Fast Rates for Offline Contextual Bandits with Forward-KL Regularization under Single-Policy Concentrability

Qingyue Zhao[1]*, **Kaixuan Ji**[1]*, **Heyang Zhao**[1], **Quanquan Gu**[1]
[1]University of California, Los Angeles
{zhaoqy24,kaixuanji,hyzhao,qgu}@cs.ucla.edu

Abstract

Kullback-Leibler (KL) regularization is ubiquitous in reinforcement learning algorithms in the form of *reverse* or *forward* KL. Recent studies have demonstrated ϵ^{-1}-type fast rates for decision making under reverse KL regularization, in contrast to the standard ϵ^{-2}-type sample complexity. However, for forward-KL-regularized objectives, existing statistical analyses are either not applicable or result in $\widetilde{O}(\epsilon^{-2})$ slow rates. We take the first step towards addressing this problem via a streamlined analysis of forward-KL-regularized offline CBs. We give the first $\widetilde{O}(\epsilon^{-1})$ upper bounds in tabular and general function approximation settings, both under notions of *single-policy concentrability*. In particular, our convex-analytical pipeline unifies these settings by exploiting the pessimism principle in a novel way and completely bypasses the proof routines in previous works based on the mean value theorem, which might be of independent interest. Moreover, we provide rate-optimal lower bounds, manifesting the tightness of our upper bounds in terms of statistical rates. Our lower bounds also demonstrate that the forward-KL-regularized sample complexity recovers the unregularized slow rate in the low-regularization regime, similarly to the reverse-KL regularization.

1 Introduction

KL-regularized objectives in the form of $J := \mathbb{E}[r] - \eta^{-1}\mathsf{KL}$ have recently become a pivotal prototype for algorithm design in decision making such as robotics (Levine and Koltun, 2013; Schulman, 2015; Haarnoja et al., 2018), and foundation model finetuning (Ouyang et al., 2022; Rafailov et al., 2023; Ji et al., 2023; Wang et al., 2023; Guo et al., 2025; Shan et al., 2025). Here, r stands for variants of reward or advantage functions, η^{-1} is the regularization intensity; and KL is typically either $\mathsf{KL}\left(\pi \| \pi^{\mathsf{ref}}\right)$ or $\mathsf{KL}\left(\pi^{\mathsf{ref}} \| \pi\right)$, where π^{ref} is the reference policy. The reverse KL regularizer $\mathsf{KL}\left(\pi \| \pi^{\mathsf{ref}}\right)$ has been arguably more popular in algorithmic formulations, which has rich connections with Gibbs distributions and entropy regularization in machine learning (Williams, 1992; McAllester, 1999; Ziebart et al., 2008; Zhang, 2023). But there are also emerging lines of works that formulate their RL fine-tuning algorithms via the forward KL regularizer $\mathsf{KL}\left(\pi^{\mathsf{ref}} \| \pi\right)$ (Ji et al., 2023; Wang et al., 2023; Shan et al., 2025), or *effectively* employ this forward counterpart at the implementation level (Guo et al., 2025; Tang and Munos, 2025; Shah et al., 2025).

There has been a line of efforts motivated by the practical relevance of KL-regularized objectives towards understanding the *data efficiency* of learning with respect to *KL-regularized performance metrics*, which dates back to at least Tiapkin et al. (2023); Xie et al. (2024); Xiong et al. (2024). In the case of reverse KL, sharpest rates in previous works have been $\mathrm{polylog}(T)$ regret (Zhao et al., 2025b; Ji et al., 2026; Nayak et al., 2025) and $\widetilde{\Theta}(\epsilon^{-1})$ sample complexity (Tiapkin et al., 2023; Zhao et al., 2025a, 2026; Foster et al., 2025; Zhang et al., 2026) in offline, online, and hybrid settings for

*Equal Contribution

Accepted at ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning..

Table 1: Comparison of regret or sample complexity upper and lower bounds for forward KL-regularized CBs. In this table, ϵ is the target suboptimality gap, and η is the inverse regularization intensity. Under function approximation, $\mathcal{G}$ is the function class, whose covering number is $N_{\mathcal{G}}$, and $D^2_{\pi^*}$ is the D^2-type single policy concentrability. In the tabular setting, $\mathcal{S}$ is the context set and $\mathcal{A}$ is the action set, and $|\mathcal{S}|$ and $|\mathcal{A}|$ denotes the corresponding cardinality. Furthermore, $C_{\varepsilon_{\text{FKL}}}$ is the density-ratio-based local forward KL-ball coverage coefficient, and C^{π^*} is the density-ratio-based single-policy concentrability. $\widetilde{O}(\cdot)$ hides logarithmic factors except $\log(N_{\mathcal{G}})$.

Type	Algorithm	Setting	Sample Complexity
Upper Bound	Forward KL-Regularized RLHF (Aminian et al., 2025)	Function Approximation	$\widetilde{O}\big(\log N_{\mathcal{G}} C_{\varepsilon_{\text{FKL}}}/\epsilon^2\big)$
	FKL-PCB (This Work)	Function Approximation	$\widetilde{O}\big(\eta C^{\pi^*} D^2_{\pi^*} \log N_{\mathcal{G}}/\epsilon\big)$
	FKL-PCB (This Work)	Tabular	$\widetilde{O}\big(\eta (C^{\pi^*})^2 \lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert/\epsilon\big)$
Lower Bound	This Work	Tabular	$\Omega(\eta\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert/\epsilon)$

contextual bandits (CBs) or Markov decision processes (MDPs). The analysis of the forward KL counterparts is significantly less explored. Zhao et al. (2026) derives $\widetilde{\Theta}(\epsilon^{-1})$ bounds for general f-divergence-regularized suboptimality for offline CBs, but they need f to be strongly convex, and is thus inapplicable to forward KL because forward KL is f-divergence with $f(x) = -\log x$, which is only strictly convex. Lee et al. (2026) derives a logarithmic regularized regret upper bound with general regularizers in a self-play setting, but the bound directly scales inversely with the minimal eigenvalue of the expected feature covariance matrix, and their regularizer needs to be strongly convex, which may not be the case for forward KL unless the sample complexity is allowed to inversely scale with the minimal mass $\min_{s,a} \pi^{\mathsf{ref}}(a|s)$ of the reference policy.[2] Aminian et al. (2025) gives the first $\widetilde{O}(\epsilon^{-2})$ upper bound for offline CBs with forward-KL regularization, where no lower bounds are known by far. Therefore, the following problem remains open **even for learning from pure** i.i.d. **data**:

What is the statistical rate of decision making with respect to forward-KL-regularized objectives?

Any meaningful progress regarding this problem in the offline setting is likely to confront the distributional shift between the i.i.d. data and the optimal policy. On one hand, *single-policy concentrability*, where the behavioral policy π^{ref} only covers the optimal policy, is shown to be a sufficient notion that permits the near-optimal $\widetilde{\Theta}(1/\epsilon)$ sample complexity in many **reverse**-KL-regularized settings where the contexts and actions are acquired offline (Zhao et al., 2026; Foster and Rakhlin, 2023). On the other hand, classic wisdom implies that the forward-KL regularizer and its reverse counterpart adapt π to the modes of π^{ref} in distinct ways (Bishop and Nasrabadi, 2006; Ji et al., 2023; GX-Chen et al., 2025): in short, (in certain parametrized cases) reverse KL is believed to be mode-seeking, which heavily penalizes $\pi > 0$ where $\pi^{\mathsf{ref}} \approx 0$; while forward KL appears to be mass-covering, which mainly penalizes $\pi \to 0$ where $\pi^{\mathsf{ref}} > 0$. The following problem then arises **on top of the statistical rate problem** above:

Is single-policy concentrability *also sufficient for rate-optimal offline learning with forward-KL regularization*?

We solve these problems for offline CBs concretely by the $\widetilde{\Theta}(\epsilon^{-1})$ sample complexity characterization[3], which is achieved under single-policy concentrability assumptions for the first time. At the core of our achievability results is a novel suboptimality decomposition for forward-KL-regularized objectives, overcoming the limitation of the traditional synergy of Taylor expansion and the mean value theorem (Zhao et al., 2025a, 2026, 2025b; Nayak et al., 2025; Ji et al., 2026) for analyzing the divergence-regularized sample complexity of bonus-based RL algorithms, which leads to a streamlined analysis of ***F**orward **KL**-Regularized **P**essimistic **C**ontextual **B**andits* (FKL-PCB) algorithms without any standalone dependency on $\min_{s,a} \pi^{\mathsf{ref}}(a|s)$. Our hardness results further demonstrate the fundamental rate limits in both the high- and low-regularization regimes.

Notation. The calligraphic $\mathcal{S}$ and $\mathcal{A}$ denote finite sets throughout this paper. $\Delta(\mathcal{A})$ is the set of probability distributions on $\mathcal{A}$ and $\Delta(\mathcal{A}|\mathcal{S})$ is the family of probability kernels from $\mathcal{S}$ to $\mathcal{A}$.

[2] See Lemma 5.1 for the concrete exposition.

[3] We omit the dependencies on quantities other than ϵ here for presentation clarity.

For $P, Q \in \Delta(\mathcal{A})$ with $P \ll Q$, the KL divergence from P to Q is denoted by $\mathsf{KL}(P\|Q) := \int \log(\mathrm{d}P/\mathrm{d}Q)\,\mathrm{d}P$, and $\mathrm{supp}(P)$ denotes the support set of P. For any subset E of a Euclidean space, $\mathrm{ri}(E)$ denotes its *relative interior*. For any extended real-valued convex function f, its *domain* is $\mathrm{dom}(f) := \{\pi : f(\pi) < +\infty\}$. Any closed convex function $\ell : \mathbb{R}^d \to \mathbb{R} \cup +\infty$ induces a Bregman divergence $\mathsf{B}_\ell(\mathbf{x}, \mathbf{y}) := \ell(\mathbf{x}) - \ell(\mathbf{y}) - \langle \nabla \ell(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$. Given $f, g : \mathcal{A} \to \mathbb{R} \cup \{+\infty\}$, $\langle f, g \rangle$ denotes the summation $\sum_{a \in \mathcal{A}} f(a)g(a)$. We use standard asymptotic notations including $O(\cdot)$, $\Omega(\cdot)$, and $\Theta(\cdot)$; for which $\widetilde{O}(\cdot), \widetilde{\Omega}(\cdot), \widetilde{\Theta}(\cdot)$ further hide $\mathrm{polylog}$ factors.

1.1 Key Related Works

We survey recent sharp sample complexity analyses of decision making with respect to regularized objectives and defer other related works on pessimism and forward KL regularization to Appendix A.

Fast Rates for Regularized Decision Making. The practical relevance of regularized objectives has motivated a broad line of sharper analyses of the statistical efficiency of decision-making under such formulations. The pioneering work of Tiapkin et al. (2023) first established an ϵ^{-1}-type fast sample complexity in the pure-exploration setting, followed by a series of works (Zhao et al., 2025b; Ji et al., 2026) that focused on regret minimization. A parallel line of research focuses on the offline or hybrid setting, starting from Zhao et al. (2025a) that obtained a $\widetilde{\Theta}(\epsilon^{-1})$ rate under concentrability of all policies. This requirement was relaxed in subsequent works (Foster et al., 2025; Kim et al., 2026; Zhao et al., 2026), in the context of adaptive reward-query, batched online learning, and pure offline learning, respectively. Similar fast convergence has also been shown in online (Nayak et al., 2025) and offline (Zhang et al., 2026) multi-agent RL, and self-play CBs with generalized bilinear function approximation (Lee et al., 2026). Despite these advances, the literature has largely focused on reverse KL or regularizers with certain strong-convexity-type properties, leaving a wide range of regularizers, such as forward KL, comparatively underexplored. Notably, Aminian et al. (2025) gave the first algorithm analysis for offline contextual dueling bandits with multiple reference models under forward-KL regularization; which, however, yields an $\widetilde{O}(\epsilon^{-2})$ slow rate.

2 Preliminaries

Consider offline CBs with context space $\mathcal{S}$, action space $\mathcal{A}$, mean reward function $r : \mathcal{A} \to [0, 1]$, context distribution $\rho \in \Delta(\mathcal{S})$, and reference policy $\pi^{\mathsf{ref}} \in \Delta(\mathcal{A}|\mathcal{S})$. The agent only learns from an i.i.d. dataset $\mathcal{D} = \{s_i, a_i, r_i\}_{i=1}^n$, where $s_i \sim \rho$, $a_i \sim \pi^{\mathsf{ref}}(\cdot|s)$ and , $r_i = r(s_i, a_i) + \varepsilon_i$ for each i. Here, $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. mean-zero 1-subgaussian noises, and π^{ref} *serves as the behavioral policy*. In particular, the class of interest satisfying all constraints above is denoted by $\mathrm{CB}(\mathcal{S}, \mathcal{A}, r, \rho, \pi^{\mathsf{ref}})$. Given the inverse regularization intensity $\eta > 0$, let the forward-KL-regularized objective be

$$J_{\mathrm{FKL}}(\pi) := J_{\mathrm{FKL}}(\pi; r) := \langle r, \pi \rangle - \eta^{-1} \mathsf{KL}\left(\pi^{\mathsf{ref}} \| \pi\right), \tag{2.1}$$

where we overload $\langle r, \pi \rangle := \mathbb{E}_{s \sim \rho} \langle r(s, \cdot), \pi(\cdot|s) \rangle$ and $\mathsf{KL}\left(\pi^{\mathsf{ref}} \| \pi\right) := \mathbb{E}_{s \sim \rho} \mathsf{KL}\left(\pi^{\mathsf{ref}}(\cdot|s) \| \pi(\cdot|s)\right)$. , The reference policy is assumed to have full support only to simplify the technical presentation.

Assumption 2.1. $\min_{s,a \in \mathcal{S} \times \mathcal{A}} \pi^{\mathsf{ref}}(a|s) > 0$.

If $\mathrm{supp}(\pi^{\mathsf{ref}}(\cdot|s) \subsetneq \mathcal{A}$, the dataset $\mathcal{D}$ will never contain actions outside the support, which makes learning even easier as we can restrict the action set for s to $\mathrm{supp}(\pi^{\mathsf{ref}}(\cdot|s)$. In other words, what really matters is whether the sample complexity will blow up if $\min_{s,a \in \mathrm{supp}(\pi^{\mathsf{ref}}(\cdot|s))} \pi^{\mathsf{ref}}(a|s)$ is super small. Assumption 2.1 has also been used in Huang et al. (2024, Lemma F.4) to derive a closed-form solution of a regularized optimal policy, and in the analyses of Wang et al. (2023); Aminian et al. (2025) implicitly in a similar way. It does *not* impose a quantitative gap between $\min_{s,a} \pi^{\mathsf{ref}}(a|s)$ and 0, which is also not needed in our analysis. The following closed-form of the forward-KL-regularized optimal policy follows from the classic KKT theory (See, e.g., Beck (2017, Appendix A) and Wang et al., 2023; Aminian et al., 2025). We provide its proof in Appendix E for completeness.

Lemma 2.2. $\forall s \in \mathcal{S}$, there exists a unique $\lambda_s > \max\{r(s, a) : a \in \mathrm{supp}(\pi^{\mathsf{ref}}(\cdot|s))\}$ such that

$$\pi^*_{\mathrm{FKL}}(\cdot|s) := \frac{\eta^{-1} \pi^{\mathsf{ref}}(\cdot|s)}{\lambda_s - r(s, \cdot)} \tag{2.2}$$

solves $\max_{\pi \in \Delta(\mathcal{A}|\mathcal{S})} J_{\mathrm{FKL}}(\pi)$. Moreover, under Assumption 2.1, π^*_{FKL} is the unique solution.

We then define the sub-optimality gap *regularized by forward KL* as

$$\mathrm{SubOpt}_{\mathrm{FKL}}(\cdot) := \mathrm{SubOpt}_{\mathrm{FKL}}(\cdot; \mathcal{S}, \mathcal{A}, r, \rho, \pi^{\mathsf{ref}}) = J_{\mathrm{FKL}}(\pi^*_{\mathrm{FKL}}) - J_{\mathrm{FKL}}(\cdot), \tag{2.3}$$

under which, the agent aims to find the *ϵ-optimal policy* $\widehat{\pi}$ with $\mathrm{SubOpt}_{\mathrm{FKL}}(\widehat{\pi}) \leq \epsilon$. For simplicity, we use π^* interchangeably with π^*_{FKL} to avoid notation clutter unless otherwise specified.

Concentrability. In offline learning and decision making, how the dataset $\mathcal{D}$ collected by π^{ref} covers a target distribution or a collection of potential target distributions largely shapes the learnability of the problem. This coverage is typically characterized by *the* concentrability in offline RL, which quantifies the ability of the behavior policy to generate diverse actions. We first introduce the density-ratio-based concentrability (Chen and Jiang, 2019; Xie et al., 2021a; Rashidinejad et al., 2021; Li et al., 2024) as follows, which characterizes how each pair of $(s, a) \in \mathcal{S} \times \mathcal{A}$ is covered by the behavior policy.

Definition 2.3 (*Density-ratio-based* concentrability)**.** Given some policy class Π and some reference policy π^{ref}, the density-ratio-based all-policy concentrability C^{Π} is defined by $C^{\Pi} := \sup_{\pi \in \Pi, s \in \mathcal{S}, a \in \mathcal{A}} \pi(a|s)/\pi^{\mathsf{ref}}(a|s)$, whose single-policy counterpart under the optimal policy π^* is $C^{\pi^*} := \sup_{s \in \mathcal{S}, a \in \mathcal{A}} \pi^*(a|s)/\pi^{\mathsf{ref}}(a|s)$.

When the optimal policy π^* is contained within the policy class Π, the relationship $C^{\pi^*} \leq C^{\Pi}$ necessarily holds. Consequently, an analytical dependency on the single-policy concentrability coefficient C^{π^*} is strictly weaker, and thus more general, than a dependency on the all-policy coefficient C^{Π}. In this work, we demonstrate that this more relaxed single-policy concentrability is sufficient for learning forward-KL-regularized bandits. This result stands in contrast to the more restrictive requirement of all-policy concentrability typically invoked in previous literature (Zhao et al., 2025a; Wu et al., 2025; Aminian et al., 2025).

2.1 Function Approximation

While the tabular setting serves as the minimal decision making formalization without additional structures, algorithms designed directly for this setting become computationally intensive or even intractable in most real-world applications due to the curse of dimensionality. To address this issue, it is generally assumed that the learner has access to some class of functions that incorporate the ground truth reward (Jin et al., 2021; Xiong et al., 2022; Di et al., 2023). In this paper, we also consider the scenarios that the reward lies in some known function class $\mathcal{G}$, as formalized by the following realizable assumption.

Assumption 2.4. For this known function class $\mathcal{G} \subseteq (\mathcal{S} \times \mathcal{A} \to [0, 1])$, $\exists g^* \in \mathcal{G}$ with $g^* = r$.

To characterize the complexity measure of the reward function class, we adopt the standard notion of covering number (Wainwright, 2019, Definition 5.1).

Definition 2.5 (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 $\|g - g'\|_\infty \leq \epsilon$. The ϵ-covering number is the smallest cardinality $N_{\mathcal{G}}(\epsilon)$ of such $\mathcal{G}(\epsilon)$.

Assumption 2.6. For any $\epsilon_c > 0$, the ϵ_c-covering number $N_{\mathcal{G}}(\epsilon_c)$ of $\mathcal{G}$ is $\mathrm{poly}(\epsilon_c^{-1})$.

Assumption 2.6 allowing $\log N_{\mathcal{G}}(\epsilon)$ to be roughly negligible is arguably mild. For example, when $\mathcal{G}$ is the class of linear functions of dimension d and radius R, the covering number is $N_{\mathcal{G}}(\epsilon) = O((1 + R\epsilon^{-1})^d)$ (Jin et al., 2020, Lemma D.6), which satisfies Assumption 2.6.

In the presence of function approximation, the density-ratio-based concentrability in Definition 2.3 does not precisely capture the quality of the offline dataset due to the absence of the function approximation structure in its definition. For example, a particular context-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ might be frequently visited by π^* but seldom by π^{ref}, leading to very large C^{π^*}. However, if there is some $(s', a') \in \mathcal{S} \times \mathcal{A}$ frequently appearing in the dataset $\mathcal{D}$ such that $g(s, a) \approx g(s', a')$ for all $g \in \mathcal{G}$, then the reward of (s, a) is still well estimated, since feature representation associated with (s, a) is frequently visited by visiting (s', a'). To capture this dependency, we introduce the following D^2-type concentrability (Gentile et al., 2022; Agarwal et al., 2023; Zhao et al., 2025a) in function approximation, which characterize the coverage of the behavior policy on the feature space.

Algorithm 1 FKL-PCB for Tabular Setting

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

Set $N(s,a) = \sum_{i=1}^{n} \mathbb{1}\{(s_i, a_i) = (s,a)\}$ for all $a \in \mathcal{A}$, $s \in \mathcal{S}$
for $s \in \mathcal{S}$, $a \in \mathcal{A}$ **do**
  if $N(s,a) = 0$ **then**
    Set the empirical reward $\bar{g}(s,a) \leftarrow 0$, penalty $b(s,a) \leftarrow 1$
  else
    Compute the empirical reward $\bar{g}(s,a) \leftarrow \sum_{i=1}^{n} r_i \mathbb{1}\{(s_i, a_i) = (s,a)\}/N(s,a)$
    Compute the penalty to be

$$b(s,a) \leftarrow \sqrt{4\log(2|\mathcal{S}||\mathcal{A}|/\delta)/N(s,a)}, \tag{3.1}$$

    Set $\widehat{g}(s,a) \leftarrow \bar{g}(s,a) - b(s,a)$
  end if
end for

Ensure: $\widehat{\pi} = \operatorname{argmax}_{\pi} \langle \widehat{g}, \pi \rangle - \eta^{-1} \mathsf{KL}\left(\pi^{\mathsf{ref}} \| \pi\right)$

Definition 2.7. Given function class $\mathcal{G} \subset (\mathcal{S} \times \mathcal{A} \to \mathbb{R})$ and policy π, their D^2-divergence is

$$D^2_{\mathcal{G}}((s,a);\pi) := \sup_{g,h \in \mathcal{G}} \frac{\left(g(s,a) - h(s,a)\right)^2}{\mathbb{E}_{(s',a') \sim \rho \times \pi}[(g(s',a') - h(s',a'))^2]}, \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

We are now ready to define the D^2-type single policy concentrability.

Assumption 2.8 (Single-policy concentrability)**.** $D^2_{\pi^*} := \mathbb{E}_{(s,a) \sim \rho \times \pi^*} D^2_{\mathcal{G}}((s,a);\pi^{\mathsf{ref}}) < \infty$.

Assumption 2.8 indicates that the errors on the context-action distributions $\rho \times \pi^*$ can be bounded by the error on the samples from $\rho \times \pi^{\mathsf{ref}}$ up to some constant. Similar to the density-ratio-based concentrabilities, here it is also true that the single-policy concentrability assumption is strictly weaker than the all-policy concentrability assumption (e.g., Zhao et al. 2026, Assumption 2.7). Remarkably, the two quantities characterizing single-policy concentrability C^{π^*} and $D^2_{\pi^*}$ cannot be bounded by each other up to constant factors in general. We provide a detailed discussion of the relation between C^{π^*} and $D^2_{\pi^*}$ in Appendix B.

3 Algorithms and Sample Complexity Upper Bounds

In this section, we present the algorithm, FKL-PCB, for learning forward-KL-regularized CBs. For ease of presentation, we present its adaptations for both tabular setting and reward function approximation respectively, in Section 3.1 and Section 3.2.

3.1 Algorithm for Tabular Setting

We first tackle the most vanilla tabular setting, where the reward function can be any map in $\mathcal{S} \times \mathcal{A} \to [0,1]$, as summarized in Algorithm 1. At a high level, the algorithm follows previous pessimism-based algorithms (Rashidinejad et al., 2021) for regularized objectives (Zhao et al., 2026). In particular, Algorithm 1 first employs a least-squares estimator to estimate the reward function. Then following Zhao et al. (2026), we subtract a penalty term to prevent over-estimation, leading to a pessimistic reward estimate. In the tabular case, we adopt the standard tabular confidence bound (Xie et al., 2021b), leading to the penalty given by (3.1). The following results show that $\widehat{g}$ is indeed a pessimistic estimate with high probability and the number of samples $N(s,a)$ does not deviate too much from its expectation $\rho(s)\pi^{\mathsf{ref}}(a|s)$.

Lemma 3.1. Given $\delta > 0$, let $\mathcal{E}_{\mathrm{tab}}(\delta)$ denote the event that the reward estimation error is uniformly controlled,

$$\mathcal{E}_{\mathrm{tab}}(\delta) := \left\{ \left|\bar{g}(s,a) - g^*(s,a)\right| \le \Gamma_n(s,a), \text{for all } (s,a) \in \mathcal{S} \times \mathcal{A} \right\}. \tag{3.2}$$

We further use $\mathcal{E}_{\mathrm{bin}}(\delta)$ to denote the event under which $N(s,a)$ does not deviate too much from the expectation, that is

$$\mathcal{E}_{\mathrm{bin}}(\delta) := \left\{ \frac{1}{N(s,a) \vee 1} \le \frac{8\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{n\rho(s)\pi^{\mathsf{ref}}(a|s)} \text{ for all } (s,a) \in \mathcal{S} \times \mathcal{A} \right\}.$$

Algorithm 2 FKL-PCB for Function Approximation

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

1: Reward estimation via least squares: $\bar{g} \in \operatorname{argmin}_{g\in\mathcal{G}} \sum_{(s_i,a_i,r_i)\in\mathcal{D}} \big(g(s_i,a_i) - r_i\big)^2$

2: Compute the pessimism penalty with $\beta = \widetilde{O}\Big(\sqrt{\log\big(2N_{\mathcal{G}}(\epsilon)/\delta\big)/n + \epsilon}\Big)$

$$\Gamma_n(s,a) = \beta D_{\mathcal{G}}\big((s,a),\pi^{\mathsf{ref}}\big), \forall (s,a) \in \mathcal{S}\times\mathcal{A} \tag{3.3}$$

3: Compute the pessimistic reward function $\widehat{g} \leftarrow \bar{g} - \Gamma_n$

Ensure: $\widehat{\pi} = \operatorname{argmax}_\pi \langle \widehat{g}, \pi\rangle - \eta^{-1}\mathsf{KL}\left(\pi^{\mathsf{ref}}\|\pi\right)$

Then the event $\mathcal{E}_{\mathrm{tab}}(\delta) \cap \mathcal{E}_{\mathrm{bin}}(\delta)$ holds with probability at least $1-\delta$.

Finally, we construct the output policy to be the maximizer of the pessimistic reward $\widehat{g}$. The following theorem provides the sample complexity guarantee of Algorithm 1.

Theorem 3.2. For sufficiently small $\epsilon \in (0,1)$, with probability at least $1-\delta$, $\widetilde{O}\big(\eta(C^{\pi^*})^2|\mathcal{S}||\mathcal{A}|\epsilon^{-1}\big)$ samples suffice to guarantee the learned policy $\widehat{\pi}$ of Algorithm 1 to be ϵ-optimal.

3.2 Algorithm for Function Approximation

While the algorithm follows the same backbone as Algorithm 1, access to the function class $\mathcal{G}$ allows us to perform least-squares estimation within $\mathcal{G}$. After the estimation, we then construct the pessimism penalty as in (3.3), following the approaches in previous works (Di et al., 2023; Zhao et al., 2026). Previous results indicate that this penalty selection is sufficient for $\widehat{g} = \bar{g} - \Gamma_n \le g^*$ with a high probability. Formally, we define the successful event $\mathcal{E}_{\mathrm{gfa}}(\delta)$ given $\delta > 0$ as

$$\mathcal{E}_{\mathrm{gfa}}(\delta) := \Big\{ \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \Big[\big|\bar{g} - g^*\big| - \Gamma_n\Big](s,a) \le 0\Big\}. \tag{3.4}$$

The following lemma indicates that $\mathcal{E}_{\mathrm{gfa}}(\delta)$ holds with probability at least $1-\delta$.

Lemma 3.3 (Lemma 2.9, Zhao et al. 2026)**.** $\forall \delta \in (0,1)$, $\mathcal{E}_{\mathrm{gfa}}(\delta)$ holds with probability at least $1-\delta$.

Finally, we construct the output policy as the maximizer of the pessimistic reward $\widehat{g}$ under the regularized objective. The following theorem provides the sample complexity guarantee of Algorithm 2.

Theorem 3.4. For sufficiently small $\epsilon \in (0,1)$, $\widetilde{O}\big(\eta C^{\pi^*} D^2_{\pi^*}\epsilon^{-1}\log N_{\mathcal{G}}(\epsilon)\big)$ samples suffice to guarantee the learned policy $\widehat{\pi}$ of Algorithm 2 to be ϵ-optimal with probability at least $1-\delta$.

Remark 3.5. Previously, Aminian et al. (2025) obtained a $\widetilde{O}\big(\log N_{\mathcal{G}} C_{\varepsilon_{\mathrm{FKL}}}/\epsilon^2\big)$ sample complexity guarantee for offline learning with forward-KL-regularized objective. In contrast, both Theorem 3.2 and Theorem 3.4 establish an ϵ^{-1}-fast rate, improving upon the previous ϵ^{-2} dependency. Moreover, the sample complexity in both Theorem 3.2 and Theorem 3.4 only relies on the notion of single policy concentrability, thereby relaxing the stronger all-policy concentrability assumption required in Aminian et al. (2025).[4] It is also worth noting that this combination of an ϵ^{-1} fast rate and relaxed concentrability requirements has recently been shown to be both necessary and sufficient for reverse-KL-regularized CBs (Zhao et al., 2026).

4 Minimax Lower Bounds

In this section, we present worst-case hardness results in the tabular setting, whose proofs are deferred to Appendix D.

Theorem 4.1. Let $\widehat{\pi} \in \Delta(\mathcal{A})$ be any estimator (with full support) from n context-action-reward pairs, then for any even $A \ge 2$, $S \ge 1$, $\eta > 0$, and $n > 16SA \vee 4\eta^2 SA$,

$$\sup_{\mathrm{CB}(\mathcal{S},\mathcal{A},r,\rho,\pi^{\mathsf{ref}}):|\mathcal{S}|=S,|\mathcal{A}|=A} \mathbb{E}_{\mathcal{D}\sim P_{\pi^{\mathsf{ref}},r}} \mathrm{SubOpt}_{\mathrm{FKL}}(\widehat{\pi};\mathcal{A},r,\pi^{\mathsf{ref}}) \gtrsim \frac{\eta SA}{n}.$$

[4] Technically, the $C_{\varepsilon_{\mathrm{FKL}}}$ in Aminian et al. (2025) requires the behavioral policy to cover every policy with a local forward-KL-ball of the reference policy well.

Theorem 4.1 indicates a minimax lower bound of $\Omega\big(SA\cdot(\eta\epsilon^{-1}\vee 4\eta^2\vee 16)\big)$ for learning an ϵ-optimal policy in the tabular setting, corroborating the near-optimality of our *rate* $\widetilde{O}(\epsilon^{-1})$ in Theorem 3.2. Moreover, Theorem 4.1 exhibits matching linear dependencies on both S and A, demonstrating that the corresponding dependencies in Theorem 3.2 are unavoidable. Currently, our hard instances are tabular CBs with $\pi^{\mathsf{ref}} = \mathsf{Unif}(\mathcal{A})$, leaving lower bounds that capture coverage dependence and settings with function approximation as important directions for future work.

The regime with large η. Note that both the fast rate upper and lower bounds are linear in η, which is a behavior similar to their counterparts for *reverse* KL (Zhao et al., 2025a, 2026). Note that the lower bound in Theorem 4.1 is only applicable for $\eta^2 \lesssim n/(SA)$, which is a *high-regularization* regime. If η is too large, the effect of regularization should intuitively vanish in this *low-regularization* regime. It is indeed the case for *reverse* KL in both offline (Zhao et al., 2026) and online (Ji et al., 2026) CBs, where the fundamental limit becomes an η-free $\Omega(\epsilon^{-2})$ slow rate if η exceeds a specific threshold. We show in Theorem D.3 that this type of "phase transition" will also occur for J_{FKL} in offline bandits, while the exactly matching phase transition threshold for η in both the upper and lower bounds is left as future work.

5 Technical Overview

We use $|\mathcal{S}| = 1$ (offline bandits) as the working example to outline the limitations (or inapplicability) of previous works and highlight the key technique for our algorithm analysis.

5.1 Limitations of Traditional Routines

In the first statistical analysis *specialized for* $\mathrm{SubOpt}_{\mathrm{FKL}}(\cdot)$, Aminian et al. (2025, Theorem 6.3) treat all terms involving forward KL with a gross bound depending on an all-policy-type concentrability coefficient $C_{\varepsilon_{\mathrm{FKL}}}$ to reduce the suboptimality to its *unregularized* counterpart, which unavoidably exhibits at least the fundamental limit of estimation, i.e., $\widetilde{O}(\epsilon^{-2})$. Another conjugate-based routine (Zhao et al., 2026) gives $\widetilde{\Theta}(\epsilon^{-1})$ sample complexity offline CBs under general f-divergence-regularized objectives with strongly convex f, which excludes $\mathsf{KL}\left(\pi^{\mathsf{ref}}\|\pi\right)$. Recently, Lee et al. (2026) consider regularized objectives in an online setting for strongly convex regularizers in a unified manner, which is, however, not ideal for forward KL in the offline case because: although $\pi \mapsto \mathsf{KL}\left(\pi\|\pi^{\mathsf{ref}}\right)$ is $\Theta(1)$-strongly-convex with respect to the L^1 distance (Polyanskiy and Wu, 2025; Zhao et al., 2026), the modulus of strong convexity for $\pi \mapsto \mathsf{KL}\left(\pi^{\mathsf{ref}}\|\pi\right)$ may depend on the condition of π^{ref} as manifested in the following lemma, whose proof is deferred to Appendix E.

Lemma 5.1. Under Assumption 2.1, if $\min_{a\in\mathcal{A}} \pi^{\mathsf{ref}}(a) = \alpha$, $\pi \mapsto \mathsf{KL}\left(\pi^{\mathsf{ref}}\|\pi\right)$ is $\Theta(\alpha)$-strongly-convex with respect to the L^1 distance as a map from $\mathrm{ri}\left(\Delta(\mathcal{A})\right)$ to $\mathbb{R}_+$.

Consequently, techniques following Lee et al. (2026) for analyzing offline algorithms may yield sample complexity bounds *proportional to* α^{-1}, which translates to the undesirable (density-ratio-based) *all*-policy concentrabilty even in multi-armed bandits.

5.1.1 The Limitation of Techniques Tailored to *Reverse* KL

Existing results under relatively weak concentrability conditions for *reverse*-KL-regularized objectives either heavily relies on specific log-linear parametric forms (Foster et al., 2025) or hinges on a curious pessimism-induced monotonic property in a Taylor-type argument based on the mean value theorem (Zhao et al., 2026). In particular, we expose here the limitation in this latter mid-point-based routine, rendering it not sharp enough for forward KL. A faithful adaptation of the spirit in Zhao et al. (2026, Section 2.4) to the nature of $\mathsf{KL}\left(\pi^{\mathsf{ref}}\|\pi\right)$ and π^*_{FKL} yields (5.1) and (5.2) in Lemma 5.2, whose proof is detailed in Appendix E.

Lemma 5.2. If $\widehat{\pi} \in \mathrm{argmax}_{\pi\in\Delta(\mathcal{A})} J_{\mathrm{FKL}}(\pi;\widehat{r})$, where $\widehat{r}:\mathcal{A}\to[0,1]$; then $\exists \bar{u}\in[0,1]$ such that under Assumption 2.1, $\mathrm{SubOpt}_{\mathrm{FKL}}(\widehat{\pi}) =$

$$\frac{\eta}{2}\left(\mathbb{E}_{a\sim\pi_{\bar{u}}}\left[\left(\frac{\pi_{\bar{u}}}{\pi^{\mathsf{ref}}}(r-\widehat{r})^2\right)(a)\right] - \left(\sum_a \left(\frac{\pi_{\bar{u}}^2}{\pi^{\mathsf{ref}}}(r-\widehat{r})\right)(a)\right)^2 \Big/ \sum_a \frac{\pi_{\bar{u}}^2}{\pi^{\mathsf{ref}}}(a)\right), \tag{5.1}$$

where $\pi_{\bar{u}} \in \mathrm{argmax}_{\pi\in\Delta(\mathcal{A})} J_{\mathrm{FKL}}(\pi;r_{\bar{u}})$ with $r_{\bar{u}} := (1-\bar{u})r + \bar{u}\widehat{r}$. (5.1) directly implies

$$\mathrm{SubOpt}_{\mathrm{FKL}}(\widehat{\pi}) \le \frac{\eta}{2}\mathbb{E}_{a\sim\pi_{\bar{u}}}\left[\left(\frac{\pi_{\bar{u}}}{\pi^{\mathsf{ref}}}(r-\widehat{r})^2\right)(a)\right]. \tag{5.2}$$

The pillar behind the proof of Zhao et al. (2026, Theorem 2.10) is that the counterpart of (5.1) or (5.2) for *reverse* KL is *non-increasing* in $\bar{u}$ on $[0,1]$ under *the premise of pessimism* (i.e., $r \geq \widehat{r}$), so that the maximizer of the corresponding RHS is $\bar{u}=0$, which helps replace the troublemaker $\pi_{\bar{u}}$ with π^*; and thus lets the RHS depend on the single-policy concentrability that only covers π^*. Unfortunately, the calculations (deferred to Appendix E) in Lemma 5.3 deny such a belief for forward KL.

Lemma 5.3. Let $F(\bar{u}) \coloneqq$ RHS of (5.1), $w_u(a) \coloneqq \frac{\pi_u(a)}{\pi^{\mathsf{ref}}(a)}$, $Z_u \coloneqq \sum_a \frac{\pi_u^2(a)}{\pi^{\mathsf{ref}}(a)}$, and the probability distribution $\mu_u(a) \coloneqq \frac{\pi_u^2(a)}{\pi^{\mathsf{ref}}(a)}/Z_u$, and $g \coloneqq r - \widehat{r}$, then $\forall u \in (0,1)$,

$$F'(u) = -\eta^2 Z_u \cdot \mathbb{E}_{a\sim\mu_u}\Big[w_u(a)\big(g(a) - \mathbb{E}_{a\sim\mu_u} g(a)\big)^3\Big]. \tag{5.3}$$

The sign of (5.3) may be hard to determine in general for $u \in (0,1)$ even if $g \geq 0$ uniformly because it is a tilted skewness of a *centered* random variable, which is strong evidence that the mean-value-type approach might not work for analyzing bandit learning under forward KL regularization.

Remark 5.4. Sharp readers may wonder whether it is possible to disregard the second term in (5.1) and take derivative of just (5.2). However, since (5.2) is a part of (5.1), the calculations in Lemma 5.3 already imply that $\widetilde{F}'(u) = -\eta^2 Z_u \mathbb{E}_{a\sim\mu_u}\big[w_u(a) g(a)^2 \big(g(a) - \mathbb{E}_{a\sim\mu_u}[g(a)]\big)\big]$ for RHS of (5.2)$=: \widetilde{F}(\bar{u})$. Since $\widetilde{F}'(u)$ turns out to be at best a *plausible* tilted covariance between g^2 and a *centered* $g - \mathbb{E}_{\mu_u} g$, its sign may still be indefinite even conditioned on $g \geq 0$.

5.2 Our Key Technique

Note that (5.1) is already an *identity*, which, together with the limitations we outline, suggests that it may not be suitable to tweak this mean-value argument based on Taylor expansion further. Instead, we prepare another identity (5.7) based on a *general* characterization (Lemma F.4) of *regularized objectives* via the dual perspective of convex conjugate (over the simplex), which is *fully detached from any specific design* of bonus terms; and craft a neat self-bounding argument (5.9) on top of it to utilize the pessimism principle *directly* in (5.10) without resorting to the aforementioned monotinicity mechanism. This suite of key techniques is elaborated in the proof of Lemma 5.5 as follows.

Lemma 5.5. Following the notation in (2.1), if $\widehat{\pi} \in \operatorname{argmax}_{\pi\in\Delta(\mathcal{A})} J_{\mathrm{FKL}}(\pi;\widehat{r})$, where $\widehat{r}: \mathcal{A} \to [0,1]$, and $r \geq \widehat{r}$ uniformly; then under Assumption 2.1,

$$\mathrm{SubOpt}_{\mathrm{FKL}}(\widehat{\pi}) = J_{\mathrm{FKL}}(\pi^*_{\mathrm{FKL}}; r) - J_{\mathrm{FKL}}(\widehat{\pi}; r) \leq 2\eta \sum_{a\in\mathcal{A}} \Big(\frac{{\pi^*_{\mathrm{FKL}}}^2}{\pi^{\mathsf{ref}}}(r-\widehat{r})^2\Big)(a). \tag{5.4}$$

Proof. Let $h(\cdot) \coloneqq \eta^{-1}\mathsf{KL}\left(\pi^{\mathsf{ref}}\|\cdot\right)$. Then h is strictly convex by Lemma 5.1, obviously continuously differentiable on $\mathrm{ri}(\Delta(\mathcal{A}))$, $\mathrm{dom}(h) \coloneqq \{\pi \in \Delta(\mathcal{A}) : h(\pi) < \infty\} = \mathrm{ri}(\Delta(\mathcal{A}))$; and the uniqueness of π_u as well as $\pi_u \in \mathrm{ri}(\Delta(\mathcal{A}))$ is guaranteed by Lemma 2.2 and Assumption 2.1. Thus, Lemmas F.3 and F.4 are applicable; under the notations of which: $\pi^*_{\mathrm{FKL}} = \pi_r$ and $\widehat{\pi} = \pi_{\widehat{r}}$; $J_{\mathrm{FKL}}(\pi^*_{\mathrm{FKL}}; r) = h^*(r)$ and $J_{\mathrm{FKL}}(\widehat{\pi};\widehat{r}) = h^*(\widehat{r})$ are their constrained convex conjugate, respectively.[5] Therefore, (5.5), which is inspired by Zhao et al. (2026, Appendix E.1), holds by definition; and $\nabla h^*(\widehat{r}) = \widehat{\pi}$ follows from Lemma F.3. In detail,

$$\begin{aligned}
\mathrm{SubOpt}_{\mathrm{FKL}}(\widehat{\pi}) &= h^*(r) - h^*(\widehat{r}) - \langle r, \widehat{\pi}\rangle + \langle \widehat{r}, \widehat{\pi}\rangle && (5.5)\\
&= \mathsf{B}_{h^*}(r, \widehat{r}) = \mathsf{B}_h(\widehat{\pi}, \pi^*) && (5.6)\\
&= \eta^{-1}\left\langle \pi^{\mathsf{ref}}, \log\frac{\pi^{\mathsf{ref}}}{\widehat{\pi}}\right\rangle - \eta^{-1}\left\langle \pi^{\mathsf{ref}}, \log\frac{\pi^{\mathsf{ref}}}{\pi^*}\right\rangle - \langle \nabla h(\pi^*), \widehat{\pi} - \pi^*\rangle\\
&= \eta^{-1}\sum_{a\in\mathcal{A}} \pi^{\mathsf{ref}}(a)\Big[\frac{\widehat{\pi}(a)}{\pi^*(a)} - 1 - \log\Big(1 - \big(1 - \frac{\widehat{\pi}(a)}{\pi^*(a)}\big)\Big)\Big]\\
&= \eta^{-1}\sum_{a\in\mathcal{A}} \pi^{\mathsf{ref}}(a)\big[-y_a - \log(1-y_a)\big] = \eta^{-1}\sum_{a\in\mathcal{A}} \pi^{\mathsf{ref}}(a)\phi(y_a), && (5.7)
\end{aligned}$$

where the first equality in (5.6) is a combination of the definition of Bregman divergence and $\nabla h^*(\widehat{r}) = \widehat{\pi}$; and the second equality in (5.6) is by applying Lemma F.4 to h; $y_a \coloneqq 1 -$

[5] Appendix F.2 review concepts related to convex conjugate.

$\widehat{\pi}(a)/\pi^*(a), \forall a \in \mathcal{A}$; and $\phi(y) := -y - \log(1-y), \forall y < 1$. On the other hand, the definition of $\widehat{\pi}$ alone already implies

$$\begin{aligned}\mathrm{SubOpt}_{\mathrm{FKL}}(\widehat{\pi}) = J(\pi^*; r) - J(\widehat{\pi}; r) &\leq J(\pi^*; r) - J(\pi^*; \widehat{r}) + J(\widehat{\pi}; \widehat{r}) - J(\widehat{\pi}; r) \\ &= \langle r - \widehat{r}, \pi^* \rangle + \langle \widehat{r} - r, \widehat{\pi} \rangle \\ &= \langle r - \widehat{r}, \pi^* - \widehat{\pi} \rangle = \sum_{a \in \mathcal{A}} \pi^*(a) g(a) y_a, \end{aligned} \tag{5.8}$$

where $g := r - \widehat{r}$. We thus define $z_a := 2\eta\pi^*(a)g(a)/\pi^{\mathsf{ref}}(a)$ to deduce from (5.8) that

$$\begin{aligned}2 \cdot \mathrm{SubOpt}_{\mathrm{FKL}}(\widehat{\pi}) &\leq \eta^{-1} \sum_{a \in \mathcal{A}} \pi^{\mathsf{ref}}(a) z_a y_a \\ &\leq \eta^{-1} \sum_{a \in \mathcal{A}} \pi^{\mathsf{ref}}(a) \big[\phi(y_a) + \phi^*(z_a)\big] \\ &= \mathrm{SubOpt}_{\mathrm{FKL}}(\widehat{\pi}) + \eta^{-1} \sum_{a \in \mathcal{A}} \pi^{\mathsf{ref}}(a) \big(z_a - \log(1 + z_a)\big), \end{aligned} \tag{5.9}$$

where the second inequality is due to the Fenchel-Young inequality, the last equality is due to Fact F.10; and the pessimism $r \geq \widehat{r}$ ensures $z_a \geq 0 > -1$, which in turn ensures the well-posedness of the constrained convex conjugate $\phi^*(z_a)$. Rearranging (5.9) yields

$$\begin{aligned}\mathrm{SubOpt}_{\mathrm{FKL}}(\widehat{\pi}) &\leq \eta^{-1} \sum_{a \in \mathcal{A}} \pi^{\mathsf{ref}}(a) \big(z_a - \log(1 + z_a)\big) \\ &\leq \eta^{-1} \sum_{a \in \mathcal{A}} \pi^{\mathsf{ref}}(a) \frac{z_a^2}{2} \\ &= 2\eta \sum_{a \in \mathcal{A}} \pi^{\mathsf{ref}}(a) \left(\frac{\pi^*(a)}{\pi^{\mathsf{ref}}(a)} g(a) \right)^2, \end{aligned} \tag{5.10}$$

where the second inequality (5.10) is due to $z_a \geq 0$ (i.e., pessimism) together with the fact that $\forall z \geq 0, z - \log(1+z) \leq z^2/2$; and the equality follows from the definition of z_a. □

Again, since the $\mathrm{SubOpt}_{\mathrm{FKL}}(\cdot)$ on general $\mathcal{S}$ with $|\mathcal{S}| \geq 1$ is the expectation of the bandit suboptimality over $s \sim \rho$, the sample complexity in Theorems 3.2 and 3.4 follow from the conjunction of a context-wise application of Lemma 5.5 and the corresponding concentration arguments conditioned on the pessimism events, respectively; as detailed in Appendix C. To see why Lemma 5.5 helps bypass the need of all-policy concentrability, we observe that the RHS of (5.4) is bounded from above in the tabular setting by $2\eta(C^{\pi^*})^2 \mathbb{E}_{\pi^{\mathsf{ref}}}(r - \widehat{r})^2$, where the squared loss admits a $\widetilde{O}(n^{-1})$ concentration with high probability; the reasoning is similar under general function approximation.

6 Conclusion, Limitations, and Future Work

We initiate the program towards exactly characterizing forward-KL regularization as a performance metric of data efficiency in RL. For offline CBs, we are the first to certify the sufficiency of single-policy concentrability for achieving the $\widetilde{O}(\epsilon^{-1})$ forward-KL-regularized sample complexity via the analysis of a minimalist pessimism-based algorithm in both the tabular and function approximation settings. The rate-optimality of our algorithm analysis is established by our new lower bounds, which also provide evidence that the sample complexity under forward KL regularization may also exhibit a phase transition from fast to slow rate as the regularization intensity decreases.

From a technical aspect, the key synergy between the conjugate-based observation and the standard linear suboptimality decomposition in our proof of the forward-KL-regularized performance difference lemma (Lemma 5.5) is neither specific to f-divergence (e.g., forward KL) nor restricted to offline CBs, and hence we believe its ideas can be extended to analyze more general regularized decision making problems, including MDPs, online learning, and learning against generic regularizers (Zhao et al., 2025b; Lee et al., 2026). Regarding message-level **limitations**, the separation between forward versus reverse KL in terms of the minimal dependencies on the coverage conditions required for rate-optimal offline learning with respect to the corresponding regularized objectives is left as future work; in particular, the *necessity* of single-policy concentrability for forward KL is still open even in offline MABs. It is even more interesting to derive a fast rate lower bound under general function approximation that match our upper bound in terms of the D^2-type notion of data coverage.

Acknowledgments and Disclosure of Funding

We thank the anonymous reviewers for their helpful comments. QZ, KJ, HZ and QG are supported in part by the National Science Foundation DMS-2323113 and IIS-2403400. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

AGARWAL, A., JIN, Y. and ZHANG, T. (2023). Vo q l: Towards optimal regret in model-free rl with nonlinear function approximation. In *The Thirty Sixth Annual Conference on Learning Theory*. PMLR.

AMINIAN, G., ASADI, A. R., SHENFELD, I. and MROUEH, Y. (2025). Kl-regularized rlhf with multiple reference models: Exact solutions and sample complexity. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

BECK, A. (2017). *First-order methods in optimization*. SIAM.

BERTSEKAS, D., NEDIC, A. and OZDAGLAR, A. (2003). *Convex analysis and optimization*, vol. 1. Athena Scientific.

BISHOP, C. M. and NASRABADI, N. M. (2006). *Pattern recognition and machine learning*, vol. 4. Springer.

CHEN, F., FOSTER, D. J., HAN, Y., QIAN, J., RAKHLIN, A. and XU, Y. (2024). 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.

CHEN, J. and JIANG, N. (2019). Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*. PMLR.

DI, Q., ZHAO, H., HE, J. and GU, Q. (2023). Pessimistic nonlinear least-squares value iteration for offline reinforcement learning. *arXiv preprint arXiv:2310.01380* .

FOSTER, D. J., MHAMMEDI, Z. and ROHATGI, D. (2025). Is a good foundation necessary for efficient reinforcement learning? the computational role of the base model in exploration. *arXiv preprint arXiv:2503.07453* .

FOSTER, D. J. and RAKHLIN, A. (2023). Foundations of reinforcement learning and interactive decision making. *arXiv preprint arXiv:2312.16730* .

GENTILE, C., WANG, Z. and ZHANG, T. (2022). Achieving minimax rates in pool-based batch active learning. In *International Conference on Machine Learning*. PMLR.

GUO, D., YANG, D., ZHANG, H., SONG, J., WANG, P., ZHU, Q., XU, R., ZHANG, R., MA, S., BI, X. ET AL. (2025). Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948* .

GX-CHEN, A., PRAKASH, J., GUO, J., FERGUS, R. and RANGANATH, R. (2025). Kl-regularized reinforcement learning is designed to mode collapse. *arXiv preprint arXiv:2510.20817* .

HAARNOJA, T., ZHOU, A., ABBEEL, P. and LEVINE, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*. PMLR.

HUANG, A., ZHAN, W., XIE, T., LEE, J. D., SUN, W., KRISHNAMURTHY, A. and FOSTER, D. J. (2024). Correcting the mythos of kl-regularization: Direct alignment without overoptimization via chi-squared preference optimization. *arXiv preprint arXiv:2407.13399v3* .

JI, H., KE, P., WANG, H. and HUANG, M. (2023). Language model decoding as direct metrics optimization. *arXiv preprint arXiv:2310.01041* .

JI, K., ZHAO, Q., ZHAO, H., DI, Q. and GU, Q. (2026). Near-optimal regret for kl-regularized multi-armed bandits. *arXiv preprint arXiv:2603.02155* .

JIN, C., YANG, Z., WANG, Z. and JORDAN, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on learning theory*. PMLR.

JIN, Y., YANG, Z. and WANG, Z. (2021). Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*. PMLR.

KIM, J., YUN, J., LEE, J. D. and JUN, K.-S. (2026). Coverage improvement and fast convergence of on-policy preference learning. *arXiv preprint arXiv:2601.08421* .

LATTIMORE, T. and SZEPESVÁRI, C. (2020). *Bandit algorithms*. Cambridge University Press.

LE CAM, L. (1973). Convergence of estimates under dimensionality restrictions. *The Annals of Statistics* 38–53.

LEE, J., HONG, M., JUN, K.-S., YUN, C. and YUN, S.-Y. (2026). Regularized online rlhf with generalized bilinear preferences. *arXiv preprint arXiv:2602.23116* .

LEVINE, S. and KOLTUN, V. (2013). Guided policy search. In *International conference on machine learning*. PMLR.

LI, G., SHI, L., CHEN, Y., CHI, Y. and WEI, Y. (2024). Settling the sample complexity of model-based offline reinforcement learning. *The Annals of Statistics* **52** 233–260.

MCALLESTER, D. A. (1999). Pac-bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*.

MILGROM, P. and SEGAL, I. (2002). Envelope theorems for arbitrary choice sets. *Econometrica* **70** 583–601.

MIN, Y., WANG, T., ZHOU, D. and GU, Q. (2021). Variance-aware off-policy evaluation with linear function approximation. *Advances in neural information processing systems* **34** 7598–7610.

NAYAK, A., YANG, T., YAGAN, O., JOSHI, G. and CHI, Y. (2025). Achieving logarithmic regret in kl-regularized zero-sum markov games. *arXiv preprint arXiv:2510.13060* .

OUYANG, L., WU, J., JIANG, X., ALMEIDA, D., WAINWRIGHT, C., MISHKIN, P., ZHANG, C., AGARWAL, S., SLAMA, K., RAY, A. ET AL. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems* **35** 27730–27744.

POLYANSKIY, Y. and WU, Y. (2025). *Information theory: From coding to learning*. Cambridge university press.

RAFAILOV, R., SHARMA, A., MITCHELL, E., MANNING, C. D., ERMON, S. and FINN, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* **36**.

RASHIDINEJAD, P., ZHU, B., MA, C., JIAO, J. and RUSSELL, S. (2021). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems* **34** 11702–11716.

ROCKAFELLAR, R. T. (1997). *Convex analysis*, vol. 28. Princeton university press.

SCHERRER, B. (2014). Approximate policy iteration schemes: A comparison. In *International Conference on Machine Learning*. PMLR.

SCHULMAN, J. (2015). Trust region policy optimization. *arXiv preprint arXiv:1502.05477* .

SHAH, V., OBANDO-CERON, J., JAIN, V., BARTOLDSON, B., KAILKHURA, B., MITTAL, S., BERSETH, G., CASTRO, P. S., BENGIO, Y., MALKIN, N. ET AL. (2025). A comedy of estimators: On kl regularization in rl training of llms. *arXiv preprint arXiv:2512.21852* .

SHAN, Z., FAN, C., QIU, S., SHI, J. and BAI, C. (2025). Forward kl regularized preference optimization for aligning diffusion policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39.

SHI, L., LI, G., WEI, Y., CHEN, Y. and CHI, Y. (2022). Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International conference on machine learning*. PMLR.

TANG, Y. and MUNOS, R. (2025). On a few pitfalls in kl divergence gradient estimation for rl. *arXiv preprint arXiv:2506.09477* .

TIAPKIN, D., BELOMESTNY, D., CALANDRIELLO, D., MOULINES, E., MUNOS, R., NAUMOV, A., PERRAULT, P., TANG, Y., VALKO, M. and MENARD, P. (2023). Fast rates for maximum entropy exploration. In *International Conference on Machine Learning*. PMLR.

WAINWRIGHT, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint*, vol. 48. Cambridge university press.

WANG, C., JIANG, Y., YANG, C., LIU, H. and CHEN, Y. (2023). Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints. *arXiv preprint arXiv:2309.16240* .

WANG, X., CUI, Q. and DU, S. (2022). On gap-dependent bounds for offline reinforcement learning. *Advances in Neural Information Processing Systems* **35** 14865–14877.

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

WU, D., SHI, C., YANG, J. and SHEN, C. (2025). Greedy sampling is provably efficient for rlhf. *arXiv preprint arXiv:2510.24700* .

XIE, T., CHENG, C.-A., JIANG, N., MINEIRO, P. and AGARWAL, A. (2021a). Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems* **34** 6683–6694.

XIE, T., FOSTER, D. J., KRISHNAMURTHY, A., ROSSET, C., AWADALLAH, A. and RAKHLIN, A. (2024). Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf. *arXiv preprint arXiv:2405.21046* .

XIE, T. and JIANG, N. (2021). Batch value-function approximation with only realizability. In *International Conference on Machine Learning*. PMLR.

XIE, T., JIANG, N., WANG, H., XIONG, C. and BAI, Y. (2021b). Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems* **34** 27395–27407.

XIONG, W., DONG, H., YE, C., WANG, Z., ZHONG, H., JI, H., JIANG, N. and ZHANG, T. (2024). Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. In *Forty-first International Conference on Machine Learning*.

XIONG, W., ZHONG, H., SHI, C., SHEN, C., WANG, L. and ZHANG, T. (2022). Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and markov game. *arXiv preprint arXiv:2205.15512* .

YIN, M. and WANG, Y.-X. (2021). Towards instance-optimal offline reinforcement learning with pessimism. *Advances in neural information processing systems* **34** 4065–4078.

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

ZANETTE, A., WAINWRIGHT, M. J. and BRUNSKILL, E. (2021). Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in neural information processing systems* **34** 13626–13640.

ZHANG, L. and BA, J. (2026). Ema policy gradient: Taming reinforcement learning for llms with ema anchor and top-k kl. *arXiv preprint arXiv:2602.04417* .

ZHANG, T. (2023). *Mathematical Analysis of Machine Learning Algorithms*. Cambridge University Press.

ZHANG, Y., CHEN, C. and JIANG, N. (2026). Beyond pessimism: Offline learning in kl-regularized games. *arXiv preprint arXiv:2604.06738* .

ZHAO, H., YE, C., GU, Q. and ZHANG, T. (2025a). Sharp analysis for KL-regularized contextual bandits and RLHF. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

ZHAO, H., YE, C., XIONG, W., GU, Q. and ZHANG, T. (2025b). Logarithmic regret for online kl-regularized reinforcement learning. In *International Conference on Machine Learning*. PMLR.

ZHAO, Q., JI, K., ZHAO, H., ZHANG, T. and GU, Q. (2026). Towards a sharp analysis of offline policy learning for f-divergence-regularized contextual bandits. In *The Fourteenth International Conference on Learning Representations*.

ZIEBART, B. D., MAAS, A. L., BAGNELL, J. A., DEY, A. K. ET AL. (2008). Maximum entropy inverse reinforcement learning. In *Aaai*, vol. 8. Chicago, IL, USA.

Additional Notations in Appendix

Boldface capital letters are reserved for matrices. Boldface lowercase letters denote vectors in Euclidean spaces or vector-valued functions. For two probability distributions P and Q, the total variation (TV) distance is denoted by $\mathsf{TV}\left(P\|Q\right) \coloneqq 0.5\int|\,\mathrm{d}P-\mathrm{d}Q|$. $\forall n\in\mathbb{N}_{++}$, $\overline{[n]}\coloneqq\{0,1,..,n-1\}$. For dimension-compatible vectors $\mathbf{x},\mathbf{y}$, $\mathbf{x}\odot\mathbf{y}$ denotes their element-wise product.

A Additional Related Works

Pessimism in Offline RL. Pessimism has become one of the central principles in offline decision making problems (Scherrer, 2014; Chen and Jiang, 2019; Xie and Jiang, 2021) for achieving statistical efficiency. In offline CBs and MDPs, pessimistic value or reward estimation has been shown to effectively mitigate the distributional shift between the behavioral policy and the target policy, leading to near-optimal sample complexity guarantees under suitable notion of single-policy-concentrability in both tabular setting (Rashidinejad et al., 2021; Yin and Wang, 2021; Wang et al., 2022; Shi et al., 2022; Li et al., 2024) and function approximation (Jin et al., 2021; Min et al., 2021; Xiong et al., 2022; Zanette et al., 2021; Di et al., 2023). More recently, pessimism-based approaches have been extended to offline decision making with KL-regularized objectives, where they play a key role in establishing fast-rate guarantees that depend only on single-policy concentrability (Zhao et al., 2026), thereby relaxing the all-policy concentrability assumptions required in prior works without algorithmic pessimism (Zhao et al., 2025a; Wu et al., 2025). Our work demonstrates that pessimism generalizes beyond standard objectives and reverse-KL-regularized objectives, and remains effective for a broader class of decision making with f-divergence regularization, including the forward-KL-regularized objective studied in this paper.

Forward-KL Regularization in RL. Forward-KL regularization has been recently utilized in variants of Direct Preference Optimization (DPO) (Rafailov et al., 2023). Wang et al. (2023) generalized DPO to f-divergence penalties, which encompasses forward KL; while other specific alignment algorithms formulated via forward-KL regularization have also been developed for large language models (LLMs) (Ji et al., 2023) and diffusion models (Shan et al., 2025). Another line of research investigates how practical gradient estimation of KL divergence in RL implicitly induces forward-KL regularization. Tang and Munos (2025) and Shah et al. (2025) theoretically and empirically demonstrated that auto-differentiating certain variance-reduced Monte Carlo estimators of reverse KL in some famous KL-regularized RL fine-tuning algorithms actually produces gradients of the forward KL in expectation. Furthermore, Zhang and Ba (2026) introduced an policy gradient framework that incorporates top-k forward-KL regularization to stabilize LLM fine-tuning. A traditional viewpoint on the distinct divergence properties of forward and reverse KL, which has usually been respectively characterized as mass-covering and mode-seeking, is systematically ablated by GX-Chen et al. (2025), who claim that diversity collapse can be an inherent consequence of the regularized RL objective itself regardless of the KL direction. On the theoretical front, the sample complexity of forward-KL-regularized RL is largely unclear. Recently, Aminian et al. (2025) provided the first $\widetilde{O}(\epsilon^{-2})$ slow-rate upper bound for offline contextual dueling bandits with multiple reference models under forward-KL regularization, leaving the achievability of ϵ^{-1}-type fast rates as an open question.

B Relation between Coverage Notions

In this section, we provide more illustrations on the relation between two coverage measures, $D^2_{\pi^*}$ and C^{π^*}, in the context of forward-KL regularized bandits. In particular, we provide two cases under linear function approximation, in one of which $D^2_{\pi^*} = \Omega(dC^{\pi^*})$ and in the other we have $D^2_{\pi^*} \ll C^{\pi^*}$, where d is the dimension of the function class. We summarized them as two propositions.

Proposition B.1. For any $C > 1$, $d \geq 3$ being odd, and $\eta \geq 2$, there exist a forward-KL-regularized linear bandit instance, such that $C^{\pi^*} = O(C)$ and $D^2_{\pi^*} = \Omega(Cd)$, leading to $D^2_{\pi^*} = \Omega(dC^{\pi^*})$.

Proof. We construct the instance as follows. Let $d = 2A + 1$ be some odd number and consider an $2A + 1$-armed bandit, such that the feature vector of the i-th arm, $\boldsymbol{\phi}(a_i) = \mathbf{e}_i \in \mathbb{R}^d$, which has 1 on its i-th entry and 0 on all other entries. The reference policy $\pi^{\mathsf{ref}}(a_i) = (2AC)^{-1}$ for $i \in [2A]$ and $\pi^{\mathsf{ref}}(a_{2A+1}) = (C-1)/C$, where $C \geq 1$ and $\eta \geq 2$. The ground truth reward function $\boldsymbol{\theta}^* = \sum_{i \leq A} \mathbf{e}_i$ and the function class is given by all $\|\boldsymbol{\theta}\|_\infty \leq 1$. By construction, we know that $\pi^*(a_i) \geq \pi^{\mathsf{ref}}(a_i)$ if and only if $i \in [A]$ and its closed form is given by

$$\pi^*(a_i) = \frac{1}{A} \frac{\sqrt{(\eta-1)^2 C^2 + 2\eta C} + (\eta - 1)C}{2\eta C} \leq \frac{2}{A}$$

which gives $C^{\pi^*} = O(C)$. Now we compute the $D^2_{\pi^*}$ of this instance. For all $i \in [A]$, we know that

$$D^2(a_i) = \sup_{\|\boldsymbol{\theta}\|_\infty \leq 2} \frac{\langle \boldsymbol{\theta}, \mathbf{e}_i \rangle^2}{\mathbb{E}_{\pi^{\mathsf{ref}}} \langle \boldsymbol{\theta}, \mathbf{e}_j \rangle^2} = 2CA = \Theta(Cd),$$

where the second equation holds with $\boldsymbol{\theta} = \mathbf{e}_i$. Taking expectation over π^*, we have

$$D^2_{\pi^*} \geq \frac{1}{2A} \sum_{i \in [A]} D^2(a_i) = \Theta(C^{\pi^*} d),$$

which concludes the proof. □

The following proposition provides another instance on which $D^2_{\pi^*} \ll C^{\pi^*}$.

Proposition B.2. For any $C \geq 2$ and $\eta \geq 2$, there exists a KL-regularized linear bandit instance, such that $C^{\pi^*} = C/2$ and $D^2_{\pi^*} = \Theta(1)$.

Proof. We consider the function class of $\boldsymbol{\theta} \in \mathbb{R}^2$ and $\|\boldsymbol{\theta}\| \leq \sqrt{2}$. The instance consists of three arms, where $\boldsymbol{\phi}(a_1) = (1, 0)$, $\boldsymbol{\phi}(a_2) = (0, 1)$, and $\boldsymbol{\phi}(a_3) = (1, 1)$. The ground truth parameter $\boldsymbol{\theta}^* = (1, 1)$. The reference policy is given by $\pi^{\mathsf{ref}}(a_1) = \pi^{\mathsf{ref}}(a_2) = 1/2 - 1/2C$ and $\pi^{\mathsf{ref}}(a_3) = 1/C$, where $C \geq 2$. We further fix any $\eta \geq 2$. A direct computation yields that

$$\pi^*(a_3) = \frac{(\eta - 1) + \sqrt{(\eta-1)^2 + 4\eta C^{-1}}}{2\eta} \geq \frac{\eta - 1}{\eta},$$

which results in $C^{\pi^*} \geq C/2$. On the other hand, we know that for $i = 1, 2$, we have $D^2(a_i) \leq \pi^{\mathsf{ref}}(a_i)^{-1} \leq 4$. As for a_3, since we have $\langle \boldsymbol{\theta}, \phi(a_3) \rangle^2 = \langle \boldsymbol{\theta}, \phi(a_1) + \phi(a_2) \rangle^2 \leq 2 \langle \boldsymbol{\theta}, \phi(a_1) \rangle^2 + 2 \langle \boldsymbol{\theta}, \phi(a_2) \rangle$, which gives that $D^2(a_3) \leq 2D^2(a_1) + 2D^2(a_2) \leq 16$. Therefore, taking expectation over π^*, we know that $D^2_{\pi^*} \leq 16$ which is a constant. □

C Missing Proofs in Section 3

C.1 Proof of Lemma 3.1

Proof of Lemma 3.1. The argument follows standard concentration analyses in the literature; we include the details for completeness. We first show that the event $\mathcal{E}_{\mathrm{tab}}(\delta)$ holds with high probability. Fix any $(s, a) \in \mathcal{S} \times \mathcal{A}$. When $N(s, a) = 0$, the claim is immediate. Otherwise, in the case $N(s, a) \geq 1$, we condition on the value of $N(s, a)$ and apply Azuma-Hoeffding's inequality (Lemma F.12). This leads to that with probability at least $1 - \delta/(2SA)$,

$$g^*(s, a) - \frac{1}{N(s, a)} \sum_{i=1}^{n} r_i \, \mathbb{1}\{(s_i, a_i) = (s, a)\} \leq \sqrt{\frac{2 \log(2SA/\delta)}{N(s, a)}} \leq b(s, a).$$

Taking a union bound over all $(s,a) \in \mathcal{S} \times \mathcal{A}$, we obtain that $\mathcal{E}_1$ holds with probability at least $1-\delta/2$. Next, we establish the second event. By Lemma F.11, for any fixed (s,a), with probability at least $1-\delta/(2SA)$,

$$\frac{1}{N \vee 1} \leq \frac{8\log(2SA/\delta)}{n\rho(s)\pi^{\mathsf{ref}}(a|s)}.$$

Taking another union bound over all $(s,a) \in \mathcal{S} \times \mathcal{A}$ shows that the event $\mathcal{E}_{\mathrm{bin}}$ holds with probability at least $1-\delta/2$. Finally, applying a union bound over the events $\mathcal{E}_{\mathrm{tab}}$ and $\mathcal{E}_{\mathrm{bin}}$ completes the proof. □

C.2 Proof of Theorem 3.4

Proof of Theorem 3.4. On event $\mathcal{E}_{\mathrm{gfa}}(\delta)$, we know that $\widehat{g}(s,a) \leq g^*(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Therefore, invoking Lemma 5.5, gives that

$$\begin{aligned}
\mathrm{SubOpt}_{\mathrm{FKL}}(\widehat{\pi}) &\leq 2\eta \mathbb{E}_{s\sim\rho}\left[\sum_{a\in\mathcal{A}}\left(\frac{{\pi^*_{\mathrm{FKL}}}^2(a|s)}{\pi^{\mathsf{ref}}(a|s)}(g^*(s,a)-\widehat{g}(s,a))^2\right)\right] \\
&= 2\eta \mathbb{E}_{s\sim\rho}\left[\sum_{a\in\mathcal{A}}\left[\left(\frac{\pi^*_{\mathrm{FKL}}(a|s)}{\pi^{\mathsf{ref}}(a|s)}\right)\pi^*_{\mathrm{FKL}}(a|s)(g^*(s,a)-\widehat{g}(s,a))^2\right]\right] \\
&\leq 2\eta C^{\pi^*}\mathbb{E}_{(s,a)\sim\rho\times\pi^*_{\mathrm{FKL}}}\left[\left(\bar{g}(s,a)-g^*(s,a)\right)^2\right] \\
&\lesssim \eta C^{\pi^*} D^2_{\pi^*} n^{-1}\log N_{\mathcal{G}}(\epsilon_c),
\end{aligned}$$

where the second inequality holds by $\pi^*_{\mathrm{FKL}}(a|s)/\pi^{\mathsf{ref}}(a|s) \leq C^{\pi^*}$ and the last inequality holds due to the definition of $D^2_{\pi^*}$. Finally, by Lemma 3.3, $\mathcal{E}_{\mathrm{gfa}}(\delta)$ holds with probability at least $1-\delta$, which finishes the proof. □

C.3 Proof of Theorem 3.2

On event $\mathcal{E}_{\mathrm{tab}}(\delta) \cap \mathcal{E}_{\mathrm{bin}}(\delta)$, we know that $\widehat{g}(s,a) \leq g^*(s,a)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Therefore, invoking Lemma 5.5, gives that

$$\begin{aligned}
\mathrm{SubOpt}_{\mathrm{FKL}}(\widehat{\pi}) &\leq 2\eta \mathbb{E}_{s\sim\rho}\left[\sum_{a\in\mathcal{A}}\left(\frac{{\pi^*_{\mathrm{FKL}}}^2(a|s)}{\pi^{\mathsf{ref}}(a|s)}(g^*(s,a)-\widehat{g}(s,a))^2\right)\right] \\
&= 2\eta \mathbb{E}_{s\sim\rho}\left[\sum_{a\in\mathcal{A}}\left[\left(\frac{\pi^*_{\mathrm{FKL}}(a|s)}{\pi^{\mathsf{ref}}(a|s)}\right)^2\pi^{\mathsf{ref}}(a|s)(g^*(s,a)-\widehat{g}(s,a))^2\right]\right] \\
&\leq 2\eta (C^{\pi^*})^2\mathbb{E}_{(s,a)\sim\rho\times\pi^{\mathsf{ref}}}\left[\left(\bar{g}(s,a)-g^*(s,a)\right)^2\right] \\
&\leq 2\eta (C^{\pi^*})^2\sum_{s\in\mathcal{S}}\rho(s)\sum_{a\in\mathcal{A}}\pi^{\mathsf{ref}}(a|s)\frac{32\log^2(2|\mathcal{S}||\mathcal{A}|/\delta)}{n\rho(s)\pi^{\mathsf{ref}}(a|s)} \\
&= \widetilde{O}\big(\eta (C^{\pi^*})^2|\mathcal{S}||\mathcal{A}|n^{-1}\big),
\end{aligned}$$

where the second inequality holds by $\pi^*_{\mathrm{FKL}}(a|s)/\pi^{\mathsf{ref}}(a|s) \leq C^{\pi*}$ and the last inequality holds due to event $\mathcal{E}_{\mathrm{bin}}(\delta)$. By Lemma 3.1, we know that $\mathcal{E}_{\mathrm{tab}}(\delta) \cap \mathcal{E}_{\mathrm{bin}}(\delta)$ holds with probability at least $1-\delta$, which finishes the proof.

D Missing Proofs in Section 4

We first present (for each given context) a suboptimality decomposition lemma applicable for general learners.[6]

Lemma D.1. Let $\widehat{\pi} \in \Delta(\mathcal{A})$ be *any* distribution with $\mathrm{supp}(\widehat{\pi}) = \mathcal{A}$, then under Assumption 2.1,

$$\mathrm{SubOpt}_{\mathrm{FKL}}(\widehat{\pi}) = \eta^{-1}\mathbb{E}_{a\sim\pi^{\mathsf{ref}}}\left[\psi\Big(\frac{\widehat{\pi}(a)}{\pi^*_{\mathrm{FKL}}(a)}\Big)\right], \tag{D.1}$$

where $\psi(x) := x - 1 - \log x$.

[6] We drop the dependency on context in Lemma D.1 as the suboptimality gaps on each $s \in \mathcal{S}$ are purely additive.

Proof of Lemma D.1. Lemma 2.2 implies that $\forall a \in \mathcal{A}$,

$$r(a) = \lambda - \eta^{-1}\frac{\pi^{\mathsf{ref}}(a)}{\pi^*(a)},$$

which, substituted into $\mathrm{SubOpt}(\cdot)$, yields

$$\begin{aligned}
\mathrm{LHS} &:= J(\pi^*; r) - J(\widehat{\pi}; r) \\
&= \langle r, \pi^* - \widehat{\pi}\rangle + \eta^{-1}\sum_{a\in\mathcal{A}} \pi^{\mathsf{ref}}(a)\log\frac{\pi^{\mathsf{ref}}(a)}{\widehat{\pi}(a)} - \eta^{-1}\sum_{a\in\mathcal{A}} \pi^{\mathsf{ref}}(a)\log\frac{\pi^{\mathsf{ref}}(a)}{\pi^*(a)} \\
&= \langle \lambda\mathbf{1} - \eta^{-1}\frac{\pi^{\mathsf{ref}}}{\pi^*}, \pi^* - \widehat{\pi}\rangle - \eta^{-1}\sum_{a\in\mathcal{A}} \pi^{\mathsf{ref}}(a)\log\frac{\widehat{\pi}(a)}{\pi^*(a)} \\
&= -\eta^{-1}\sum_{a\in\mathcal{A}} \pi^{\mathsf{ref}}(a)\Big(1 - \frac{\widehat{\pi}(a)}{\pi^*(a)}\Big) - \eta^{-1}\sum_{a\in\mathcal{A}} \pi^{\mathsf{ref}}(a)\log\frac{\widehat{\pi}(a)}{\pi^*(a)} \\
&= \eta^{-1}\mathbb{E}_{a\sim\pi^{\mathsf{ref}}}\left[\frac{\widehat{\pi}(a)}{\pi^*(a)} - 1 - \log\frac{\widehat{\pi}(a)}{\pi^*(a)}\right],
\end{aligned}$$

where the penultimate equality is because both π^* and $\widehat{\pi}$ are probability distributions. □

Remark D.2. Although the *pattern* of Lemma D.1 fits into (5.6) exactly, their applicable scopes are different because the first equality of (5.6) is a specialization of Lemma F.4 to forward KL, which only needs mild regularity conditions of the regularizer h but requires $\widehat{\pi}$ to be the maximizer of $J_{\mathrm{FKL}}(\cdot; \widehat{r})$. On the other hand, Lemma D.1 is tailored to forward KL but $\widehat{\pi}$ here can be essentially an arbitrary probability distribution.

D.1 The Fast Rate Lower Bound

Proof of Theorem 4.1. Let $K := A/2$. We consider MABs with context set $\mathcal{S} = \overline{[S]}$ and action set $\mathcal{A} = \overline{[A]}$. We fix $\rho = \mathsf{Uniform}(\mathcal{S})$ $\pi^{\mathsf{ref}} = \mathsf{Uniform}(\mathcal{A})$, where every reward follows a Bernoulli distribution. We consider the subset of instances with the following parameterization. Let $\mathcal{V} = \{\pm 1\}^{S\times K}$ and we use $v_{s,k}$ to denote its (s,k)-th entry given any $v \in \mathcal{V}$. The reward distribution given $v \in \mathcal{V}$ is given as follows:

$$r_v(s, 2k) = 0.5 - v_k c,\ r_v(s, 2k+1) = 0.5 + v_k c; \quad \forall\, (s,k) \in \overline{[S]} \times \overline{[K]},$$

where $c \in (0, 0.25)$ is to be specified.
We now compute the optimal policy π_v for each instance parameterized by $v \in \mathcal{V}$. Fix any s, we know by Lemma 2.2 that $\pi_v^*(\cdot|s) = (\eta A)^{-1}/(\lambda - r_v(s,\cdot))$; and thus all instances share the same λ over all the contexts $s \in \mathcal{S}$ and λ is determined by

$$\sum_{k=0}^{K-1}\Big(\frac{(\eta A)^{-1}}{\lambda - 0.5 - c} + \frac{(\eta A)^{-1}}{\lambda - 0.5 + c}\Big) = 1 \Longrightarrow \lambda - 0.5 = \frac{1 + \sqrt{1 + 4\eta^2 c^2}}{2\eta},$$

which is because $\lambda > \max_{v,a} r_v(a)$. Let $\iota = \lambda - 0.5$, then $\iota = \eta(\iota^2 - c^2)$, and thus we set $\kappa_{v,s,a} := (-1)^{a-2\lfloor a/2\rfloor}\cdot v_{s,\lfloor a/2\rfloor}$ and $\omega := c/\iota = 2\eta c/\big(1 + \sqrt{1+4\eta^2c^2}\big)$ to obtain

$$\begin{aligned}
\pi_v^*(a|s) &= \frac{1}{\eta A}\cdot\frac{1}{\iota + (-1)^{a-2\lfloor a/2\rfloor}\cdot v_{s,\lfloor a/2\rfloor}\cdot c} \\
&= \frac{\iota - \kappa_{v,s,a}c}{\eta A(\iota^2 - c^2)} \\
&= A^{-1}(1 - \kappa_{v,s,a}c/\iota) \\
&= \frac{1 - \kappa_{v,s,a}\omega}{A}.
\end{aligned}$$

Note that Lemma D.1 implies

$$\mathrm{SubOpt}(\widehat{\pi}; r_v) = \sum_{s=0}^{S-1}\sum_{k=0}^{K-1}\frac{1}{\eta SA}\left[\psi\Big(\frac{\widehat{\pi}(2k|s)}{\pi_v^*(2k|s)}\Big) + \psi\Big(\frac{\widehat{\pi}(2k+1|s)}{\pi_v^*(2k+1|s)}\Big)\right]$$

$$= \sum_{s=0}^{S-1} \sum_{k=0}^{K-1} \underbrace{\frac{1}{\eta SA} \left[\psi\Big(\frac{A\widehat{\pi}(2k|s)}{1 - v_k \omega} \Big) + \psi\Big(\frac{A\widehat{\pi}(2k+1|s)}{1 + v_k \omega} \Big) \right]}_{=:\Psi_{s,k}(\widehat{\pi}; v)}.$$

Then for $v \sim_{s,j} v'$ (i.e., $v, v' \in \{\pm 1\}^{S \times K}$ differ only in the (s, j)-th coordinate),

$$\begin{aligned} &\Psi_{s,j}(\widehat{\pi}; v) + \Psi_{s,j}(\widehat{\pi}; v') \\ &\quad = \frac{1}{\eta SA} \bigg(\psi\Big(\frac{A\widehat{\pi}(2j|s)}{1 - \omega} \Big) + \psi\Big(\frac{A\widehat{\pi}(2j|s)}{1 + \omega} \Big) + \psi\Big(\frac{A\widehat{\pi}(2j+1|s)}{1 - \omega} \Big) + \psi\Big(\frac{A\widehat{\pi}(2j+1|s)}{1 + \omega} \Big) \bigg). \end{aligned} \tag{D.2}$$

Now we specify $c = \sqrt{SA/n}$, which is less than 0.25 by design. Recall that $\psi(x) = x - 1 - \log x$, elementary calculus shows that $o \mapsto \psi\big(o/(1+\omega)\big) + \psi\big(o/(1-\omega)\big)$ is minimized at $o = 1 - \omega^2$, and thus

$$\begin{aligned} \text{(D.2)} &\geq \frac{2}{\eta SA} \log \frac{1}{1 - \omega^2} \\ &= \frac{2}{\eta SA} \cdot \log \frac{1 + \sqrt{1 + 4\eta^2 c^2}}{2} \\ &\geq \frac{2}{\eta SA} \cdot \frac{\sqrt{1 + 4\eta^2 c^2} - 1}{4} \\ &= \frac{\sqrt{1 + 4\eta^2 c^2} - 1}{2\eta SA} \\ &\geq \frac{2\eta c^2}{SA(1 + \sqrt{17})} \\ &= \Omega(\eta n^{-1}), \end{aligned} \tag{D.3}$$

where the second inequality is because $\eta^{-1} > c/2$ and $\log(1 + x) \geq x/2$ for $x \in [0, 2]$, and the last inequality is again due to $\eta^{-1} > c/2$.

Now, given any reward function r, let $\mathbb{P}_{\pi^{\mathsf{ref}}, \mathtt{r}}$ be the distribution of $(a, \mathtt{r})$ for $(s, a) \sim \rho \times \pi^{\mathsf{ref}}$ and $\mathtt{r} \sim \mathsf{Bernoulli}(r(a))$. We further use $\mathbb{P}_{\pi^{\mathsf{ref}}, v}$ for $v \in \mathcal{V}$ to denote $\mathbb{P}_{\pi^{\mathsf{ref}}, r}$ if r is parameterized by $v \in \mathcal{V}$. The divergence decomposition lemma (Lattimore and Szepesvári, 2020, Lemma 15.1) then implies that for any $(s, j) \in [S] \times [K]$ and $v \sim_{s,j} v'$,

$$\begin{aligned} \mathsf{KL}\left(\mathbb{P}^{\otimes n}_{\pi^{\mathsf{ref}}, v} \| \mathbb{P}^{\otimes n}_{\pi^{\mathsf{ref}}, v'} \right) &= \frac{n}{SA} \mathsf{KL}\left(\mathrm{Bern}(r_v(2j) \| \mathrm{Bern}(r_{v'}(2j)) \right) \qquad \text{(D.4)} \\ &\quad + \frac{n}{SA} \mathsf{KL}\left(\mathrm{Bern}(r_v(2j+1) \| \mathrm{Bern}(r_{v'}(2j+1)) \right) \\ &\leq \frac{128}{3} \frac{nc^2}{SA} = \frac{128}{3}, \qquad \text{(D.5)} \end{aligned}$$

where we utilize $\mathsf{KL}\left(\mathrm{Bern}(p) \| \mathrm{Bern}(q) \right) \leq (p - q)^2 / [q(1 - q)]$ and $c \in [0, 0.25]$. By Lemma F.14, (D.3) and (D.5) together imply the desired lower bound. □

D.2 Lower Bound for Two-Armed Bandits: Phase Transition

Theorem D.3. Let $\widehat{\pi} \in \Delta(\mathcal{A})$ be any estimator from n action-reward pairs, then for any $n > 16$ and $\eta > 0$,

$$\sup_{(\mathcal{A}, r, \pi^{\mathsf{ref}}) \in \mathrm{MAB}_2} \mathbb{E}_{\mathcal{D} \sim P_{\pi^{\mathsf{ref}}, r}} \mathrm{SubOpt}_{\mathrm{FKL}}(\widehat{\pi}; \mathcal{A}, r, \pi^{\mathsf{ref}}) \gtrsim \eta n^{-1} \wedge n^{-1/2}. \tag{D.6}$$

Proof. We construct two MABs sharing $\mathcal{A} = \{0, 1\}$ and $\pi^{\mathsf{ref}} = \mathsf{Uniform}(\mathcal{A})$,[7] where the rewards follow Bernoulli distribution with means specified as

$$\acute{r} = (0.5 + c, 0.5 - c), \grave{r} = (0.5 - c, 0.5 + c). \tag{D.7}$$

[7] In this proof, we drop notations that are clear in the context to make the presentation less heavy.

Here, $c \in (0, 0.25)$ will be specified later. By Lemma 2.2, $\pi^*(\cdot) = (2\eta)^{-1}\big(\lambda - r(\cdot)\big)^{-1}$, which means the two instances $\acute{r}$ and $\grave{r}$ share the same λ, i.e., the λ in both cases satisfies

$$(\lambda - 0.5 - c)(\lambda - 0.5 + c) = \frac{2\lambda - 1}{2\eta}. \tag{D.8}$$

To get an eligible distribution π^*, we require $\lambda > 0.5 + c$ according to (2.2), which means (D.8) yields

$$2\lambda - 1 = \eta^{-1} + \sqrt{\eta^{-2} + 4c^2}. \tag{D.9}$$

Thus, for any estimator $\widehat{\pi} \in \Delta(\{0, 1\})$,

$$\begin{aligned}
\mathrm{SubOpt}(\widehat{\pi}; \acute{r}) + \mathrm{SubOpt}(\widehat{\pi}; \grave{r}) &= \langle \acute{r}, \acute{\pi}^* \rangle + \langle \grave{r}, \grave{\pi}^* \rangle - \overbrace{\langle \acute{r} + \grave{r}, \widehat{\pi} \rangle}^{=1} + \overbrace{2\eta^{-1}\mathsf{KL}\left(\pi^{\mathsf{ref}} \| \widehat{\pi}\right)}^{\geq 0} \\
&\quad - \eta^{-1}\mathsf{KL}\left(\pi^{\mathsf{ref}} \| \acute{\pi}^*\right) - \eta^{-1}\mathsf{KL}\left(\pi^{\mathsf{ref}} \| \grave{\pi}^*\right) \\
&\geq 2\langle \acute{r}, \acute{\pi}^* \rangle - 1 - 2\eta^{-1}\mathsf{KL}\left(\pi^{\mathsf{ref}} \| \acute{\pi}^*\right) \\
&= 2\Big((0.5 + c)\acute{\pi}^*(0) + (0.5 - c)\acute{\pi}^*(1)\Big) - 1 \\
&\quad - \eta^{-1}\Big(\log \eta(\lambda - 0.5 - c) + \log \eta(\lambda - 0.5 + c)\Big) \\
&= \frac{c\eta^{-1}}{\eta(\lambda - 0.5 - c)} - \frac{c\eta^{-1}}{\eta(\lambda - 0.5 + c)} \\
&\quad - \eta^{-1}\Big(\log \eta(\lambda - 0.5 - c) + \log \eta(\lambda - 0.5 + c)\Big) \\
&= \frac{4c^2}{2\lambda - 1} - \eta^{-1} \log \frac{\eta(2\lambda - 1)}{2} \qquad \text{(D.10)} \\
&= \frac{4c^2}{\eta^{-1} + \sqrt{\eta^{-2} + 4c^2}} - \eta^{-1} \log \Big(1 + \frac{\sqrt{1 + 4c^2\eta^2} - 1}{2}\Big). \qquad \text{(D.11)}
\end{aligned}$$

where (D.10) is due to (D.8) and (D.11) follows from (D.9). Given any mean function r, let $P_{\pi^{\mathsf{ref}}, r}$ be the distribution of $(a, \mathtt{r})$ for $a \sim \pi^{\mathsf{ref}}$ and $\mathtt{r} \sim \mathsf{Bernoulli}(r(a))$. The tensorization property of KL (Polyanskiy and Wu, 2025, Chapter 2) then implies

$$\mathsf{KL}\left(P^{\otimes n}_{\pi^{\mathsf{ref}}, \acute{r}} \| P^{\otimes n}_{\pi^{\mathsf{ref}}, \grave{r}}\right) = n\mathsf{KL}\left(P_{\pi^{\mathsf{ref}}, \acute{r}} \| P_{\pi^{\mathsf{ref}}, \grave{r}}\right) = 2nc \log \frac{1 + 2c}{1 - 2c} \leq 16nc^2, \tag{D.12}$$

where the inequality follows from the requirement $c \in (0, 0.25)$. Since $n > 16$. we can set $c = n^{-1/2}$.

Case $\eta^{-1} \leq c/2 = 0.5n^{-1/2}$**.** Using standard $\sqrt{x + y} \leq \sqrt{x} + \sqrt{y}$ and $\log(1 + x) \leq x$ for $x, y \geq 0$,

$$\text{(D.11)} \geq \frac{4c^2}{2\eta^{-1} + 2c} - \eta^{-1}\log(1 + c\eta) \geq \frac{2c^2}{\eta^{-1} + c} - c \geq c/3,$$

which, combined with Lemma F.13 and (D.12), yields

$$\text{LHS of (D.6)} \gtrsim n^{-1/2}.$$

Case $\eta^{-1} > c/2 = 0.5n^{-1/2}$**.** Using $1 = \sup_{x>0} x^{-1}\log(1 + x)$,

$$\begin{aligned}
\text{(D.11)} &= \frac{2}{1 + \sqrt{1 + 4c^2\eta^2}}\left(2\eta c^2 - \eta c^2 \cdot \frac{\log\left(1 + \frac{\sqrt{1+4c^2\eta^2}-1}{2}\right)}{\frac{\sqrt{1+4c^2\eta^2}-1}{2}}\right) \\
&\geq \frac{2}{1 + \sqrt{17}}(2\eta c^2 - \eta c^2) \geq \eta c^2/3,
\end{aligned}$$

which, combined with Lemma F.13 and (D.12), yields

$$\text{LHS of (D.6)} \gtrsim \eta n^{-1}.$$

□

E Other Missing Proofs

Proof of Lemma 2.2. Consider the primal problem[8]

$$\begin{aligned} &\min_{\pi} -\langle r, \pi\rangle + \eta^{-1}\mathsf{KL}\left(\pi^{\mathsf{ref}}\|\pi\right) \\ &\text{s.t. } \langle \mathbf{1}, \pi\rangle = 1, -\pi \le \mathbf{0}; \end{aligned} \tag{P}$$

whose Lagrangian with dual variables (ν, λ) is

$$\mathcal{L}(\pi, \nu, \lambda) := -\langle r, \pi\rangle + \eta^{-1}\mathsf{KL}\left(\pi^{\mathsf{ref}}\|\pi\right) + \lambda(\langle \mathbf{1}, \pi\rangle - 1) - \langle \nu, \pi\rangle.$$

Since the only inequality constraint in (P) is affine, Slater's condition is satisfied, which implies strong duality. Thus, π solves (P) as long as (π, ν, λ) solves $\min_{\pi\in\mathbb{R}^{\mathcal{A}}} \max_{\nu\in\mathbb{R}^{\mathcal{A}}_{+}, \lambda\in\mathbb{R}} \mathcal{L}(\pi, \nu, \lambda)$, for the latter of which the KKT conditions are necessary:

$$\nabla_{\pi}\mathcal{L}(\pi, \nu, \lambda) = \mathbf{0}, \tag{E.1}$$

$$\pi \ge \mathbf{0}, \langle \mathbf{1}, \pi\rangle = 1, \tag{E.2}$$

$$\nu \ge \mathbf{0}, \pi \odot \nu = \mathbf{0}. \tag{E.3}$$

(E.1) implies

$$\forall a \in \pi(a) = \frac{\eta^{-1}\pi^{\mathsf{ref}}(a)}{\lambda - r(a) - \nu(a)}.$$

Thus, if $a \in \operatorname{supp}(\pi^{\mathsf{ref}})$, $\nu(a) = 0$ by (E.3); otherwise, $\pi(a) = \pi^{\mathsf{ref}}(a) = 0$. Thus, λ solves

$$\sum_{a\in\mathcal{A}} \frac{\eta^{-1}\pi^{\mathsf{ref}}(a)}{\lambda - r(a)} = 1. \tag{E.4}$$

Since $\lambda \mapsto \eta^{-1}\pi^{\mathsf{ref}}(a)/\big(\lambda - r(a)\big)$ is monotone and continuous for $\lambda > \max_{a\in\operatorname{supp}(\pi^{\mathsf{ref}})} r(a)$, whose range is $(0, +\infty)$; such a λ solving (E.4) is unique. When $\operatorname{supp}(\pi^{\mathsf{ref}}) = \mathcal{A}$, (P) is at least a strict convex minimization problem by Lemma 5.1, and hence the solution is unique. □

Proof of Lemma 5.1. Pick an arbitrary $\pi \in \operatorname{ri}\big(\Delta(\mathcal{A})\big)$, let $F(\cdot) = \mathsf{KL}\left(\pi^{\mathsf{ref}}\|\cdot\right)$. Then

$$\nabla F(\pi) = -\frac{\pi^{\mathsf{ref}}}{\pi} \Rightarrow \nabla^2 F(\pi) = \operatorname{diag}\Big\{\frac{\pi^{\mathsf{ref}}}{\pi^2}\Big\}. \tag{E.5}$$

Here, both divisions are element-wise. $\forall \mu, \nu \in \Delta(\mathcal{A})$,

$$\begin{aligned} (\mu - \nu)^{\top}\nabla^2 F(\pi)(\mu - \nu) &= \sum_{a\in\mathcal{A}} \frac{\pi^{\mathsf{ref}}(a)\big(\mu(a) - \nu(a)\big)^2}{\big(\pi(a)\big)^2} \\ &\ge \alpha \sum_{a\in\mathcal{A}} \frac{\big(\mu(a) - \nu(a)\big)^2}{\big(\pi(a)\big)^2} \\ &\ge \alpha \frac{\Big(\sum_{a\in\mathcal{A}} |\mu(a) - \nu(a)|\Big)^2}{\sum_{a\in\mathcal{A}} \big(\pi(a)\big)^2} \\ &\ge 4\alpha \cdot \big(\mathsf{TV}\left(\mu\|\nu\right)\big)^2, \end{aligned}$$

where the second inequality follows from Cauchy–Schwarz.[9] Here, we recall that TV is equivalent to the L^1 distance up to a multiplicative constant 2, and thus the conclusion follows. □

[8] We drop again the context $s \in \mathcal{S}$ and its conditioning in this proof since the optimal solution for each context is fully decoupled.

[9] Technically, $\dim \Delta(\mathcal{A}) = A - 1 < A$; and thus by $\nabla F(\pi)$ (resp. $\nabla^2 F(\pi)$) for $\pi \in \operatorname{ri}(\Delta(\mathcal{A}))$ in this proof, we actually mean the Euclidean gradient (resp. Hessian) of the analytical continuation of F defined as $F_{\text{ext}}(\mu) := \langle \pi^{\mathsf{ref}}, \log \frac{\pi^{\mathsf{ref}}}{\mu}\rangle$ for $\mu \in \mathbb{R}^{\mathcal{A}}_{++}$ and $+\infty$ otherwise. A similar convention follows in the proof of Lemma 2.2.

Proof of Lemma 5.2. For $u \in [0,1]$, $r_u := (1-u)r + u\widehat{r}$, and $\pi_u = \operatorname{argmax}_{\pi \in \Delta(\mathcal{A})} J_{\mathrm{FKL}}(\pi; r_u)$, the latter of which exists and is unique by Lemma 2.2. Let

$$V(u) := J_{\mathrm{FKL}}(\pi_u; r_u), \quad \Phi(u) := J_{\mathrm{FKL}}(\pi_u; r).$$

Lemma 2.2 and the definition of J_{FKL} guarantee that both $V(\cdot)$ and $u \mapsto J_{\mathrm{FKL}}(\pi, r_u)$ is differentiable given any $\pi \in \Delta(\mathcal{A})$. Therefore, by Lemma F.2, $\forall u \in (0,1)$,

$$V'(u) = \Big\langle \frac{\mathrm{d}}{\mathrm{d}u} r_u, \pi_u \Big\rangle = \langle \widehat{r} - r, \pi_u \rangle,$$

which, substituted to $\Phi(u) = V(u) + \langle r - r_u, \pi_u \rangle = V(u) + u\langle \widehat{r} - r, \pi_u \rangle$, yields

$$\begin{aligned} \forall u \in (0,1), \Phi'(u) &= V'(u) + \frac{\mathrm{d}}{\mathrm{d}u}\Big(u\langle r - \widehat{r}, \pi_u \rangle\Big) \\ &= \langle \widehat{r} - r, \pi_u \rangle + \langle r - \widehat{r}, \pi_u \rangle + u\Big\langle r - \widehat{r}, \frac{\mathrm{d}}{\mathrm{d}u}\pi_u \Big\rangle = u\Big\langle r - \widehat{r}, \frac{\mathrm{d}}{\mathrm{d}u}\pi_u \Big\rangle. \end{aligned} \tag{E.6}$$

By definition,

$$\text{LHS of (5.1)} = \Phi(0) - \Phi(1) = -\int_0^1 \Phi'(u)\,\mathrm{d}u = \int_0^1 u\Big\langle \widehat{r} - r, \frac{\mathrm{d}}{\mathrm{d}u}\pi_u \Big\rangle \mathrm{d}u. \tag{E.7}$$

Let $h(\cdot) := \eta^{-1}\mathsf{KL}\left(\pi^{\mathsf{ref}} \| \cdot\right)$. Since h is strictly convex and continuously differentiable on $\mathrm{ri}(\Delta(\mathcal{A}))$, $\mathrm{dom}(h) := \{\pi \in \Delta(\mathcal{A}) : h(\pi) < \infty\} = \mathrm{ri}(\Delta(\mathcal{A}))$, and the uniqueness of π_u is guaranteed by Lemma 2.2, applying Lemma F.3 to h yields that $\forall u \in (0,1)$, $\pi_u = \nabla h^*(r_u)$, where the notation h^* is consistent with Lemma F.3. Consequently,

$$\frac{\mathrm{d}}{\mathrm{d}u}\pi_u = \frac{\mathrm{d}}{\mathrm{d}u}\nabla h^*(r_u) = \left[\nabla^2 h^*(r_u)\right] \cdot \frac{\mathrm{d}}{\mathrm{d}u} r_u = \left[\nabla^2 h^*(r_u)\right](\widehat{r} - r), \tag{E.8}$$

where $\widehat{r} - r$ is again viewed as a vector. Substituting (E.8) into (E.7) yields

$$\begin{aligned} \text{LHS of (5.1)} &= \int_0^1 u(\widehat{r} - r)^\top \left[\nabla^2 h^*(r_u)\right](\widehat{r} - r)\,\mathrm{d}u \\ &= \int_0^1 u(r - \widehat{r})^\top \left[\nabla^2 h(\pi_u)\right]^{-1}(r - \widehat{r})\,\mathrm{d}u - \int_0^1 u \frac{\left(\mathbf{1}^\top \left[\nabla^2 h(\pi_u)\right]^{-1}(r - \widehat{r})\right)^2}{\mathbf{1}^\top \left[\nabla^2 h(\pi_u)\right]^{-1}\mathbf{1}}\,\mathrm{d}u \\ &= \eta \int_0^1 \left(\sum_a \left(\frac{\pi_u^2}{\pi^{\mathsf{ref}}}(r - \widehat{r})^2\right)(a) - \frac{\left(\sum_a \left(\frac{\pi_u^2}{\pi^{\mathsf{ref}}}(r - \widehat{r})\right)(a)\right)^2}{\sum_a \frac{\pi_u^2}{\pi^{\mathsf{ref}}}(a)} \right) u\,\mathrm{d}u, \end{aligned}$$

where $\frac{\pi_u^2}{\pi^{\mathsf{ref}}}(r - \widehat{r})^2$ is viewed as a function, the second equality follows from Lemma F.7 and Assumption 2.1 and the last equality follows from direct calculations detailed in (E.5). The final result thus follows from the *second* mean value theorem for definite integrals.

□

Proof of Lemma 5.3. Recall that $F(\bar{u})$ denotes (5.1), $\widetilde{F}(\bar{u})$ denotes (5.2); $w_u(a) := \frac{\pi_u(a)}{\pi^{\mathsf{ref}}(a)}$, $Z_u := \sum_a \frac{\pi_u^2(a)}{\pi^{\mathsf{ref}}(a)}$, and $\mu_u(a) := \frac{\pi_u^2(a)}{\pi^{\mathsf{ref}}(a)}/Z_u$, and $g := r - \widehat{r}$. First, the chain rule directly gives

$$\widetilde{F}'(u) = \eta \left\langle g \odot w_u, g \odot \frac{\mathrm{d}}{\mathrm{d}u}\pi_u \right\rangle. \tag{E.9}$$

We follow the same pipeline as that in the proof of Lemma 5.2 to calculate $\frac{\mathrm{d}}{\mathrm{d}u}\pi_u$. Specifically, we again substitute $\left[\nabla^2 h(\pi_u)\right]^{-1} = \eta \cdot \mathrm{diag}(\pi_u^2/\pi^{\mathsf{ref}})$ into Lemma F.7, and then substitute the result into (E.8) to obtain

$$\eta^{-1}\frac{\mathrm{d}}{\mathrm{d}u}\pi_u = Z_u\big(-\mu_u \odot g + \langle \mu_u, g \rangle\, \mu_u\big). \tag{E.10}$$

We set $X(a) := -w_u(a) \cdot g(a)$ and substitute (E.10) into (E.9) to obtain

$$\eta^{-2}\widetilde{F}'(u) = \mathbb{E}_{a \sim \pi^{\mathsf{ref}}} X(a)^3 - \Big(\mathbb{E}_{a \sim \pi_u} X(a)^2\Big)\mathbb{E}_{a \sim \pi_u} X(a)/Z_u. \tag{E.11}$$

Another application of the chain rule yields that $\widetilde{F}'(u) - F'(u)$

$$= \frac{\mathrm{d}}{\mathrm{d}u}\left(\frac{\eta Z_u}{2} \cdot \langle \mu_u(a), g(a)\rangle^2\right)$$
$$= -\frac{2\eta \mathbb{E}_{a\sim\pi_u} X(a)}{Z_u}\left\langle w_u, g \odot \frac{\mathrm{d}}{\mathrm{d}u}\pi_u\right\rangle - \frac{\eta}{Z_u^2}\left(\mathbb{E}_{a\sim\pi_u} X(a)\right)^2 \left\langle w_u, \frac{\mathrm{d}}{\mathrm{d}u}\pi_u\right\rangle. \tag{E.12}$$

Substituting (E.10) into (E.12) yields

$$\eta^{-2} \cdot (\text{E.12}) = \frac{\mathbb{E}_{a\sim\pi_u} X(a)}{Z_u} \cdot \left(\mathbb{E}_{a\sim\pi_u} X(a)^2 - \mathbb{E}_{a\sim\mu_u} X(a) \mathbb{E}_{a\sim\pi_u} X(a)\right)$$
$$- \left(\mathbb{E}_{a\sim\pi_u} X(a)/Z_u\right)^2 \cdot \left(Z_u \cdot \mathbb{E}_{a\sim\mu_u} X(a) - \langle w_u, \mu_u\rangle \cdot \mathbb{E}_{a\sim\pi_u} X(a)\right),$$

which, added to (E.11), yields

$$\eta^{-2} F'(u) = \mathbb{E}_{a\sim\pi^{\mathsf{ref}}} X(a)^3 - \frac{3}{Z_u}\mathbb{E}_{a\sim\pi_u} X(a)^2 \mathbb{E}_{a\sim\pi_u} X(a)$$
$$+ \frac{3}{Z_u}\left(\mathbb{E}_{a\sim\pi_u} X(a)\right)^2 \mathbb{E}_{a\sim\mu_u} X(a) - \left(\frac{\mathbb{E}_{a\sim\pi_u} X(a)}{Z_u}\right)^3 \cdot \langle w_u, \mu_u\rangle Z_u. \tag{E.13}$$

$\eta^2 \cdot$ (E.13) is exactly $-\eta^2 Z_u \cdot \mathbb{E}_{a\sim\mu_u}\left[w_u(a) \cdot \left(g(a) - \mathbb{E}_{a\sim\mu_u} g(a)\right)^3\right]$.

□

F Technical Tools

F.1 Differentiating maximal functions

We adapt simplified versions of Bertsekas et al. (2003, Proposition 4.5.1) and Milgrom and Segal (2002, Theorem 1) as follows.

Lemma F.1 (Danskin's Theorem, Bertsekas et al. 2003)**.** Let $\Pi \subset \mathbb{R}^m$ be compact, $G : \mathbb{R}^n \times \Pi \to \mathbb{R}$ be continuous, and $\forall \pi \in \Pi, r \mapsto G(r, \pi)$ be convex. For $g(\cdot) := \max_{\pi\in\Pi} G(\cdot, \pi)$, if $\operatorname{argmax}_{\pi\in\Pi} G(\widetilde{r}, \pi) = \{\widetilde{\pi}\}$ is a singleton and $r \mapsto G(r, \widetilde{\pi})$ is differentiable at $r = \widetilde{r}$, then $\nabla g(\widetilde{r})$ exists and $\nabla g(\widetilde{r}) = \nabla_1 G(\widetilde{r}, \widetilde{\pi})$, where ∇_1 is the gradient with respect to the first input.

Lemma F.2 (The Envelope Theorem, Milgrom and Segal 2002)**.** Let Π be a set and $J : \Pi \times [0, 1] \to \mathbb{R}$ be a function. Let $V(\cdot) := \sup_{\pi\in\Pi} J(\pi, \cdot)$. For any $u \in (0, 1)$ and $\widetilde{\pi} \in \Pi$, if $J(\widetilde{\pi}, u) = V(u)$, $V'(u)$ exists, and $\partial_u J(\widetilde{\pi}, u)$ exists; then $V'(u) = \partial_u J(\widetilde{\pi}, u)$.

F.2 Properties of convex conjugate over the simplex

The *convex conjugate*, whose supremum is taken over the entire Euclidean space, has been a well-established notion (Rockafellar, 1997); it has various handy properties for the analysis of bandit learning (see, e.g., Lattimore and Szepesvári (2020, Chapter 26)). We adapt some of them to the case where the supremum is taken over the probability simplex $\Delta(\mathcal{A})$ for the sake of rigorousness.

Lemma F.3. Suppose a closed convex function $h : \Delta(\mathcal{A}) \to \mathbb{R} \cup \{+\infty\}$ is continuously differentiable and strictly convex on $\operatorname{ri}(\Delta(\mathcal{A}))$, and $\operatorname{dom}(h) \supset \operatorname{ri}(\Delta(\mathcal{A}))$. Suppose for any $r \in \mathbb{R}^{\mathcal{A}}$, $\pi_r \in \operatorname{ri}(\Delta(\mathcal{A}))$ exists, where $\pi_r := \operatorname{argmax}_{\pi\in\Delta(\mathcal{A})} \langle r, \pi\rangle - h(\pi)$. Then $h^*(\cdot) := \max_{\pi\in\Delta(\mathcal{A})} \langle \cdot, \pi\rangle - h(\pi)$ is differentiable and $\nabla h^*(r) = \pi_r$.

Proof of Lemma F.3. First, note that π_r in the relative interior is uniquely well-defined by the existence assumption and strict convexity. Also, h induces another proper, closed, and convex function $\widetilde{h} : \mathbb{R}^{\mathcal{A}} \to (-\infty, +\infty]$ defined as

$$\widetilde{h}(\pi) := \begin{cases} h(\pi), & \pi \in \Delta(\mathcal{A}); \\ +\infty, & \text{otherwise}; \end{cases}$$

which is also bounded from below (Beck, 2017, Theorem 2.12). Here, the (unconstrained) conjugate of $\widetilde{h}$ evaluates to h^* because

$$\widetilde{h}^*(r) = \sup_{\pi\in\mathbb{R}^{\mathcal{A}}} \langle r, \pi\rangle - \widetilde{h}(\pi) = \max_{\pi\in\Delta(\mathcal{A})} \langle r, \pi\rangle - h(\pi) = h^*(r),$$

which, together with the compactness of $\Delta(\mathcal{A})$, implies that h^* is finite everywhere. Also, since $h^* = \widetilde{h}^*$ implies that h^* is convex (Beck, 2017, Theorem 4.3). Therefore, the (assumed) uniqueness of $\pi_r = \operatorname{argmax}_{\mu \in \mathbb{R}^{\mathcal{A}}} \langle r, \mu \rangle - \widetilde{h}(\mu)$ applying to the conjugate subgradient theorem (Beck, 2017, Theorem 4.20) yields that

$$\pi_r \in \partial \widetilde{h}^*(r) = \partial h^*(r) = \{\pi_r\}. \tag{F.1}$$

Since we already shown h^* is finite everywhere, applying Beck (2017, Theorem 3.33) to (F.1) yields that h^* is differentiable and $\nabla h^*(r) = \pi_r$. □

Lemma F.4. Under the conditions (and following the notation) in Lemma F.3, $\forall r, r' \in \mathbb{R}^{\mathcal{A}}$,

$$\mathsf{B}_{h^*}(r, r') = \mathsf{B}_h(\pi_{r'}, \pi_r),$$

where the Bregman divergence induced by h is defined as $\mathsf{B}_h(\pi_{r'}, \pi_r) := h(\pi_{r'}) - h(\pi_r) - \langle \nabla h(\pi_r), r' - r \rangle$ and similarly for h^*.

Proof. Direct calculates yield

$$\begin{aligned}
\text{LHS} &= h^*(r) - h^*(r') - \langle \nabla h^*(r'), r - r' \rangle \\
&= h^*(r) - h^*(r') - \langle \pi_{r'}, r - r' \rangle \\
&= \big(\langle r, \pi_r \rangle - h(\pi_r)\big) - \big(\langle r', \pi_{r'} \rangle - h(\pi_{r'})\big) - \langle \pi_{r'}, r \rangle + \langle \pi_{r'}, r' \rangle \\
&= h(\pi_{r'}) - h(\pi_r) - \langle r, \pi_{r'} - \pi_r \rangle,
\end{aligned} \tag{F.2}$$

where the second and third equalities follow from Lemma F.3. Because $\pi_r \in \operatorname{ri}(\Delta(\mathcal{A}))$ maximizes $\langle r, \cdot \rangle - h(\cdot)$ over $\Delta(\mathcal{A})$, the first-order optimality condition (of the Lagrangian) under the normalization constraint (i.e., probabilities sum to one) implies that $\exists \lambda \in \mathbb{R}$ such that

$$\forall a \in \mathcal{A}, r(a) - \frac{\partial h}{\partial \pi(a)}(\pi_r) = \lambda,$$

where we note that there is no Lagrangian multiplier (dual variable) for the non-negativity constraint due to complementary slackness and the assumption that $\pi_r \in \operatorname{ri}(\Delta(\mathcal{A}))$. Thus, $r - \nabla h(\pi_r) \in \operatorname{span}(\mathbf{1})$.[10] This relation implies $\langle r, \pi_{r'} - \pi_r \rangle = \langle \nabla h(\pi_r), \pi_{r'} - \pi_r \rangle$ and in turn implies

$$\text{(F.2)} = h(\pi_{r'}) - h(\pi_r) - \langle \nabla h(\pi_r), \pi_{r'} - \pi_r \rangle = \text{RHS}.$$

□

Remark F.5. To be conceptually minimal (without the introduction any convex-analytical definitions of *gradient* for functions defined on low-dimensional convex subsets), we understand the "continuous differentiability" for h in Lemmas F.3 and F.4 as: h admits a *continuation* $h_{\text{ext}} : \mathbb{R}_+^{\mathcal{A}} \to (-\infty, +\infty]$ such that the *Euclidean gradient* $\mu \mapsto \nabla h_{\text{ext}}(\mu)$ is continuous. The continuation h_{ext} is not necessarily unique, but this non-uniqueness does not hurt in the proof of either Lemma F.3 or Lemma F.4 because if there are two different valid continuations $h_{\text{ext}}^{(1)}$ and $h_{\text{ext}}^{(2)}$, they must agree on $\Delta(\mathcal{A})$, viz., their directional derivatives along any path strictly inside $\Delta(\mathcal{A})$ must be equal, i.e., $\forall \pi \in \operatorname{ri}(\Delta(\mathcal{A})), \langle \nabla h_{\text{ext}}^{(1)}(\pi), \mathbf{t} \rangle = \langle \nabla h_{\text{ext}}^{(2)}(\pi), \mathbf{t} \rangle$ as long as the tangent direction $\mathbf{t}$ satisfies $\langle \mathbf{t}, \mathbf{1} \rangle = 0$; in which case $\nabla h_{\text{ext}}^{(1)}(\pi) - \nabla h_{\text{ext}}^{(2)}(\pi) \in \operatorname{span}(\mathbf{1})$, and thus the first-order quantity $\langle \nabla h_{\text{ext}}(\pi), \pi' - \pi \rangle$ between any two probability vectors π' and π does NOT vary with the choice of continuation. We thus only write ∇h in place of ∇h_{ext} in these lemma statements and proofs.

Remark F.6. The $\operatorname{dom}(h) \supset \operatorname{ri}(\Delta(\mathcal{A}))$ in Lemma F.3 implies that we allow h to blow up on the boundary of $\Delta(\mathcal{A})$, which is the case for forward KL $h(\cdot) = \eta^{-1}\mathsf{KL}\left(\pi^{\mathsf{ref}} \| \cdot\right)$ even under Assumption 2.1. This is the reason that prevents us from using Danskin's theorem (Lemma F.1) to directly prove Lemma F.3.

Lemma F.7. Under the conditions (and following the notation) in Lemma F.3, if h is also twice continuously differentiable on $\operatorname{ri}(\Delta(\mathcal{A}))$ and there is a $r_\flat$ such that the Hessian matrix $\nabla^2 h(\pi_{r_\flat}) \succ \mathbf{0}_{A \times A}$, then h^* is also twice continuously differentiable on $\operatorname{ri}(\Delta(\mathcal{A}))$ and

$$\left[\nabla^2 h^*(r_\flat)\right] = \left[\nabla^2 h(\pi_{r_\flat})\right]^{-1} - \frac{\left[\nabla^2 h(\pi_{r_\flat})\right]^{-1} \mathbf{1}\mathbf{1}^\top \left[\nabla^2 h(\pi_{r_\flat})\right]^{-1}}{\mathbf{1}^\top \left[\nabla^2 h(\pi_{r_\flat})\right]^{-1} \mathbf{1}}, \tag{F.3}$$

[10] Recall that this $\operatorname{span}(\mathbf{1})$ is inherent, as elaborated in Remark F.5.

where $\mathbf{1} \in \mathbb{R}^{\mathcal{A}}$ is the all-one vector. Consequently, let $\mathbf{P} \coloneqq \mathbf{I} - A^{-1}\mathbf{1}\mathbf{1}^\top$,[11] (F.3) implies

$$\left[\nabla^2 h^*(r_\flat)\right] \cdot \left[\nabla^2 h(\pi_{r_\flat})\right] \cdot \mathbf{P} = \mathbf{P}.$$

Proof. Let $\mathbf{H} \coloneqq \left[\nabla^2 h(\pi_{r_\flat})\right]$. Then $\mathbf{H} \succ \mathbf{0}_{A\times A}$ by assumption. By Lemma F.3, h^* is a closed convex function that is differentiable everywhere and $\nabla h^*(r) = \pi_r$, and thus Lemma F.9 implies that $r \mapsto \nabla h^*(r)$ is continuous in r. Since $\pi_{r_\flat} \in \operatorname{ri}\left(\Delta(\mathcal{A})\right)$, there is a open neighborhood $\mathcal{U} \subset \operatorname{ri}\left(\Delta(\mathcal{A})\right)$ of $r_\flat$ such that $\forall r \in \mathcal{U}$, the non-negativity constraint is inactive due to complementary slackness. And thus the first-order optimality condition and the primal feasibility $\mathbf{1}^\top \pi = 0$ are necessary and sufficient for characterizing the maximizer π_r (unique by assumption) for every $r \in \mathcal{U}$, i.e., there exists a unique Lagrangian multiplier λ_r for each $r \in \mathcal{U}$ such that

$$\begin{cases} \nabla h(\pi_r) - r - \lambda_r \mathbf{1} & = \mathbf{0}, \\ \langle \mathbf{1}, \pi_r \rangle - 1 & = 0. \end{cases}$$

Let the mapping $\mathbf{f} : \mathbb{R}^A \times \mathbb{R} \times \mathbb{R}^A \to \mathbb{R}^{A+1}$ be

$$\mathbf{f}(\pi, \lambda, r) \coloneqq \begin{bmatrix} \nabla h(\pi) - r - \lambda \mathbf{1} \\ \langle \mathbf{1}, \pi \rangle - 1. \end{bmatrix}$$

Then $\mathbf{f}(\pi_{r_\flat}, \lambda_{r_\flat}, r_\flat) = \mathbf{0}_{A+1}$. The Jacobian matrix $\mathbf{J}_\flat$ of $\mathbf{f}$ with respect to the variables (π, λ) (i.e., excluding the explicit dependency of r) evaluated at $(\pi_{r_\flat}, \lambda_{r_\flat}, r_\flat)$ is thus

$$\mathbf{J}_\flat = \begin{bmatrix} \mathbf{H} & -\mathbf{1}_A \\ \mathbf{1}_A^\top & 0 \end{bmatrix}.$$

Note that for any $[\mathbf{a}^\top \ b]^\top \in \mathbb{R}^{A+1}$,

$$\begin{bmatrix} \mathbf{a}^\top & b \end{bmatrix} \begin{bmatrix} \mathbf{H} & -\mathbf{1}_A \\ \mathbf{1}_A^\top & 0 \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ b \end{bmatrix} = \mathbf{a}^\top \mathbf{H} \mathbf{a}.$$

Thus, the positive definiteness assumption on $\mathbf{H}$ implies $\mathbf{J}_\flat \succ \mathbf{0}$, and thus $\mathbf{J}_\flat$ is non-singular; applying the implicit function theorem on which gives that $r \mapsto (\pi_r, \lambda_r)$ is continuously differentiable in an open neighborhood $\widetilde{\mathcal{U}} \subset \operatorname{ri}\left(\Delta(\mathcal{A})\right)$ of $r_\flat$. Recall that $\nabla h^*(r) = \pi_r$, we further obtain that $r \mapsto \nabla h^*(r)$ is continuously differentiable locally, viz., h^* is twice continuously differentiable at $r_\flat$. Therefore, it is legal to calculate $\nabla^2 h^*(r_\flat)$. Recall that for $r \in \mathcal{U}$, $\mathbf{f}(\pi_r, \lambda_r, r) \equiv \mathbf{0}_{A+1}$, we use the chain rule for taking the derivative on both sides of $\mathbf{f}(\pi_r, \lambda_r, r) = \mathbf{0}_{A+1}$ with respect to r and evaluate the derivative at $r = r_\flat$ to obtain

$$\begin{bmatrix} \mathbf{H} & -\mathbf{1} \\ \mathbf{1}^\top & 0 \end{bmatrix} \begin{bmatrix} \nabla_r \pi_{r_\flat} \\ \nabla_r \lambda_{r_\flat} \end{bmatrix} + \begin{bmatrix} -\mathbf{I}_{A\times A} \\ \mathbf{0}_A^\top \end{bmatrix} = \mathbf{0}_{(A+1)\times A},$$

which implies $\mathbf{1}^\top \nabla_r \pi_{r_\flat} = \mathbf{0}_A^\top$ and $\mathbf{H}\nabla_r \pi_{r_\flat} - \mathbf{1}\nabla_r \lambda_{r_\flat} = \mathbf{I}_{A\times A}$.[12] Thus, $\nabla_r \pi_{r_\flat} = \mathbf{H}^{-1} + \mathbf{H}^{-1}\mathbf{1}\nabla_r \lambda_{r_\flat}$, substituting which into $\mathbf{1}^\top \nabla_r \pi_{r_\flat} = \mathbf{0}^\top$ yields

$$\mathbf{0}_A = \mathbf{1}^\top \left(\mathbf{H}^{-1} + \mathbf{H}^{-1}\mathbf{1}\nabla_r \lambda_{r_\flat}\right) \Longrightarrow \nabla_r \lambda_{r_\flat} = -\frac{\mathbf{1}^\top \mathbf{H}^{-1}}{\mathbf{1}^\top \mathbf{H}^{-1}\mathbf{1}}.$$

Finally,

$$\nabla^2 h^*(r_\flat) = \nabla_r \pi_{r_\flat} = \mathbf{H}^{-1} + \mathbf{H}^{-1}\mathbf{1}\nabla_r \lambda_{r_\flat} = \mathbf{H}^{-1} - \frac{\mathbf{H}^{-1}\mathbf{1}\mathbf{1}^\top \mathbf{H}^{-1}}{\mathbf{1}^\top \mathbf{H}^{-1}\mathbf{1}}.$$

□

Remark F.8. Lemma F.7 is a fine-grained version of Zhao et al. (2026, Lemma E.4), which is used in the current paper only to demonstrate why the mean-value-type argument fails to achieve single-policy concentrability under forward-KL regularization, but can be of independent interest. The counterpart of Lemma F.4 for *unconstrained* convex conjugate has been established before (see, e.g., Lattimore and Szepesvári (2020, Lemma 26.4)).

[11] $\mathbf{P}$ is the the orthogonal projection matrix onto the tangent space of $\Delta(\mathcal{A})$ at any $\pi \in \Delta(\mathcal{A})$.

[12] In this equation we use the row-layout for notational simplicity, i.e., $\nabla_r \lambda_{r_\flat} \in \mathbb{R}^{1\times A}$.

F.3 Miscellanies

Lemma F.9 (Rockafellar 1997, Corollary 25.5.1)**.** Suppose h is a proper convex function that is finite and differentiable on an open convex set C, then f is continuously differentiable on C.

Fact F.10. $\phi(y) = -y - \log(1-y)$ is convex on $(-\infty, 1)$, whose (constrained) convex conjugate is

$$\phi^*(z) := \sup_{y<1} yz - \phi(y) = \begin{cases} z - \log(1+z), & z > -1; \\ +\infty, & z \le -1. \end{cases}$$

And the Fenchel-Young inequality $yz \le \phi(y) + \phi^*(z)$ holds for any $y < 1$ and $z \in \mathbb{R}$.

Proof. Direct calculation gives that $\phi'(y) = -1 + (1-y)^{-1}$, thus $\phi''(y) = (1-y)^{-2} > 0$ for all $y < 1$, concluding that ϕ is convex. Now fix any $z > -1$, let $\phi_z(y) = yz - \phi(y)$ for $y < 1$, then

$$\phi_z'(y) = 1 + z - \frac{1}{1-y}.$$

Therefore ϕ_z takes maximum at $z(1+z)^{-1}$. Plugging in this value of y gives that $\phi^*(z) = z - \log(1+z)$. Finally, for any $y < 1$ and $z \ge 0$, $\phi^*(z) \ge yz - \phi(y)$, resulting in $yz \le \phi(y) + \phi^*(z)$. □

Lemma F.11 (Lemma A.1, Xie et al. 2021b)**.** Suppose $N \sim \mathsf{Bin}(n, p)$ where $n \ge 1$ and $p > 0$, then

$$\mathbb{P}\Big[N \ge \frac{np}{2}\Big] \ge 1 - \exp(-np/8).$$

Equivalently, with probability at least $1 - \delta$,

$$\frac{p}{N \vee 1} \le \frac{8\log(1/\delta)}{n}.$$

Lemma F.12 (Azuma-Hoeffding's inequality)**.** Let $Z_0, Z_1, \cdots, Z_n$ be a martingale sequence of random variables such that for all i, there exists a constant c_i such that $|Z_i - Z_{i-1}| < c_i$, then

$$\mathbb{P}[Z_n - Z_0 \ge t] \le \exp\Big(-\frac{t^2}{2\sum_{i=1}^n c_i^2}\Big).$$

In particular, if $\sup_i c_i \le M$, then $\forall \delta \in (0,1)$, the following inequality holds with probability at least $1 - \delta$:

$$Z_n - Z_0 \le M\sqrt{2n\log(1/\delta)}.$$

Lemma F.13 (Le Cam's two-point method, Le Cam 1973; Yu 1997)**.** Let $\mathcal{R}$ be the set of instances, Π be the set of estimators, and $L : \Pi \times \mathcal{R} \to \mathbb{R}_+$ be a loss function. For $\acute{r}, \grave{r} \in \mathcal{R}$, suppose $\exists c > 0$ such that

$$\inf_{\pi \in \Pi} L(\pi, \acute{r}) + L(\pi, \grave{r}) \ge c,$$

then

$$\inf_{\pi \in \Pi} \sup_{r \in \mathcal{R}} \mathbb{E}_{\mathcal{D} \sim P_r} L\big(\pi(\mathcal{D}), r\big) \ge \frac{c}{4} \cdot \exp\big(-\mathsf{KL}\left(P_{\acute{r}} \| P_{\grave{r}}\right)\big),$$

where the trajectory distribution of π interacting with instance r is denoted by P_r.

Lemma F.14 (Assouad's Lemma, Yu 1997; Chen et al. 2024)**.** 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

$$\theta \sim_j \theta' \Rightarrow \inf_{\pi \in \Pi} L_j(\pi, r_\theta) + L_j(\pi, r_{\theta'}) \ge c \tag{F.4}$$

for some $c > 0$, then

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

where the trajectory distribution of π interacting with instance $r \in \mathcal{R}$ is denoted by P_r.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We clearly state the problem of forward-KL-regularized contextual bandits settled in this paper and the corresponding sample complexity results, at least informally, together with the implications of our bounds. We also discuss in the introduction the role of our work in the line of efforts towards finally figuring out sample complexity of decision making problem with regularized objectives.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We detail in the last paragraph of our conclusion that the necessity of single-policy concentrability for learning forward-KL-regularized contextual bandits is still open.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: Every theorem and lemma in this paper has a complete and rigorous proof, either in the main text or deferred to appendix.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [N/A]

Justification: This paper is purely theoretical by the deadline of submission.

Guidelines:

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

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

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

   Answer: [N/A]

   Justification: This paper is purely theoretical by the deadline of submission.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [N/A]

   Justification: This paper is purely theoretical by the deadline of submission.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [N/A]

   Justification: This paper is purely theoretical by the deadline of submission.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

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

8. **Experiments compute resources**

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

   Answer: [N/A]

   Justification: This paper is purely theoretical by the deadline of submission.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: This theoretical research is fully in line with the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [No]

    Justification: The purpose of this research is to advance the theory of reinforcement learning. It may have downstream impacts, none of which is worth a specific discussion.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: We believe that our theoretical analysis poses no such risk.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [N/A]

    Justification: We do not use any existing assets.

    Guidelines:

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

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

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

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

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

    Answer: [N/A]

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

    Guidelines:

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

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

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

    Answer: [N/A]

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

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: We only use LLM for writing purpose and the core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

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