta^2}{\alpha} = \frac{\eta\delta^2}{\alpha}.
\end{aligned} \tag{C.5}$$

Now we take expectation over all possible contexts and recall that $\|v-v'\|_1 \geq S/2$ for $v \neq v'$, we know that for any $r_1 \neq r_2 \in \mathcal{G}$

$$\text{SubOpt}(\widehat{\pi};r_1) + \text{SubOpt}(\widehat{\pi};r_2) \geq \frac{\eta\delta^2}{2\alpha}$$

Given any mean reward function $r \in \mathcal{G}$, let P_r be the distribution of $(s,a,\mathtt{r})$ when $s \sim \rho$, $a \sim \pi^{\mathsf{ref}}(\cdot|s)$, and $\mathtt{r} \sim \mathsf{Bern}(r(s,a))$. Suppose $P_{\mathcal{D}_r}$ is the distribution of the dataset given mean reward function r, then $\mathsf{KL}\left(P_{\mathcal{D}_{r_1}}\|P_{\mathcal{D}_{r_2}}\right) = n\mathsf{KL}\left(P_{r_1}\|P_{r_2}\right)$ for any pair of $r_1,r_2 \in \mathcal{G}$. Now we invoke Fano's inequality (Lemma F.2) to obtain

$$\begin{aligned}
\inf_{\pi}\sup_{\text{inst}\in\text{CB}_{\mathcal{G}}}\text{SubOpt}(\widehat{\pi};\text{inst}) &\geq \frac{\eta\delta^2}{4\alpha}\left(1 - \frac{\max_{r_1\neq r_2\in\mathcal{G}}\mathsf{KL}\left(P_{\mathcal{D}_{r_1}}\|P_{\mathcal{D}_{r_2}}\right) + \log 2}{\log|\mathcal{G}|}\right) \\
&\geq \frac{\eta\delta^2}{4\alpha}\left(1 - \frac{64n\delta^2 + 8\log 2}{S}\right),
\end{aligned}$$

Algorithm 3 Offline KL-Regularized Pessimistic Contextual Dueling Bandit (KL-PCDB)

Require: regularization η, reference policy π^{ref}, function class $\mathcal{G}$, offline dataset $\mathcal{D} = \{(s_i, a_i^1, a_i^2, y_i)\}_{i=1}^n$

1: Compute the maximum likelihood estimator of the reward function

$$\bar{g} = \operatorname*{argmin}_{g \in \mathcal{G}} \sum_{i=1}^n \Big[y_i \log \sigma\Big(\big[g(s_i, a_i^1) - g(s_i, a_i^2) \big] \Big) + (1 - y_i) \log \sigma\Big(\big[g(s_i, a_i^2) - g(s_i, a_i^1) \big] \Big) \Big]$$

2: Let $\widehat{g}(s,a) = \bar{g}(s,a) - \Gamma_n(s,a)$, where $\Gamma_n(s,a)$ is the bonus term in (D.1)

Ensure: $\widehat{\pi}(a|s) \propto \pi^{\mathsf{ref}}(a|s) \exp\big(\eta \cdot \widehat{g}(s,a)\big)$

where the second inequality holds due to $\mathsf{KL}(\mathsf{Bern}(p) \| \mathsf{Bern}(q)) \le (p-q)^2/[q(1-q)]$. Let $\delta = 16\sqrt{\alpha \eta^{-1} n^{-1}}$, then we obtain that for all π we have

$$\sup_{\mathrm{inst} \in \mathrm{CB}_{\mathcal{G}}} \mathrm{SubOpt}(\widehat{\pi}; \mathrm{inst}) \gtrsim \frac{\eta S}{\alpha n},$$

which finishes the proof in that $\log_2 |\mathcal{G}| = S$. □

D Missing Details from Section 4

In this section, we provide additional details that has been omitted in Section 4.

D.1 Algorithms for KL-regularized Contextual Dueling Bandits

We elucidate KL-PCDB for offline KL-regularized contextual dueling bandits, whose pseudocode is summarized in Algorithm 3. KL-PCDB first estimate the ground truth function g^* on offline dataset with maximum likelihood estimator (MLE) to estimate a function $\bar{g} \in \mathcal{G}$. After that, analogous to Algorithm 1, we adopt the principle of pessimism in the face of uncertainty. Specifically, we define the penalty term

$$\Gamma_n(s,a) = \beta \sqrt{D^2_{\mathcal{G}}((s,a), \pi^{\mathsf{ref}})}, \tag{D.1}$$

where

$$\beta^2 = 128 \log(2\mathcal{N}_{\mathcal{G}}(\epsilon_c)/\delta)/3n + 18\epsilon_c = \widetilde{O}(n^{-1}) \tag{D.2}$$

and then subtract it from the MLE $\bar{g}$ to obtain a pessimistic estimator $\widehat{g}$. KL-PCB then output the policy $\widehat{\pi}$, maximizing the estimated objective

$$\widehat{J}(\pi) = \mathbb{E}_{(s,a) \sim \rho \times \pi} \bigg[\widehat{g}(s,a) - \eta^{-1} \log \frac{\pi(a|s)}{\pi^{\mathsf{ref}}(a|s)} \bigg],$$

the maximizer of which is in closed form as the counterpart of (2.2).

$$\widehat{\pi}(a|s) \propto \pi^{\mathsf{ref}}(a|s) \exp\big(\eta \cdot \widehat{g}(s,a)\big).$$

We provide the following theoretical guarantees for Algorithm 3.

Theorem D.1. Under Assumption 4.3, if we set Γ_n according to D.1, then for sufficiently small $\epsilon \in (0,1)$, with probability at least $1 - \delta$, $n = \widetilde{O}\big(\eta D^2_{\pi^*} \epsilon^{-1}\big)$ is sufficient to guarantee the output policy $\widehat{\pi}$ of Algorithm 3 to be ϵ-optimal.

Remark D.2. Zhao et al. (2024) achieved an $\widetilde{O}(\epsilon^{-1})$ sample complexity under Assumption 4.2. Comparing to Zhao et al. (2024), KL-PCDB achieves the same $\widetilde{O}(\epsilon^{-1})$ sample complexity but only requiring Assumption 4.3, which is weaker than Assumption 4.2.

The following theorem provides the sample complexity lower bound for KL-regularized dueling contextual bandits.

Theorem D.3. For any sufficiently small $\epsilon \in (0,1), \eta > 0$, $1 \le C^* \le \exp(\eta/2)/2$, and any algorithm Alg, there is a KL-regularized contextual dueling bandit instance with single-policy concentrability $D^2_{\pi^*} \le 2C^*$ such that Alg requires at least $\Omega\big(\min\{\eta C^* \log \mathcal{N}_{\mathcal{G}}(\epsilon_c)/\epsilon, \log \mathcal{N}_{\mathcal{G}}(\epsilon_c)(C^*)^2/\epsilon^2\}\big)$ samples to return an ϵ-optimal policy.

Remark D.4. Theorem D.3 shows that when ϵ is sufficiently small, any algorithm for offline KL-regularized contextual dueling bandits requires at least $\Omega(\eta D^2_{\pi^*} \log \mathcal{N}_{\mathcal{G}}(\epsilon) \epsilon^{-1})$ samples to output an ϵ-optimal policy, which matches the sample complexity upper bound in Theorem D.1, indicating that KL-PCB is nearly optimal.

Algorithm 4 Offline f-Divergence Regularized Contextual Dueling Bandits (f-CDB)

Require: regularization η, reference policy π^{ref}, function class $\mathcal{G}$, offline dataset $\mathcal{D} = \{(s_i, a_i^1, a_i^2, y_i)\}_{i=1}^n$

1: Compute the maximum likelihood estimator of the reward function

$$\bar{g} = \operatorname*{argmin}_{g \in \mathcal{G}} \sum_{i=1}^n \Big[y_i \log \sigma\Big(\big[g(s_i, a_i^1) - g(s_i, a_i^2)\big]\Big) + (1 - y_i) \log \sigma\Big(\big[g(s_i, a_i^2) - g(s_i, a_i^1)\big]\Big)\Big].$$

2: Compute the optimal policy with respect to reward $\bar{g}$

$$\widehat{\pi}(\cdot|s) \leftarrow \operatorname*{argmax}_{\pi(\cdot|s) \in \Delta(\mathcal{A})} \sum_{a \in \mathcal{A}} \pi(a|s)\bar{g}(s,a) + \eta^{-1} D_f\big(\pi(\cdot|s) \| \pi^{\mathsf{ref}}(\cdot|s)\big)$$

Ensure: $\widehat{\pi}(a|s)$

D.2 Algorithm and Results for f-divergence Regularized Contextual Dueling Bandit

In this subsection, we present an offline learning algorithm for f-divergence regularized contextual dueling bandit, f-CDB, in Algorithm 4. f-CDB first leverages maximum likelihood estimator to find a function $\bar{g} \in \mathcal{G}$ that minimizes its risk on the offline dataset. Then the algorithm constructs the output policy $\widehat{\pi}$ that maximizes the f-divergence regularized objective induced by $\bar{g}$. Similar to Algorithm 2, we do not require any pessimism in f-CDB. The following theorem provides an upper bound of Algorithm 4.

Theorem D.5. For any sufficiently small $\epsilon \in (0,1)$, and $\eta, \alpha > 0$, with probability at least $1-\delta$, $n = \widetilde{O}\big(\alpha^{-1}\eta \log \mathcal{N}(\epsilon)\epsilon^{-1}\big)$ is sufficient to guarantee that the output policy $\widehat{\pi}$ of Algorithm 4 is ϵ-optimal.

The following theorem provides a lower bound for offline f-divergence regularized contextual dueling bandit with strongly convex f.

Theorem D.6. For any $\epsilon \in (0,1)$, $\alpha, \eta > 0$, and offline RL algorithm Alg, there is an α-strongly convex f and f-divergence regularized contextual dueling bandit instance such that Alg requires at least $\Omega\big(\alpha^{-1}\eta \log \mathcal{N}(\epsilon)\epsilon^{-1}\big)$ samples to return an ϵ-optimal policy.

Remark D.7. Theorem D.6 indicates that, when ϵ is sufficiently small, to produce an ϵ-optimal policy, any algorithm for offline f-regularized contextual bandits with strongly convex f requires at least $\widetilde{\Omega}(\alpha^{-1}\eta\epsilon^{-1})$ samples. This lower bound matches the sample-complexity upper bound in Theorem D.5, indicating that Algorithm 4 is nearly optimal.

E Missing Proof from Appendix D

E.1 Proof of Theorem D.1

The proof follows the proof in Section 2. At the beginning, we first define the event $\mathcal{E}(\delta)$ given $\delta > 0$ as

$$\mathcal{E}(\delta) := \Big\{\exists\, b : \mathcal{S} \to [-1,1], \forall (s,a) \in \mathcal{S} \times \mathcal{A}, \big|\bar{g}(s,a) - b(s) - g^*(s,a)\big| \le \Gamma_n(s,a)\Big\}. \tag{E.1}$$

Here, Γ_n is defined in (D.1). We abuse the notation and define $b(\cdot)$ as

$$b = \operatorname*{argmin}_{\mathcal{B}} \sup_{(s,a) \in \mathcal{S} \times \mathcal{A}} \Phi_b(s,a) - \Gamma_n(s,a), \tag{E.2}$$

where $\Phi_b(s,a) = \big|\bar{g}(s,a) - b(s) - g^*(s,a)\big|$ and when $\mathcal{E}$ holds, for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, we have $\Phi_b(s,a) \le \Gamma_n(s,a)$ This indicates that the least square estimation $\bar{g}$ obtained in Line 3 of Algorithm 3, after adjusted by some bias function b, is close to the true function g^*. The following lemma shows that this event holds with high probability.

Lemma E.1. For any $\delta > 0$, $\mathbb{P}(\mathcal{E}(\delta)) \ge 1 - \delta$.

Proof. From Lemma F.1, we have that with probability at least $1 - \delta$, it holds that

$$\mathbb{E}_{s' \sim \rho} \operatorname{Var}_{a' \sim \pi^{\mathsf{ref}}(\cdot|s')} \big[\bar{g}(s',a') - g^*(s',a')\big] \le O\bigg(\frac{1}{n} \log(\mathcal{N}_{\mathcal{G}}(\epsilon_c)/\delta) + \epsilon_c\bigg). \tag{E.3}$$

It further holds true that for some $b : \mathcal{S} \to \mathbb{R}$

$$D^2_{\mathcal{G}}((s,a),\pi^{\mathsf{ref}})) \cdot \mathbb{E}_{s\sim\rho} \operatorname{Var}_{a\sim\pi^{\mathsf{ref}}(\cdot|s)} \left[\bar{g}(s,a) - g^*(s,a)\right] \geq \left(\bar{g}(s,a) - b(s) - g^*(s,a)\right)^2. \tag{E.4}$$

Substituting (E.3) into (E.4), we have

$$\begin{aligned}
&\inf_b \left(\bar{g}(s,a) - b(s) - g^*(s,a)\right)^2 && \text{(E.5)}\\
&= \inf_b \frac{\left(\bar{g}(s,a) - b(s) - g^*(s,a)\right)^2}{\mathbb{E}_{s'\sim\rho} \operatorname{Var}_{a'\sim\pi^{\mathsf{ref}}(\cdot|s')} \left[\bar{g}(s',a') - g^*(s',a')\right]} \mathbb{E}_{s'\sim\rho} \operatorname{Var}_{a'\sim\pi^{\mathsf{ref}}(\cdot|s')} \left[\bar{g}(s',a') - g^*(s',a')\right]\\
&\leq D^2_{\mathcal{G}}((s,a),\pi^{\mathsf{ref}}) \mathbb{E}_{\pi^{\mathsf{ref}}}\left[\left(\bar{g}(s,a) - b(s) - g^*(s,a)\right)^2\right] && \text{(E.6)}\\
&\leq D^2_{\mathcal{G}}((s,a),\pi^{\mathsf{ref}}) O\left(\frac{1}{n}\log(\mathcal{N}_{\mathcal{G}}(\epsilon_c)/\delta) + \epsilon_c\right), && \text{(E.7)}
\end{aligned}$$

where the first inequality holds due to the definition of $D^2_{\mathcal{G}}((s,a),\pi^{\mathsf{ref}})$ and the last inequality holds due to Lemma F.1. $\square$

We overload the following quantities. For any $\gamma \in [0,1]$ and $(s,a) \in \mathcal{S} \times \mathcal{A}$, we define

$$g_\gamma(s,a) := \gamma(\widehat{g}(s,a) - b(s)) + (1-\gamma) g^*(s,a).$$

Furthermore, we introduce the following quantities

$$\begin{aligned}
\pi_\gamma(\cdot|\cdot) &= \pi_{g_\gamma}(\cdot|\cdot) \propto \pi^{\mathsf{ref}}(\cdot|\cdot) \exp\left(\eta g_\gamma(\cdot,\cdot)\right),\\
G(\gamma) &:= \mathbb{E}_{\rho\times\pi_\gamma}\left[\left(\widehat{g}(s,a) - b(s) - g^*(s,a)\right)^2\right],
\end{aligned}$$

where $b(\cdot)$ is defined in (E.2). We still have the monotonicity of the function $G(\gamma)$, which is characterized by the following lemma.

Lemma E.2. On event $\mathcal{E}(\delta)$, $0 \in \operatorname{argmax}_{\gamma\in[0,1]} G(\gamma)$.

Proof. For simplicity, we use $\triangle(s,a)$ to denote $\widehat{g}(s,a) - b(s) - g^*(s,a)$ in *this* proof. Then on event $\mathcal{E}(\delta)$, we know that $\triangle(s,a) \leq 0$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Taking derivatives of G w.r.t., γ directly, we conclude that for all $\gamma \in [0,1]$,

$$\begin{aligned}
G'(\gamma) &= \eta \mathbb{E}_\rho \mathbb{E}_{a\sim\pi_\gamma}\left[\triangle^2(s,a)\left(\triangle(s,a) - \mathbb{E}_{a'\sim\pi_\gamma}[\triangle(s,a')]\right)\right]\\
&= \eta \mathbb{E}_\rho\Big[\mathbb{E}_{\pi_\gamma}\left[\triangle^3(s,a)\right] - \mathbb{E}_{\pi_\gamma}\left[\triangle^2(s,a)\right]\mathbb{E}_{\pi_\gamma}\left[\triangle(s,a)\right]\Big]\\
&\leq 0,
\end{aligned}$$

where $\mathbb{E}_\rho$ is the shorthand of $\mathbb{E}_{s\sim\rho}$, $\mathbb{E}_{\pi_\gamma}$ is the shorthand of $\mathbb{E}_{a\sim\pi_\gamma}$ and the inequality holds conditioned on the event $\mathcal{E}(\delta)$ due to Lemma 2.13. $\square$

Finally, we have the proposition that adding some bias term $b : \mathcal{S} \to \mathbb{R}$ does not affect the resulting policy.

Proposition E.3. Let $b : \mathcal{S} \to \mathbb{R}$ be some bias function, then for all $g \in \mathcal{G}$ we have $J(\pi_g) = J(\pi_{g-b})$, where $(g-b)(s,a) = g(s,a) - b(s)$.

Proof. For any fixed state $s \in \mathcal{S}$, we have for any $a \in \mathcal{A}$ that,

$$\begin{aligned}
\pi_g(a|s) &= \frac{\pi^{\mathsf{ref}}(a|s)\exp\left(\eta g(s,a)\right)}{\sum_{a'\in\mathcal{A}} \pi^{\mathsf{ref}}(a'|s)\exp\left(\eta g(s,a')\right)}\\
&= \frac{\pi^{\mathsf{ref}}(a|s)\exp\left(\eta g(s,a)\right)\exp\left(-\eta b(s)\right)}{\sum_{a'\in\mathcal{A}} \pi^{\mathsf{ref}}(a'|s)\exp\left(\eta g(s,a')\right)\exp\left(-\eta b(s)\right)}\\
&= \frac{\pi^{\mathsf{ref}}(a|s)\exp\left(\eta[g(s,a) - b(s)]\right)}{\sum_{a'\in\mathcal{A}} \pi^{\mathsf{ref}}(a'|s)\exp\left(\eta[g(s,a') - b(s)]\right)}\\
&= \pi_{g-b}(a|s),
\end{aligned}$$

which indicates that $\pi_g = \pi_{g-b}$. This immediately leads to $J(\pi_g) = J(\pi_{g-b})$. $\square$

Now we are ready to prove Theorem D.1.

Proof of Theorem D.1. We proceed the proof under the event $\mathcal{E}(\delta)$. By Proposition E.3, we know that

$$\begin{aligned} J(\pi^*) - J(\widehat{\pi}) &= J(\pi^*) - J(\pi_{\widehat{g}}) \\ &= J(\pi^*) - J(\pi_{\widehat{g}-b}). \end{aligned}$$

Consequently, there exist some $\gamma \in [0, 1]$ and $b : \mathcal{S} \to [-1, 1]$ such that

$$\begin{aligned} J(\pi^*) - J(\widehat{\pi}) &= J(\pi^*) - J(\pi_{\widehat{g}-b}) \\ &\le \eta \mathbb{E}_{\rho\times\pi_\gamma}\left[\left(\widehat{g}(s,a) - b(s) - g^*(s,a)\right)^2\right] \\ &= \eta G(\gamma), \end{aligned} \tag{E.8}$$

where the inequality holds due to Lemma 2.16. Under event $\mathcal{E}(\delta)$, we know that $\widehat{g}(s,a) - b(s) \le g^*(s,a)$. Together with Lemma E.2, we obtain $G(\gamma) \le G(0)$. Therefore, we know that

$$\begin{aligned} J(\pi^*) - J(\widehat{\pi}) &\le G(0) \\ &= \eta \mathbb{E}_{\rho\times\pi^*}\left[\left(\widehat{g}(s,a) - b(s) - g^*(s,a)\right)^2\right] \\ &\le 4\eta \mathbb{E}_{\rho\times\pi^*}\left[\Gamma_n^2(s,a)\right] \\ &= 4\eta\beta^2 \mathbb{E}_{\rho\times\pi^*}\left[D_{\mathcal{G}}^2((s,a); \pi^{\mathsf{ref}})\right] \\ &= \widetilde{O}\left(\eta D_{\pi^*}^2 \log \mathcal{G}(\epsilon_c) n^{-1}\right), \end{aligned} \tag{E.9}$$

where the inequality holds due to the definition of $\mathcal{E}(\delta)$. Plugging (E.9) into (E.8), we know that $J(\pi^*) - J(\widehat{\pi})$ has upper bound $\widetilde{O}(D_{\pi^*}^2 n^{-1})$. By Lemma E.1, event $\mathcal{E}$ with probability at least $1-\delta$, which concludes the proof. □

E.2 Proof of Theorem D.3

Proof of Theorem D.3. The proof is similar to the proof of Theorem 2.10. Consider the following family of contextual dueling bandit instances with $S := |\mathcal{S}|, A := |\mathcal{A}| < \infty$ and reward in some function class $\mathcal{G}$.

$$\mathrm{CDB} := \{(\mathcal{S}, \mathcal{A}, \rho, r, \pi^{\mathsf{ref}}, \eta) : r \in \mathcal{G}, \rho \in \Delta(\mathcal{S}), \pi^{\mathsf{ref}} \in \Delta(\mathcal{A}|\mathcal{S})\}. \tag{E.10}$$

Fixing any $S \ge 1$, $\eta > 4\log 2$ and $C^* \in (2, \exp(\eta/4)]$, we aim to prove that, for any estimator $\mathcal{D} \mapsto \widehat{\pi} \in \Delta(\mathcal{A}|\mathcal{S})$, for any $n \ge 16SC^*$, there exist some function class $\mathcal{G}$, such that $\exists$ inst $= (\mathcal{S}, \mathcal{A}, \rho, r, \pi^{\mathsf{ref}}, \eta) \in$ CDB with single-policy concentrability $C^{\pi^*} \le 2C^*$, regularization coefficient η, $|\mathcal{S}| = S = \Theta(\log|\mathcal{G}|)$, and

$$\inf_{\mathrm{inst}\in\mathrm{CDB}} \mathrm{SubOpt}_{\mathrm{RKL}}(\widehat{\pi}; \mathrm{inst}) \gtrsim \min\{\eta SC^* n^{-1}, (SC^*)^{1/2} n^{-1/2}\}. \tag{E.11}$$

Since $\log|\mathcal{G}| \ge \log \mathcal{N}_{\mathcal{G}}(\epsilon)$ for any $\epsilon \in (0,1)$, the above bound yields the desired result.
We construct the same reward function class as in the proof of Theorem 2.10. In particular, we set $\mathcal{S} = [S]$, $\mathcal{A} = \{\pm 1\}$, $\rho = \mathsf{Unif}(\mathcal{S})$, and the reference policy to be

$$\forall s \in \mathcal{S}, \pi^{\mathsf{ref}}(-1|s) = C^{-1}, \pi^{\mathsf{ref}}(+1|s) = 1 - C^{-1};$$

where $C = C^*$. Then the total sub-optimality of any $\pi \in \Delta(\mathcal{A}|\mathcal{S})$ given any reward function $r : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ is

$$\mathrm{SubOpt}_{f\mathrm{div}}(\pi; r) = \frac{1}{S}\sum_{s=1}^{S} \mathrm{SubOpt}_{f\mathrm{div}}\left(\pi(\cdot|s); r(s,\cdot)\right). \tag{E.12}$$

We further let $\alpha = \eta^{-1}\log(C-1) \Leftrightarrow C - 1 = \exp(\eta\alpha)$. We construct 2^S Bernoulli reward functions, in particular, $\forall \tau \in \{\pm 1\}^S$, the mean function r_τ of the reward (indexed by τ) is defined as

$$r_\tau(s, -1) = 0.5 + \tau_s\delta, r_\tau(s, +1) = 0.5 - \alpha.$$

Then, following the derivation of (B.10) and (B.11), we know that $\max_{\tau\in\{\pm 1\}^S} C^{\pi^*_\tau} \le C^*$, and $\forall s \in \mathcal{S}, \forall \tau, \tau' \in \{\pm 1\}^{\mathcal{S}}$ with $\tau \sim_s \tau'$,

$$\mathrm{SubOpt}_s(\widehat{\pi};\tau) + \mathrm{SubOpt}_s(\widehat{\pi};\tau') \ge \frac{\eta\delta^2}{8} \wedge \frac{3\delta}{10}. \tag{E.13}$$

Let P_r be the distribution of (s, a^1, a^2, y) for $s \sim \rho$, $a^1, a^2 \overset{\text{i.i.d.}}{\sim} \pi^{\mathsf{ref}}(\cdot|s)$ and $y \sim \mathsf{Bern}(\sigma(r(s,a^1) - r(s,a^2)))$. Now we set $\delta = \sqrt{S/n}$ and conclude that for $\tau \sim \tau'$ with $\tau_s = -\tau'_s$,

$$\begin{aligned}
&\mathsf{KL}\left(P_{r_\tau} \| P_{r_{\tau'}}\right) \\
&= \frac{(C-1)}{SC^2} \sum_{s',a^1,a^2} \mathsf{KL}\left(\mathsf{Bern}(\sigma(r_\tau(s',a^1) - r(s',a^2))) \| \mathsf{Bern}(\sigma(r_{\tau'}(s',a^1) - r(s',a^2)))\right) \\
&= \frac{2(C-1)}{SC^2}\Big(\mathsf{KL}\left(\mathsf{Bern}(\sigma(\alpha+\delta)) \| \mathsf{Bern}(\sigma(\alpha-\delta))\right) \vee \mathsf{KL}\left(\mathsf{Bern}(\sigma(\alpha-\delta)) \| \mathsf{Bern}(\sigma(\alpha+\delta))\right)\Big).
\end{aligned}$$

Since $\alpha, \delta \in (0, 1/2)$, by the fact $\mathsf{KL}(P\|Q) \le 2Q_{\min}^{-1}\mathsf{TV}\left(P\|Q\right)^2$ (see e.g., Polyanskiy & Wu (2025, Section 7.6)), we know that

$$\begin{aligned}
\mathsf{KL}\left(P_{r_\tau} \| P_{r_{\tau'}}\right) &\le \frac{2(C-1)}{SC^2}\frac{4}{1+\exp(\alpha+\delta)}\left(\frac{1}{1+\exp(\alpha-\delta)} - \frac{1}{1+\exp(\alpha+\delta)}\right)^2 \\
&\le \frac{4}{3SC}\frac{\exp(2\alpha)(\exp(\delta)-\exp(-\delta))^2}{(1+\exp(\alpha-\delta))^4} \\
&\le \frac{4e}{3SC}(\exp(\delta)-\exp(-\delta))^2 \\
&\le 36S^{-1}C^{-1}\delta^2,
\end{aligned} \tag{E.14}$$

where the second and third inequality hold due to $\alpha, \delta \le 1/2$, and last inequality follows from $\exp(x) - \exp(-x) \le 3x$ for $x \in [0, 1/2]$. Now we set $\delta = \sqrt{SC/n} \le 1/4$. We substitute (E.13) into Assouad's Lemma (Lemma F.3) and obtain that

$$\begin{aligned}
\inf_{\text{inst}\in\text{CDB}} \mathrm{SubOpt}_{\text{RKL}}(\widehat{\pi};\text{inst}) &\ge \frac{1}{4}S \cdot \frac{1}{S} \cdot \left(\frac{\eta\delta^2}{8} \wedge \frac{3\delta}{10}\right) \cdot \min_{\tau\sim\tau'} \exp\Big(-\mathsf{KL}\left(P_{\mathcal{D}_\tau} \| P_{\mathcal{D}_{\tau'}}\right)\Big) \\
&= \frac{1}{4}\left(\frac{\eta\delta^2}{8} \wedge \frac{3\delta}{10}\right)\exp\Big(-n\mathsf{KL}\Big(P_{r_\tau} \| P_{r_{\mathcal{D}_{\tau'}}}\Big)\Big) \\
&\ge \frac{\exp(-36)}{32}\min\{\eta CSn^{-1}, S^2C^2n^{-2}\},
\end{aligned}$$

where the $1/S$ comes from the denominator of (E.12) and the second inequality follows from (E.14). □

E.3 Proof of Theorem D.5

Proof of Theorem D.5. The proof is similar to the proof of Theorem 3.2. Recall that $b(\cdot)$ defined in (E.2), we know that

$$\begin{aligned}
\widehat{\pi} &= \operatorname*{argmax}_{\pi\in\Delta^d}\Big\{\mathbb{E}_{(s,a)\sim\rho\times\pi}[\bar{g}(s,a)] - \eta^{-1}\mathbb{E}_{s\sim\rho}\big[D_f(\pi||\pi^{\mathsf{ref}})\big]\Big\} \\
&= \operatorname*{argmax}_{\pi\in\Delta^d}\Big\{\mathbb{E}_{(s,a)\sim\rho\times\pi}[\bar{g}(s,a) - b(s)] - \eta^{-1}\mathbb{E}_{s\sim\rho}\big[D_f(\pi||\pi^{\mathsf{ref}})\big]\Big\}.
\end{aligned}$$

We have the following sub-optimality decomposition

$$\begin{aligned}
J(\pi^*) - J(\widehat{\pi}) &= \mathbb{E}_{s\sim\rho}\Big[\mathbb{E}_{a\sim\pi^*}[g^*(s,a)] - \mathbb{E}_{a\sim\widehat{\pi}}[g^*(s,a)] - \eta^{-1}\big[D_f(\pi^*\|\pi^{\mathsf{ref}}) - D_f(\widehat{\pi}\|\pi^{\mathsf{ref}})\big]\Big] \\
&= \mathbb{E}_{s\sim\rho}\big[H_s^*(g^*) - H_s^*(\bar{g}-b) - \langle\widehat{\pi}, g^* - \bar{g} + b\rangle\,\big] \\
&= \mathbb{E}_{s\sim\rho}\big[H_s^*(g^*) - H_s^*(\bar{g}-b) - \langle\nabla H_s^*(\bar{g}-b), g^* - \bar{g} + b\rangle\,\big] \\
&= \mathbb{E}_{s\sim\rho}[(g^* - \bar{g} + b)^\top \nabla^2 H_s^*(\widetilde{g})(g^* - \bar{g} + b)],
\end{aligned}$$

where $\widetilde{g} = \gamma g^* + (1-\gamma)\bar{g}$ and $\gamma \in [0,1]$, $(\bar{g} - b)(s,a) = \bar{g}(s,a) - b(s)$ and the last equation holds due to Taylor's expansion. Now, for any $\delta \in (0,1)$ and $\epsilon_c > 0$, with probability at least $1-\delta$

$$\begin{aligned}
J(\pi^*) - J(\widehat{\pi}) &= \mathbb{E}_{s\sim\rho}[(g^* - \bar{g} + b)^\top \nabla^2 H_s^*(\widetilde{g})(g^* - \bar{g} + b)] \\
&\leq \alpha^{-1}\eta\mathbb{E}_{s\sim\rho}\Big[(g^* - \bar{g} + b)^\top \mathrm{diag}\big(\pi^{\mathsf{ref}}(a_1|s), \cdots, \pi^{\mathsf{ref}}(a_d|s)\big)(g^* - \bar{g} + b)\Big] \\
&= \alpha^{-1}\eta\mathbb{E}_{(s,a)\sim\rho\times\pi^{\mathsf{ref}}}\left[\big(g^*(s,a) - \bar{g}\big(s,a\big) + b(s)\big)^2\right] \\
&\leq \alpha^{-1}\eta\left(\frac{128}{3n}\log(2\mathcal{N}_{\mathcal{G}}(\epsilon_c)/\delta) + 18\epsilon_c\right),
\end{aligned}$$

where the first inequality holds due to Lemma C.4 and last inequality holds due to equation (E.3). Setting $\epsilon_c = O(n^{-1})$ completes the proof. □

E.4 Proof of Theorem D.6

Proof of Theorem D.6. We still consider the contextual dueling bandit instance class defined in (E.10). We show that given any positive α, η, for any $n \geq S \cdot \max\{16, \eta^2\alpha^{-2}\}$, there exists $f : \mathbb{R} \to \mathbb{R}$ such that f is α-strongly convex, $\log|\mathcal{G}| = \Theta(S)$ and

$$\inf_{\widehat{\pi}\in\widehat{\Pi}(\mathcal{D})} \sup_{\mathrm{inst}\in\mathrm{CDB}} \mathrm{SubOpt}_{f\mathrm{div}}(\widehat{\pi}; \mathrm{inst}) \gtrsim \frac{\eta S}{\alpha n}, \tag{E.15}$$

where $\mathcal{D} = \{(s_i, a_i^1, a_i^2, y_i)\}_{i=1}^n$ is the offline preference dataset, all (possibly randomized) maps from which to $\Delta(\mathcal{A}|\mathcal{S})$ is denoted by $\widehat{\Pi}(\mathcal{D})$. Since $S = \Theta(\log|\mathcal{G}|) \gtrsim \log\mathcal{N}_{\mathcal{G}}(\epsilon_c)$ for all $\epsilon_c \in (0,1)$, we can conclude the theorem.
Let $\mathcal{S} = [S]$, $\mathcal{A} = \{\pm 1\}$, $\rho = \mathsf{Unif}(\mathcal{S})$ and $\pi^{\mathsf{ref}}(\cdot|s) = \mathsf{Unif}(\mathcal{A})$ for any $s \in \mathcal{S}$. Then the total sub-optimality of any $\pi \in \Delta(\mathcal{A}|\mathcal{S})$ given any reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is

$$\mathrm{SubOpt}_{f\mathrm{div}}(\pi; r) = \frac{1}{S}\sum_{s=1}^{S} \mathrm{SubOpt}_{f\mathrm{div}}\big(\pi(\cdot|s); r(s,\cdot)\big). \tag{E.16}$$

We still consider the reward function class $\mathcal{G}$ indexed by $\{\pm1\}^{\mathcal{S}}$. For all $\tau \in \{\pm1\}^{\mathcal{S}}$ the reward instance "shaped" by τ is

$$r_\tau(s,a) = \frac{1}{2} + a\tau_s \cdot \sqrt{\frac{S}{n}}, \tag{E.17}$$

where $a\tau_s = \pm1$ because $a \in \mathcal{A} = \{\pm1\}$. We thereby refer $\tau \sim \tau'$ to any pair in $\{\pm1\}^{\mathcal{S}}$ that differs only in one coordinate. $\forall \tau, \tau' \in \{\pm1\}^{\mathcal{S}}$, if $\tau \sim \tau'$, then suppose $\tau_s = -\tau'_s$, we have

$$\mathrm{SubOpt}_{f\mathrm{div}}(\pi(\cdot|s); r_\tau(s,\cdot)) + \mathrm{SubOpt}_{f\mathrm{div}}(\pi(\cdot|s); r_{\tau'}(s,\cdot)) \geq \frac{\eta S}{\alpha n}, \tag{E.18}$$

where the inequality follows from exactly the same calculation in equation (C.5) by setting $f(x) = \alpha(x-1)^2/2$.[3] Let P_r be the distribution of (s, a^1, a^2, y) for $s \sim \rho$, $a^1, a^2 \overset{\text{i.i.d.}}{\sim} \pi^{\mathsf{ref}}(\cdot|s)$ and $y \sim \mathsf{Bern}(\sigma(r(s,a^1) - r(s,a^2)))$. Then we denote $\delta = \sqrt{S/n}$ and conclude that for $\tau \sim \tau'$ with $\tau_s = -\tau'_s$,

$$\begin{aligned}
\mathsf{KL}\left(P_{r_\tau} \| P_{r_{\tau'}}\right) &= \frac{1}{SA^2}\sum_{s',a^1,a^2} \mathsf{KL}\left(\mathsf{Bern}(\sigma(r_\tau(s',a^1) - r(s',a^2)) \| \mathsf{Bern}(\sigma(r_{\tau'}(s',a^1) - r(s',a^2))\right) \\
&= \frac{1}{4S}\Big(\mathsf{KL}\left(\mathsf{Bern}(\sigma(2\delta)) \| \mathsf{Bern}(\sigma(-2\delta))\right) + \mathsf{KL}\left(\mathsf{Bern}(\sigma(-2\delta)) \| \mathsf{Bern}(\sigma(2\delta))\right)\Big) \\
&\leq \frac{1}{4S}\Big((\exp(-2\delta) - 1)^2 + (\exp(2\delta) - 1)^2\Big) \\
&\leq \frac{1}{2S}(\exp(2\delta) - 1)^2 \leq \frac{36\delta^2}{2S} = \frac{18}{n},
\end{aligned} \tag{E.19}$$

[3] Recall that in this case $D_f = \chi^2$, where $2\chi^2(\mu\|\nu) + 1 = \sum_{a\in\mathcal{A}}[\mu(a)]^2/\nu(a)$.

where the last inequality follows from $\exp(x) - 1 \le 3x$ for $x \in [0, 0.5]$ and $\delta = \sqrt{S/n} \le 0.25$ by assumption. Therefore, we substitute (E.18) into Assouad's Lemma (Lemma F.3) to obtain

$$\text{LHS of (E.15)} \ge \frac{1}{S} \cdot S \cdot \frac{\eta S}{\alpha n} \cdot \frac{1}{4} \cdot \min_{\tau \sim \tau'} \exp\Big(- \mathsf{KL}\left(P_{\mathcal{D}_\tau} \| P_{\mathcal{D}_{\tau'}}\right)\Big)$$
$$= 0.25 \cdot \frac{\eta S}{\alpha n} \cdot \exp\Big(- n\mathsf{KL}\left(P_{r_\tau} \| P_{r_{\mathcal{D}_{\tau'}}}\right)\Big) \ge \frac{\eta S}{\alpha n} \cdot \frac{1}{3} \cdot \exp(-18) \gtrsim \frac{\eta S}{\alpha n}, \tag{E.20}$$

where the $1/S$ comes from the denominator of (E.16) and the second inequality follows from (E.19). □

F Technical Results and Standard results

Lemma F.1 (Zhao et al. 2024, Lemma D.4)**.** Consider a offline dataset $\{(s_i, a_i^1, a_i^2, y_i)\}_{i=1}^n$ generated from the product of the context distribution $\rho \in \Delta(\mathcal{S})$, policy $\pi \in \Delta(\mathcal{A}|\mathcal{S})$, and the Bradley-Terry Model defined in Section 4.1. Suppose $\bar{g}$ is the result of MLE estimation of Algorithm 3, and we further define $b(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[\bar{g}(s,a) - g^*(s,a)\right]$, then with probability at least $1 - 2\delta$, we have

$$\mathbb{E}_{s,a\sim\rho\times\pi}\left[\left(\bar{g}(s,a) - g^*(s,a) - b(s)\right)^2\right] \le O\left(\frac{1}{n}\log(\mathcal{N}_{\mathcal{G}}(\epsilon_c)/\delta) + \epsilon_c\right).$$

Lemmas F.2 and F.3 are two standard reductions (Le Cam, 1973; Yu, 1997; Polyanskiy & Wu, 2025). See, e.g., Chen et al. (2024, Section 3) for a general proof.

Lemma F.2 (Fano's inequality)**.** Fix any $\mathcal{R} := \{r_1, \cdots, r_S\}$ and policy class Π, let $L : \Pi \times \mathcal{R} \to \mathbb{R}_+$ be some loss function. Suppose there exist some constant $c > 0$ such that the following condition holds:

$$\min_{i \ne j} \min_{\pi \in \Pi} L(\pi, r_i) + L(\pi, r_j) \ge c.$$

Then we have

$$\inf_{\pi\in\Pi}\sup_{r\in\mathcal{R}} L(\pi, r) \ge \frac{c}{2}\left(1 - \frac{\max_{i\ne j}\mathsf{KL}(P_{r_i}\|P_{r_j}) + \log 2}{\log S}\right),$$

where P_r is the distribution of dataset given model $r \in \mathcal{R}$.

Lemma F.3 (Assouad's Lemma)**.** Let $\mathcal{R}$ be the set of instances, Π be the set of estimators, $\Theta := \{\pm 1\}^S$ for some $S > 0$, and $\{L_j\}_{j=1}^S$ be S functions from $\Pi \times \mathcal{R}$ to $\mathbb{R}_+$. Suppose $\{r_\theta\}_{\theta\in\Theta} \subset \mathcal{R}$ and the loss function is

$$L(\pi, r) := \sum_{j=1}^S L_j(\pi, r), \forall(\pi, r) \in \Pi \times \mathcal{R}.$$

We denote $\theta \sim_j \theta'$ if they differ only in the j-th coordinate. Further assume that

$$\theta \sim_j \theta' \Rightarrow \inf_{\pi\in\Pi} L_j(\pi, r_\theta) + L_j(\pi, r_{\theta'}) \ge c \tag{F.1}$$

for some $c > 0$, then

$$\inf_{\pi\in\Pi}\sup_{r\in\mathcal{R}} L(\pi, r) \ge S \cdot \frac{c}{4} \min_{\exists j:\theta\sim_j\theta'} \exp\Big(- \mathsf{KL}\left(P_{r_\theta}\|P_{r_{\theta'}}\right)\Big),$$

where P_r denotes the distribution of the dataset given $r \in \mathcal{R}$.

The following Lemma F.4 is due to Gilbert (1952); Varshamov (1957).

Lemma F.4. There exists a subset $\mathcal{V}$ of $\{-1, 1\}^S$ such that (1) $|\mathcal{V}| \ge \exp(S/8)$ and (2) for any $v, v' \in \mathcal{V}, v \ne v_j$, one has $\|v - v'\|_1 \ge S/2$.