On the Optimal Sample Complexity of Offline Multi-Armed Bandits with KL Regularization

Kaixuan Ji [*1] **Qiwei Di** [*1] **Heyang Zhao** [1] **Qingyue Zhao** [1] **Quanquan Gu** [1]

Abstract

Kullback-Leibler (KL) regularization is widely used in offline decision-making and offers several benefits, motivating recent work on the sample complexity of offline learning with respect to *KL-regularized performance metrics*. Nevertheless, the exact sample complexity of KL-regularized offline learning remains far from fully characterized. In this paper, we study this question in the setting of multi-armed bandits (MABs). We provide a sharp analysis of KL-PCB (Zhao et al., 2026), showing that it achieves a sample complexity of $\widetilde{O}(\eta SAC^{\pi^*}/\epsilon)$ under large regularization $\eta = \widetilde{O}(\epsilon^{-1})$, and $\widetilde{O}(SAC^{\pi^*}/\epsilon^2)$ under small regularization $\eta = \widetilde{\Omega}(\epsilon^{-1})$, where S is the number of contexts, A is the number of arms, C^{π^*} is the policy coverage coefficient at the optimal policy π^*, ϵ is the desired suboptimality, and $\widetilde{O}$ and $\widetilde{\Omega}$ hide all poly-logarithmic factors. We further provide a pair of sharper sample complexity lower bounds that match the upper bounds across the entire range of regularization strengths. Overall, our results provide a nearly complete characterization of offline multi-armed bandits with KL regularization.

1. Introduction

Offline reinforcement learning (RL) algorithms that learn from pre-collected data without interactive data collection have recently become appealing in both embodied settings (Levine et al., 2018; 2020) and language modeling (Rafailov et al., 2023; Ethayarajh et al., 2024; Meng et al., 2024; Rafailov et al., 2024) due to their computational and memory efficiency, ease of implementation, and strong safety guarantees, especially when the interaction is risky. In this paradigm, algorithms typically employ divergence constraints that keep the learned policy close to a reference policy π^{ref} (Wu et al., 2019; Kumar et al., 2020; Rafailov et al., 2023). Among these, the reverse Kullback–Leibler (KL) divergence regularizer $\mathsf{KL}\left(\widehat{\pi}\|\pi^{\mathsf{ref}}\right)$ has become a popular and effective choice, especially for fine-tuning large language models (Rafailov et al., 2023; Ethayarajh et al., 2024; Meng et al., 2024; Lai et al., 2024; Rafailov et al., 2024).

The prevalence of KL regularization has motivated a line of statistical analysis of offline learning with respect to the *KL-regularized performance metric* (Zhao et al., 2025a; 2026; Wu et al., 2025a). Zhao et al. (2025a) provide the first pair of n^{-1}-type upper and lower bounds of the suboptimality relative to the optimal *KL-regularized* policy. However, the upper bound requires a strict uniform coverage condition, while the lower bound does not characterize the dependence on concentrability. Zhao et al. (2026) take a step further by showing that *single-policy concentrability* is necessary and sufficient for achieving the n^{-1}-type fast convergence rate. Nevertheless, there is a discrepancy in the notion of concentrability between the upper and lower bounds, which implies looseness in the worst case (e.g., in tabular settings). Moreover, both Zhao et al. (2025a; 2026) achieve the n^{-1}-type rates only when the regularization is sufficiently strong. More recently, Wu et al. (2025a) derive algorithms for this setting without pessimism, at the cost of an exponentially worse dependence on the inverse regularization intensity. Therefore, to the best of our knowledge, the following question remains open even for the simplest offline-learning models.

What is the sample complexity of offline decision making with KL regularization?

In this paper, we provide a nearly comprehensive answer to this question for MABs[1] (up to logarithmic factors). We dichotomize the problem into two regimes based on the strength of KL regularization. In each regime, we show

[1]Department of Computer Science, University of California, Los Angeles, CA 90095, USA. Correspondence to: Quanquan Gu <qgu@cs.ucla.edu>.

ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

[1]We consider a slightly generalized setting similar to the problem described in Rashidinejad et al. 2021, Section 4, which allows the presence of different contexts.

that the sample complexity of KL-PCB[2] (Zhao et al., 2026) matches the corresponding statistical limit, thereby providing a near-complete characterization of offline learning for multi-armed bandits with KL regularization. Our main contributions are summarized as follows:

- We provide a sharp analysis of KL-PCB (Zhao et al., 2026) in the multi-armed bandit setting. In particular, for KL-PCB, we provide a $\widetilde{O}(\eta SAC^{\pi^*}/\epsilon)$ sample complexity upper bound under large regularization $\eta = \widetilde{O}(\epsilon^{-1})$, and a $\widetilde{O}(SAC^{\pi^*}/\epsilon^2)$ sample complexity upper bound in the small regularization regime $\eta = \widetilde{\Omega}(\epsilon^{-1})$. Along with the sharp analysis, we further establish two lower bounds that match the upper bounds up to logarithmic factors in both regimes, indicating that KL-PCB is near-optimal.

- We construct two distinct types of hard instances designed to exploit the unique structure of KL regularization under different regimes. Specifically, in the small regularization regime, we show that its statistical limit is closely related to the hardness of learning MABs with multiple optima (Degenne & Koolen, 2019; Zhu & Nowak, 2020; De Heide et al., 2021) in the offline setting. In the presence of multiple optima, the direct application of the standard lower bound techniques: Le Cam's method (Assouad's method) and Fano's methods, is not sharp enough. We address this with a novel pairing-and-counting technique which may be of independent interest beyond our setting.

- As a by-product, we identify the problem of learning offline MABs with multiple optima. Restricting to a uniform behavior policy, we provide an $\widetilde{\Omega}(K^2/(A\epsilon^2))$ sample complexity lower bound, where A is the number of arms and K is the number of suboptimal arms. We further provide the sample complexity upper bound of a minimalist algorithm, empirical best-arm selection, for this setting, which certifies that the $\widetilde{\Omega}(K^2/(A\epsilon^2))$ sample complexity lower bound is tight up to logarithmic factors.

For ease of comparison, we summarize our results on KL-regularized MABs, together with representative results from related works, in Table 1.

Notation. The sets $\mathcal{S}$ and $\mathcal{A}$ are assumed to be finite throughout the paper. For nonnegative sequences $\{x_n\}$ and $\{y_n\}$, we write $x_n = O(y_n)$ if $\limsup_{n\to\infty} x_n/y_n < \infty$, $x_n = o(y_n)$ if $\limsup_{n\to\infty} x_n/y_n = 0$, $y_n = \Omega(x_n)$ (interchangeably written as $y_n \gtrsim x_n$) if $x_n = O(y_n)$, and $y_n = \Theta(x_n)$ if $x_n = O(y_n)$ and $x_n = \Omega(y_n)$. We further employ $\widetilde{O}(\cdot), \widetilde{\Omega}(\cdot)$, and $\widetilde{\Theta}(\cdot)$ to hide polylog factors. For finite $\mathcal{X}$ and $\mathcal{Y}$, we denote the family of probability kernels from $\mathcal{X}$ to $\mathcal{Y}$ by $\Delta(\mathcal{Y}|\mathcal{X})$. For a pair of probability measures $P \ll Q$ on the same space, we use $\mathsf{KL}(P\|Q) := \int \log(\mathrm{d}P/\mathrm{d}Q)\,\mathrm{d}P$ to denote their KL divergence. We denote by $\mathsf{Unif}(\mathcal{X})$ the uniform distribution on finite set $\mathcal{X}$. For some $\pi \in \Delta(\mathcal{X})$ with full support and $\mathcal{Y} \subseteq \mathcal{X}$, we use $\pi|_{\mathcal{Y}} \in \Delta(\mathcal{Y})$ to denote the distribution such that $(\pi|_{\mathcal{Y}})(y) \propto \pi(y)$. We denote $[N] := \{1, \cdots, N\}$ for any positive integer N. Boldfaced lowercase letters are reserved for vectors. For $\mathbf{x}, \mathbf{y} \in \mathcal{X}^n$, we use $d_H(\mathbf{x}, \mathbf{y})$ to denote their Hamming distance. For $x, y \in \mathbb{R}$, we denote $x \vee y = \max\{x, y\}$. We use $\mathsf{Bern}(p)$ to denote the Bernoulli distribution with mean p, and $\mathsf{Bin}(n, p)$ for the binomial distribution with n trials and success rate p. For probability measure P, we use $\mathrm{supp}(P)$ to denote the support of P.

2. Related Work

KL-Regularized Bandit and RL. Several recent studies (Xie et al., 2025; Xiong et al., 2024; Zhao et al., 2025a; Foster et al., 2025) investigated the sample complexity of KL-regularized objective, which provably enjoys an $\mathcal{O}(\epsilon^{-1})$ rate. This fast rate was first demonstrated by Tiapkin et al. (2023) in the setting of pure-exploration for maximum-entropy RL. For the setting of online regret minimization, Zhao et al. (2025b) obtained a $\widetilde{O}(\eta d_{\mathcal{R}} \log N_{\mathcal{R}} \log T)$ regret upper bound under function approximation, which was then improved to the near-optimal $\widetilde{O}(\eta A \log^2 T \wedge \sqrt{AT})$ rate when specialized to multi-armed bandits (Ji et al., 2026). In the pure offline setting, Zhao et al. (2025a) established the optimal sample complexity $\mathcal{O}(\epsilon^{-1})$, albeit under the strong assumption that the behavior policy π^{ref} provides uniform coverage over the entire function class for all policies. This requirement was subsequently removed by Zhao et al. (2026) via pessimism, which yields an $\widetilde{O}(\eta D^2_{\pi^*} \log N_{\mathcal{R}} \epsilon^{-1})$ upper bound and an $\Omega(\eta C^{\pi^*} \log N_{\mathcal{R}} \epsilon^{-1})$ lower bound. Wu et al. (2025a) shows that one can also achieve $\widetilde{O}(\eta \exp(\eta) \log N_{\mathcal{R}} \epsilon^{-1})$ upper bound with greedy sampling. Similar fast rates have also been established in privacy-sensitive (Wu et al., 2025b; Weng et al., 2025) and competitive multi-agent (Nayak et al., 2025; Zhang et al., 2026) settings, as well as under hybrid query protocols with linear function approximation (Foster et al., 2025)

3. Preliminaries

In this section, we introduce the multi-armed (contextual) bandit with a KL-regularized objective, defined by a tuple $(\mathcal{S}, \mathcal{A}, r^*, \eta, \pi^{\mathsf{ref}})$. Here $\mathcal{S}$ is the context space, $\mathcal{A}$ is the action space and $r^* : \mathcal{S} \times \mathcal{A} \to [0, 1]$ is the reward function. In the offline setting, the agent only has access to an i.i.d. dataset $\mathcal{D} = \{(s_i, a_i, r_i)\}_{i=1}^n$, where s_i is a context sampled from a fixed distribution $\rho \in \Delta(\mathcal{S})$, $a_i \in \mathcal{A}$ is

[2]Despite a minor adaptation to the multi-armed setting, our algorithm largely follows the original KL-PCB in Zhao et al. (2026); thus, we still refer to our algorithm as KL-PCB.

Table 1. Comparison of sample complexity upper and lower bounds for offline KL-regularized bandits. In this table, n is the sample size, η is the regularization parameter and C^{π^*} is the (density ratio-based) single-policy coverage coefficient. For the multi-armed setting, S is the size of context set and A is the size of action set. For the function approximation setting, D^2 is the (D^2-based) all-policy concentrability, $D^2_{\pi^*}$ is the (D^2-based) single-policy concentrability at the optimal policy π^* and $N_{\mathcal{R}}$ is the covering number of the function class. When reducing to the multi-armed setting, $D^2 = D^2_{\pi^*} = \Theta(SAC^{\pi^*})$ in the worst case, $\log N_{\mathcal{R}} = \Theta(SA)$ for the upper bound and should be interpreted as $\Theta(S)$ when it appears in the lower bounds (see Remark 5.3). $\widetilde{O}(\cdot)$ and $\widetilde{\Omega}(\cdot)$ hide logarithmic factors.

Type	Algorithm	Setting	Large Regularization $\eta^2 = \widetilde{O}(n(SAC^{\pi^*})^{-1})$	Small Regularization $\eta^2 = \widetilde{\Omega}(n(SAC^{\pi^*})^{-1})$
Upper Bound	TMPS (Zhao et al., 2025a)	Function Approximation	$\widetilde{O}\left(\frac{\eta D^2 \log N_{\mathcal{R}}}{n}\right)$	N/A
	KL-PCB (Zhao et al., 2026)	Function Approximation	$\widetilde{O}\left(\frac{\eta D^2_{\pi^*} \log N_{\mathcal{R}}}{n}\right)$	N/A
	Greedy Sampling (Wu et al., 2025a)	Preference Feedback	$\widetilde{O}\left(\frac{\eta \exp(\eta) \log N_{\mathcal{R}}}{n}\right)$	N/A
	KL-PCB (This Work)	Multi-armed	$\widetilde{O}\left(\frac{\eta SAC^{\pi^*}}{n}\right)$	$\widetilde{O}\left(\sqrt{\frac{SAC^{\pi^*}}{n}}\right)$
Lower Bound	Zhao et al. (2025a)	Function Approximation	$\Omega\left(\frac{\eta \log N_{\mathcal{R}}}{n}\right)$	N/A
	Zhao et al. (2026)	Function Approximation	$\Omega\left(\frac{\eta C^{\pi^*} \log N_{\mathcal{R}}}{n}\right)$	$\Omega\left(\sqrt{\frac{C^{\pi^*} \log N_{\mathcal{R}}}{n}}\right)$
	This Work	Multi-armed	$\widetilde{\Omega}\left(\frac{\eta SAC^{\pi^*}}{n}\right)$	$\widetilde{\Omega}\left(\sqrt{\frac{SAC^{\pi^*}}{n}}\right)$

the action sampled from a *behavior policy*, and r_i is the observed reward given by $r_i = r^*(s_i, a_i) + \varepsilon_i$. We assume that ε_i is 1-sub-Gaussian (Lattimore & Szepesvári, 2020, Definition 5.2). The learner's goal is to output a policy $\pi \in \Delta(\mathcal{A}|\mathcal{S})$ that maximizes the following KL-regularized objective

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

where π^{ref} is a known reference policy and η^{-1} is proportional to the regularization intensity. For simplicity, we assume that π^{ref} is also the behavior policy that generates the dataset $\mathcal{D}$. This type of "behavior regularization" has also been studied in previous works (Zhan et al., 2022; Zhao et al., 2026). The unique optimal policy π^*, defined by $\pi^* := \mathrm{argmax}_{\pi\in\Delta(\mathcal{A}|\mathcal{S})} J(\pi)$, admits the following closed form (see, e.g., Zhang 2023, Proposition 7.16):

$$\pi^*(\cdot|s) \propto \pi^{\mathsf{ref}}(\cdot|s)\exp\left(\eta \cdot r^*(s,\cdot)\right), \forall s \in \mathcal{S}. \tag{3.2}$$

For any policy π, we define the *suboptimality gap* as

$$\mathrm{SubOpt}(\pi) := J(\pi^*) - J(\pi). \tag{3.3}$$

A policy π is said to be ϵ-optimal if $\mathrm{SubOpt}(\pi) \le \epsilon$. The goal of offline learning is to output an ϵ-optimal policy based on the dataset $\mathcal{D}$.

Concentrability. In offline RL, a standard assumption concerns the coverage of the behavior policy, which serves as a measure of the dataset's quality. Specifically, it assesses whether the dataset provides sufficient support for distributions induced by other comparator policies.

Definition 3.1 (Single-Policy Concentrability)**.** Given a reference policy π^{ref}, we define the concentrability of the optimal policy π^* with respect to π^{ref} as

$$C^{\pi^*} := \sup_{s\in\mathcal{S}, a\in\mathcal{A}} \frac{\pi^*(a|s)}{\pi^{\mathsf{ref}}(a|s)}.$$

Compared with the more stringent *uniform coverage* assumptions (Sidford et al., 2018; Agarwal et al., 2020; Wang et al., 2021; Di et al., 2024; Zhao et al., 2025a), which require the data distribution of *any* policy to be sufficiently covered by the dataset, the *single-policy* concentrability framework (Rashidinejad et al., 2021; Uehara & Sun, 2021; Xie et al., 2021a; Rashidinejad et al., 2022; Cheng et al., 2022; Ozdaglar et al., 2023; Zhao et al., 2026) considered here relaxes the coverage constraint to only the distribution induced by the optimal policy, and thus typically corresponds to a much smaller concentrability coefficient. As indicated by Zhao et al. (2026), this single-policy concentrability is both necessary and sufficient for learning offline KL-regularized bandits.

4. Algorithm and Sample Complexity Analysis

In this section, we introduce a variant of KL-PCB (Zhao et al., 2026), an algorithm for learning offline MABs with

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

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 $\bar{r}(s,a) \leftarrow 0$, penalty $b(s,a) \leftarrow 1$
  else
    Compute the empirical reward $\bar{r}(s,a) \leftarrow \frac{1}{N(s,a)} \sum_{i=1}^{n} r_i \mathbb{1}\{(s_i,a_i) = (s,a)\}$
    Compute the penalty to be $b(s,a) \leftarrow \sqrt{\frac{4\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{N(s,a)}}$, let $\widehat{r}(s,a) \leftarrow \bar{r}(s,a) - b(s,a)$
  end if
end for
Ensure: $\widehat{\pi}(a|s) \propto \pi^{\mathsf{ref}}(a|s) \exp\big(\eta \cdot \widehat{r}(s,a)\big)$

the KL-regularized objective. The algorithm is summarized in Algorithm 1. In particular, for each $(s,a) \in \mathcal{S} \times \mathcal{A}$, we first compute the empirical average reward $\bar{r}(s,a)$ as an estimate of the ground-truth reward function $r^*(s,a)$. Based on $\bar{r}(s,a)$, we then construct a pessimistic estimation of r^*, which enables the algorithm to get rid of all-policy coverage as shown in Zhao et al. (2026). Compared to Zhao et al. (2026), we apply the standard pessimism in MABs literature (Rashidinejad et al., 2021; Xie et al., 2021b) given by

$$b(s,a) = \sqrt{\frac{4\log(2SA/\delta)}{N(s,a) \vee 1}},$$

while Zhao et al. (2026) uses reward function approximation, where the constructed pessimistic penalty has to encompass the entire function class and hence over-penalizes the reward when specialized to multi-armed cases.

We then obtain our pessimistic estimation of $r^*(s,a)$, $\widehat{r}(s,a) = \bar{r}(s,a) - b(s,a)$. Specifically, the following results show that $\widehat{r}$ is a pessimistic estimate with high probability and that the number of samples $N(s,a)$ does not deviate too much from its expectation $\rho(s)\pi^{\mathsf{ref}}(a|s)$.

Lemma 4.1. *Given $\delta > 0$, let $\mathcal{E}_1(\delta)$ denote the event that the estimation is indeed pessimistic*

$$\mathcal{E}_1(\delta) := \Big\{ \big|\bar{r}(s,a) - r^*(s,a)\big| \le b(s,a),\ \forall\,(s,a) \in \mathcal{S} \times \mathcal{A} \Big\}. \tag{4.1}$$

We further use $\mathcal{E}_2(\delta)$ to denote the event under which $N(s,a)$ does not deviate too much from the expectation, that is

$$\mathcal{E}_2(\delta) := \left\{ \frac{1}{N(s,a) \vee 1} \le \Gamma(s,a),\ \forall\,(s,a) \in \mathcal{S} \times \mathcal{A} \right\}.$$

where

$$\Gamma(s,a) = \frac{8\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{n\rho(s)\pi^{\mathsf{ref}}(a|s)}.$$

Then the event $\mathcal{E}_1(\delta) \cap \mathcal{E}_2(\delta)$ holds with probability at least $1-\delta$.

After obtaining the pessimistic estimate $\widehat{r}$, KL-PCB then outputs the policy using the closed-form solution $\widehat{\pi}(\cdot|s) \propto \pi^{\mathsf{ref}}(\cdot|s)\exp(\eta\widehat{r}(s,\cdot))$ for all $s \in \mathcal{S}$, based on the pessimistic reward $\widehat{r}$. The upper bound of the suboptimality of Algorithm 1 is given by the following theorem.

Theorem 4.2. *With probability at least $1-\delta$, the suboptimality gap of the output of Algorithm 1, depending on the regularization level, can be bounded as follows, depending on the regularization level:*

- *For large regularization $\eta^2 = \widetilde{O}\big(n(SAC^{\pi^*})^{-1}\big)$, the output policy $\widehat{\pi}$ obeys*

$$\mathrm{SubOpt}(\widehat{\pi}) = \widetilde{O}\big(\eta SAC^{\pi^*} n^{-1}\big).$$

- *For small regularization $\eta^2 = \widetilde{\Omega}\big(n(SAC^{\pi^*})^{-1}\big)$, the output policy $\widehat{\pi}$ obeys*

$$\mathrm{SubOpt}(\widehat{\pi}) = \widetilde{O}\big(\sqrt{SAC^{\pi^*} n^{-1}}\big).$$

Theorem 4.2 exhibits a "phase transition" as regularization varies. When η is small, the curvature introduced by the regularization term determines the problem, leading to an $\mathcal{O}(\epsilon^{-1})$ rate, matching the results in previous literature (Zhao et al., 2025a; 2026; Foster et al., 2025). On the other hand, when η is large, the reward estimation error dominates the suboptimality gap; therefore, the problem is similar to its counterpart with the standard objective. This yields a rate of $\mathcal{O}(\epsilon^{-2})$, recovering the rate under standard objective (Rashidinejad et al., 2021).

Remark 4.3. Previously, Zhao et al. (2026) obtained a sample complexity of $\eta D^2_{\pi^*}\epsilon^{-1}\log N_{\mathcal{R}}(\epsilon)$ under function approximation, where $D^2_{\pi^*}$ is the D^2-type single policy concentrability and $N_{\mathcal{R}}$ is the covering number of the function class. When restricted to the multi-armed contextual

bandit setting, $D^2_{\pi^*} = \Theta(SAC^{\pi^*})$ in the worst case[3] and $\log \mathcal{N}_{\mathcal{G}}(\epsilon) = \widetilde{\Theta}(SA)$, which gives an $\widetilde{O}(\eta S^2 A^2 C^{\pi^*} \epsilon^{-1})$ sample complexity. Compared to their sample complexity, it can be seen that the sample complexity obtained by our version of KL-PCB is strictly better.

5. Hardness Results

Following Rashidinejad et al. (2021), we define the instance class with bounded concentrability. For any instance $(r, \eta, \pi^{\mathsf{ref}})$, let π^*_r denote the corresponding optimal policy (3.2). The class of instances with concentrability at most C^* is

$$\mathrm{MAB}(C^*) := \left\{ (r, \eta, \pi^{\mathsf{ref}}) \middle| \sup_{s,a} \frac{\pi^*_r(a|s)}{\pi^{\mathsf{ref}}(a|s)} \le C^* \right\}.$$

Our goal is to characterize the minimax risk over this class:

$$\inf_{\mathtt{Alg}} \sup_{(r,\eta,\pi^{\mathsf{ref}}) \in \mathrm{MAB}(C^*)} \mathbb{E}_{\mathcal{D}} \big[\mathrm{SubOpt}\big(\mathtt{Alg}(\mathcal{D})\big) \big],$$

where the expectation is over the randomness of the dataset $\mathcal{D}$ generated under behavior policy π^{ref}. Our first theorem establishes the suboptimality gap lower bound in the large regularization regime $n = \widetilde{\Omega}(\eta^2 SAC^{\pi^*})$.

Theorem 5.1. *For any $S \ge 1$, $A \ge 3$, $\eta > 4 \log 2$, $C^* \in (2, \exp(\eta/4)]$, when the regularization level is large, i.e., $\eta^2 = \widetilde{O}\big(n(SAC^{\pi^*})^{-1}\big)$, one has*

$$\inf_{\mathtt{Alg}} \sup_{\mathrm{MAB}(C^*)} \mathbb{E}_{\mathcal{D}} \big[\mathrm{SubOpt}\big(\mathtt{Alg}(\mathcal{D})\big) \big] \gtrsim \frac{\eta S A C^*}{n \log A}.$$

The following theorem provides a suboptimality lower bound for the small regularization regime $n = \widetilde{O}(\eta^2 SAC^{\pi^*})$. Together with Theorem 5.1, these two theorems provide a comprehensive characterization of the statistical limit of offline learning for multi-armed contextual bandits with KL regularization.

Theorem 5.2. *For any $S > 1$, $A > 3$, $C^* \in (4, \exp(\eta/2)]$, when the regularization level is small, i.e., $\eta^2 = \widetilde{\Omega}\big(n(SAC^{\pi^*})^{-1}\big)$, one has*

$$\inf_{\mathtt{Alg}} \sup_{\mathrm{MAB}(C^*)} \mathbb{E}_{\mathcal{D}} \big[\mathrm{SubOpt}\big(\mathtt{Alg}(\mathcal{D})\big) \big] \gtrsim \sqrt{\frac{SAC^*}{n \log A}}.$$

Theorem 5.2 shows that, when $n = \widetilde{O}(\eta^2 SAC^{\pi^*})$, the suboptimality gap of any algorithm is lower bounded by $\widetilde{\Omega}(\sqrt{SAC^{\pi^*} n^{-1}})$. On the other hand, when $n = \widetilde{\Omega}(\eta^2 SAC^{\pi^*})$, Theorem 5.1 shows that for any algorithm, its suboptimality is lower bounded by $\widetilde{\Omega}(SAC^{\pi^*} n^{-1})$. Together with the upper bounds characterized in Theorem 4.2, these lower bounds show that KL-PCB is nearly minimax optimal in both the large regularization and small regularization regimes.

Remark 5.3. Previously, Zhao et al. (2026) established an $\Omega(\eta C^{\pi^*} \log N_{\mathcal{R}} n^{-1})$ lower bound for large regularization and an $\Omega(\sqrt{C^{\pi^*} \log N_{\mathcal{R}} n^{-1}})$ lower bound for small regularization by considering a collection of 2-armed contextual bandits. Regarding their construction, $\log N_{\mathcal{R}} = \Theta(S)$, leading to $\Omega(\eta S C^{\pi^*} n^{-1})$ for large regularization and $\Omega(\sqrt{S C^{\pi^*} n^{-1}})$ for small regularization. As a comparison, we provide an $\widetilde{\Omega}(\eta SAC^{\pi^*} n^{-1})$ lower bound for large regularization in Theorem 5.1 and an $\widetilde{\Omega}(\sqrt{SAC^{\pi^*} n^{-1}})$ lower bound in Theorem 5.2 for small regularization. Both results strictly improve upon previous lower bounds. For both Theorem 5.1 and Theorem 5.2, the restriction that $C^* \le \exp(\mathrm{poly}(\eta))$ is inevitable, since we always have $C^{\pi^*} \le \exp(\eta)$ in reverse KL regularized bandits with bounded rewards. Such a constraint has also appeared in previous works (Zhao et al., 2026; Foster et al., 2025).

6. Proof Overview of Hardness Results

6.1. Mechanism behind Hard Instance Construction

The regularized objective $J(\pi) = \langle r, \pi \rangle - \eta^{-1} \mathsf{KL}(\pi \| \pi^{\mathsf{ref}})$ combines a linear term with a KL penalty term. Under large regularization, i.e., small η, the curvature of the dominating regularizer leads to a ϵ^{-1}-type behavior. In contrast, when η is sufficiently large, the effect of regularization becomes negligible and the objective *plausibly* approaches an unregularized one, leading to a ϵ^{-2}-type rate similarly to that of learning standard MABs. This transition of dependency on ϵ is characterized in Zhao et al. (2026), through a construction of two-point type hard instances.

However, to capture the correct dependency of A, we need a more curated construction, which requires a more careful investigation of the different behaviors of the regularized objective with different regularization strength. We begin with the more tractable large-regularization regime, where η is close to zero. In this regime, the KL regularizer dominates the objective, forcing the output policy $\widehat{\pi}$ to remain close to π^*, and thus necessitating accurate estimation of the *reward across all arms*. Our hard instance, therefore, focuses on making reward estimation difficult simultaneously over a sufficient large set of $\Omega(A)$ arms. Similar constructions have been employed in the online setting to establish the fast $\log T$-type lower bounds (Ji et al., 2026) with correct dependency on A.

The construction in the small-regularization regime is more subtle. We first briefly discuss why previous two-point-type hard instances (e.g., Lattimore & Szepesvári 2020, Theorem 15.2; Rashidinejad et al. 2021, Theorem 4; Zhao et al. 2026, Theorem 2.11; Ji et al. 2026, Theorem 5.1)

[3] We refer the readers to Appendix B for elaboration.

for establishing lower bounds in offline MABs, fails to manifest the A dependency. In particular, this approach considers two instances that differ only on a single arm $\widetilde{a}$, whose rewards differ by ϵ. Such a construction also requires $\widetilde{a}$ to be the *unique* optimal arm in at least one instance, implying $\pi^{\mathsf{ref}}(\widetilde{a})^{-1} = C^{\pi^*}$. To distinguish the two instances, any strategy must pay $\Omega(C^{\pi^*}\epsilon^{-2})$ samples[4], recovering the *optimal* $\widetilde{\Theta}(C^{\pi^*}/\epsilon^2)$ rate for the standard objective (Rashidinejad et al., 2021). Importantly, under this construction, scaling up the number of arms A does not change the rate, since $C^{\pi^*} = \pi^{\mathsf{ref}}(\widetilde{a})^{-1}$ increases simultaneously (e.g., when $\pi^{\mathsf{ref}} = \mathsf{Unif}(\mathcal{A})$), and thus the bound remains A-independent. For the KL-regularized objective with $\eta \to \infty$, the KL penalty vanishes and the regularized objective reduces to the standard objective. Since Zhao et al. (2026) use the same type of construction for the KL-regularized objective, the $\Omega(C^{\pi^*}/\epsilon^2)$ lower bound in Zhao et al. (2026) coincides with that of Rashidinejad et al. (2021). This observation suggests that, without modifying the underlying two-point-type argument, extending the two-armed construction of Zhao et al. (2026) to A-armed instances would still yield a $\widetilde{\Theta}(SC^{\pi^*}/\epsilon^2)$ lower bound, leaving a gap of a factor A compared to the $\widetilde{O}(SAC^{\pi^*}/\epsilon^2)$ upper bound.

The key obstacle is that, when there is a *unique* optimal arm, the scaling of C^{π^*} with A is identical for the standard objective and for the KL-regularized objective in the limit $\eta \to \infty$. In contrast, when multiple optima exist, the relationship between C^{π^*} and A differs fundamentally between the two objectives. To illustrate this, we analyze the behavior of the optimal policy on an A-armed MAB, $\pi^{\mathsf{ref}} = \mathsf{Unif}(\mathcal{A})$, with K *suboptimal* arms, where the optimal set of arms is $\mathcal{A}^* = \{a : r(a) = \max_{a'} r(a')\}$. Under standard objective,choosing a deterministic policy π that places all mass on some $a^* \in \mathcal{A}^*$ gives $C^{\pi^*} = A$. On the other hand, under the KL-regularized objective, even as $\eta \to \infty$, the regularization still forces $\pi^* \to \pi^{\mathsf{ref}}|_{\mathcal{A}^*}$, rather than any arbitrary distribution over $\mathcal{A}^*$. Consequently, the concentrability $C^{\pi^*} = \sup_a \pi^*(a)/\pi^{\mathsf{ref}}(a) = A/(A-K)$ depends on the number of optimal arms. This distinction leads to the following tradeoff between concentrability and hardness of learning. On one hand, since we only require the concentrability of the constructed instances to be bounded by a target constant C^*, the reduced value $A/(A-K)$ (as opposed to A) leaves more room to skew the behavior policy, potentially making the learning problem harder. On the other hand, the presence of multiple optimal arms makes the problem intrinsically easier, as identifying *any* optimal arm suffices when $\eta \to \infty$. Carefully balancing this tradeoff is crucial for establishing tight lower bounds in this regime.

In summary, with large regularization, our construction prioritizes the hardness of reward estimation across all arms. With small regularization, we leverage the hardness result for learning MABs with multiple optima in Section 7 in order to balance the tradeoff between statistical indistinguishability and concentrability, which leads to a tight lower bound detailed in Section 6.3.

[4]Generally, $\widetilde{\Theta}(\Delta^{-2})$ samples are necessary and sufficient to distinguish two distributions whose means differ by Δ.

6.2. Proof Outline of Theorem 5.1

In this section, we outline the proof of Theorem 5.1. We first consider the simple case $\pi^{\mathsf{ref}} = \mathsf{Unif}(\mathcal{A})$ and $S = 1$. At a high level, in the large regularization regime, since the KL-regularization term dominates, the learner is required to accurately estimate the reward on all arms to obtain a near-optimal policy. Motivated by this, we design the following class of hard instances, which involves many arms with unknown rewards. Let $A_0 = \Omega(A)$ with $A_0 \le A/2$, and let $\mathcal{V} := \{\pm 1\}^{A_0}$. We consider a class of instances $\{([1],[A], r_{\boldsymbol{\mu}}, \eta, \pi^{\mathsf{ref}})\}_{\boldsymbol{\mu}\in\mathcal{V}}$, where the reward $r_{\boldsymbol{\mu}}$ given $\boldsymbol{\mu}$ is defined as

$$r_{\boldsymbol{\mu}}(i) = \begin{cases} 1/2 + \boldsymbol{\mu}_i\delta & \text{if } i \in [A_0], \\ 1/2 & \text{otherwise,} \end{cases}$$

Within this class of instances, the learner has to determine the rewards on all of the first $[A_0]$ arms to achieve a near-optimal policy. In fact, we can prove that under this regime, estimation errors on each arm accumulate. In other words, the total cost is $\Omega(m\eta\delta^2/A)$ if the learner makes m estimation errors, given that an error on one arm generally costs $\Omega(\eta\delta^2/A)$. In particular, given any $\boldsymbol{\lambda}, \boldsymbol{\mu} \in \mathcal{V}$, we can prove

$$\mathrm{SubOpt}_{\boldsymbol{\lambda}}(\widehat{\pi}) + \mathrm{SubOpt}_{\boldsymbol{\mu}}(\widehat{\pi}) \gtrsim d_H(\boldsymbol{\mu}, \boldsymbol{\lambda})\eta\delta^2/A. \quad (6.1)$$

Now we take $\delta \sim \sqrt{An^{-1}}$. Since each arm only has n/A samples on average, it is not enough to reliably distinguish the reward on any of the first A_0 arms. Therefore, the learner is expected to make at least $A_0/2$ mistakes. Formally, based on (6.1), we can prove the following inequality through a Fano or Assouad-type argument.

$$\inf_{\mathtt{Alg}} \sup_{\boldsymbol{\lambda}\in\mathcal{V}} \mathbb{E}_{\widehat{\pi}\sim\boldsymbol{\lambda}}\big[\mathrm{SubOpt}_{\boldsymbol{\lambda}}(\widehat{\pi})\big] \gtrsim A_0\eta\delta^2/A \gtrsim \eta A n^{-1}. \quad (6.2)$$

Finally, scaling up the bound in (6.2) by a factor $S \cdot C^{\pi^*}$ via considering instances with S contexts and unbalanced π^{ref} leads to desired $\widetilde{\Omega}(\eta SAC^{\pi^*}n^{-1})$ lower bound.

6.3. Proof Outline of Theorem 5.2

In this section, we provide an overview of the proof of Theorem 5.2. We first focus on the simplest setting with $\pi^{\mathsf{ref}} = \mathsf{Unif}(\mathcal{A})$ and $|\mathcal{S}| = 1$. For this regime, we need the result for learning MABs with multiple optima to help us balance the tradeoff between statistical indistinguishability

and concentrability. The result is stated as follows, and we defer the proof idea to Section 7 and the formal statement to Appendix C.

Lemma 6.1 (Informal)**.** *Given any $A > K > 0$ and any sufficiently large n, for any algorithm, there exists an* unregularized *multi-armed bandit with $|\mathcal{A}| = A$, $\pi^{\mathsf{ref}} = \mathsf{Unif}(\mathcal{A})$ and K suboptimal arms, whose definition is nearly the same as the regularized setting in Section 3, except that the notion of suboptimality is defined with respect to the standard objective, i.e., $J(\pi) \coloneqq \mathbb{E}_{a\sim\pi}[r(a)]$, on which the algorithm results in $\widetilde{\Omega}\big(K/\sqrt{nA}\big)$ suboptimality. Moreover, this result is tight up to logarithmic factors.*

We consider the following class of hard instances, which are composed of $(A-K)$ optima and parameterized as follows

$$\boldsymbol{\mu} \in \mathcal{V}_K := \left\{ \boldsymbol{\mu} \in \{\pm 1\}^A \middle| \sum_{i=1}^{A} \mathbb{1}(\boldsymbol{\mu}_i = -1) = K \right\}. \tag{6.3}$$

For any $\boldsymbol{\mu} \in \mathcal{V}_K$, the corresponding instance is given by $([A], r_{\boldsymbol{\mu}}, \mathsf{Unif}(\mathcal{A}))$ where $r_{\boldsymbol{\mu}}(i) = \boldsymbol{\mu}_i\delta$, with $\delta > 0$ to be specified later. Since the KL-regularized objective approaches its unregularized counterpart in the small regularization regime, we have

$$\mathbb{E}_{\boldsymbol{\lambda}}\big[\mathrm{SubOpt}_{\boldsymbol{\lambda}}(\widehat{\pi})\big] \gtrsim \delta - \mathbb{E}_{\boldsymbol{\lambda}}\big[\delta\,\langle \boldsymbol{\lambda}, \widehat{\pi}\rangle\,\big], \tag{6.4}$$

as long as the right-hand side of (6.4) is sufficiently large. In fact, the right-hand side of (6.4) is exactly the sub-optimality gap of instance $\boldsymbol{\lambda}$ under the standard objective. Consequently, invoking Lemma 6.1 [5] gives a lower bound of RHS of (6.4). Specifically, for any $1 \le K \le A-1$ there exists an $\boldsymbol{\lambda} \in \mathcal{V}_K$ such that

$$\mathbb{E}_{\boldsymbol{\lambda}}\big[\mathrm{SubOpt}_{\boldsymbol{\lambda}}(\widehat{\pi})\big] \gtrsim \delta - \mathbb{E}_{\boldsymbol{\lambda}}\big[\delta\,\langle \boldsymbol{\lambda}, \widehat{\pi}\rangle\,\big] \gtrsim \sqrt{\frac{K^2}{An}}.$$

On the other hand, recall that we have $C^{\pi^*} = A/(A-K)$ in this case. Therefore, rewriting the lower bound with the problem parameter C^{π^*} results in

$$\mathbb{E}_{\boldsymbol{\lambda}}\big[\mathrm{SubOpt}_{\boldsymbol{\lambda}}(\widehat{\pi})\big] \gtrsim \sqrt{\frac{C^{\pi^*}}{n} \cdot \frac{K^2(A-K)}{A^{-2}}}.$$

Now we select K to balance the tradeoff between $(A-K)/A$ and K^2/A. A direct calculation shows that $K^2(A-K)$ takes the maximum at $K = 2A/3$, at which $C^{\pi^*} = \Theta(1)$, yielding the following lower bound:

$$\mathbb{E}_{\boldsymbol{\lambda}}\big[\mathrm{SubOpt}_{\boldsymbol{\lambda}}(\widehat{\pi})\big] \gtrsim \sqrt{\frac{A}{n}}. \tag{6.5}$$

[5] The hard instances in (6.3) coincide with the hard instances in the proof of Lemma 6.1, enabling us to directly apply Lemma 6.1 here.

Finally, scaling up the bound in (6.5) by a factor of $\sqrt{S \cdot C^{\pi^*}}$ via considering instances with S contexts and unbalanced π^{ref} leads to the desired $\widetilde{\Omega}(\sqrt{SAC^{\pi^*}n^{-1}})$ lower bound.

7. Proof Overview of Lemma 6.1

In this section, we present the proof overview of Lemma 6.1. For the sake of simplicity, we restrict our output policy $\widehat{\pi}$ to be deterministic in this section and therefore write $\widehat{\pi} = a$ if $\widehat{\pi}(a) = 1$. For the general proof, please refer to Appendix C.

Hard Instances Construction. We first introduce the class of bandit instances and related notation. The considered class of instances admits the following reward parameterization:

$$\boldsymbol{\mu} \in \mathcal{V}_K := \left\{ \boldsymbol{\mu} \in \{\pm 1\}^A \middle| \sum_{i=1}^{A} \mathbb{1}(\boldsymbol{\mu}_i = -1) = K \right\},$$

i.e., exactly K entries of $\boldsymbol{\mu}$ are -1. For any $\boldsymbol{\mu} \in \mathcal{V}_K$, the corresponding instance is given by $([A], r_{\boldsymbol{\mu}}, \mathsf{Unif}(\mathcal{A}))$ where $r_{\boldsymbol{\mu}}(i) = \boldsymbol{\mu}_i\delta$, with $\delta > 0$ to be specified later. We use $a^*(\boldsymbol{\mu}) \coloneqq \{i : \boldsymbol{\mu}_i = 1\}$ to denote the set of optimal arms of $\boldsymbol{\mu}$. We write $\boldsymbol{\mu} \sim_{i,j} \boldsymbol{\lambda}$ if $a^*(\boldsymbol{\mu}) \triangle a^*(\boldsymbol{\lambda}) = \{i, j\}$ with $i \in a^*(\boldsymbol{\mu})$ and $j \in a^*(\boldsymbol{\lambda})$. We use $\mathbb{P}_{\boldsymbol{\mu}}$ to denote the distribution induced by $\boldsymbol{\mu}$, and let $\mathcal{E}_i = \{\widehat{\pi} = i\}$.

For each pair of hard-to-distinguish instances that differ on only two arms, i.e., $\boldsymbol{\mu} \sim_{i,j} \boldsymbol{\lambda}$, the induced distributions $\mathbb{P}_{\boldsymbol{\lambda}}$ and $\mathbb{P}_{\boldsymbol{\mu}}$ are close, and thus the probability of selecting $\widehat{\pi} = i$ is also similar under the two instances. However, $\widehat{\pi} = i$ incurs a cost of δ under $\boldsymbol{\lambda}$, while it is optimal under $\boldsymbol{\mu}$. Specifically, we invoke the following change-of-measure proposition.

Proposition 7.1 (Proposition 18, Degenne & Koolen 2019)**.** *Consider two distributions $\mathbb{P}$ and $\mathbb{Q}$, denoting the log-likelihood ratio with $L = \log(\mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q})$, then for any measurable event $\mathcal{E}$ and threshold $\gamma \in \mathbb{R}$,*

$$\mathbb{P}(\mathcal{E}) \le e^{\gamma}\mathbb{Q}(\mathcal{E}) + \mathbb{P}\{L > \gamma\}.$$

Applying Proposition 7.1. Now we apply Proposition 7.1 by setting $\mathbb{P} = \mathbb{P}_{\boldsymbol{\mu}}$, $\mathbb{Q} = \mathbb{P}_{\boldsymbol{\lambda}}$ and $\mathcal{E} = \mathcal{E}_i$, which gives the following inequality

$$\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_i) \ge e^{-\gamma}\left[\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_i) - \mathbb{P}_{\boldsymbol{\mu}}\left(\log\frac{\mathrm{d}\mathbb{P}_{\boldsymbol{\mu}}}{\mathrm{d}\mathbb{P}_{\boldsymbol{\lambda}}} > \gamma\right)\right].$$

It remains to select the correct γ to be an upper bound of the log-likelihood ratio with high probability so that we can safely get rid of the $\mathbb{P}_{\boldsymbol{\mu}}\big[\mathrm{d}\mathbb{P}_{\boldsymbol{\mu}}/\mathrm{d}\mathbb{P}_{\boldsymbol{\lambda}} > \gamma\big]$ term. Noticing that $\mathbb{E}_{\boldsymbol{\mu}}\big[\log(\mathrm{d}\mathbb{P}_{\boldsymbol{\mu}}/\mathrm{d}\mathbb{P}_{\boldsymbol{\lambda}})\big] = \mathsf{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}\|\mathbb{P}_{\boldsymbol{\lambda}}\right)$, by standard concentration, we know that $\mathbb{P}_{\boldsymbol{\mu}}\big[\log(\mathrm{d}\mathbb{P}_{\boldsymbol{\mu}}/\mathrm{d}\mathbb{P}_{\boldsymbol{\lambda}}) > \gamma\big]$ will be small if $\gamma \gtrsim \mathsf{KL}\left(\mathbb{P}_{\boldsymbol{\mu}}\|\mathbb{P}_{\boldsymbol{\lambda}}\right)$. Actually, a slightly more delicate selection of γ enables $\mathbb{P}_{\boldsymbol{\mu}}\big[\log(\mathrm{d}\mathbb{P}_{\boldsymbol{\mu}}/\mathrm{d}\mathbb{P}_{\boldsymbol{\lambda}}) > \gamma\big] =$

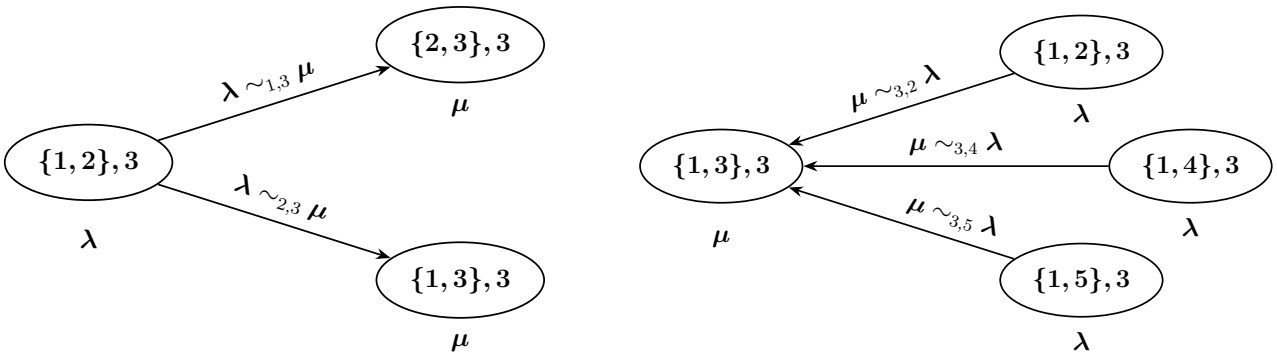

Figure 1. An illustration of the counting argument for the case $A = 5$ and $K = 3$. The set $\{g, h\}$ denotes the instance with its g-th and h-th arm being optimal. Each node, labeled by the tuple $(\{g, h\}, k)$, corresponds to the term $\mathbb{P}_{\{g,h\}}(\mathcal{E}_k)$ in (7.1) and each arrow indicates an instantiation of (7.1). Figure 1a demonstrates that each of $\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_i)$ s.t. $i \notin a^*(\boldsymbol{\lambda})$ can be bounded by $A - K$ possible $\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_i)$ s.t. $i \in a^*(\boldsymbol{\mu})$, thus appears $A - K$ times. Likewise, Figure 1b shows that each $\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_i)$ on the RHS of (7.1) appears K times in total.

$o(A^{-1})$, leading to

$$\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_i) \geq \exp\big(-\mathsf{KL}\left(\mathbb{P}_{\boldsymbol{\mu}} \| \mathbb{P}_{\boldsymbol{\lambda}}\right)\big)\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_i) - o(A^{-1}).$$

Moreover, to make $\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_i)$ sufficiently large, we set $\delta = \widetilde{O}(\sqrt{An^{-1}})$ to make $\mathsf{KL}\left(\mathbb{P}_{\boldsymbol{\mu}} \| \mathbb{P}_{\boldsymbol{\lambda}}\right) = O(1)$. Consequently, we see that the error probability is at least the same order as choosing correctly:

$$\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_i) \gtrsim \mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_i) - o(A^{-1}) \tag{7.1}$$

Aggregating Everything Together. To obtain the minimax lower bound from (7.1), we apply an "averaging hammer" argument (Shamir, 2015; Lattimore & Szepesvári, 2020) over all possible pairs $\boldsymbol{\mu} \sim_{i,j} \boldsymbol{\lambda}$, where the multiplicative factor before each term is given by the following counting.

- For any $\boldsymbol{\lambda}$, fix some $i \notin a^*(\boldsymbol{\lambda})$. We can find $j \in a^*(\boldsymbol{\lambda})$ with $A - K$ choices, and swapping $\boldsymbol{\lambda}_i$ and $\boldsymbol{\lambda}_j$ determines a unique $\boldsymbol{\mu}$. Conversely, for any $\boldsymbol{\mu}$, fix $i \in a^*(\boldsymbol{\mu})$. We have K choices of $j \notin a^*(\boldsymbol{\mu})$ that determine a unique $\boldsymbol{\lambda}$. Therefore, the factor before $\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_i)$ is $A - K$ and $\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_i)$ is K. See Figure 1 for a visual illustration.
- Then we count the total number of pairs. We start by constructing a tuple $\boldsymbol{\mu} \sim_{i,j} \boldsymbol{\lambda}$ with selecting K entries to be -1 to form $\boldsymbol{\mu}$, giving $\binom{A}{K}$ choices. Then we select $i \in a^*(\boldsymbol{\mu})$ from $A - K$ choices, and subsequently $j \notin a^*(\boldsymbol{\mu})$, leading to another K choices. Together, $\boldsymbol{\mu}, i, j$ determines $\boldsymbol{\lambda}$. Applying the multiplication principle, we obtain the $K(A-K)\binom{A}{K}$ factor in front of the constant term.

Applying the counting leads to the following inequality:

$$\begin{aligned}
&(A-K)\sum_{\boldsymbol{\lambda}\in\mathcal{V}_K}\sum_{i\notin a^*(\boldsymbol{\lambda})}\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_i) \\
&\gtrsim K\sum_{\boldsymbol{\mu}\in\mathcal{V}_K}\sum_{i\in a^*(\boldsymbol{\mu})}\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_i) - K(A-K)\binom{A}{K}o\Big(\frac{1}{A}\Big) \\
&\gtrsim K\sum_{\boldsymbol{\mu}\in\mathcal{V}_K}\Big(1-\sum_{j\notin a^*(\boldsymbol{\mu})}\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_j)\Big) - K\binom{A}{K}o(1).
\end{aligned}$$

Using a change of variable $(\boldsymbol{\mu}, j) \to (\boldsymbol{\lambda}, i)$ on the right-hand side and moving it to the left gives

$$\sum_{\boldsymbol{\lambda}\in\mathcal{V}_K}\sum_{i\notin a^*(\boldsymbol{\lambda})}\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_i) \gtrsim \frac{K}{A}\binom{A}{K}(1-o(1)) \gtrsim \frac{K}{A}\binom{A}{K}.$$

Therefore, there exists a $\boldsymbol{\lambda} \in \mathcal{V}_K$ such that

$$\mathrm{SubOpt}_{\boldsymbol{\lambda}}(\widehat{\pi}) = \delta\sum_{i\notin a^*(\boldsymbol{\lambda})}\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_i) \gtrsim \sqrt{\frac{K^2}{An}},$$

where we use $\delta = \widetilde{O}(\sqrt{An^{-1}})$. This finishes the proof of Lemma 6.1.

8. Conclusion and Future Work

In this paper, we study offline learning in KL-regularized MABs. In particular, we provide a sharp sample complexity analysis of a variant of KL-PCB (Zhao et al., 2026) across different regularization magnitudes, showing that KL-PCB achieves a sample complexity of $\widetilde{O}(\eta SAC^{\pi^*}\epsilon^{-1})$ under large regularization and a $\widetilde{O}(SAC^{\pi^*}\epsilon^{-2})$ under small regularization. We further provide two lower bounds in the corresponding regimes that match the upper bounds up to logarithmic factors, certifying that KL-PCB is nearly minimax optimal, and thereby providing a comprehensive understanding of the problem. Currently, our analysis is limited to the multi-armed setting. Extending our results and techniques to KL-regularized objectives with richer structures, like linear or general function approximation, remains an interesting direction for future work.

Impact Statement

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

References

Agarwal, A., Kakade, S., and Yang, L. F. Model-based reinforcement learning with a generative model is minimax optimal. In *Conference on Learning Theory*, pp. 67–83. PMLR, 2020.

Assouad, P. Deux remarques sur l'estimation. *Comptes rendus des séances de l'Académie des sciences. Série 1, Mathématique*, 296(23):1021–1024, 1983.

Audibert, J.-Y. and Bubeck, S. Best arm identification in multi-armed bandits. In *COLT-23th Conference on learning theory-2010*, pp. 13–p, 2010.

Aziz, M., Anderton, J., Kaufmann, E., and Aslam, J. Pure exploration in infinitely-armed bandit models with fixed-confidence. In *Algorithmic Learning Theory*, pp. 3–24. PMLR, 2018.

Azuma, K. Weighted sums of certain dependent random variables. *Tohoku Mathematical Journal, Second Series*, 19(3):357–367, 1967.

Bondy, J. A. and Murty, U. S. R. *Graph theory with applications*. north-Holland, 1979.

Burnetas, A. N. and Katehakis, M. N. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996.

Chaudhuri, A. R. and Kalyanakrishnan, S. Pac identification of a bandit arm relative to a reward quantile. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 1777–1783, 2017.

Chaudhuri, A. R. and Kalyanakrishnan, S. Quantile-regret minimisation in infinitely many-armed bandits. In *UAI*, pp. 425–434, 2018.

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

Cheng, C.-A., Xie, T., Jiang, N., and Agarwal, A. Adversarially trained actor critic for offline reinforcement learning. In *International Conference on Machine Learning*, pp. 3852–3878. PMLR, 2022.

Cover, T. M. *Elements of information theory*. John Wiley & Sons, 1999.

De Heide, R., Cheshire, J., Ménard, P., and Carpentier, A. Bandits with many optimal arms. *Advances in Neural Information Processing Systems*, 34:22457–22469, 2021.

Degenne, R. and Koolen, W. M. Pure exploration with multiple correct answers. *Advances in Neural Information Processing Systems*, 32, 2019.

Di, Q., Zhao, H., He, J., and Gu, Q. Pessimistic nonlinear least-squares value iteration for offline reinforcement learning. In *The Twelfth International Conference on Learning Representations*, 2024.

Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., and Kiela, D. Model alignment as prospect theoretic optimization. In *International Conference on Machine Learning*, pp. 12634–12651. PMLR, 2024.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*, 2021.

Foster, D. J., Mhammedi, Z., and Rohatgi, D. Is a good foundation necessary for efficient reinforcement learning? the computational role of the base model in exploration. In *The Thirty Eighth Annual Conference on Learning Theory*, pp. 2026–2142. PMLR, 2025.

Freedman, D. A. On tail probabilities for martingales. *the Annals of Probability*, pp. 100–118, 1975.

Garivier, A. and Kaufmann, E. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, pp. 998–1027. PMLR, 2016.

Garivier, A. and Kaufmann, E. Nonasymptotic sequential tests for overlapping hypotheses applied to near-optimal arm identification in bandit models. *Sequential Analysis*, 40(1):61–96, 2021.

Garivier, A., Ménard, P., and Stoltz, G. Explore first, exploit next: The true shape of regret in bandit problems. *Mathematics of Operations Research*, 44(2):377–399, 2019.

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

Gu, Y., Han, Y., and Qian, J. Evolution of information in interactive decision making: A case study for multi-armed bandits. *arXiv preprint arXiv:2503.00273*, 2025.

Guruswami, V., Rudra, A., and Sudan, M. Essential coding theory. *Draft available at http://cse. buffalo. edu/faculty/atri/courses/coding-theory/book*, 11, 2019.

Ji, K., Zhao, Q., Zhao, H., Di, Q., and Gu, Q. Near-optimal regret for kl-regularized multi-armed bandits. *arXiv preprint arXiv:2603.02155*, 2026.

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

Kumar, A., Zhou, A., Tucker, G., and Levine, S. Conservative q-learning for offline reinforcement learning. *Advances in neural information processing systems*, 33: 1179–1191, 2020.

Lai, T. L. and Robbins, H. Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1): 4–22, 1985.

Lai, X., Tian, Z., Chen, Y., Yang, S., Peng, X., and Jia, J. Step-dpo: Step-wise preference optimization for long-chain reasoning of llms. *arXiv preprint arXiv:2406.18629*, 2024.

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

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

Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J., and Quillen, D. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *The International journal of robotics research*, 37(4-5):421–436, 2018.

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

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

Mannor, S. and Tsitsiklis, J. N. The sample complexity of exploration in the multi-armed bandit problem. *Journal of Machine Learning Research*, 5(Jun):623–648, 2004.

Meng, Y., Xia, M., and Chen, D. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

Nayak, A., Yang, T., Yagan, O., Joshi, G., and Chi, Y. Achieving logarithmic regret in kl-regularized zero-sum markov games. *arXiv preprint arXiv:2510.13060*, 2025.

Ozdaglar, A. E., Pattathil, S., Zhang, J., and Zhang, K. Revisiting the linear-programming framework for offline rl with general function approximation. In *International Conference on Machine Learning*, pp. 26769–26791. PMLR, 2023.

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

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

Rafailov, R., Hejna, J., Park, R., and Finn, C. From r to q^*: Your language model is secretly a q-function. *arXiv preprint arXiv:2404.12358*, 2024.

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

Rashidinejad, P., Zhu, H., Yang, K., Russell, S., and Jiao, J. Optimal conservative offline rl with general function approximation via augmented lagrangian. *arXiv preprint arXiv:2211.00716*, 2022.

Shamir, O. A variant of azuma's inequality for martingales with subgaussian tails. *arXiv preprint arXiv:1110.2392*, 2011.

Shamir, O. On the complexity of bandit linear optimization. In *Conference on Learning Theory*, pp. 1523–1551. PMLR, 2015.

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

Sidford, A., Wang, M., Wu, X., Yang, L. F., and Ye, Y. Near-optimal time and sample complexities for solving discounted markov decision process with a generative model. *arXiv preprint arXiv:1806.01492*, 2018.

Tiapkin, D., Belomestny, D., Calandriello, D., Moulines, E., Munos, R., Naumov, A., Perrault, P., Tang, Y., Valko, M., and Menard, P. Fast rates for maximum entropy exploration. In *International Conference on Machine Learning*, pp. 34161–34221. PMLR, 2023.

Tsybakov, A. Introduction to nonparametric estimation. In *Springer Series in Statistics*, 2008.

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

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

Wang, R., Foster, D., and Kakade, S. M. What are the statistical limits of offline rl with linear function approximation? In *International Conference on Learning Representations*, 2021.

Wang, Z., Zhou, D., Lui, J. C., and Sun, W. Model-based RL as a minimalist approach to horizon-free and second-order bounds. In *The Thirteenth International Conference on Learning Representations*, 2025.

Weng, W., He, Y., and Zhou, X. Improved bounds for private and robust alignment. *arXiv preprint arXiv:2512.23816*, 2025.

Wu, D., Shi, C., Yang, J., and Shen, C. Greedy sampling is provably efficient for RLHF. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Wu, Y., Tucker, G., and Nachum, O. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*, 2019.

Wu, Y., Thareja, R., Vepakomma, P., and Orabona, F. Offline and online kl-regularized rlhf under differential privacy. *arXiv preprint arXiv:2510.13512*, 2025b.

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

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

Xie, T., Foster, D. J., Krishnamurthy, A., Rosset, C., Awadallah, A. H., and Rakhlin, A. Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025.

Xiong, W., Zhong, H., Shi, C., Shen, C., Wang, L., and Zhang, T. Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and markov game. In *International Conference on Learning Representations (ICLR)*, 2023.

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

Yan, Y., Li, G., Chen, Y., and Fan, J. The efficacy of pessimism in asynchronous q-learning. *IEEE Transactions on Information Theory*, 69(11):7185–7219, 2023.

Yin, M., Duan, Y., Wang, M., and Wang, Y.-X. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. In *International Conference on Learning Representation*, 2022.

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

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

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

Zhang, Y., Chen, C., and Jiang, N. Beyond pessimism: Offline learning in kl-regularized games. *arXiv preprint arXiv:2604.06738*, 2026.

Zhao, H., Ye, C., Gu, Q., and Zhang, T. Sharp analysis for KL-regularized contextual bandits and RLHF. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025a.

Zhao, H., Ye, C., Xiong, W., Gu, Q., and Zhang, T. Logarithmic regret for online KL-regularized reinforcement learning. In *Forty-second International Conference on Machine Learning*, 2025b.

Zhao, Q., Ji, K., Zhao, H., Zhang, T., and Gu, Q. Towards a sharp analysis of offline policy learning for f-divergence-regularized contextual bandits. In *The Fourteenth International Conference on Learning Representations*, 2026.

Zhu, Y. and Nowak, R. On regret with multiple best arms. *Advances in Neural Information Processing Systems*, 33: 9050–9060, 2020.

A. Additional Related Work

Pessimism in Offline RL. The principle of pessimism, under which the learner acts conservatively in the face of uncertainty, plays a crucial role in offline RL. In the tabular setting, pessimism is applied in both model-based (Rashidinejad et al., 2022; Xie et al., 2021b) and model-free (Shi et al., 2022; Yan et al., 2023) algorithm design, based on which the optimal sample complexity is achieved (Li et al., 2024). Under reward or value function approximation, the pessimism principle is also widely adopted (Jin et al., 2021; Yin et al., 2022; Xie et al., 2021a; Uehara & Sun, 2021; Wang et al., 2025) and achieves optimal sample complexity in both linear (Xiong et al., 2023) and general function approximation (Di et al., 2024). Under KL-regularized objectives, pessimism is again provably beneficial and admits a sharp analysis (Zhao et al., 2026).

Classical Minimax Lower Bounds Techniques. Information-theoretic techniques for proving non-asymptotic worst-case hardness results in *offline* learning have been well-established (Cover, 1999; Tsybakov, 2008; Polyanskiy & Wu, 2025), among which Le Cam's two-point method, Assouad's lemma (Assouad, 1983), and Fano's method (Cover, 1999) are arguably the most commonly used (Yu, 1997). These tools have also been extended to counterparts that mainly aim to accommodate the blessing or cost of exploration in lower bounds for *interactive* protocols (Lattimore & Szepesvári, 2020; Foster et al., 2021; Chen et al., 2024; Gu et al., 2025), which are beyond our scope. Specialized to our setting, the proof of our *worst case* lower bound (Theorem 5.2) for *noninteractive* learning of KL-regularized multi-armed contextual bandits builds upon ideas that are closely related to measure tilting techniques in Degenne & Koolen (2019); Garivier & Kaufmann (2021), where they are used to establish instance-dependent lower bounds. These techniques originate from a broader line of bandit literature that employs variants of change-of-measure arguments to prove instance-dependent hardness results in online settings or asymptotic regimes (Lai & Robbins, 1985; Mannor & Tsitsiklis, 2004; Audibert & Bubeck, 2010; Garivier & Kaufmann, 2016; Degenne & Koolen, 2019; Garivier & Kaufmann, 2021).

B. Relation Between Coverage Notions

In this section, we provide a detailed discussion about the relation between the coverage notions C^{π^*} and $D^2_{\pi^*}$ in Zhao et al. (2026) when specialized to our multi-armed setting. We first recall the definition of D^2-divergence (Zhao et al., 2026, Definition 2.5):

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

where $\mathcal{R}$ is the function class. When specialized to multi-armed setting, $\mathcal{R}$ corresponds to all possible bounded reward functions over $\mathcal{S}\times\mathcal{A}$. Moreover, the single-policy $\mathcal{D}^2$-concentrability is defined as $D^2_{\pi^*} := \mathbb{E}_{(s,a)\sim\rho\times\pi^*} D^2_{\mathcal{R}}((s,a);\pi^{\mathsf{ref}})$ (Zhao et al., 2026, Definition 2.7).

Specifically, we provide an example in which $D^2_{\pi^*} = \Omega(SAC^{\pi^*})$.

Proposition B.1. *For any $S \geq 1$, $A \geq 3$, $\eta > 0$, $2 < C \leq \exp(\eta)$, there exists an offline KL-regularized multi-armed bandit on which $C^{\pi^*} = C/2$ and $D^2_{\pi^*} = \Omega(SAC^{\pi^*})$.*

Proof. We consider the following offline KL-regularized multi-armed bandit $(\mathcal{S},\mathcal{A},r,\eta,\pi^{\mathsf{ref}})$, where $\mathcal{S}=[S]$, $\mathcal{A}=[A+1]$. For all $s\in\mathcal{S}$, the reference policy $\pi^{\mathsf{ref}}(\cdot|s)$ is given by:

$$\pi^{\mathsf{ref}}(i|s) = \frac{1}{AC}, \forall\, i\leq A;\ \pi^{\mathsf{ref}}(A+1|s) = \frac{C-1}{C},$$

where $2<C\leq\exp(\eta)$. We further set the reward as

$$r(s,i)=1, \forall\, i\in[A];\ r(s,A+1)=1-\alpha,$$

for all $s\in\mathcal{S}$ and $\alpha=\eta^{-1}\log(C-1)$. Based on the setup, since $\pi^*(\cdot|s)\propto\pi^{\mathsf{ref}}(\cdot|s)\exp(\eta r(s,\cdot))$, a direct calculation yields that for any $s\in\mathcal{S}$, the optimal policy $\pi^*(\cdot|s)$ is given by

$$\pi^*(i|s)=\frac{1}{2A}, \forall\, i\in[A];\ \pi^*(A+1|s)=\frac{1}{2}.$$

Recall that $C^{\pi^*} = \sup_{s,a} \pi^*(a|s)/\pi^{\mathsf{ref}}(a|s)$. It is easy to see that the supremum is attained at any $i \in [A]$ and $s \in \mathcal{S}$, which leads to

$$C^{\pi^*} = \frac{\pi^*(1|s)}{\pi^{\mathsf{ref}}(1|s)} = \frac{C}{2}.$$

Now we come to $D^2_{\pi^*}$. When specialized to multi-armed setting, $\mathcal{R}$ corresponds to all possible reward functions. Therefore, for any $(s,a) \in \mathcal{S} \times \mathcal{A}$, the supremum is achieved at $g = 1$ at (s,a) otherwise 0 and $h = 0$, which gives

$$D^2_{\mathcal{R}}((s,a); \pi^{\mathsf{ref}}) = \frac{1}{\rho(s)\pi^{\mathsf{ref}}(a|s)}.$$

Recall that $D^2_{\pi^*} := \mathbb{E}_{(s,a)\sim\rho\times\pi^*} D^2_{\mathcal{R}}((s,a); \pi^{\mathsf{ref}})$, we know that

$$D^2_{\pi^*} = \mathbb{E}_{(s,a)\sim\rho\times\pi^*}\left[\frac{1}{\rho(s)\pi^{\mathsf{ref}}(a|s)}\right] \geq \sum_{s\in\mathcal{S}}\sum_{a\in[A]} \frac{\pi^*(a|s)}{\pi^{\mathsf{ref}}(a|s)} = \frac{SAC}{2} = SAC^{\pi^*},$$

which finishes the proof. □

C. Offline Multi-armed Bandits with Multiple Optima

C.1. Problem Setup

In this section, we provide a formal setup of the problem of offline multi-armed bandits with multiple optima. Recall that the problem of interest is offline MABs, which is denoted by a tuple $(\mathcal{A}, r, \pi^{\mathsf{ref}})$. Here, $\mathcal{A}$ is the action space with $|\mathcal{A}| = A$, and $r : \mathcal{A} \to [0,1]$ is the reward function. In the offline setting, the agent only has access to a dataset $\mathcal{D} = \{(a_i, r_i)\}_{i=1}^n$. Here $a_i \in \mathcal{A}$ is the action taken independently from the behavior policy π^{ref}, and r_i is the observed reward given by $r_i = r(a_i) + \varepsilon_i$, where ε_i is independent 1-sub-Gaussian. The goal is to find a policy that maximizes the following objective

$$J(\pi) := \mathbb{E}_{a\sim\pi}[r(a)].$$

The optimal policy π^* is defined to be the policy that maximizes the above objective. A policy π is said to be ϵ-optimal if $J(\pi^*) - J(\pi) \leq \epsilon$.

In this work, we focus on the case where we might have multiple arms on which the reward takes the maximum. Specifically, we use $\mathcal{A}^*$ to denote the actions that achieve the maximal reward, i.e., $\mathcal{A}^* = \{a^* \in \mathcal{A} : \ r(a^*) = \max_{a\in\mathcal{A}} r(a)\}$ and $\mathcal{A}_{\text{sub}} = \mathcal{A} \setminus \mathcal{A}^*$. We use K to denote the cardinality of $\mathcal{A}_{\text{sub}}$ and consider the non-degenerate case $1 \leq K \leq A-1$. In this section, we restrict our behavior policy to $\pi^{\mathsf{ref}} = \mathsf{Unif}(\mathcal{A})$.

C.2. Hardness Results

In this section, we provide the following hardness result for this problem and the proof is deferred to Appendix F.1.

Theorem C.1 (Formal Version of Lemma 6.1)**.** *Given any $A > K > 0$ and any sufficiently large n, for any algorithm, there exists a multi-armed bandit with $|\mathcal{A}| = A$, uniform π^{ref} and K suboptimal arms on which the algorithm results in $\Omega\big(K/\sqrt{nA\log A}\big)$ suboptimality.*

Previously, Rashidinejad et al. (2021) considered a class of hard instances with a single optimal arm and used them to establish the near-optimal lower bound $\Omega(C^{\pi^*}/\epsilon^2)$, which reduces to $\Omega(A/\epsilon^2)$ under a uniform behavior policy. This result is recovered (up to logarithmic factors) as a special case of Theorem C.1 by setting $K = A-1$. In contrast, Theorem C.1 covers the entire range $1 \leq K \leq A-1$, and thus constitutes a nontrivial generalization of the previous result.

C.3. Algorithm

In this section, we show that the rate in Theorem C.1 cannot be improved up to minor factors. Actually, this rate can be achieved through the simplest algorithm that first computes the empirical mean and then outputs the maximizer, as shown in Algorithm 2.

Theorem C.2. *For sufficiently small ϵ, $n = \widetilde{O}(K^2A^{-1}\epsilon^{-2})$ samples suffice to guarantee that the expected suboptimality gap of the output policy of Algorithm 2 is less than ϵ.*

Algorithm 2 Empirical Reward Maximization for Multi-armed Bandit with Multiple Optima

Require: Offline dataset $\mathcal{D} = \{(a_i, r_i)\}_{i=1}^n$ sampled from $\pi^{\mathsf{ref}} = \mathsf{Unif}(\mathcal{A})$
1: **for** all $a \in \mathcal{A}$ **do**
2: Count the number of visits $N(a) = \sum_{i=1}^n \mathbb{1}\{a_i = a\}$
3: **if** $N(a) = 0$ **then**
4: Set the empirical reward $\bar{r}(a) \leftarrow 0$
5: **else**
6: Compute the empirical reward $\bar{r}(a) \leftarrow \sum_{i=1}^n r_i \mathbb{1}\{a_i = a\}/N(a)$
7: **end if**
8: **end for**
Ensure: $\widehat{a} = \mathrm{argmax}_{a \in \mathcal{A}} \bar{r}(a)$.

Remark C.3. It is well known that the optimal sample complexity of offline contextual MABs scales as $\widetilde{O}(SC^{\pi^*}/\epsilon^2)$ (Rashidinejad et al., 2021; Li et al., 2024). When specialized to our setup, $S = 1$ and $C^{\pi^*} = A$, leading to the $O(A/\epsilon^2)$ sample complexity. In comparison, Theorem C.2 establishes an $\widetilde{O}(K^2A^{-1}\epsilon^{-2})$ rate, which can be much smaller than $O(A/\epsilon^2)$ when K is small. This comparison highlights that the existence of many optimal arms substantially simplifies offline bandit learning.

Theorem C.1 indicates an $\widetilde{\Omega}\big(K^2A^{-1}\epsilon^{-2}\big)$ sample complexity lower bound for achieving ϵ-optimality. On the other hand, Theorem C.2 provides a simple algorithm attaining a sample complexity of $\widetilde{O}(K^2A^{-1}\epsilon^{-2})$, confirming that the lower bound in Theorem C.1 is tight up to logarithmic factors. Despite the sharpened results, Algorithm 2, together with Theorem C.2 and Theorem C.1, are restricted to uniform behavior policy. Extending these results to any π^{ref} remains an interesting direction for future work.

Proof Overview of Theorem C.2. We provide an intuition for understanding the rate given by Theorem C.2 and defer the full proof to Appendix F.2. When there are only K suboptimal arms, the sum of probabilities of each bad arm being selected by Algorithm 2 is no greater than K/A. Consequently, even if each bad arm incurs a suboptimality gap as large as $A\epsilon/K$, their aggregate contribution to the expected suboptimality remains on the order of ϵ.

This observation implies that we only need to reliably distinguish those arms with a suboptimality gap larger than $A\epsilon/K$. To identify such an arm, following the standard Δ^{-2} sample complexity, we need $\widetilde{O}\big((A\epsilon/K)^{-2}\big) = \widetilde{O}(K^2A^{-2}\epsilon^{-2})$ samples from that arm. Since the samples are drawn uniformly over all arms, each arm receives roughly a $1/A$ fraction of the total samples. Therefore, a total of $\widetilde{O}(K^2A^{-1}\epsilon^{-2})$ samples suffices for ϵ-optimality.

C.4. Additional Related Work on Bandits with Multiple Optima

Previous works have studied MABs with multiple optimal arms in both regret minimization and pure exploration settings. In the pure exploration setting, the asymptotically optimal rate for finite-armed bandits was established by Degenne & Koolen (2019) for the fixed confidence setting. In the regret minimization setting, if the algorithm is restricted to be uniformly fast convergent (Garivier et al., 2019, Definition 1), that is, to perform well across all problem instances, then the regret is given by $\sum_{a \in \mathcal{A} \setminus \mathcal{A}^*} \Delta_a^{-1} \log T$ asymptotically as $T \to \infty$, which roughly grows linearly with the number of suboptimal arms K (Lai & Robbins, 1985; Burnetas & Katehakis, 1996). This uniform convergence requirement is relaxed in a line of work on bandits with a large or even infinite number of arms ($A \gg T$ or $A = \infty$) (Chaudhuri & Kalyanakrishnan, 2017; 2018; Aziz et al., 2018; De Heide et al., 2021; Zhu & Nowak, 2020), where the dependence on K would otherwise render the bounds vacuous as $A, K \to \infty$. When specialized back to the finite-armed setting, the asymptotic regret obtained becomes $A \log T \log(\Delta^{-1})((A-K)\Delta)^{-1}$ (De Heide et al., 2021), which is tight up to the $\log(\Delta^{-1})$ factor. From a minimax perspective, it is possible to achieve an $O(\sqrt{AT/(A-K)})$ regret (Zhu & Nowak, 2020), which is optimal when $K \geq A/2$, as implied by the two-point argument in Lattimore & Szepesvári (2020, Theorem 15.2), as remarked in Zhu & Nowak (2020). Despite these advances, the sample complexity of this problem in the pure offline setting remains largely underexplored. Furthermore, in both online and offline regimes, the minimax optimal rate for the regime $K < A/2$, where the Le Cam two-point method becomes ineffective, is still not well understood.

D. Proof of Theorems in Section 4

D.1. Proof of Lemma 4.1

Proof of Lemma 4.1. The proof is standard in previous literature (e.g., Lemma B.1 in Xie et al. (2021b)), and we provide the detailed proof here for completeness. We first prove that $\mathcal{E}_1$ holds with high probability. Fixing a pair (s,a), the event holds trivially when $N(s,a)=0$. When $N(s,a)\geq 1$, conditioning on $N(s,a)$ and applying Azuma-Hoeffding's inequality (Lemma G.2), we know that with probability at least $1-\delta/(2SA)$, we have

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

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 $1-\delta/2$. For the second event, we directly invoke Lemma G.1 and obtain that 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 union bound over all $(s,a)\in\mathcal{S}\times\mathcal{A}$ gives that $\mathcal{E}_2$ holds with probability at least $1-\delta/2$. Taking union bound once more over $\mathcal{E}_1$ and $\mathcal{E}_2$ finishes the proof. □

D.2. Proof of Theorem 4.2

In this section, we present the proof of Theorem 4.2. First, we need the following regret decomposition lemma for regularized objective.

Lemma D.1 (Lemma 2.16, Zhao et al. 2026)**.** *Let $r:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ be any reward function, and $\pi_r(\cdot|s)\propto\pi^{\mathsf{ref}}(\cdot|s)\exp(\eta r(s,\cdot))$ be the optimal policy under r, then there exists some $\gamma\in[0,1]$ such that if we denote $r_\gamma=\gamma r+(1-\gamma)r^*$ and $\pi_\gamma(\cdot|s)\propto\pi^{\mathsf{ref}}(\cdot|s)\exp(\eta r_\gamma(s,\cdot))$, then*

$$J(\pi^*)-J(\pi_r)\leq\eta\mathbb{E}_{(s,a)\sim\rho\times\pi_\gamma}\left[\left(r^*(s,a)-r(s,a)\right)^2\right].$$

Moreover, Lemma 2.15 in Zhao et al. (2026) shows that when $r\leq r^*$, we can take $\gamma=0$. For simplicity, we combine the original Lemma 2.15 in Zhao et al. (2026) with Lemma D.1 and state the following Lemma.

Lemma D.2. *Let $r:\mathcal{S}\times\mathcal{A}\to\mathbb{R}$ be any reward function satisfying $r(s,a)\leq r^*(s,a)$ for any $(s,a)\in\mathcal{S}\times\mathcal{A}$, and let $\pi_r(\cdot|s)\propto\pi^{\mathsf{ref}}(\cdot|s)\exp(\eta r(s,\cdot))$ be the optimal policy under r, then*

$$J(\pi^*)-J(\pi_r)\leq\eta\mathbb{E}_{(s,a)\sim\rho\times\pi^*}\left[\left(r^*(s,a)-r(s,a)\right)^2\right].$$

Now we are ready to prove Theorem 4.2.

Proof of Theorem 4.2. We derive the proof conditioning on event $\mathcal{E}_1(\delta)\cap\mathcal{E}_2(\delta)$. $\mathcal{E}_1(\delta)$ ensures that for all s,a, $\widehat{r}(s,a)\leq r^*(s,a)$. We first prove the fast-rate bound, where the proof is almost identical to the proof in Zhao et al. (2026) and we provide the full proof for completeness. First, by Lemma D.2, we have the following regret decomposition

$$J(\pi^*)-J(\widehat{\pi})\leq\eta\mathbb{E}_{(s,a)\sim\rho\times\pi^*}\left[\left(\widehat{r}(s,a)-r^*(s,a)\right)^2\right]\leq 4\eta\mathbb{E}_{(s,a)\sim\rho\times\pi^*}[b^2(s,a)],$$

where the second inequality holds due to $|\widehat{r}(s,a)-r^*(s,a)|\leq 2b(s,a)$ on $\mathcal{E}_1$. Plugging in the exact construction of $b(s,a)$, we know that

$$\begin{aligned}J(\pi^*)-J(\widehat{\pi})&\leq 4\eta\sum_{s\in\mathcal{S}}\rho(s)\sum_{a\in\mathcal{A}}\pi^*(a|s)\frac{4\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{N(s,a)}\\&\leq 4\eta\sum_{s\in\mathcal{S}}\rho(s)\sum_{a\in\mathcal{A}}\pi^*(a|s)\frac{32\log^2(2|\mathcal{S}||\mathcal{A}|/\delta)}{n\rho(s)\pi^{\mathsf{ref}}(a|s)}\end{aligned}$$

$$\leq 128\eta C^{\pi^*}\log^2(2|\mathcal{S}||\mathcal{A}|/\delta)\sum_{(s,a)\in\mathcal{S}\times\mathcal{A}}\rho(s)\pi^{\mathsf{ref}}(a|s)\frac{1}{\rho(s)\pi^{\mathsf{ref}}(a|s)}$$
$$= \widetilde{O}(\eta C^{\pi^*}SAn^{-1}),$$

where the second inequality holds due to event $\mathcal{E}_2$ and the last inequality holds due to the definition $\sup_{s,a}\pi^*(a|s)/\pi^{\mathsf{ref}}(a|s) = C^{\pi^*}$.

We then prove the slow-rate bound, which follows the standard argument in the analysis. In particular, we have

$$\begin{aligned}
J(\pi^*) - J(\widehat{\pi}) &\leq \mathbb{E}_{s,a\sim\rho\times\pi^*}[r^*(s,a) - \widehat{r}(s,a)] \\
&\leq 4\sum_{s\in\mathcal{S}}\rho(s)\sum_{a\in\mathcal{A}}\pi^*(a|s)\sqrt{\frac{\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{N(s,a)}} \\
&\leq 16\sum_{s\in\mathcal{S}}\rho(s)\sum_{a\in\mathcal{A}}\pi^*(a|s)\frac{\log(2|\mathcal{S}||\mathcal{A}|/\delta)}{\sqrt{n\rho(s)\pi^{\mathsf{ref}}(a|s)}} \\
&= 16\log(2|\mathcal{S}||\mathcal{A}|/\delta)\sum_{s\in\mathcal{S}}\rho(s)\sum_{a\in\mathcal{A}}\pi^*(a|s)\frac{1}{\sqrt{n\rho(s)\pi^*(a|s)}}\sqrt{\frac{\pi^*(a|s)}{\pi^{\mathsf{ref}}(a|s)}} \\
&\leq 16\log(2|\mathcal{S}||\mathcal{A}|/\delta)\sqrt{C^{\pi^*}n^{-1}}\sum_{s\in\mathcal{S}}\rho(s)\sum_{a\in\mathcal{A}}\pi^*(a|s)\frac{1}{\sqrt{\rho(s)\pi^*(a|s)}} \\
&= \widetilde{O}\big(\sqrt{SAC^{\pi^*}n^{-1}}\big),
\end{aligned}$$

where the second inequality holds due to event $\mathcal{E}_1(\delta)$, the third holds due to event $\mathcal{E}_2(\delta)$, and the last holds due to the Cauchy-Schwarz inequality. By Lemma 4.1, we know that $\mathcal{E}_1(\delta)\cap\mathcal{E}_2(\delta)$ holds with probability at least $1-\delta$, which finishes the proof. □

E. Proof of Theorems in Section 5

E.1. Proof of Theorem 5.1

Proof of Theorem 5.1. We first fix the context size $S \geq 1$, action set size $A+1 \geq 3$, regularization parameter $\eta > 4\log 2$, upper bound of coverage coefficient $C^* \in (2, \exp(\eta/4)]$ and any $n \geq \eta^2 SC^*A(1024\log A)^{-1}$, we consider the family of contextual bandits with $S := |\mathcal{S}|$, $A := |\mathcal{A}| - 1 < \infty$ and reward function in some function class $\mathcal{V}$ composed of function $\mathcal{S}\times\mathcal{A}\to[0,1]$ as follows.

$$\mathrm{CB}_{\mathcal{V}} := \{(\mathcal{S},\mathcal{A},\rho,r,\pi^{\mathsf{ref}},\eta) : r\in\mathcal{V}, \rho\in\Delta(\mathcal{S}), \pi^{\mathsf{ref}}\in\Delta(\mathcal{A}|\mathcal{S})\}. \tag{E.1}$$

We set $\mathcal{S} = [S]$, $\mathcal{A} = [A+1]$, $\rho = \mathsf{Unif}(\mathcal{S})$, and the reference policy to be

$$\forall s\in\mathcal{S}, \pi^{\mathsf{ref}}(i|s) = (CA)^{-1}\ \forall i\in[A], \pi^{\mathsf{ref}}(A+1|s) = 1 - C^{-1},$$

where $C \geq 1$ is a parameter to be specified later. Now we construct our reward functions. We first leverage Lemma G.7 with $\Sigma = \{-1,+1\}$ and obtain a set $\mathcal{U}\subset\{-1,+1\}^A$ such that (1) $|\mathcal{U}| \geq \exp(A/8)$ and (2) for any $\boldsymbol{\mu},\boldsymbol{\mu}'\in\mathcal{U}, \boldsymbol{\mu}\neq\boldsymbol{\mu}'$, one has $\|\boldsymbol{\mu}-\boldsymbol{\mu}'\|_1 \geq A/2$. We once again apply the general version of Lemma G.7. By setting $\Sigma = \mathcal{U}$, we can obtain a subset $\mathcal{V}\subset\mathcal{U}^S$ such that

1. $\log_{|\mathcal{U}|}|\mathcal{V}| \geq S/8$, which gives $\log|\mathcal{V}| \geq \log_{|\mathcal{U}|}|\mathcal{V}|\log|\mathcal{U}| \geq SA/64$
2. For any $\boldsymbol{\nu},\boldsymbol{\nu}'\in\mathcal{V}, \boldsymbol{\nu}\neq\boldsymbol{\nu}'$, one has $d_H(\boldsymbol{\nu}-\boldsymbol{\nu}') \geq S/2$

Now we construct our Bernoulli reward function whose means are given as follows.

$$\mathcal{G} = \{r_{\boldsymbol{\nu}}(s,i) = 1/2 + \boldsymbol{\nu}_{s,i}\delta\ \forall i\in[A], r_{\boldsymbol{\nu}}(s,A+1) = 1/2-\alpha,\ \forall s\in\mathcal{S}|\boldsymbol{\nu}\in\mathcal{V}\},$$

where $\boldsymbol{\nu}_{s,i}$ is the i-th coordinate of the s-th entry of $\boldsymbol{\nu}$. In other words, for any two reward function r_1 and r_2 in $\mathcal{G}$, there exist $\Omega(S)$ contexts such that for these contexts, there exist $\Omega(A)$ arms such that the two rewards on these arms are different.

Now we consider two different reward functions $r_1, r_2 \in \mathcal{G}$ for a certain shared context s such that in this context the two reward functions differ on at least $A/2$ arms. In the following, we drop the subscripts for context s when there is no ambiguity. Without loss of generality, we assume that r_1 and r_2 are distributed as follows.

$$\begin{aligned}
&r_1(i) = 1/2 + \delta, r_2(i) = 1/2 - \delta,\ \forall i \in [1, l];\\
&r_1(i) = 1/2 - \delta, r_2(i) = 1/2 + \delta,\ \forall i \in [l+1, m];\\
&r_1(i) = r_2(i) = 1/2 + r^*(i), r^*(i) \in \{\pm\delta\},\ \forall i \in [m+1, A],
\end{aligned}$$

where $0 \le l \le m$ and $m \ge A/2$ are some integers. Let π_1^* and π_2^* be the optimal policy regarding r_1 and r_2, we have

$$\pi_1^*(i) = \frac{\pi^{\mathsf{ref}}(i)\exp(\eta r_1(i))}{\sum_{j=1}^{A+1}\pi^{\mathsf{ref}}(j)\exp(\eta r_1(j))},\ \pi_2^*(i) = \frac{\pi^{\mathsf{ref}}(i)\exp(\eta r_2(i))}{\sum_{j=1}^{A+1}\pi^{\mathsf{ref}}(j)\exp(\eta r_2(j))}.$$

Now we assign $C^* = C$ and $\alpha = \eta^{-1}\log(C-1)$. Therefore, we know that

$$C^{\pi_1^*} = \max_i \frac{\exp(\eta r_1(i))}{\sum_{j=1}^{A+1}\pi^{\mathsf{ref}}(j)\exp(\eta r_1(j))} \le \frac{\exp(\eta\delta)}{\exp(-\eta\delta)/C + (C-1)\exp(-\eta\alpha)/C} \le C = C^*. \tag{E.2}$$

The argument above holds similarly for $C^{\pi_2^*}$, which gives that for all reward function $r \in \mathcal{G}$, we have $C^{\pi_r^*} \le C^*$. Now we consider the suboptimality gap $\mathrm{SubOpt}_1(\widehat{\pi}) = \mathrm{SubOpt}(\widehat{\pi}; r_1)$ and $\mathrm{SubOpt}_2(\widehat{\pi}) = \mathrm{SubOpt}(\widehat{\pi}; r_2)$. By Lemma G.9, we know that

$$\mathrm{SubOpt}_1(\widehat{\pi}) + \mathrm{SubOpt}_2(\widehat{\pi}) = \eta^{-1}\mathsf{KL}\left(\widehat{\pi}\|\pi_1^*\right) + \eta^{-1}\mathsf{KL}\left(\widehat{\pi}\|\pi_2^*\right).$$

Similar to Zhao et al. (2026, (B.9)), $\mathrm{SubOpt}_1(\widehat{\pi})+\mathrm{SubOpt}_2(\widehat{\pi})$ is minimized by $\bar{\pi}$ with $\bar{\pi}(i) \propto \sqrt{\pi_1^*(i)\pi_2^*(i)}$. Consequently, we have

$$\mathrm{SubOpt}_1(\widehat{\pi}) + \mathrm{SubOpt}_2(\widehat{\pi}) \ge \mathrm{SubOpt}_1(\bar{\pi}) + \mathrm{SubOpt}_2(\bar{\pi}) = \eta^{-1}\mathsf{KL}\left(\bar{\pi}\|\pi_1^*\right) + \eta^{-1}\mathsf{KL}\left(\bar{\pi}\|\pi_2^*\right),$$

which can be reformulated as

$$\begin{aligned}
&\mathrm{SubOpt}_1(\bar{\pi}) + \mathrm{SubOpt}_2(\bar{\pi})\\
&= 2\eta^{-1}\sum_{a\in\mathcal{A}}\bar{\pi}(a)\log\frac{\sqrt{\sum_{i=1}^{A+1}\pi^{\mathsf{ref}}(i)\exp(\eta r_1(i))}\sqrt{\sum_{j=1}^{A+1}\pi^{\mathsf{ref}}(j)\exp(\eta r_2(j))}}{\sum_{k=1}^{A+1}\pi^{\mathsf{ref}}(k)\exp\left(\eta\big(r_1(k)+r_2(k)\big)/2\right)}\\
&= 2\eta^{-1}\log\frac{\sqrt{\sum_{i=1}^{A+1}\pi^{\mathsf{ref}}(i)\exp(\eta r_1(i))}\sqrt{\sum_{j=1}^{A+1}\pi^{\mathsf{ref}}(j)\exp(\eta r_2(j))}}{\sum_{k=1}^{A+1}\pi^{\mathsf{ref}}(k)\exp\left(\eta\big(r_1(k)+r_2(k)\big)/2\right)}\\
&= \eta^{-1}(X_1 + X_2),
\end{aligned}$$

where

$$\begin{aligned}
X_1 &= \log\frac{\sum_{j=1}^{A+1}\pi^{\mathsf{ref}}(j)\exp(\eta r_1(j))}{\sum_{i=1}^{A+1}\pi^{\mathsf{ref}}(i)\exp\left(\eta\big(r_1(i)+r_2(i)\big)/2\right)};\\
X_2 &= \log\frac{\sum_{j=1}^{A+1}\pi^{\mathsf{ref}}(j)\exp(\eta r_2(j))}{\sum_{i=1}^{A+1}\pi^{\mathsf{ref}}(i)\exp\left(\eta\big(r_1(i)+r_2(i)\big)/2\right)}.
\end{aligned}$$

Now we compute the first term X_1.

$$X_1 = \log\frac{\sum_{j=1}^{l}\pi^{\mathsf{ref}}(j)e^{\eta\delta} + \sum_{j=l+1}^{m}\pi^{\mathsf{ref}}(j)e^{-\eta\delta} + \sum_{j=m+1}^{A}\pi^{\mathsf{ref}}(j)e^{\eta r^*(j)} + \pi^{\mathsf{ref}}(A+1)e^{-\eta\alpha}}{\sum_{j=1}^{m}\pi^{\mathsf{ref}}(j) + \sum_{j=m+1}^{A}\pi^{\mathsf{ref}}(j)e^{\eta r^*(j)} + \pi^{\mathsf{ref}}(A+1)e^{-\eta\alpha}}$$

$$
\begin{aligned}
&= \log \frac{(CA)^{-1} l e^{\eta\delta} + (CA)^{-1}(m-l)e^{-\eta\delta} + (CA)^{-1}\sum_{j=m+1}^{A} e^{\eta r^*(j)} + C^{-1}(C-1)e^{-\eta\alpha}}{(CA)^{-1}m + (CA)^{-1}\sum_{j=m+1}^{A} e^{\eta r^*(j)} + C^{-1}(C-1)e^{-\eta\alpha}} \\
&= \log \frac{A^{-1} l \exp(\eta\delta) + A^{-1}(m-l)\exp(-\eta\delta) + A^{-1}\sum_{j=m+1}^{A}\exp(\eta r^*(j)) + (C-1)e^{-\eta\alpha}}{A^{-1}m + \underbrace{A^{-1}\sum_{j=m+1}^{A}\exp(\eta r^*(j)) + (C-1)e^{-\eta\alpha}}_{M}} \\
&= \log \frac{A^{-1} l \exp(\eta\delta) + A^{-1}(m-l)\exp(-\eta\delta) + M}{A^{-1}m + M}.
\end{aligned}
$$

Following a similar argument, we know that

$$
X_2 = \log \frac{A^{-1}(m-l)\exp(\eta\delta) + A^{-1} l \exp(-\eta\delta) + M}{A^{-1}m + M}
$$

By Jensen's inequality, we know that $X_1 + X_2$ takes the minimum with respect to l when $l = 0$ or $l = m$. Therefore we have

$$
\begin{aligned}
X_1 + X_2 &\geq \log \frac{A^{-1}m\exp(\eta\delta) + M}{A^{-1}m + M} + \log \frac{A^{-1}m\exp(-\eta\delta) + M}{A^{-1}m + M} \\
&= \log \frac{M^2 + A^{-2}m^2 + MA^{-1}m(\exp(\eta\delta) + \exp(-\eta\delta))}{(A^{-1}m + M)^2} \\
&= \log\left(1 + \frac{2MA^{-1}m}{(A^{-1}m + M)^2}\left(\frac{\exp(\eta\delta) + \exp(-\eta\delta)}{2} - 1\right)\right) \\
&\geq \log\left(1 + \frac{M}{(1+M)^2}\left(\frac{\exp(\eta\delta) + \exp(-\eta\delta)}{2} - 1\right)\right),
\end{aligned}
$$

where the last inequality holds due to $A/2 \leq m \leq A$. Therefore, we have

$$
\mathrm{SubOpt}_1(\bar{\pi}) + \mathrm{SubOpt}_2(\bar{\pi}) \geq \eta^{-1}\log\left(1 + \frac{M}{(1+M)^2}\left(\frac{\exp(\eta\delta) + \exp(-\eta\delta)}{2} - 1\right)\right). \tag{E.3}
$$

Now we aim to lower bound (E.3). We pick $\delta = \sqrt{SAC^* n^{-1}\log^{-1} A}/32$. We know that

$$
n \geq \frac{\eta^2 SC^* A}{1024\log A} \quad \Rightarrow \quad \eta\delta \leq 1.
$$

Now, the lower bound of M is straightforward as follows

$$
M = A^{-1}\sum_{j=m+1}^{A}\exp(\eta r^*(j)) + (C-1)e^{-\eta\alpha} = A^{-1}\sum_{j=m+1}^{A}\exp(\eta r^*(j)) + 1 \geq 1.
$$

On the other hand, recall that we can upper bound M as follows:

$$
M = A^{-1}\sum_{j=m+1}^{A}\exp(\eta r^*(j)) + 1 \leq 1 + \frac{1}{2}\exp(\eta\delta).
$$

By the fact that $f(x) = x(1+x)^{-2}$ is monotonically decreasing when $x \geq 1$, we know that

$$
\frac{M}{(1+M)^2} \geq \frac{1}{(2 + \exp(\eta\delta)/2)^2} \geq \frac{1}{16}.
$$

Recall that $(\mathrm{e}^x + \mathrm{e}^{-x})/2 - 1 = x^2 \sum_{k=0}^{\infty} \frac{x^{2k}}{(2k+2)!} \geq x^2/2$ for all $x \in \mathbb{R}$, which implies that

$$\begin{aligned}\mathrm{SubOpt}_1(\widehat{\pi}) + \mathrm{SubOpt}_2(\widehat{\pi}) &\geq \eta^{-1} \log\left(1 + \frac{MA^{-1}m}{(A^{-1}m + M)^2}\eta^2\delta^2\right) \\ &\geq \eta^{-1}\log\left(1 + \frac{1}{16}\eta^2\delta^2\right).\end{aligned}$$

Since $\eta\delta \leq 1$, we know that $\eta^2\delta^2/16 \leq 1$. By the fact that $\log(1+x) \geq x/2$ for all $x \in [0,1]$, we know that

$$\mathrm{SubOpt}_1(\widehat{\pi}) + \mathrm{SubOpt}_2(\widehat{\pi}) \geq \eta^{-1}\frac{\eta^2\delta^2}{32} \geq \frac{1}{32}\eta\delta^2.$$

Now we consider all contexts $s \in \mathcal{S}$. Consider two different reward functions r_1 and r_2, then for any $\widehat{\pi} : \mathcal{S} \to \Delta(\mathcal{A})$

$$\begin{aligned}&\mathrm{SubOpt}(\widehat{\pi}, r_1) + \mathrm{SubOpt}(\widehat{\pi}, r_2) \\ &= \frac{1}{S}\sum_{s\in\mathcal{S}} \big(\mathrm{SubOpt}_s(\widehat{\pi}(\cdot|s), r_1(s,\cdot)) + \mathrm{SubOpt}_s(\widehat{\pi}(\cdot|s), r_2(s,\cdot))\big) \\ &\geq \frac{1}{64}\eta\delta^2,\end{aligned}$$

where the last inequality holds due to our construction that on at least $S/2$ contexts r_1 and r_2 differs on at least $A/2$ actions. Now we invoke Fano's inequality (Lemma G.6) to obtain

$$\begin{aligned}&\inf_{\pi}\sup_{\mathrm{inst}\in\mathrm{CB}_{\mathcal{G}}} \mathrm{SubOpt}(\widehat{\pi}; \mathrm{inst}) \\ &\geq \frac{1}{128}\eta\delta^2\left(1 - \frac{\max_{r_1\neq r_2\in\mathcal{G}} \mathsf{KL}\left(P_{\mathcal{D}_{r_1}} \| P_{\mathcal{D}_{r_2}}\right) + \log 2}{\log|\mathcal{V}|}\right) \\ &\geq \frac{1}{128}\eta\delta^2\left(1 - \frac{512n\delta^2 + 64\log 2}{C^*SA}\right),\end{aligned}$$

where we use the condition that $\delta \leq 1/4$ and $\mathsf{KL}\left(\mathsf{Bern}(p)\|\mathsf{Bern}(q)\right) \leq (p-q)^2/\big(q(1-q)\big)$. Recall that $\delta = \sqrt{SAC^*n^{-1}\log^{-1}A}/32$, and thus for any $\widehat{\pi}$ we have

$$\sup_{\mathrm{inst}\in\mathrm{CB}_{\mathcal{G}}} \mathrm{SubOpt}(\widehat{\pi}; \mathrm{inst}) \gtrsim \frac{\eta C^* SA}{n\log A},$$

which finishes the proof. □

E.2. Proof of Theorem 5.2

Proof of Theorem 5.2. The proof largely extends from the proof of Theorem C.1. Given the context size S, size of action set $|\mathcal{A}| = 2A+1$ and upper bound of coverage coefficient $C \in [2, e^{\eta/2}/2]$, we consider the KL-regularized contextual bandit (with Gaussian reward) class parameterized by $\boldsymbol{\mu} \in \mathbb{R}^{S\times 2A}$, such that for any $s \in \mathcal{S}$, $\boldsymbol{\mu}(s) \in \{1/2, 1/2+\delta\}^{2A}$ and half of the entries of $\boldsymbol{\mu}(s)$ are $1/2+\delta$ and half of the entries are $1/2$. The multi-armed contextual bandit corresponding to $\boldsymbol{\mu}$ is given as follows. For each context $s \in \mathcal{S}$, the mean reward is given by $r_{\boldsymbol{\mu}}(\cdot, s) = \boldsymbol{\mu}(s) \oplus \{1/2 - \alpha\}$, that is, the expected reward of first $2A$ arms are given by $\boldsymbol{\mu}(s)$ and those on last arm is given by $1/2 - \alpha$, where α will be specified later. The reference policy is given by

$$\pi^{\mathsf{ref}}(a|s) = \frac{1}{2AC} \text{ if } a \in [2A] \text{ else } \pi^{\mathsf{ref}}(a|s) = \frac{C-1}{C}.$$

Given a context s and instance $\boldsymbol{\mu}$, we use $a^*(\boldsymbol{\mu}(s))$ and $\boldsymbol{\mu}^*(s)$ (interchangeably) to denote all the set of optimal arms of $\boldsymbol{\mu}$ under context s. Given a context s, we use the notation $\boldsymbol{\mu}(s) \sim_{i,j} \boldsymbol{\lambda}(s)$ if $a^*(\boldsymbol{\mu}(s)) \triangle a^*(\boldsymbol{\lambda}(s)) = \{i, j\}$ with $i \in a^*(\boldsymbol{\mu}(s))$ and $j \in a^*(\boldsymbol{\lambda}(s))$. We use $\mu(s)_i$ to denote the i-th entry of $\boldsymbol{\mu}(s)$, and when no ambiguity arises we also use $\mu(s)_i$ to denote

the distribution with mean $\mu(s)_i$. Furthermore, we use the notation $\boldsymbol{\mu} \sim_{s,(i,j)} \boldsymbol{\lambda}$ if $\boldsymbol{\mu}(s) \sim_{i,j} \boldsymbol{\lambda}(s)$ and $\boldsymbol{\mu}(s') = \boldsymbol{\lambda}(s')$ for all $s' \neq s$. Given an algorithm and a dataset, suppose the final output policy is $\widehat{\pi} \in \Delta(\mathcal{A}|\mathcal{S})$. We use $\widehat{a}$ to denote a random variable sampled from $\widehat{\pi}(\cdot|s)$, that is, $\widehat{a} \sim \widehat{\pi}(\cdot|s)$, for a given s. We use $\widehat{\pi} \sim \boldsymbol{\mu}$ if $\widehat{\pi}$ is obtained by running some given algorithm on instance $\boldsymbol{\mu}$. Let $\alpha = \eta^{-1}\log(C-1) + \eta^{-1}\log 2$. Then, for each context, the optimal policy on one optimum is given by

$$\pi^*(a^*) = \frac{e^{\eta\delta}/(2CA)}{(2C)^{-1}e^{\eta\delta} + (2C)^{-1} + (C-1)C^{-1}e^{-\eta\alpha}} = A^{-1}\frac{e^{\eta\delta}}{e^{\eta\delta}+2},$$

which gives that $C^{\pi^*} \leq 2C$.

We consider two instances that $\boldsymbol{\mu} \sim_{s,(i,j)} \boldsymbol{\lambda}$. We show that if $\mathbb{P}_{\boldsymbol{\mu},\widehat{\pi}(\cdot|s)}[\widehat{a}=i]$ turns out to be large, then $\mathbb{P}_{\boldsymbol{\lambda},\widehat{\pi}(\cdot|s)}[\widehat{a}=i]$, on which $\widehat{\pi}$ makes a mistake, is also somewhat large. In particular, by Proposition 7.1, for any event $\mathcal{E}$,

$$\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}) \geq e^{-\gamma}\left[\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}) - \mathbb{P}_{\boldsymbol{\mu}}\left(\log\frac{\mathrm{d}\mathbb{P}_{\boldsymbol{\mu}}}{\mathrm{d}\mathbb{P}_{\boldsymbol{\lambda}}} > \gamma\right)\right]. \tag{E.4}$$

When the noise is standard Gaussian with variance 1, given a trajectory in which the number of i-th arm's occurrence under context s is $N_{n,s,i}$ and their reward empirical mean is $\widehat{r}_{s,i}$, the log-likelihood ratio can be computed as follows:

$$\log\frac{\mathrm{d}\mathbb{P}_{\boldsymbol{\mu}}}{\mathrm{d}\mathbb{P}_{\boldsymbol{\lambda}}} = \sum_{x\in\mathcal{S}}\sum_{k\in\mathcal{A}} N_{n,x,k}\mathsf{KL}\left(\mu(x)_k\|\lambda(x)_k\right) + \sum_{x\in\mathcal{S}}\sum_{k\in\mathcal{A}} N_{n,x,k}(\mu(x)_k - \lambda(x)_k)(\widehat{r}_{x,k} - \mu_{x,k}),$$

For the first term, each item in the summation is a Bernoulli random variable with mean $(SCA)^{-1}$. We apply the following Bernstein-type inequality (Lemma G.3):

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\mu}}\Big[\sum_{x\in\mathcal{S}}\sum_{k\in\mathcal{A}} N_{n,x,k}\mathsf{KL}\left(\mu(x)_k\|\lambda(x)_k\right) > 2\mathsf{KL}(\mathbb{P}_{\boldsymbol{\mu}}\|\mathbb{P}_{\boldsymbol{\lambda}})\Big] \\
&\leq \exp\left(-\frac{4n^2(SCA)^{-2}}{4n(SCA)^{-1} + 4n(SCA)^{-1}/3}\right) \\
&\leq \exp\left(-\frac{n}{2SCA}\right).
\end{aligned} \tag{E.5}$$

For the second term, we first show that $N_{n,s,i} + N_{n,s,j} \leq 4n(SCA)^{-1}$ with high probability, which again follows from applying Bernstein's inequality

$$\mathbb{P}_{\boldsymbol{\mu}}[N_{n,s,i} + N_{n,s,j} > 4n(SCA)^{-1}] \leq \exp\left(-\frac{n}{2SCA}\right).$$

Let $T = N_{n,s,i} + N_{n,s,j}$ and $\mathcal{E}_{T,n} \coloneqq \{T \leq 4n(SCA)^{-1}\}$. Applying Hoeffding's inequality (Lemma G.2) gives

$$\begin{aligned}
&\mathbb{P}_{\boldsymbol{\mu}}\Big[\sum_{x\in\mathcal{S}}\sum_{k\in\mathcal{A}} N_{n,x,k}(\mu(x)_k - \lambda(x)_k)(\widehat{r}_{x,k} - \mu_{x,k}) > \beta\Big] \\
&\leq \mathbb{P}_{\boldsymbol{\mu}}\Big[\mathcal{E}_{T,n}^{\mathsf{C}}\Big] + \mathbb{P}_{\boldsymbol{\mu}}\Big[\Big\{\sum_{x\in\mathcal{S}}\sum_{k\in\mathcal{A}} N_{n,x,k}(\mu(x)_k - \lambda(x)_k)(\widehat{r}_{x,k} - \mu_{x,k}) > \beta\Big\}\bigcap\mathcal{E}_{T,n}\Big] \\
&\leq \exp\left(-\frac{n}{2SCA}\right) + \exp\left(-\frac{SCA\beta^2}{4n\delta^2}\right).
\end{aligned} \tag{E.6}$$

Set $\gamma \leftarrow 2\mathsf{KL}(\mathbb{P}_{\boldsymbol{\mu}}\|\mathbb{P}_{\boldsymbol{\lambda}}) + \beta$ in Equation (E.4), then a union bound over Equation (E.6) and Equation (E.5) gives

$$\mathbb{P}_{\boldsymbol{\mu}}\left[\log\frac{\mathrm{d}\mathbb{P}_{\boldsymbol{\mu}}}{\mathrm{d}\mathbb{P}_{\boldsymbol{\lambda}}} > 2\mathsf{KL}(\mathbb{P}_{\boldsymbol{\mu}}\|\mathbb{P}_{\boldsymbol{\lambda}}) + \beta\right] \leq 2\exp\left(-\frac{n}{2SCA}\right) + \exp\left(-\frac{SCA\beta^2}{4n\delta^2}\right).$$

Combining everything together, we obtain that

$$\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}) \geq \exp\left(-2\mathsf{KL}(\mathbb{P}_{\boldsymbol{\mu}}\|\mathbb{P}_{\boldsymbol{\lambda}}) - \beta\right)\left[\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}) - 2\exp\left(-\frac{n}{2SCA}\right) - \exp\left(-\frac{SCA\beta^2}{4n\delta^2}\right)\right]$$

$$= \exp\Big(-\frac{2n\delta^2}{SCA} - \beta\Big)\Big[\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}) - 2\exp\Big(-\frac{n}{2SCA}\Big) - \exp\Big(-\frac{SCA\beta^2}{4n\delta^2}\Big)\Big]. \tag{E.7}$$

Since $n \geq SAC\log A$ by assumption, we obtain under the premise of $\delta \asymp \sqrt{SAC(n\log A)^{-1}}$ and $\beta = 4$ that

$$2\exp\Big(-\frac{n}{2SCA}\Big) \leq \exp\Big(-\frac{SCA\beta^2}{4n\delta^2}\Big).$$

Let $\mathcal{E}_{s,i} = \{\widehat{a} = i : \widehat{a} \sim \widehat{\pi}(\cdot|s)\}$. For any $s \in \mathcal{S}, i \in [2A]$, define the bipartite graph $G_{s,i}$ with left vertices $L_{s,i} = \{\boldsymbol{\mu} \in \binom{[2A]}{A}^S : i \in \boldsymbol{\mu}(s)\}$ and right vertices $R_{s,i} = \{\boldsymbol{\lambda} \in \binom{[2A]}{A}^S : i \notin \boldsymbol{\lambda}(s)\}$; and connect $(\mu, \lambda) \in L_{s,i} \times R_{s,i}$ if $\exists j \in [2A]\backslash\{i\}$ such that $\boldsymbol{\mu} \sim_{s,(i,j)} \boldsymbol{\lambda}$. Then by construction, each $\mu \in L_{s,i}$ and $\lambda \in R_{s,i}$ has exactly A neighbors, i.e., $G_{s,i}$ is a A-regular bipartite graph, which admits a perfect matching $M_{s,i} \subset L_{s,i} \times R_{s,i}$ by Lemma G.8.[6] Note that any pair $(\boldsymbol{\mu}, \boldsymbol{\lambda}) \in M_{s,i}$ fits the template Equation (E.7), which implies

$$\begin{aligned}
&\sum_{(\boldsymbol{\mu},\boldsymbol{\lambda})\in M_{s,i}} \mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_{s,i}) \\
&\geq \sum_{(\boldsymbol{\mu},\boldsymbol{\lambda})\in M_{s,i}} \exp\Big(-\frac{2n\delta^2}{SCA} - \beta\Big)\Big[\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_{s,i}) - 2\exp\Big(-\frac{n}{2SCA}\Big) - \exp\Big(-\frac{SCA\beta^2}{4n\delta^2}\Big)\Big].
\end{aligned}$$

Therefore, additionally summing over $s \in \mathcal{S}$ and $i \in [2A]$, we obtain

$$\begin{aligned}
\sum_i \sum_s \sum_{\boldsymbol{\lambda}: i\in[2A]\backslash\boldsymbol{\lambda}^*(s)} \mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_{s,i}) \geq \sum_i \sum_s \sum_{\boldsymbol{\mu}: i\in\boldsymbol{\mu}^*(s)} &\exp\Big(-\frac{2n\delta^2}{SCA} - \beta\Big)\times \\
&\Big[\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_{s,i}) - 2\exp\Big(-\frac{n}{2SCA}\Big) - \exp\Big(-\frac{SCA\beta^2}{4n\delta^2}\Big)\Big].
\end{aligned}$$

Interchanging the order of summations on both sides yields

$$\begin{aligned}
&\sum_{\boldsymbol{\lambda}} \sum_{s\in\mathcal{S}} \sum_{i\in[2A]\backslash\boldsymbol{\lambda}^*(s)} \mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_{s,i}) \\
&\geq \exp\Big(-\frac{2n\delta^2}{SCA} - \beta\Big)\Big[\sum_{\boldsymbol{\mu}} \sum_{s\in\mathcal{S}} \sum_{i\in\boldsymbol{\mu}^*(s)} \mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_{s,i}) - SA\binom{2A}{A}^S 2\exp\Big(-\frac{SCA\beta^2}{4n\delta^2}\Big)\Big] \\
&= \exp\Big(-\frac{2n\delta^2}{SCA} - \beta\Big)\times \\
&\quad\Big[\sum_{\boldsymbol{\lambda}} \sum_{s\in\mathcal{S}} \Big(1 - \sum_{i\notin\boldsymbol{\lambda}^*(s)} \mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_{s,i})\Big) - SA\binom{2A}{A}^S 2\exp\Big(-\frac{SCA\beta^2}{4n\delta^2}\Big)\Big],
\end{aligned}$$

which further implies that

$$\sum_{\boldsymbol{\lambda}} \sum_{s\in\mathcal{S}} \sum_{i\notin a^*(\boldsymbol{\lambda}(s))} \mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_{s,i}) \geq \exp\Big(-\frac{2n\delta^2}{SCA} - \beta\Big)\binom{2A}{A}^S\Big[1 - 2A\exp\Big(-\frac{SCA\beta^2}{4n\delta^2}\Big)\Big]S.$$

As a consequence, we know that there exists some $\boldsymbol{\lambda}$ such that

$$\sum_{s\in\mathcal{S}} \sum_{i\notin a^*(\boldsymbol{\lambda}(s))} \mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_{s,i}) \geq \exp\Big(-\frac{2n\delta^2}{SCA} - \beta\Big)\Big[1 - 2A\exp\Big(-\frac{SCA\beta^2}{4n\delta^2}\Big)\Big]S.$$

Now we set $\delta = \sqrt{SCA(n\log A)^{-1}}$ and let $\beta = 4$, then we have

$$\exp\Big(-\frac{2n\delta^2}{SCA} - \beta\Big) = \exp\Big(-\frac{2}{\log A} - 4\Big) \geq \exp(-6),$$

[6] We suppress the last arm with reward $1/2 - \alpha$ in this argument to avoid notation clutter. And hence for each $G_{s,i}$, $|L_{s,i}| = |R_{s,i}| = \binom{2A}{A}^S \cdot \binom{2A-1}{A-1}$.

$$\exp\left(-\frac{SCA\beta^2}{4n\delta^2}\right) = \exp\big(-4\log A\big) = A^{-4}.$$

Thus gives

$$\sum_{s\in\mathcal{S}}\sum_{i\notin a^*(\boldsymbol{\lambda}(s))} \mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_{s,i}) \geq \frac{S}{2}\exp(-6) \geq S\exp(-7).$$

We notice that $\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_{s,i}) = \mathbb{E}_{\widehat{\pi}\sim\boldsymbol{\lambda},\widehat{a}\sim\widehat{\pi}(\cdot|s)}[\mathbb{1}\{\widehat{a}=i\}] = \mathbb{E}_{\boldsymbol{\lambda}}[\widehat{\pi}(i|s)]$, which gives that

$$\begin{aligned}\mathbb{E}_{\widehat{\pi}\sim\boldsymbol{\lambda}}\mathbb{E}_{s\sim\mathsf{Unif}(\mathcal{S})}\left[\sum_{a\notin a^*(\boldsymbol{\lambda}(s))}\widehat{\pi}(a|s)\right] &= \frac{1}{S}\sum_{s\in\mathcal{S}}\sum_{i\notin a^*(\boldsymbol{\lambda}(s))}\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_{s,i}) \\ &\geq \exp(-7). \end{aligned} \tag{E.8}$$

Now, we consider the suboptimality gap for the instance $\boldsymbol{\lambda}$ satisfying (E.8). Let $\pi^*_{\boldsymbol{\lambda}}$ be the optimal policy under the instance $\boldsymbol{\lambda}$, i.e., $\pi^*_{\boldsymbol{\lambda}}(\cdot|s) \propto \pi^{\mathsf{ref}}(\cdot|s)\exp\big(\eta r_{\boldsymbol{\lambda}}(\cdot,s)\big)$. For any context $s\in\mathcal{S}$, let $a^*_{\boldsymbol{\lambda}}(s)$ be the set of optimal arms for the instance $\boldsymbol{\lambda}$. We will omit the subscript $\boldsymbol{\lambda}$ when it will not cause any confusion. Recall that $\widehat{\pi}$ is the policy produced by the algorithm after interacting with the problem instance. We write $\widehat{\pi}\sim\boldsymbol{\lambda}$ to indicate that the policy is obtained from interaction with instance $\boldsymbol{\lambda}$. Thus, for any given algorithm, the expected suboptimality gap under instance $\boldsymbol{\lambda}$ can be written as $\mathbb{E}_{\widehat{\pi}\sim\boldsymbol{\lambda}}\big[\mathrm{SubOpt}(\widehat{\pi},\pi^*_{\boldsymbol{\lambda}})\big]$. Using Lemma G.9 and the data processing inequality (Polyanskiy & Wu, 2025, Theorem 2.17), we know that

$$\begin{aligned}\eta\mathbb{E}_{\widehat{\pi}\sim\boldsymbol{\lambda}}\big[\mathrm{SubOpt}(\widehat{\pi},\pi^*_{\boldsymbol{\lambda}})\big] &= \mathbb{E}_{\widehat{\pi}\sim\boldsymbol{\lambda}}\big[\mathsf{KL}(\widehat{\pi}\|\pi^*_{\boldsymbol{\lambda}})\big] \\ &\geq \mathbb{E}_{\widehat{\pi}\sim\boldsymbol{\lambda}}\mathbb{E}_{s\sim\mathsf{Unif}(\mathcal{S})}\left[\widehat{\pi}(\widetilde{a}|s)\log\frac{\widehat{\pi}(\widetilde{a}|s)}{\pi^*_{\boldsymbol{\lambda}}(\widetilde{a}|s)} + \widehat{\pi}(a^*|s)\log\frac{\widehat{\pi}(a^*|s)}{\pi^*_{\boldsymbol{\lambda}}(a^*|s)}\right].\end{aligned}$$

where for any policy π, we use the shorthand notation $\pi(\widetilde{a}|s) \coloneqq \sum_{a\notin a^*}\pi(a|s)$ and $\pi(a^*|s) \coloneqq \sum_{a\in a^*}\pi(a|s)$ to denote the total probability assigned to suboptimal and optimal arms, respectively. Note that for any $s\in\mathcal{S}$, we have $\pi^*_{\boldsymbol{\lambda}}(\widetilde{a}|s) = 2(2+e^{\eta\delta})^{-1}$ and $\pi^*_{\boldsymbol{\lambda}}(a^*|s) = e^{\eta\delta}(2+e^{\eta\delta})^{-1}$, which are independent of s. Hence, we drop the dependence on s and write $\pi^*_{\boldsymbol{\lambda}}(\widetilde{a})$, $\pi^*_{\boldsymbol{\lambda}}(a^*)$ without ambiguity. Applying Jensen's inequality to the convex function $f(x) = x\log(x/a)$, where a is a constant, we have

$$\begin{aligned}\eta\mathbb{E}_{\widehat{\pi}\sim\boldsymbol{\lambda}}\big[\mathrm{SubOpt}(\widehat{\pi},\pi^*_{\boldsymbol{\lambda}})\big] \geq{}& \mathbb{E}_{\widehat{\pi}\sim\boldsymbol{\lambda}}\mathbb{E}_{s\sim\mathsf{Unif}(\mathcal{S})}[\widehat{\pi}(\widetilde{a}|s)]\log\frac{\mathbb{E}_{\widehat{\pi}\sim\boldsymbol{\lambda}}\mathbb{E}_{s\sim\mathsf{Unif}(\mathcal{S})}[\widehat{\pi}(\widetilde{a}|s)]}{\pi^*_{\boldsymbol{\lambda}}(\widetilde{a})} \\ &+ \mathbb{E}_{\widehat{\pi}\sim\boldsymbol{\lambda}}\mathbb{E}_{s\sim\mathsf{Unif}(\mathcal{S})}[\widehat{\pi}(a^*|s)]\log\frac{\mathbb{E}_{\widehat{\pi}\sim\boldsymbol{\lambda}}\mathbb{E}_{s\sim\mathsf{Unif}(\mathcal{S})}[\widehat{\pi}(a^*|s)]}{\pi^*_{\boldsymbol{\lambda}}(a^*)}.\end{aligned} \tag{E.9}$$

Therefore, the suboptimality gap can be lower bounded by the KL divergence between two Bernoulli variables, with mean $\mathbb{E}_{\widehat{\pi}\sim\boldsymbol{\lambda}}\mathbb{E}_{s\sim\mathsf{Unif}(\mathcal{S})}[\widehat{\pi}(a^*|s)]$ and $\pi^*_{\boldsymbol{\lambda}}(\widetilde{a})$. Our next step is to show that $\mathbb{E}_{\widehat{\pi}\sim\boldsymbol{\lambda}}\mathbb{E}_{s\sim\mathsf{Unif}(\mathcal{S})}[\widehat{\pi}(\widetilde{a}|s)]$ is larger than $\pi^*_{\boldsymbol{\lambda}}(\widetilde{a}) = 2(2+e^{\eta\delta})^{-1}$. Specifically, we have

$$n \leq \frac{\eta^2 SAC}{81\log A} \Rightarrow \eta\sqrt{\frac{SAC}{n\log A}} = \eta\delta \geq 9.$$

Therefore, we know that

$$\pi^*(\widetilde{a}|s) = \frac{2}{2+e^{\eta\delta}} \leq 2\exp(-9) \leq \exp(-7) \overset{\textit{Equation (E.8)}}{\leq} \mathbb{E}_{\widehat{\pi}\sim\boldsymbol{\lambda}}\mathbb{E}_{s\sim\mathsf{Unif}(\mathcal{S})}[\widehat{\pi}(\widetilde{a}|s)].$$

As a result, we can plug (E.8) into Equation (E.9) and compute as follows,

$$\begin{aligned}&\eta\mathbb{E}_{\widehat{\pi}\sim\boldsymbol{\lambda}}\big[\mathrm{SubOpt}(\widehat{\pi},\boldsymbol{\lambda})\big] \\ &\geq \mathsf{KL}\Big(\mathrm{Bern}\big(\mathbb{E}_{\widehat{\pi}\sim\boldsymbol{\lambda}}\mathbb{E}_{s\sim\mathsf{Unif}(\mathcal{S})}[\widehat{\pi}(\widetilde{a}|s)]\big)\;\Big\|\;\mathrm{Bern}\big(\pi^*(\widetilde{a}|\cdot)\big)\Big) \\ &\geq \mathsf{KL}\Big(\mathrm{Bern}\big(\exp(-7)\big)\;\Big\|\;\mathrm{Bern}\big(\pi^*(\widetilde{a}|\cdot)\big)\Big)\end{aligned}$$

$$\geq \underbrace{\exp\big(-7\big)\log\frac{\exp\big(-7\big)}{2(2+\exp(\eta\delta))^{-1}}}_{I_1} + \underbrace{\log\frac{1-\exp\big(-7\big)}{1-2(2+\exp(\eta\delta))^{-1}}}_{I_2}, \tag{E.10}$$

where the second inequality holds due to $\mathsf{KL}\left(\mathrm{Bern}(p_1)\|\mathrm{Bern}(q)\right) \geq \mathsf{KL}\left(\mathrm{Bern}(p_2)\|\mathrm{Bern}(q)\right)$ whenever $1 > p_1 \geq p_2 \geq q > 0$ and the third inequality holds due to $1-\exp(-7) \leq 1$ and $1-\exp(-7) \leq 1-2(2+\exp(\eta\delta)^{-1}$. For the first term I_1, we have

$$\begin{aligned} I_1 &= \exp\big(-7\big)\log\frac{\exp\big(-7\big)}{2(2+\exp(\eta\delta))^{-1}} \\ &\geq \exp\big(-7\big)\log\frac{\exp\big(-7\big)\exp(\eta\delta)}{2} \\ &\geq \exp\big(-7\big)\big(\eta\delta - \log 2 - 7\big). \end{aligned} \tag{E.11}$$

For the second term I_2, we have

$$\begin{aligned} I_2 &= -\log\frac{1-2(2+\exp(\eta\delta))^{-1}}{1-\exp\big(-7\big)} \\ &= -\log\left(1+\frac{\exp\big(-7\big)-2(2+\exp(\eta\delta))^{-1}}{1-\exp\big(-7\big)}\right) \\ &\geq -\frac{\exp\big(-7\big)-2(2+\exp(\eta\delta))^{-1}}{1-\exp\big(-7\big)} \\ &\geq -2\exp\big(-7\big), \end{aligned} \tag{E.12}$$

where the first inequality holds due to $\log(1+x) \leq x$, and the last inequality is by $1-\exp\big(-7\big) \geq 1/2$ and $2(2+\exp(\eta\delta))^{-1} \geq 0$. Substituting Equations (E.11) and (E.12) into Equation (E.10) yields

$$\mathbb{E}_{\widehat{\pi}\sim\boldsymbol{\lambda}}\big[\mathrm{SubOpt}(\widehat{\pi},\boldsymbol{\lambda})\big] \geq \eta^{-1}\exp\big(-7\big)\cdot\big(\eta\delta - 7 - 2 - \log 2\big)$$

Since $n \leq \eta^2 SAC/(400\log A)$, we know that

$$n \leq \frac{\eta^2 SAC}{400\log A} \Rightarrow \eta\sqrt{\frac{SAC}{n\log A}} = \eta\delta \geq 20,$$

which provides that $\eta\delta/2 \geq 10 \geq 9+\log 2$. Subsequently, we have

$$\begin{aligned} \mathbb{E}_{\widehat{\pi}\sim\boldsymbol{\lambda}}\big[\mathrm{SubOpt}(\widehat{\pi},\boldsymbol{\lambda})\big] &\geq \eta^{-1}\exp\big(-7\big)\cdot\big(\eta\delta - 7 - 2 - \log 2\big) \\ &\geq \eta^{-1}\exp\big(-7\big)\cdot\frac{\eta\delta}{2} \\ &\gtrsim \sqrt{\frac{SAC}{n\log A}}, \end{aligned}$$

which concludes the proof. □

F. Proof of Theorems in Appendix C

F.1. Proof of Theorem C.1

Proof of Theorem C.1. We consider the multi-armed bandit class parameterized by mean vectors:

$$\boldsymbol{\mu} \in \mathcal{V}_K := \left\{\boldsymbol{\mu} \in \{\pm 1\}^A \,\middle|\, \sum_{i=1}^{A} \mathbb{1}(\boldsymbol{\mu}_i = -1) = K\right\},$$

i.e., exactly K entries of $\boldsymbol{\mu}$ are -1. Unless stated otherwise, we assume that noise of rewards is standard Gaussian. For any $\boldsymbol{\mu} \in \mathcal{V}_K$, the corresponding instance is given by $([A], r_{\boldsymbol{\mu}}, \mathcal{D})$ where the reward is given by $r_{\boldsymbol{\mu}}(i) = \boldsymbol{\mu}_i\delta$, with $\delta > 0$ to be specified later. For any $\boldsymbol{\mu}$, we denote by $a^*(\boldsymbol{\mu})$ the set of optimal actions (those with reward δ). For two instances $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$, we write $\boldsymbol{\mu} \sim_{i,j} \boldsymbol{\lambda}$ if $a^*(\boldsymbol{\mu}) \triangle a^*(\boldsymbol{\lambda}) = \{i, j\}$ with $i \in a^*(\boldsymbol{\mu})$ and $j \in a^*(\boldsymbol{\lambda})$, i.e., instances $\boldsymbol{\mu}$ and $\boldsymbol{\lambda}$ differ only in that action i is optimal under $\boldsymbol{\mu}$ while action j is optimal under $\boldsymbol{\lambda}$. When there is no ambiguity, we also write $\boldsymbol{\mu}_i$ for the distribution with mean $\boldsymbol{\mu}_i\delta$ and standard Gaussian noise. Given an algorithm and a dataset, let $\widehat{\pi}$ denote the output policy. We write $\widehat{a} \sim \widehat{\pi}$ for a random action sampled from $\widehat{\pi}$, and $\widehat{\pi} \sim \boldsymbol{\mu}$ to indicate that $\widehat{\pi}$ is produced by running the algorithm on instance $\boldsymbol{\mu}$.

Consider two instances that $\boldsymbol{\mu} \sim_{i,j} \boldsymbol{\lambda}$. We show that if $\mathbb{P}_{\boldsymbol{\mu}}[\widehat{\pi} = i]$ is large then $\mathbb{P}_{\boldsymbol{\lambda}}[\widehat{\pi} = i]$ is also large, resulting in an error on $\boldsymbol{\lambda}$. By Proposition 7.1, we know that for any event $\mathcal{E}$,

$$\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}) \geq e^{-\gamma}\left[\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}) - \mathbb{P}_{\boldsymbol{\mu}}\left[\log\frac{\mathrm{d}\mathbb{P}_{\boldsymbol{\mu}}}{\mathrm{d}\mathbb{P}_{\boldsymbol{\lambda}}} > \gamma\right]\right]. \tag{F.1}$$

We first compute the term $\log \mathrm{d}\mathbb{P}_{\boldsymbol{\mu}}/\,\mathrm{d}\mathbb{P}_{\boldsymbol{\lambda}}$. Specifically, consider a dataset $\mathcal{D} = (a_1, r_1, \ldots, a_n, r_n)$. Let $p_{\boldsymbol{\mu}}(\mathcal{D})$ denote the joint density of $\mathcal{D}$ under the bandit instance $\boldsymbol{\mu}$, and define $p_{\boldsymbol{\lambda}}(\mathcal{D})$ analogously. Then,

$$\frac{\mathrm{d}\mathbb{P}_{\boldsymbol{\mu}}}{\mathrm{d}\mathbb{P}_{\boldsymbol{\lambda}}} = \frac{p_{\boldsymbol{\mu}}(\mathcal{D})}{p_{\boldsymbol{\lambda}}(\mathcal{D})} = \prod_{i=1}^{n}\frac{p_{\boldsymbol{\mu}}(a_i, r_i)}{p_{\boldsymbol{\lambda}}(a_i, r_i)},$$

where the last equation holds because the samples in $\mathcal{D}$ are independent and identically distributed. Moreover, using Bayes' rule and the fact that the reward distribution is standard Gaussian, we have

$$\begin{aligned}
\log\prod_{i=1}^{n}\frac{p_{\boldsymbol{\mu}}(a_i, r_i)}{p_{\boldsymbol{\lambda}}(a_i, r_i)} &= \log\prod_{i=1}^{n}\frac{p_{\boldsymbol{\mu}}(r_i|a_i)}{p_{\boldsymbol{\lambda}}(r_i|a_i)} \\
&= \sum_{k\in\mathcal{A}}\sum_{i=1}^{n}\mathbb{1}(a_i = k)\log\frac{p_{\boldsymbol{\mu}}(r_i|a_i)}{p_{\boldsymbol{\lambda}}(r_i|a_i)} \\
&= \sum_{k\in\mathcal{A}}\sum_{i=1}^{n}\mathbb{1}(a_i = k)\log\frac{\exp[-(r_i - \mu_k)^2/2]}{\exp[-(r_i - \lambda_k)^2/2]} \\
&= \frac{1}{2}\sum_{k\in\mathcal{A}}\sum_{i=1}^{n}\mathbb{1}(a_i = k)\Big[(r_i - \lambda_k)^2 - (r_i - \mu_k)^2\Big] \\
&= \frac{1}{2}\sum_{k\in\mathcal{A}}\sum_{i=1}^{n}\mathbb{1}(a_i = k)\Big[2(r_i - \mu_k)(\mu_k - \lambda_k) + (\mu_k - \lambda_k)^2\Big].
\end{aligned}$$

Therefore, the logarithmic Radon–Nikodym derivative can be represented as

$$\begin{aligned}
\log\frac{\mathrm{d}\mathbb{P}_{\boldsymbol{\mu}}}{\mathrm{d}\mathbb{P}_{\boldsymbol{\lambda}}} &= \frac{1}{2}\sum_{k\in\mathcal{A}}\sum_{i=1}^{n}\mathbb{1}(a_i = k)\Big[2(r_i - \mu_k)(\mu_k - \lambda_k) + (\mu_k - \lambda_k)^2\Big] \\
&= \underbrace{\frac{1}{2}\sum_{k\in\mathcal{A}}N(k)(\mu_k - \lambda_k)^2}_{I_1} + \underbrace{\sum_{k\in\mathcal{A}}N(k)(\mu_k - \lambda_k)(\widehat{r}_k - \mu_k)}_{I_2},
\end{aligned}$$

where $N(k) = \sum_{i=1}^{n}\mathbb{1}(a_i = k)$ denotes the number of occurrences of arm k, and $\widehat{r}_k = \sum_{i=1}^{n}\mathbb{1}(a_i = k)r_i/N(k)$ denotes the empirical mean reward of arm k. For I_1, we first recall our choice of $\boldsymbol{\mu} \sim_{i,j} \boldsymbol{\lambda}$, such that the two instances only differ in two arms, with $|\mu_k - \lambda_k| = 2\delta$ if and only if $k \in \{i, j\}$. Therefore, we have

$$I_1 = 2\delta^2\big[N(i) + N(j)\big].$$

Using standard concentration inequalities (Lemma G.5), we show that $N(i) + N(j) = O(n/A)$ with high probability. Specifically, by Lemma G.5, we have

$$\mathbb{P}[N(i) \geq 2n/A] \leq \exp\Big(-\frac{n}{3A}\Big),$$

$$\mathbb{P}[N(j) \geq 2n/A] \leq \exp\Big(-\frac{n}{3A}\Big),$$

Then, taking a union bound, we have

$$\mathbb{P}[N(i)+N(j) \geq 4n/A] \leq 2\exp\Big(-\frac{n}{3A}\Big).$$

Let $\mathcal{E}_{i,j;n} := \{N(i)+N(j) \leq 4n/A\}$. We have seen $\mathbb{P}[\mathcal{E}_{i,j;n}^{\mathsf{C}}] \leq 2\exp(-n/[3A])$. Moreover,

$$\mathbb{P}[I_1 \geq 8n\delta^2/A] \leq 2\exp\Big(-\frac{n}{3A}\Big). \tag{F.2}$$

For I_2, we have

$$\begin{aligned} I_2 &= \sum_{k\in\mathcal{A}} N(k)(\mu_k-\lambda_k)(\widehat{r}_k-\mu_k) \\ &= 2\delta\sum_{s=1}^{n}\big[\,\mathbb{1}(a_s=i)(r_s-\mu_i) - \mathbb{1}(a_s=j)(r_s-\mu_j)\big], \end{aligned}$$

where the last term is the summation of $[N(i)+N(j)]$ independent Gaussians. Note that the selected actions and the noise of rewards are independent. We can use Hoeffding's inequality (Lemma G.4), conditioned on the event $\mathcal{E}_{i,j;n}$. Thus, we have

$$\mathbb{P}_{\boldsymbol{\mu}}\Big[\{I_2 \geq 2\beta\delta\}\bigcap\mathcal{E}_{i,j;n}\Big] \leq \exp\Big(-\frac{A\beta^2}{8n}\Big).$$

Using the union bound, we have

$$\begin{aligned} \mathbb{P}_{\boldsymbol{\mu}}\big[I_2 \geq 2\beta\delta\big] &\leq \mathbb{P}_{\boldsymbol{\mu}}[\mathcal{E}_{i,j;n}^{\mathsf{C}}] + \mathbb{P}_{\boldsymbol{\mu}}\Big[\{I_2 \geq 2\beta\delta\}\bigcap\mathcal{E}_{i,j;n}\Big] \\ &\leq 2\exp\Big(-\frac{n}{3A}\Big) + \exp\Big(-\frac{A\beta^2}{8n}\Big). \end{aligned} \tag{F.3}$$

Combining (F.2) and (F.3), we get

$$\begin{aligned} \mathbb{P}_{\boldsymbol{\mu}}\left[\log\frac{\mathrm{d}\mathbb{P}_{\boldsymbol{\mu}}}{\mathrm{d}\mathbb{P}_{\boldsymbol{\lambda}}} > \frac{8n\delta^2}{A}+2\beta\delta\right] &\leq \mathbb{P}_{\boldsymbol{\mu}}\left[I_1 \leq \frac{8n\delta^2}{A}\right] + \mathbb{P}_{\boldsymbol{\mu}}[I_2 \leq 2\beta\delta] \\ &\leq 4\exp\left(-\frac{n}{3A}\right) + \exp\left(-\frac{A\beta^2}{8n}\right) \\ &\leq 5\exp\left(-\frac{A\beta^2}{8n}\right), \end{aligned}$$

where the last inequality holds *under the premise of* $n \geq \beta A$. Now we plug in the above inequality back to the change-of-measure argument (F.1) with $\gamma = 8n\delta^2/A + 2\beta\delta$. Then, we obtain that

$$\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}) \geq \exp\big(-(8n\delta^2/A+2\beta\delta)\big)\left[\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}) - 5\exp\left(-\frac{A\beta^2}{8n}\right)\right].$$

Let $\mathcal{E}_i = \{\widehat{a}=i : \widehat{a}\sim\widehat{\pi}(\mathcal{D})\}$, we traverse over every $\boldsymbol{\mu}$, $\boldsymbol{\lambda}$, i and j such that $\boldsymbol{\mu}\sim_{i,j}\boldsymbol{\lambda}$ and sum up the inequality above. For each $(\boldsymbol{\lambda}, i\notin a^*(\boldsymbol{\lambda}))$, there exists $(A-K)$ j such that the reward on i and j to be switched to get $\boldsymbol{\mu}$ such that $\boldsymbol{\mu}\sim_{i,j}\boldsymbol{\lambda}$. Therefore, each $(\boldsymbol{\lambda}, i\notin a^*(\boldsymbol{\lambda}))$ appears $(A-K)$ times. Similarly, each $(\boldsymbol{\mu}, j\in a^*(\boldsymbol{\mu}))$ appears K times and we have $\binom{A}{K}(A-K)K$ inequalities in total. These give the multiplicative factors in front of each term:

$$\begin{aligned} &(A-K)\sum_{\boldsymbol{\lambda}}\sum_{i\notin a^*(\boldsymbol{\lambda})}\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_i) \\ &\geq \exp\left(-\frac{8n\delta^2}{A}-2\beta\delta\right)\left[K\sum_{\boldsymbol{\mu}}\sum_{i\in a^*(\boldsymbol{\mu})}\mathbb{P}_{\boldsymbol{\mu}}(\mathcal{E}_i) - 5K(A-K)\binom{A}{K}\exp\left(-\frac{A\beta^2}{8n}\right)\right] \end{aligned}$$

$$\geq \exp\bigg(-\frac{8n\delta^2}{A}-2\beta\delta\bigg)K\bigg[\sum_{\boldsymbol{\lambda}}\bigg(1-\sum_{i\notin a^*(\boldsymbol{\lambda})}\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_i)\bigg)-5(A-K)\binom{A}{K}\exp\bigg(-\frac{A\beta^2}{8n}\bigg)\bigg],$$

which implies that

$$\begin{aligned}
&\sum_{\boldsymbol{\lambda}}\sum_{i\notin a^*(\boldsymbol{\lambda})}\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_i)\\
&\quad\geq \frac{K\exp\big(-8n\delta^2/A-2\beta\delta\big)}{A-K+K\exp\big(-8n\delta^2/A-2\beta\delta\big)}\binom{A}{K}\bigg[1-5(A-K)\exp\bigg(-\frac{A\beta^2}{8n}\bigg)\bigg]\\
&\quad\geq \exp\bigg(-\frac{8n\delta^2}{A}-2\beta\delta\bigg)\frac{K}{A}\binom{A}{K}\bigg[1-5(A-K)\exp\bigg(-\frac{A\beta^2}{8n}\bigg)\bigg],
\end{aligned}$$

where the second inequality holds due to $\exp\big(-8n\delta^2/A-2\beta\delta\big)\leq 1$. Now we select $\beta=4\sqrt{n\log A/A}$ and $\delta=\sqrt{A(n\log A)^{-1}}$, we have

$$\begin{aligned}
\exp\bigg(-\frac{8n\delta^2}{A}-2\beta\delta\bigg)&=\exp\bigg(-\frac{8}{\log A}-8\bigg)\geq e^{-16},\\
\exp\bigg(-\frac{A\beta^2}{8n}\bigg)&=\exp\bigg(-2\log A\bigg)=A^{-2}.
\end{aligned}$$

Consequently, we get

$$\sum_{\boldsymbol{\lambda}}\sum_{i\notin a^*(\boldsymbol{\lambda})}\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_i)\gtrsim\frac{K}{A}\binom{A}{K}(1-5(A-K)A^{-2})\gtrsim\frac{K}{A}\binom{A}{K}.$$

Therefore, we know that there exists an $\boldsymbol{\lambda}\in\mathcal{V}_K$ such that

$$\sum_{i\notin a^*(\boldsymbol{\lambda})}\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_i)\gtrsim\frac{K}{A}\binom{A}{K}(1-5(A-K)A^{-2})\gtrsim\frac{K}{A}.$$

Recall that $\mathbb{E}_{\boldsymbol{\lambda}}[\mathrm{SubOpt}(\widehat{\pi})]=\delta\sum_{i\notin a^*(\boldsymbol{\lambda})}\mathbb{P}_{\boldsymbol{\lambda}}(\mathcal{E}_i)$, we obtain that there exists an $\boldsymbol{\lambda}\in\mathcal{V}_K$ such that

$$\mathbb{E}_{\boldsymbol{\lambda}}[\mathrm{SubOpt}(\widehat{\pi})]\gtrsim\frac{K\delta}{A}\geq\frac{K}{\sqrt{nA\log A}},$$

which finishes the proof. □

F.2. Proof of Theorem C.2

Proof of Theorem C.2. Without loss of generality, we consider the optimal reward $r^*=0$ here. Then, by our bounded reward assumption, we have $r(a)\geq -1$ for all $a\in\mathcal{A}$. We assume $\mathcal{A}=[A]$ and $\mathcal{A}^*=[A-K]$. We further denote the suboptimality gap on arm k to be $2\delta_k$, or $r(k)=-2\delta_k$, where $1/2>\delta_k>0$ for all $k>A-K$. Thus, the suboptimality gap of output policy $\widehat{\pi}=\delta(\widehat{a})$ is $2\delta_{\widehat{a}}$ and therefore the expected suboptimality gap is given by

$$\mathbb{E}[\mathrm{SubOpt}(\pi)]=2\sum_{k>A-K}\mathbb{P}[\widehat{a}=k]\delta_k.$$

Our goal is to show that $\mathbb{E}[\mathrm{SubOpt}(\pi)]\leq 2\epsilon$ as long as

$$n\geq 8A\log\bigg(\frac{8A}{\epsilon}\bigg)+\frac{64K^2\log A}{A\epsilon^2}=\widetilde{O}(K^2/A\epsilon^2).$$

We first show that if $\delta_k=O(A\epsilon/K)$, then the suboptimality gap on these arms accumulates to at most an order of ϵ, thanks to having sufficiently many optimal arms. In particular, since our samples a_i and reward noise ϵ_i are drawn independently

and symmetrically with respect to $\mathcal{A}$, we know that for all $i > A-K$ and $j \le A-K$, the probability that $\bar{r}(i)$ takes the maximum is no greater than the probability that $\bar{r}(j)$ takes the maximum, since $r(i) < r(j)$. This implies that

$$\sum_{i>A-K} \mathbb{P}[\widehat{a}=i] \le \frac{K}{A}.$$

Therefore, the total suboptimality gap on those suboptimal arm k with $\delta_k \le A\epsilon/(4K)$ results in at most $\epsilon/4$ suboptimality gap.

$$\begin{aligned}\mathbb{E}\big[\mathrm{SubOpt}(\pi)\big] &= 2\sum_{k>A-K} \mathbb{P}[\widehat{a}=k]\,\mathbb{1}\{\delta_k \le A\epsilon/(4K)\}\delta_k \\ &\quad + 2\sum_{k>A-K} \mathbb{P}[\widehat{a}=k]\,\mathbb{1}\{\delta_k > A\epsilon/(4K)\}\delta_k \\ &\le \frac{\epsilon}{2} + \underbrace{2\sum_{k>A-K} \mathbb{P}[\widehat{a}=k]\,\mathbb{1}\{\delta_k > A\epsilon/(4K)\}\delta_k}_{X}, \end{aligned} \tag{F.4}$$

Therefore, we only need to consider the expected suboptimality incurred by those arm where $\delta_k = \Omega(\epsilon A/K)$, denoted by X in (F.4). For each of these arms, since the reward gap is large enough, the probability that one of them is selected as the maximum of empirical mean is small. Specifically, we consider a superset of the event $\widehat{a}=k$ for a given k:

$$\mathcal{E}_k = \{\bar{r}(k) \ge -\delta_k\} \cup \big\{\min_{i\in\mathcal{A}^*} \bar{r}(i) \le \delta_k\big\} \subseteq \{\bar{r}(k) \ge -\delta_k\} \cup \{\bar{r}(1) \le \delta_k\}.$$

It is easy to see that $\{\widehat{a}=k\} \subseteq \mathcal{E}_k$, since on the complement of $\mathcal{E}_k$, $\bar{r}(k) < -\delta_k$, but on the first arm (which is optimal) we have $\bar{r}(1) > \delta_k > \bar{r}(k)$, so k cannot be selected. Now, let $\mathcal{X}_k = \{\bar{r}(k) \ge -\delta_k\}$ and $\mathcal{U} = \{\bar{r}(1) \le \delta_k\}$. Then, $\mathbb{P}[\widehat{a}=k] \le \mathbb{P}(\mathcal{X}_k \cup \mathcal{U}) \le \mathbb{P}(\mathcal{X}_k) + \mathbb{P}(\mathcal{U})$. We first bound the probability of $\mathcal{U}$. Since $N(1) \sim \mathsf{Bin}(n, 1/A)$, by Lemma G.1, we have

$$\mathbb{P}\Big[N(1) \le \frac{n}{2A}\Big] \le \exp\Big(-\frac{n}{8A}\Big).$$

Conditioning on $N(1) \ge n/(2A)$, by Lemma G.2, we know that

$$\mathbb{P}[\bar{r}(1) \le -\delta_k] \le \exp\left(-\frac{N(1)^2\delta_k^2}{2N(1)}\right) \le \exp\left(-\frac{n^2\delta_k^2}{4A}\right).$$

Taking a union bound, we know that

$$\mathbb{P}(\mathcal{U}) \le \exp\left(-\frac{n}{8A}\right) + \exp\left(-\frac{n^2\delta_k^2}{4A}\right).$$

Noticing that $\mathcal{X}_k$ and $\mathcal{U}$ are symmetric, we can repeat the above argument to obtain an upper bound of $\mathbb{P}[\mathcal{X}_k]$:

$$\mathbb{P}(\mathcal{X}_k) \le \exp\left(-\frac{n}{8A}\right) + \exp\left(-\frac{n^2\delta_k^2}{4A}\right).$$

Taking a union bound over $\mathcal{X}_k$ and $\mathcal{U}$, we obtain that

$$\mathbb{P}(\mathcal{E}_k) \le 2\exp\left(-\frac{n}{8A}\right) + 2\exp\left(-\frac{n^2\delta_k^2}{4A}\right).$$

Now we are ready to bound the second term X in (F.4). Specifically, we have

$$\begin{aligned} X &= 2\sum_{k>A-K} \mathbb{P}[\widehat{a}=k]\,\mathbb{1}\{\delta_k > A\epsilon/(4K)\}\delta_k \\ &\le 4\sum_{k>A-K} \mathbb{1}\{\delta_k > A\epsilon/(4K)\}\left[\exp\left(-\frac{n}{8A}\right) + \exp\left(-\frac{n^2\delta_k^2}{4A}\right)\right]\delta_k \end{aligned}$$

$$= \underbrace{4 \sum_{k>A-K} \mathbb{1}\{\delta_k > A\epsilon/(4K)\} \exp\Big(- \frac{n}{8A}\Big)\delta_k}_{\text{(i)}}$$
$$+ \underbrace{4 \sum_{k>A-K} \mathbb{1}\{\delta_k > A\epsilon/(4K)\} \exp\Big(- \frac{n^2\delta_k^2}{4A}\Big)\delta_k}_{\text{(ii)}}.$$

For the first term (i), our selection of n ensures $n \geq 8A\log(8A/\epsilon)$, resulting in $\exp(-n/[8A]) \leq \epsilon/(8K)$. Consequently, (i) can be bounded as

$$\text{(i)} \leq 4 \sum_{k>A-K} \delta_k \frac{\epsilon}{8K} \leq \frac{\epsilon}{4}, \tag{F.5}$$

where the last inequality holds due to $\delta_k \leq 1/2$. For the second term (ii), we rewrite $\delta_k = C_k A\epsilon/(4K)$, where $C_k \geq 1$. For fixed k, we have

$$\exp\Big(- \frac{n^2\delta_k^2}{4A}\Big)\delta_k = \exp\Big(- \frac{n^2 A C_k^2 \epsilon^2}{64K^2}\Big)\frac{C_k A\epsilon}{4K}.$$

By our choice of n, we have $n \geq 64K^2\log A/(A\epsilon^2)$, which gives

$$\exp\Big(- \frac{n^2 A C_k^2 \epsilon^2}{64K^2}\Big)\frac{C_k A\epsilon}{4K} \leq \exp\big(- C_k^2 \log A\big)\frac{C_k A\epsilon}{4K} = \frac{\epsilon}{4K} C_k A^{1-C_k^2} \leq \frac{\epsilon}{4K},$$

where the last inequality holds due to xA^{1-x^2} is monotonically decreasing w.r.t. x in $[1,\infty)$ when $A \geq 2$. Therefore, we know that

$$\text{(ii)} = 4 \sum_{k>A-K} \mathbb{1}\{\delta_k > A\epsilon/(4K)\} \exp\Big(- \frac{n^2\delta_k^2}{4A}\Big)\delta_k \leq \epsilon. \tag{F.6}$$

Therefore, combining (F.4), (F.5) and (F.6), we obtain that $\mathrm{SubOpt}(\pi) \leq 2\epsilon$ given

$$n = 8A\log\Big(\frac{8A}{\epsilon}\Big) + \frac{64K^2\log A}{A\epsilon^2} = \widetilde{O}\Big(\frac{K^2}{A\epsilon^2}\Big),$$

which finishes the proof. □

G. Auxiliary Lemmas

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

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

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

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

Lemma G.2 (Azuma-Hoeffding's inequality, Azuma 1967). *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*

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

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

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

Remarkably, these inequalities also holds for martingale differences with sub-gaussian tails and we refer to Shamir (2011) for a detailed derivation.

Lemma G.3 (Freedman's inequality, Freedman 1975)**.** *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 $|Z_i - Z_{i-1}| < M$, and $\mathbb{E}[(Z_i - Z_{i-1})^2] \le \sigma_i^2$, then*

$$\mathbb{P}[Z_n - Z_0 \ge t] \le \exp\left(- \frac{t^2}{2\sum_{i=1}^n \sigma_i^2 + 2Mt/3} \right).$$

Lemma G.4 (Hoeffding's inequality)**.** *Suppose that $\{\boldsymbol{X}_i, i = 1, \ldots, n\}$ are independent. Each $\boldsymbol{X}_i$ has mean μ_i and sub-Gaussian parameter σ_i. Then, for all $t \ge 0$, we have*

$$\mathbb{P}\left[\sum_{i=1}^n (\boldsymbol{X}_i - \mu_i) \ge t \right] \le \exp\left\{ \frac{-t^2}{2\sum_{i=1}^n \sigma_i^2} \right\}.$$

Lemma G.5 (Chernoff Bounds)**.** *Assume that $\{\boldsymbol{X}_i, i = 1, \ldots, n\}$ are i.i.d random variables. Moreover, $\mathbb{E}[\boldsymbol{X}_1] = \mu$, $\boldsymbol{X}_i \in [0, 1]$. Then, for all $\epsilon > 0$,*

$$\mathbb{P}\left[\frac{\sum_{i=1}^n \boldsymbol{X}_i}{n} \ge (1+\epsilon)\mu \right] \le \exp\left[\frac{-n\mu\epsilon^2}{2+\epsilon} \right],$$
$$\mathbb{P}\left[\frac{\sum_{i=1}^n \boldsymbol{X}_i}{n} \le (1-\epsilon)\mu \right] \le \exp\left[\frac{-n\mu\epsilon^2}{2} \right].$$

Lemma G.6 is a standard reduction to Fano's inequality (Le Cam, 1973; Yu, 1997; Polyanskiy & Wu, 2025). See, e.g., Chen et al. (2024, Section 3) for a general proof.

Lemma G.6 (Fano's inequality)**.** *Fix any $\mathcal{R} \coloneqq \{r_1, \cdots, r_S\}$ and policy class Π, and let $L : \Pi \times \mathcal{R} \to \mathbb{R}_+$ be some loss function. Suppose there exists a constant $c > 0$ such that the following condition holds:*

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

Then we have

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

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

The following Lemma G.7 is due to Gilbert (1952); Varshamov (1957), which is a classical result in coding theory. We refer the readers to Theorem 4.2.1 in Guruswami et al. (2019) for more details.

Lemma G.7. *Suppose Σ is a set of characters with $|\Sigma| = q$ where $q \ge 2$ is a prime power and $N > 0$ is some natural number. Then there exists a subset $\mathcal{V}$ of Σ^N such that (1) for any $v, v' \in \mathcal{V}, v \ne v'$, one has $d_H(v, v') \ge N/2$ and (2) $\log_q |\mathcal{V}|/N \ge 1 - H_q(1/2) = \Theta(1) \ge 1/8$, where d_H is the Hamming distance (i.e., the number of different entries) and the entropy function H is given by*

$$H_q(x) = x\frac{\log(q-1)}{\log q} - x\frac{\log x}{\log q} - (1-x)\frac{\log(1-x)}{\log q}.$$

In particular, when $q = 2$, this means that there exists a subset $\mathcal{V}$ of $\{-1, 1\}^S$ such that (1) $|\mathcal{V}| \ge \exp(S/8)$ and (2) for any $v, v' \in \mathcal{V}, v \ne v'$, one has $\|v - v'\|_1 \ge S/2$.

Recall that a regular bipartite graph is a bipartite graph in which all vertices have the same degree.

Lemma G.8 (Bondy & Murty 1979, Corollary 5.2)**.** *Every regular bipartite graph with non-empty edge set has a perfect matching.*

We adapt the following folklore, see, e.g., Zhao et al. (2026, (D.10)).

Lemma G.9. *Let $\mathcal{A}$ be a finite action set, $\eta > 0$, and $r : \mathcal{A} \to \mathbb{R}$ be some reward function. Let $\pi^{\mathsf{ref}} \in \Delta(\mathcal{A})$ be some reference policy and $\pi^* \in \Delta(\mathcal{A})$ be the optimal policy under r, i.e., $\pi^*(a) \propto \pi^{\mathsf{ref}}(a)\exp(\eta r(a))$ for all $a \in \mathcal{A}$. For any policy π, the suboptimality gap between π and π^* under the KL-regularized objective is given by $\mathrm{SubOpt}(\pi, \pi^*) = \eta^{-1}\mathsf{KL}(\pi \| \pi^*)$.*