Sharp Analysis for KL-Regularized Contextual Bandits and RLHF

Heyang Zhao
University of California, Los Angeles
Los Angeles, CA 90095
hyzhao@cs.ucla.edu

Chenlu Ye
University of Illinois Urbana-Champaign
Urbana, IL 61801
chenluy3@illinois.edu

Quanquan Gu
University of California, Los Angeles
CA 90095, USA
qgu@cs.ucla.edu

Tong Zhang
University of Illinois Urbana-Champaign
Urbana, IL 61801
tongzhang@tongzhang-ml.org

Abstract

Reverse-Kullback-Leibler (KL) regularization has emerged to be a predominant technique to enhance policy optimization in reinforcement learning (RL) and reinforcement learning from human feedback (RLHF), which forces the learned policy to stay close to a reference policy. While the effectiveness of KL-regularization has been empirically demonstrated in various practical scenarios, current theoretical analyses of KL-regularized RLHF still yield the same $\mathcal{O}(1/\epsilon^2)$ sample complexity as ones without KL-regularization. To understand the fundamental distinction between objectives with KL-regularization and ones without KL-regularization, we are the first to theoretically demonstrate the power of KL-regularization by providing a sharp analysis for KL-regularized contextual bandits and RLHF, revealing an $\mathcal{O}(1/\epsilon)$ sample complexity when ϵ is sufficiently small. We also prove matching lower bounds for both settings. More specifically, we study how the coverage of the reference policy affects the sample complexity of KL-regularized online contextual bandits and RLHF. We show that with sufficient coverage from the reference policy, a simple two-stage mixed sampling algorithm can achieve an $\mathcal{O}(1/\epsilon)$ sample complexity with only an additive dependence on the coverage coefficient, thus proving the benefits of online data even without explicit exploration. Our results provide a comprehensive understanding of the roles of KL-regularization and data coverage in online decision making, shedding light on the design of more efficient algorithms.

1 Introduction

The KL-regularized contextual bandit problem (Langford and Zhang, 2007; Xiong et al., 2024a) has raised tremendous interest recently because of the significant development of the post-training stage in large language models (LLMs) and diffusion models from preference feedback (Christiano et al., 2017; Ziegler et al., 2019; Ouyang et al., 2022; Bai et al., 2022; Rafailov et al., 2024), which is called *Reinforcement Learning from Human Feedback* (RLHF). RLHF aims to optimize the policy by aligning it with human feedback, exhibiting impressive capabilities in applications such as Chatgpt (Achiam et al., 2023), Claude (Anthropic, 2023), Gemini (Team et al., 2023), and LLaMA-3 (Meta, 2024).

In RLHF, we treat the language model as a policy that takes a prompt x and produces a response a conditioned on x, optimizing the policy by aligning it with human feedback. There are mainly two

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

kinds of feedback: absolute rating and preference comparison. For absolute rating, the collection typically involves human annotators to provide rating scores like 1 to 5 (Wang et al., 2024a,b) for responses or hard 0-1 scores for math reasoning tasks since the reasoning tasks often have gold standard answers (Cobbe et al., 2021; Hendrycks et al., 2021; Xiong et al., 2024b). Although discrete feedback is believed to be more intuitive for human users and easier to collect, it also poses more challenges for the RLHF algorithms to effectively leverage the feedback signals since the reward signals are not directly observed. In practice, the learning process typically involves (a) constructing a reward model based on the maximum likelihood estimation (MLE) of *Bradley-Terry* (BT) (Bradley and Terry, 1952a) model from the preference feedback; (b) applying RL algorithms like PPO (Schulman et al., 2017b) to train the language model so that it maximizes the reward signals with KL regularization (Ouyang et al., 2022; Bai et al., 2022; Touvron et al., 2023).

Since the human feedback data only covers a tiny fraction of possible interactions, if we optimize the model purely for the reward without constraints, it might learn behaviors that work well for the training feedback but fail catastrophically on slightly different inputs. For example, the policy may generate disproportionate bold words or emoji to please the learned reward (Zhang et al., 2024). Hence, the KL-regularization between the learned policy and a reference policy (the pre-trained model after supervised fine-tuning) plays a fundamental role in RLHF to avoid overfitting. There is a line of theoretical RLHF work that modeled the problem as a reverse-KL regularized contextual bandit (Xiong et al., 2024a; Ye et al., 2024a; Zhong et al., 2024; Wu et al., 2024; Xie et al., 2024). However, they adopt the techniques from contextual bandits and neglect the power of reverse-KL-regularization, thus obtaining almost the same $\mathcal{O}(1/\epsilon^2)$ [1] sample complexity as learning objectives without KL-regularization. Therefore, the question of

whether there exists a fundamental distinction between bandit learning objectives with and without KL-regularization

is still largely under-explored.

Additionally, an emerging line of offline RLHF literature highlights the coverage of the reference policy π_0. The coverage of π_0 refers to the ability of the model to generate diverse responses for a wide range of prompts. A model with good coverage can generalize well to unseen contexts and actions, which is essential for the learned reward function to also generalize well. In practice, this is evidenced by the fact that the simple best-of-n sampling based on π_0 is competitive with the well-tuned PPO algorithm for general open-ended conversation tasks (Dong et al., 2023), and the fact that the π_0 can solve a majority of the math problems with multiple responses (Shao et al., 2024; Nakano et al., 2021). While the coverage of π_0 is recognized as a key factor in offline RLHF, its impact on the sample complexity of online RLHF is still largely unknown. Thus, it is natural to ask:

If online RLHF is theoretically more efficient than offline RLHF under strong coverage of π_0?

In this paper, we answer the above questions by (1) providing a novel fine-grained decomposition for the suboptimiality of objective functions, which adapts to the strongly convex optimization landscape of the reverse-KL regularization and obtains a sharper sample complexity than the existing results, and (2) proposing an efficient 2-stage mixed sampling strategy for online RLHF with good coverage of π_0, which achieves sample complexity with only an additive dependence on the coverage coefficient. In contrast, the existing RLHF algorithms typically require a multiplicative dependence on the coverage coefficient.

1.1 Our Contributions

In this work, we make a first attempt to illustrate the statistical benefits of KL-regularization for contextual bandits and RLHF. Our main contributions are summarized as follows:

- In Section 2, we study the contextual bandit problem with KL-regularization, which also serves as a mathematical formulation for RLHF with absolute-rating feedback. We provide a lower bound for the KL-regularized contextual bandit problem, which indicates that the sample complexity of the problem is $\Omega(\eta \log N_{\mathcal{R}}(\epsilon)/\epsilon)$ when ϵ is sufficiently small, where $N_{\mathcal{R}}(\epsilon)$ is the covering number of the reward function class and η is the KL-regularization coefficient.

[1] For simplicity, we omit here the dependencies on quantities other than ϵ.

- We povide a novel analysis to upper bound the suboptimality gap of the KL-regularized objective in contextual bandits, and propose a simple two-stage mixed sampling strategy to achieve a sample complexity of $\mathcal{O}(\max(\eta^2 D^2, \eta/\epsilon) \log N_{\mathcal{R}}(\epsilon/\delta))$ when the reward scale is a constant, where D is the coverage coefficient of the reference policy π_0 and δ is the confidence parameter. To the best of our knowledge, this is the first work to provide an $\mathcal{O}(1/\epsilon)$ sample complexity for KL-regularized contextual bandits.
- In Section 3, we extend our analysis to RLHF. We rigorously demonstrate that KL-regularization is essential for more efficient policy learning in RLHF with preference data. We further propose a two-stage mixed sampling strategy for online RLHF with good coverage of π_0, which achieves a sample complexity of $\mathcal{O}(\max(\eta^2 D^2, \eta/\epsilon) \log N_{\mathcal{R}}(\epsilon/\delta))$ when the reward scale is a constant.
- In Appendix C, we consider the setting where the coverage condition is not satisfied. We show that an active querying subroutine can be used to improve the coverage of the collected data in the first stage, which can be combined with our two-stage mixed sampling strategy to achieve a sample complexity of $\mathcal{O}(d \log N_{\mathcal{R}}(\epsilon/\delta)\eta^3/\epsilon)$, where d is the eluder dimension (Russo and Van Roy, 2013) of the reward function class.

1.2 Previous Understanding of KL Regularization in RL

Our analysis of KL-regularized contextual bandits and RLHF also contributes to the theoretical understanding of the impact of KL-regularization in RL since contextual bandits can be viewed as a simplified version of Markov decision processes (MDPs). In RL, KL-regularization has been widely used to stabilize the learning process and prevent the policy from deviating too far from the reference policy. Here, we provide a brief overview of the existing understanding of KL-regularization in decision-making problems. From the perspective of policy optimization, KL-regularization captures entropy regularization as a special case [2], which is also an extensively used technique in RL literature (Sutton, 2018; Szepesvári, 2022). There is a large body of literature that has explored the benefits of entropy regularization or KL-regularization in RL (Schulman et al., 2015; Fox et al., 2016; Schulman et al., 2017a; Haarnoja et al., 2017, 2018; Ahmed et al., 2019). Most related to our work, Ahmed et al. (2019) provided a comprehensive understanding of the role of entropy regularization in RL, showing that entropy regularization can improve the training efficiency and stability of the policy optimization process by changing the optimization landscape through experiments on continuous control tasks (Brockman, 2016). Theoretically, Neu et al. (2017) provided a unified view of entropy regularization as approximate variants of Mirror Descent or Dual Averaging, and left the statistical justification for using entropy regularization in RL as an open question. Geist et al. (2019) provided a framework for analyzing the error propagation in regularized MDPs, which also focused on the proof of the convergence for the policy optimization methods with regularization and lacked a sharp sample complexity analysis.

2 KL-Regularized Contextual Bandits

In this section, we formally define the KL-regularized contextual bandit problem and provide a lower bound for the sample complexity of the problem. We then propose a novel two-stage mixed sampling strategy for online regularized bandits with good coverage of the reference policy π_0.

2.1 Problem Setup

In the contextual bandit setting, in each round t, the agent observes a context $x_t \in \mathcal{X}$ generated from a distribution d_0 and chooses an action $a_t \in \mathcal{A}$. The agent receives a stochastic reward $r_t \in \mathbb{R}$ depending on the context x_t and the action a_t. The goal is to maximize the expected cumulative reward over T rounds.

The learner has access to a family of reward functions $R(\theta, x, a)$ parameterized by $\theta \in \Theta$, such that there exists $\theta_* \in \Theta$ satisfying $\mathbb{E}[r_t | x_{1:t}, a_{1:t}] = R(\theta_*, x_t, a_t)$. WLOG, we assume that the reward feedback r_t at all rounds is a non-negative real number bounded by B. We consider a KL-regularized

[2] We can regard the entropy regularization as a special case of KL-regularization by setting the reference policy as the uniform distribution.

objective as follows:

$$Q(\pi) = \mathbb{E}_{x\sim d_0}\mathbb{E}_{a\sim\pi(\cdot|x)}\left[R(\theta_*, x, a) - \eta^{-1}\log\frac{\pi(a|x)}{\pi_0(a|x)}\right], \tag{2.1}$$

where π_0 is a known fixed policy, and $\eta > 0$ is a hyperparameter that controls the trade-off between maximizing rewards and staying close to the reference policy π_0.

Remark 2.1. In RLHF with the absolute-rating feedback, we can directly measure the quality of the responses by querying absolute reward value. For instance, in the NVIDIA Helpsteer project (Wang et al., 2023b, 2024c), human labelers are required to provide absolute score in five attributes: helpfulness, correctness, coherence, complexity, and verbosity.

The absolute-rating feedback is directly modeled as the stochastic reward in the contextual bandit setting (Wang et al., 2024a; Xiong et al., 2024b). Under the online RLHF setting, in each round t, the learner observes a prompt x_t (modeled as the context) and chooses a response a_t (modeled as the action). The learner then updates the model (policy) based on the absolute-rating feedback.

Remark 2.2. It is worth noting that entropy or Kullback-Leibler (KL) regularization is also widely used in contextual bandits (Berthet and Perchet, 2017; Wu et al., 2016) and deep RL algorithms (Schulman et al., 2015; Fox et al., 2016; Schulman et al., 2017a; Haarnoja et al., 2017, 2018), where KL-divergence regularization is a popular technique for preventing drastic updates to the policy. Algorithms such as Trust Region Policy Optimization (TRPO) (Schulman et al., 2015) explicitly incorporate KL-regularization to limit the policy updates during optimization, ensuring that the updated policy does not deviate too much from the current policy. This constraint promotes stable and reliable learning, particularly in high-dimensional state-action spaces. Additionally, KL-regularization is central to Proximal Policy Optimization (PPO) (Schulman et al., 2017a), where a penalty term involving KL-divergence ensures updates remain within "trust region".

Reward function class We consider a function class $\mathcal{R} = \{R(\theta, \cdot, \cdot)|\theta \in \Theta\}$ and for the realizability, we assume that the ground truth reward function $R(\theta_*, x, a)$ is in the function class $\mathcal{R}$. Then, we define the covering number of $\mathcal{R}$ as follows.

Definition 2.3 (ϵ-cover and covering number)**.** Given a function class $\mathcal{F}$, for each $\epsilon > 0$, an ϵ-cover of $\mathcal{F}$ with respect to $||\cdot||_\infty$, denoted by $\mathcal{C}(\mathcal{F}, \epsilon)$, satisfies that for any $f \in \mathcal{F}$, we can find $f' \in \mathcal{C}(\mathcal{F}, \epsilon)$ such that $||f - f'||_\infty \leq \epsilon$. The ϵ-covering number, denoted as $N_{\mathcal{F}}(\epsilon)$, is the smallest cardinality of such $\mathcal{C}(\mathcal{F}, \epsilon)$.

Planning oracle Given a reward model, we can learn the policy by optimizing the KL-regularized objective in (2.1). To simplify the analysis, we assume that there exists a planning oracle, which in empirical can be efficiently approximated by rejection sampling (Liu et al., 2023), Gibbs sampling (Xiong et al., 2024a), and iterative preference learning with a known reward (Dong et al., 2024).

Definition 2.4 (Policy Improvement Oracle)**.** For a reward function $R(\theta, \cdot, \cdot) \in \mathcal{R}$ and a reference policy π_0, for any prompt $x \sim d_0$, we can compute:

$$\pi_\theta^\eta(\cdot|x) := \underset{\pi(\cdot|x)\in\Delta(\mathcal{A})}{\operatorname{argmax}} \mathbb{E}_{a\sim\pi(\cdot|x)}\left[R(\theta, x, a) - \eta^{-1}\log\frac{\pi(a|x)}{\pi_0(a|x)}\right] \propto \pi_0(\cdot|x)\cdot\exp\left(\eta R(\theta, x, \cdot)\right).$$

Hence, the comparator policy is the solution to the oracle given the true reward function $R(\theta_*, \cdot, \cdot)$: $\pi^*(\cdot|\cdot) \propto \pi_0(\cdot|\cdot)\cdot\exp(\eta R(\theta_*, \cdot, \cdot))$. The **goal** is to minimize the sub-optimality of our learned policy $\widehat{\pi}$ with respect to π^*: $Q(\pi^*) - Q(\widehat{\pi})$.

Coverage conditions It is crucial to assume that our data-collector policy π_0 possesses good coverage, which can ensure that the learned reward function can generalize well to unseen contexts (prompts) and actions (responses), and thus can enable us to approximate the optimal policy.

Definition 2.5 (Data Coverage)**.** Given a reference policy π_0, D^2 is the minimum positive real number satisfying $\forall\ (x, a) \in \mathcal{X}\times\mathcal{A},\ \pi(a|x) > 0$, we have for any pair of $\theta, \theta' \in \Theta$,

$$\frac{[R(\theta', x, a) - R(\theta, x, a)]^2}{\mathbb{E}_{x'\sim d_0, a'\sim\pi_0(\cdot|x')}[(R(\theta', x', a') - R(\theta, x', a'))^2]} \leq D^2.$$

The coverage coefficient D measures how well the in-sample error induced by distribution $d_0 \times \pi_0$ can characterize the out-of-sample error. This concept is adapted from the F-design for online RL under general function approximation (Agarwal et al., 2024), and follows the coverage coefficient for offline RL (Di et al., 2023; Ye et al., 2024b), and the eluder dimension (Wang et al., 2020; Ye et al., 2023; Agarwal et al., 2023; Zhao et al., 2023a) for online RL. Take the linear model as an example, where the reward function is embedded into a d-dimensional vector space: $R(\theta, x, a) = \theta^\top \phi(x, a)$ for $\theta \in \mathbb{R}^d$. Let the covariance matrix $\Sigma = \mathbb{E}_{x\sim d_0}\mathbb{E}_{a\sim\pi_0(\cdot|x)}\phi(x, a)\phi(x, a)^\top$. Then, the coverage condition turns into $\sup_{\theta,\theta'\in\Theta} \frac{|(\theta'-\theta)^\top \phi(x,a)|^2}{(\theta'-\theta)^\top \Sigma(\theta'-\theta)} \leq \|\phi(x, a)\|^2_{\Sigma^{-1}} \leq D^2$, where the first inequality uses the Cauchy-Schwarz inequality. Hence, this quantity measures how much does the reference policy covers all directions of the feature space, and we can show that there exists π_0 such that $D^2 = O(d)$ through G-optimal design (Zhang, 2023; Lattimore and Szepesvári, 2020).

2.2 Lower Bound

In this subsection, we provide a lower bound for the KL-regularized contextual bandit problem.

Theorem 2.6. For any $\epsilon \in (0, 1/256), \eta > 4$, and any algorithm A, there exists a KL-regularized contextual bandit problem with reward function class $\mathcal{R}$ and $O(N_{\mathcal{R}}(\epsilon))$ data coverage coefficient (as defined in Definition 2.5) such that A requires at least $\Omega\big(\min(\frac{\eta \log N_{\mathcal{R}}(\epsilon)}{\epsilon}, \frac{\log N_{\mathcal{R}}(\epsilon)}{\epsilon^2})\big)$ rounds to achieve a suboptimality gap of ϵ.

Remark 2.7. The lower bound in Theorem 2.6 indicates that the sample complexity of the KL-regularized contextual bandit problem is $\Omega(\eta \log N_{\mathcal{R}}(\epsilon)/\epsilon)$ when ϵ is sufficiently small. In our proof, the KL-regularization term shifts the local landscape of the objective function, which prevents us to directly apply the standard bandit analysis, and thus requires a novel analysis to derive the new lower bound. This $\Omega(\eta \log N_{\mathcal{R}}(\epsilon)/\epsilon)$ lower bound suggests that the KL-regularized contextual bandit problem potentially enjoy a lower sample complexity compared to the standard contextual bandit.

2.3 The Proposed Algorithm

We present our algorithm in Algorithm 1 for the KL-regularized contextual bandit problem, which serves as a theoretical model for online RLHF with absolute-rating feedback. The algorithm consists of two stages:

- In the first stage, we sample m contexts (prompts) and actions (answers) from the foundation model π_0 and observe the corresponding rewards (absolute ratings). These ratings can be regarded as noisy observations of the underlying reward function $R(\theta_*, x, a)$. In line 6, we compute an estimate of the reward function $\widehat{\theta}_0$ using least squares regression based on the collected data. In line 7, we apply the planning oracle to obtain the policy $\pi^\eta_{\widehat{\theta}_0}$ which maximizes the following KL-regularized estimated objective in Definition 2.4 with reward function $R(\theta, \cdot, \cdot) = R(\widehat{\theta}_0, \cdot, \cdot)$.
- In the second stage, we utilize the trained policy $\pi^\eta_{\widehat{\theta}_0}$ to sample n contexts (prompts) and actions (responses). With the intermediate policy $\pi^\eta_{\widehat{\theta}_0}$, we can collect new data $\{(x_i, a_i, r_i)\}^n_{i=1}$ which is more aligned with the data distribution induced by the optimal policy π_*. In line 13, the algorithm combines data from both stages $\{(x_i, a_i, r_i)\}^n_{i=1}$ and $\{(x^0_i, a^0_i, r^0_i)\}^m_{i=1}$ to compute a refined least squares estimate $\widehat{\theta}$ of the reward function, minimizing the sum of squared errors across both datasets. By aggregating the two datasets together, there is an overlap between the data to compute $\widehat{\theta}$ and $\widehat{\theta}_0$, so that the output policy $\pi^\eta_{\widehat{\theta}}$ is well covered by the intermediate policy $\pi^\eta_{\widehat{\theta}_0}$.

2.4 Theoretical Guarantees

Review of previous analysis The previous analysis (e.g., Xiong et al., 2024a) basically follows the techniques of bandits and neglects the significance of KL-regularization. For simplicity, We use short-hand notation $R(\theta, x, \pi) = \mathbb{E}_{a\sim\pi(\cdot|x)}R(\theta, x, a)$ and denote $\mathrm{KL}(\pi(\cdot|x)\|\pi'(\cdot|x))$ by $\mathrm{KL}(\pi\|\pi')$ when there is no confusion. We make the estimation on a dataset $\{(x_i, a_i, r_i) : x_i \sim d_0, a_i \sim \pi_0(\cdot|x_i)\}^n_{i=1}$: $\pi^\eta_{\widehat{\theta}} = \mathrm{argmax}_{\pi\in\Pi} \mathbb{E}_{x\sim d_0}[R(\widehat{\theta}, x, \pi) - \eta^{-1}\mathrm{KL}(\pi\|\pi_0)]$, and has a small in-sample-

Algorithm 1 Two-stage Mixed-Policy Sampling (TMPS)

Input: η, ϵ, π_0, Θ.

▷ Stage 1: Use policy π_0 to achieve sufficient data coverage

for $i = 1, \ldots, m$ **do**

  Sample context $x_i^0 \sim d_0$ and action $a_i^0 \sim \pi_0(\cdot|x_i^0)$.

  Observe reward $r_i^0 = R(\theta_*, x_i^0, a_i^0) + \epsilon_i^0$, where ϵ_i^0 is the random noise.

end for

Compute the least square estimate of the reward function based on $D_0 = \{(x_i^0, a_i^0, r_i^0)\}_{i=1}^m$:

$$\widehat{\theta}_0 \leftarrow \operatorname*{argmin}_{\theta \in \Theta} \sum_{i=1}^{m} (R(\theta, x_i^0, a_i^0) - r_i^0)^2.$$

Apply the planning oracle to compute $\pi_{\widehat{\theta}_0}^{\eta}(\cdot|\cdot) \propto \pi_0(\cdot|\cdot) \exp\big(\eta R(\widehat{\theta}_0, \cdot, \cdot)\big)$.

▷ Stage 2: Use policy $\pi_{\widehat{\theta}_0}^{\eta}$ to sample new responses

for $i = 1, \ldots, n$ **do**

  Sample context $x_i \sim d_0$ and action $a_i \sim \pi_{\widehat{\theta}_0}^{\eta}(\cdot|x_i)$.

  Observe reward $r_i = R(\theta_*, x_i, a_i) + \epsilon_i$, where ϵ_i is the random noise.

end for

Compute the least square estimate of the reward function using $\{(x_i, a_i, r_i)\}_{i=1}^n$ together with D_0:

$$\widehat{\theta} \leftarrow \operatorname*{argmin}_{\theta \in \Theta} \sum_{i=1}^{m} (R(\theta, x_i^0, a_i^0) - r_i^0)^2 + \sum_{i=1}^{n} (R(\theta, x_i, a_i) - r_i)^2.$$

Output $\pi_{\widehat{\theta}}^{\eta}(\cdot|\cdot) \propto \pi_0(\cdot|\cdot) \exp\big(\eta R(\widehat{\theta}, \cdot, \cdot)\big)$.

error: $\mathbb{E}_{x\sim d_0}\mathbb{E}_{a\sim\pi_o(\cdot|x)}\big[(R(\widehat{\theta}, x, a) - R(\theta_*, x, a))^2\big] = O(1/n)$. The sub-optimality is decomposed as:

$$\begin{aligned}
Q(\pi^*) - Q(\pi_{\widehat{\theta}}^{\eta}) &= \mathbb{E}_{x\sim d_0}\big[R(\theta_*, x, \pi^*) - R(\widehat{\theta}, x, \pi^*)\big] + \mathbb{E}_{x\sim d_0}\big[R(\widehat{\theta}, x, \pi_{\widehat{\theta}}^{\eta}) - R(\theta_*, x, \pi_{\widehat{\theta}}^{\eta})\big] \\
&\quad + \mathbb{E}_{x\sim d_0}\big[R(\widehat{\theta}, x, \pi^*) - \eta^{-1}\mathrm{KL}(\pi^*\|\pi_0)\big] - \mathbb{E}_{x\sim d_0}\big[R(\widehat{\theta}, x, \pi_{\widehat{\theta}}^{\eta}) - \eta^{-1}\mathrm{KL}(\pi_{\widehat{\theta}}^{\eta}\|\pi_0)\big] \\
&\le \mathbb{E}_{x\sim d_0}\big[R(\theta_*, x, \pi^*) - R(\widehat{\theta}, x, \pi^*) + R(\widehat{\theta}, x, \pi_{\widehat{\theta}}^{\eta}) - R(\theta_*, x, \pi_{\widehat{\theta}}^{\eta})\big],
\end{aligned}$$

where the inequality holds since $\pi_{\widehat{\theta}}^{\eta}$ is the maximum.

Then, the suboptimality can be further bounded by using the coverage condition (Definition 2.10) and concentration inequalities:

$$\begin{aligned}
Q(\pi^*) - Q(\pi_{\widehat{\theta}}^{\eta}) &\le 2C_{\mathrm{GL}}\mathbb{E}_{x\sim d_0}\mathbb{E}_{a\sim\pi_0(\cdot|x)}\big[|R(\theta_*, x, a) - R(\widehat{\theta}, x, a)|\big] \\
&\le 2C_{\mathrm{GL}}\sqrt{\mathbb{E}_{a\sim\pi_0(\cdot|x)}\big[(R(\theta_*, x, a) - R(\widehat{\theta}, x, a))^2\big]} = O(C_{\mathrm{GL}}/\sqrt{n}).
\end{aligned}$$

Hence, they need $\Theta(C_{\mathrm{GL}}^2/\epsilon^2)$ sample complexity to ensure $O(\epsilon)$ sub-optimility.

Sharper results and analysis

Theorem 2.8. Suppose that Assumption 2.5 holds. For any $\delta \in (0, 1/5)$, $\epsilon > 0$ and constant $c_{m,n} > 0$, if we set $m = \Theta(\eta^2 D^2 \cdot B^2 \log(2N_{\mathcal{R}}(\epsilon_c)/\delta))$ and $n = \Theta(\eta/\epsilon \cdot B^2 \log(N_{\mathcal{R}}(\epsilon_c)/\delta))$ and $\epsilon_c = \min\{\frac{\epsilon}{2(1+c_{m,n}^{-1})B}, \frac{1}{8(1+c_{m,n})B\eta^2 D^2}\}$, then with probability at least $1 - 5\delta$ the output policy of Algorithm 1 $\pi_{\widehat{\theta}}^{\eta}$ is $\mathcal{O}(\epsilon)$ optimal.

Theorem 2.8 shows that the sample complexity of Algorithm 1 is $\mathcal{O}(\eta/\epsilon \log N_{\mathcal{R}}(\epsilon/\delta))$ when the reward scale is a constant and ϵ is sufficiently small. The result indicates that the proposed two-stage mixed sampling strategy can achieve a suboptimality gap of ϵ with only an additive dependence on the coverage coefficient D^2.

To illustrate the novel techniques to obtain the sharper bound, we highlight the crucial points in the sequel and defer the detailed proof to Appendix E.2.

Part I: Decomposition of the suboptimality gap The most challenging part is how to proceed with the suboptimality gap based on the strong convexity of the objective Q with the KL-regularization. Given the closed-form solution of $\pi^*(a|x) = \pi_0(a|x)\exp(\eta R(\theta, x, a))/Z^\eta_{\theta_*}(x)$ and $\pi^\eta_{\widehat{\theta}}(a|x) = \pi_0(a|x)\exp(\eta R(\widehat{\theta}, x, a))/Z^\eta_{\widehat{\theta}}(x)$, where $Z^\eta_\theta(x) = \sum_{a\in\mathcal{A}} \pi_0(a|x)\cdot\exp(\eta R(\theta, x, a))$ denotes the normalization constant, we can write the suboptimality as

$$\begin{aligned}
&\mathbb{E}_{\pi^*}\left[R(\theta_*, x, a) - \frac{1}{\eta}\log\frac{\pi^*(a|x)}{\pi_0(a|x)}\right] - \mathbb{E}_{\pi^\eta_{\widehat{\theta}}}\left[R^*(x, a) - \frac{1}{\eta}\log\frac{\pi^\eta_{\widehat{\theta}}(a|x)}{\pi_0(a|x)}\right] \\
&\quad= \frac{1}{\eta}\mathbb{E}_{\pi^*}\left[\log\frac{\pi_0(a|x)\exp(\eta R(\theta_*, x, a))}{\pi^*(a|x)}\right] - \frac{1}{\eta}\mathbb{E}_{\pi^\eta_{\widehat{\theta}}}\left[\log\frac{\pi_0(a|x)\exp(\eta R(\theta_*, x, a))}{\pi^\eta_{\widehat{\theta}}(a|x)}\right] \\
&\quad= -\frac{1}{\eta}\big(J(x;\widehat{\theta}) - J(x;\theta_*)\big),
\end{aligned}$$

where we define $J(x;\theta) = \log Z^\eta_\theta(x) - \eta\mathbb{E}_{\pi^\eta_\theta}[R(\theta, x, a) - R(\theta_*, x, a)]$, and the last equation is deduced by taking the distribution of π^* and $\pi^\eta_{\widehat{\theta}}$ in the terms.

Thus, the suboptimality is expressed by the gap between $\widehat{\theta}$ and θ_* with respect to the function J. By taking the first-order Taylor expansion with respect to $\{\Delta(x, a) = R(\widehat{\theta}, x, a) - R(\theta_*, x, a) : a \in \mathcal{A}\}$, we can prove the following lemma.

Lemma 2.9. For any estimator $\widehat{\theta} \in \Theta$, and the policy $\pi^\eta_{\widehat{\theta}}$ satisfying Definiton 2.4, we have

$$\begin{aligned}
Q(\pi^*) - Q(\pi^\eta_{\widehat{\theta}}) &= \eta\mathbb{E}_{x\sim d_0}\Big[\sum_{a\in\mathcal{A}}\pi^\eta_f(a|x)\Delta^2(x, a) - \sum_{a_1,a_2\in\mathcal{A}}\pi^\eta_f(a_1|x)\pi^\eta_f(a_2|x)\Delta(x, a_1)\Delta(x, a_2)\Big] \\
&\le \eta\mathbb{E}_{x\sim d_0}\Big[\sum_{a\in\mathcal{A}}\pi^\eta_f(a|x)\Delta^2(x, a)\Big], \qquad (2.2)
\end{aligned}$$

where $f(\cdot,\cdot) = \gamma R(\widehat{\theta}, \cdot, \cdot) + (1-\gamma)R(\theta_*, \cdot, \cdot)$ $(\gamma \in (0, 1))$ the inequality uses the fact that second term on the right-hand side of the equality is $(\sum_{a\in\mathcal{A}}\pi^\eta_f(a|x)\Delta(x, a))^2 \ge 0$.

The proof of this lemma is provided in Appendix E.2.

Part II: Utilizing the coverage condition In Algorithm 1, with the coverage condition (Definition 2.5) and the concentration inequalities, if the sample size $m = \Theta(\eta^2 D^2)$, we can prove that $\|R(\widehat{\theta}, \cdot, \cdot) - R(\theta_*, \cdot, \cdot)\|_\infty \le \eta^{-1}$ and $\|R(\widehat{\theta}_0, \cdot, \cdot) - R(\theta_*, \cdot, \cdot)\|_\infty \le \eta^{-1}$, which implies the whole-policy coverage condition for all contexts: $\|\pi^\eta_f(\cdot|\cdot)/\pi^\eta_{\widehat{\theta}_0}(\cdot|\cdot)\|_\infty \le e^4$. Therefore, substituting it back into (2.2) leads to

$$Q(\pi^*) - Q(\pi^\eta_{\widehat{\theta}}) \lesssim \eta\mathbb{E}_{x\sim d_0}\mathbb{E}_{a\sim\pi^\eta_{\widehat{\theta}_0}}\Big[\big(R(\widehat{\theta}, x, a) - R(\theta_*, x, a)\big)^2\Big].$$

Note that $\widehat{\theta}$ in the RHS is computed using the data sampled from $\pi^\eta_{\widehat{\theta}_0}$. By setting $n = \Theta(\eta/\epsilon)$, we obtain that $\pi^\eta_{\widehat{\theta}}$ is $O(\epsilon)$ optimal.

2.5 Result under Local-Coverage Condition

In this subsection, we consider another coverage conditions appearing in previous work as described in Definition 2.11.

Definition 2.10 (Global-Policy Coverage)**.** Given a reference policy π_0, C_{GL} is the minimum positive real number satisfying that for any $\pi : \mathcal{X} \to \Delta(\mathcal{A})$

$$\sup_{x\sim d_0, a\in\mathcal{A}}\frac{\pi(a|x)}{\pi_0(a|x)} \le C_{\mathrm{GL}}.$$

Definition 2.11 (Local KL-ball Coverage, Song et al. 2024)**.** Given a reference policy π_0, for a positive constant $\rho_{\mathrm{KL}} < \infty$, and all policy satisfying that $\mathbb{E}_{x\sim d_0}[\mathrm{KL}(\pi, \pi_0)] \le \rho_{\mathrm{KL}}$, we define

$$\sup_{x\sim d_0, a\in\mathcal{A}}\frac{\pi(a|x)}{\pi_0(a|x)} := C_{\rho_{\mathrm{KL}}}.$$

Remark 2.12 (Relation between Local and Global Coverage Conditions)**.** The local-coverage condition (Definition 2.11) is more precise because compared to the global conditions targeting all possible policies, it only constrains the coverage to a KL-ball. In Song et al. (2024), because of the specific form of the oracle (Definition 2.4), the considered policy class is $\Pi = \{\pi(\cdot|\cdot) \propto \pi_0(\cdot|\cdot)\exp(\eta R(\theta, \cdot, \cdot)) : R(\theta, \cdot, \cdot) \in \mathcal{R}\}$. Thus, they only need to assume that the condition hold for $\rho = 2\eta B$, indicating that $C_{\rho_{\mathrm{KL}}} \leq C_{\mathrm{GL}}$. On the other hand, the data coverage condition (Definition 2.5) is measured on the level of reward functions instead of policies. In this sense, the data coverage condition and local-coverage condition do not encompass each other.

Corollary 2.13. Let $C_{\rho_{\mathrm{KL}}}$ be in Definition 2.11 where $\rho_{\mathrm{KL}} = 2\eta B$. For any $\delta \in (0, 1/6)$ and $\epsilon > 0$, if we set $n = c_{m,n}m = \Theta(C_{\rho_{\mathrm{KL}}}\eta/\epsilon \cdot B\log(N_{\mathcal{R}}(\epsilon_c)/\delta))$ (where constant $c_{m,n} > 0$, $\epsilon_c = \epsilon/(2(1 + c_{m,n}^{-1})B)$) then with probability at least $1 - 6\delta$ the output policy of Algorithm 3 $\pi_{\widehat{\theta}}^{\eta}$ is $O(\epsilon)$ optimal.

In comparison with the sample complexity $\Theta(\eta^2 D^2 + \eta/\epsilon)$ under data coverage in Theorem 2.8, the order $\Theta(C_{\rho_{\mathrm{KL}}}\eta/\epsilon)$ depends on a local coverage coefficient $C_{\rho_{\mathrm{KL}}}$, but has a multiplicative dependence on the coverage coefficient instead of additive dependence. Whether the additive dependence can be achieved under the local-coverage condition is left as future work. Moreover, we compare this result with Theorem 4.2 in Song et al. (2024) and suppose that the in-sample-error $\epsilon_{\mathrm{reward}}$ of Song et al. (2024) is $O(1/n)$, their sample complexity is $\Theta(C_{\rho_{\mathrm{KL}}}^2/\epsilon^2)$, which is looser than ours $\Theta(C_{\rho_{\mathrm{KL}}}\eta/\epsilon)$ when $\eta = o(C_{\rho_{\mathrm{KL}}}/\epsilon)$.

3 Reinforcement Learning from Preference Feedback

In this section, we consider the problem of aligning the language model with preference feedback.

3.1 Problem Setup

In each round, we can sample a pair of actions (responses) a_1, a_2 and query a preference oracle to get the preference label $y \in \{0, 1\}$, where $y = 1$ means that the user prefers a_1 over a_2. Specifically, when receiving a prompt $x \in \mathcal{X}$, and two actions (responses) $a^1, a^2 \in \mathcal{A}$ from some LLM policy $\pi(\cdot|x)$, a preference oracle will give feedback y defined as follows:

Definition 3.1 (Preference Oracle)**.** A Preference Oracle is a function $\mathbb{P} : \mathcal{X} \times \mathcal{A} \times \mathcal{A} \to \{0, 1\}$. Given a context $x \in \mathcal{X}$ and two actions $a_1, a_2 \in \mathcal{A}$, the oracle can be queried to obtain a preference signal $y \sim Bernoulli(\mathbb{P}(x, a_1, a_2))$, where $y = 1$ indicates that a_1 is preferred to a_2 in the context x, and $y = 0$ indicates the opposite.

To learn the preference, we follow Ouyang et al. (2022); Zhu et al. (2023); Rafailov et al. (2024); Liu et al. (2023); Xiong et al. (2024a) and assume that the preference oracle is measured by the difference of ground-truth reward functions $R(\theta_*, x, a)$, which is named the Bradley-Terry (BT) model (Bradley and Terry, 1952b).

Definition 3.2 (Bradley-Terry Model)**.** Given a context $x \in \mathcal{X}$ and two actions $a_1, a_2 \in \mathcal{A}$, the probability of a_1 being preferred to a_2 is modeled as

$$\mathbb{P}(x, a_1, a_2) = \frac{\exp(R(\theta_*, x, a_1))}{\exp(R(\theta_*, x, a_1)) + \exp(R(\theta_*, x, a_2))} = \sigma(R(\theta_*, x, a_1) - R(\theta_*, x, a_2)), \quad (3.1)$$

where $\sigma(u) = (1 + e^{-u})^{-1}$ is the sigmoid function.

The RLHF training always follows the fine-tuning process, which yields a reference policy π_0. When performing RLHF on specific tasks, to avoid overfitting, we impose KL-regularization to the learned reward model when optimizing the policy. Hence, our objective function is also (2.1).

To learn the reward function, we introduce the following assumption to ensure the existence of an MLE estimation oracle that can globally maximize the likelihood of the BT model over all possible reward functions.

Definition 3.3 (MLE estimation oracle)**.** Given a set of context-action pairs $\{(x_i, a_i^1, a_i^2, y_i)\}_{i=1}^n$ generated from the BT model, can output the parameter $\widehat{\theta}$ such that

$$\widehat{\theta} = \operatorname*{argmax}_{\theta \in \Theta} \sum_{i=1}^{n} \underbrace{y_i \cdot \log \sigma(R(\theta, x_i, a_i^1) - R(\theta, x_i, a_i^2)) + (1 - y_i) \cdot \log \sigma(R(\theta, x_i, a_i^2) - R(\theta, x_i, a_i^1))}_{\mathcal{L}(\theta | x_i, a_i^1, a_i^2, y_i)}.$$

Following the previous analysis for RLHF (Xiong et al., 2024a), we assume the existence of a policy improvement oracle (Definition 2.4) that can compute the optimal policy $\pi^{\eta}_{\widehat{\theta}}$ based on the reward function $\widehat{\theta}$.

Remark 3.4. We learn the reward function since we can always control the reward (like clipping and normalization) to ensure that the reward function is always bounded by B. The bounded assumption does not apply for direct preference learning like DPO (Rafailov et al., 2024) since there is no intrinsic policy function class encompassing the soundness (Song et al., 2024), thus increasing the cases of overfitting.

In RLHF setting, we cannot directly observe or estimate absolute reward values. Consequently, the most intuitive estimation approach is to focus on relative rewards: for any context x and actions a^1, a^2, our estimated difference $f(s, a^1) - f(s, a^2)$ should closely approximate the true reward difference $r(s, a^1) - r(s, a^2)$. Therefore, we extend the data coverage condition to the RLHF setting as follows.

Definition 3.5 (Data Coverage)**.** Given a reference policy π_0, D^2 is the minimum positive real number satisfying $\forall\ (x, a) \in \mathcal{X} \times \mathcal{A}$, $\pi(a|x) > 0$, we have for any pair of $\theta, \theta' \in \Theta$, there exists $b : \mathcal{X} \to [-B, B]$ such that

$$\frac{|R(\theta', x, a) - R(\theta, x, a) - b(x)|^2}{\mathbb{E}_{x' \sim d_0} \operatorname{Var}_{a' \sim \pi_0(\cdot|x')}[R(\theta', x', a') - R(\theta, x', a')]} \leq D^2.$$

3.2 Theoretical Guarantees

Lower bound We provide a lower bound for the RLHF problem with preference feedback. The lower bound is derived by constructing a hard instance where the reward function is difficult to estimate from the preference feedback.

Theorem 3.6. For any $\epsilon \in (0, 1/256), \eta > 4$, and any algorithm A, there exists a KL-regularized preference learning problem with $O(N_{\mathcal{R}}(\epsilon))$ coverage coefficient and reward function class $\mathcal{R}$ such that A requires at least $\Omega\big(\min(\frac{\eta \log N_{\mathcal{R}}(\epsilon)}{\epsilon}, \frac{\log N_{\mathcal{R}}(\epsilon)}{\epsilon^2})\big)$ samples to achieve a suboptimality gap of ϵ.

We present a two-stage mixed-policy sampling algorithm for RLHF, which can be seen as an extention of Algorithm 1. Due to space limit, we defer it to Algorithm 3 in Appendix D.

Upper bound We provide the theoretical guarantees for Algorithm 3 in the following theorem.

Theorem 3.7. Suppose that Assumption 2.5 holds. For any $\delta \in (0, 1/6)$ and $\epsilon > 0$, if we set

$$m = \Theta(\eta^2 D^2 \cdot e^B \log(N_{\mathcal{R}}(\epsilon_c)/\delta)) \text{ and } n = \Theta(\eta/\epsilon \cdot e^B \log(N_{\mathcal{R}}(\epsilon_c)/\delta))$$

where $\epsilon_c = \min\{\frac{\epsilon}{2(1+c_{m,n}^{-1})e^B}, \frac{1}{8(1+c_{m,n})e^B \eta^2 D^2}\}$ then with probability at least $1 - 6\delta$ the output policy of Algorithm 3 $\pi^{\eta}_{\widehat{\theta}}$ is $O(\epsilon)$ optimal.

Remark 3.8 (Comparison with Hybrid Framework)**.** We compare our two-stage mixed sampling method with hybrid RL. From the algorithmic perspective, a hybrid RL algorithm first learns from an offline dataset and then requires sufficient online iterations to ensure the performance (Xiong et al., 2024a). For example, for a finite action space with A actions, the number of online iterations should be $\Theta(A)$. In contrast, our method only requires two iterations of sampling from mixed policy and interacting with the environment. Moreover, the results of hybrid literature depend on both the coverage coefficient and the structure complexity of the function class (like the dimension for a linear function class or eluder dimension (Russo and Van Roy, 2013)). Our result only needs the coverage condition of the reference policy. More importantly, we obtain a sharper bound on the sample complexity and derive the additive dependence on the coverage coefficient.

Remark 3.9. Although the coefficient e^B appearing in sample size m, n can be exponentially large, this term is caused by the non-linearity of the link function for the preference model, and is common in RLHF literature (Zhu et al., 2023; Xiong et al., 2024a; Ye et al., 2024a; Song et al., 2024).

Theorem 3.7 shows that the sample complexity of Algorithm 3 is $\mathcal{O}(\eta/\epsilon \log N_{\mathcal{R}}(\epsilon/\delta))$ when the reward scale is a constant and ϵ is sufficiently small. The result indicates that the proposed two-stage mixed sampling strategy can achieve a suboptimality gap of ϵ with only an additive dependence on the coverage coefficient D^2.

Besides, the algorithm only requires sampling from the reference policy π_0 and the intermediate policy $\pi^{\eta}_{\widehat{\theta}_0}$, which is more aligned with the practical scenarios where the preference feedback is collected from the human users and it is expensive to collect the data while the language model is being updated. Our result implies that we may achieve a near-optimal sample complexity by simply leveraging an intermediate policy to collect more data, and the training process of the reward model and the policy (language model) can be highly decoupled.

Upper bound for local coverage We also show the result under the local-coverage assumption (Definition 2.11) as follows.

Corollary 3.10. Let $C_{\rho_{\mathrm{KL}}}$ be in Definition 2.11 where $\rho = 2\eta B$. For any $\delta \in (0, 1/6)$ and $\epsilon > 0$, if we set $n = c_{m,n} m = \Theta(C_{\rho_{\mathrm{KL}}} \eta/\epsilon \cdot e^B \log(N_{\mathcal{R}}(\epsilon_c)/\delta))$ (where constant $c_{m,n} > 0$, $\epsilon_c = \frac{\epsilon}{2(1+c_{m,n}^{-1})e^B}$) then with probability at least $1 - 6\delta$ the output policy of Algorithm 3 $\pi^{\eta}_{\widehat{\theta}}$ is $O(\epsilon)$ optimal.

4 Conclusions and Future Work

We have presented a comprehensive theoretical analysis of the role of reverse-KL regularization in decision-making models including contextual bandits and reinforcement learning from preference feedback, highlighting its significance in terms of sample complexity. Our results provide new insights into the power of regularization extending beyond its traditional role of mitigating errors from the current critic (or reward) model. Additionally, we examined the role of data coverage in both contextual bandits and RLHF. Our analysis shows that with sufficient coverage from the reference policy, a mixed sampling strategy can achieve a sample complexity that exhibits only an additive dependence on the coverage coefficient without the need for explicit exploration or additional structural assumptions. For future directions, it is interesting to study if the sharp bound exists for the KL-regularized Markov Decision Process (MDP) problem. Additionally, it is also worthwhile to explore how to achieve a sample complexity bound with only additive dependence on the global and local coverage conditions in Definitions 2.10 and 2.11.

Acknowledgment

We thank the anonymous reviewers and area chair for their helpful comments. HZ and QG are supported in part by the National Science Foundation DMS-2323113 and IIS-2403400. This work is also partially supported by NSF grant No. 2416897 and ONR grant No. N000142512318. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

References

ACHIAM, J., ADLER, S., AGARWAL, S., AHMAD, L., AKKAYA, I., ALEMAN, F. L., ALMEIDA, D., ALTENSCHMIDT, J., ALTMAN, S., ANADKAT, S. ET AL. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* .

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

AGARWAL, A., KAKADE, S. M., LEE, J. D. and MAHAJAN, G. (2020). Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*. PMLR.

AGARWAL, A., KAKADE, S. M., LEE, J. D. and MAHAJAN, G. (2021). On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research* **22** 1–76.

AGARWAL, A., QIAN, J., RAKHLIN, A. and ZHANG, T. (2024). The non-linear f-design and applications to interactive learning. In *Forty-first International Conference on Machine Learning*.

AHMED, Z., LE ROUX, N., NOROUZI, M. and SCHUURMANS, D. (2019). Understanding the impact of entropy on policy optimization. In *International conference on machine learning*. PMLR.

ANTHROPIC, A. (2023). Introducing claude.

AZAR, M. G., GUO, Z. D., PIOT, B., MUNOS, R., ROWLAND, M., VALKO, M. and CALANDRIELLO, D. (2024). A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

BAI, Y., JONES, A., NDOUSSE, K., ASKELL, A., CHEN, A., DASSARMA, N., DRAIN, D., FORT, S., GANGULI, D., HENIGHAN, T. ET AL. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* .

BERTHET, Q. and PERCHET, V. (2017). Fast rates for bandit optimization with upper-confidence frank-wolfe. *Advances in Neural Information Processing Systems* **30**.

BRADLEY, R. A. and TERRY, M. E. (1952a). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* **39** 324.

BRADLEY, R. A. and TERRY, M. E. (1952b). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* **39** 324–345.

BROCKMAN, G. (2016). Openai gym. *arXiv preprint arXiv:1606.01540* .

CALANDRIELLO, D., GUO, D., MUNOS, R., ROWLAND, M., TANG, Y., PIRES, B. A., RICHEMOND, P. H., LAN, C. L., VALKO, M., LIU, T. ET AL. (2024). Human alignment of large language models through online preference optimisation. *arXiv preprint arXiv:2403.08635* .

CHEN, Z., DENG, Y., YUAN, H., JI, K. and GU, Q. (2024). Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335* .

CHRISTIANO, P. F., LEIKE, J., BROWN, T., MARTIC, M., LEGG, S. and AMODEI, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems* **30**.

COBBE, K., KOSARAJU, V., BAVARIAN, M., CHEN, M., JUN, H., KAISER, L., PLAPPERT, M., TWOREK, J., HILTON, J., NAKANO, R. ET AL. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* .

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

DONG, H., XIONG, W., GOYAL, D., ZHANG, Y., CHOW, W., PAN, R., DIAO, S., ZHANG, J., SHUM, K. and ZHANG, T. (2023). Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767* .

DONG, H., XIONG, W., PANG, B., WANG, H., ZHAO, H., ZHOU, Y., JIANG, N., SAHOO, D., XIONG, C. and ZHANG, T. (2024). Rlhf workflow: From reward modeling to online rlhf. *arXiv preprint arXiv:2405.07863* .

FOSTER, D. J., KAKADE, S. M., QIAN, J. and RAKHLIN, A. (2021). The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487* .

FOX, R., PAKMAN, A. and TISHBY, N. (2016). Taming the noise in reinforcement learning via soft updates. In *32nd Conference on Uncertainty in Artificial Intelligence 2016, UAI 2016*. Association For Uncertainty in Artificial Intelligence (AUAI).

GEIST, M., SCHERRER, B. and PIETQUIN, O. (2019). A theory of regularized markov decision processes. In *International Conference on Machine Learning*. PMLR.

GOU, Z., SHAO, Z., GONG, Y., YELONG SHEN, YANG, Y., HUANG, M., DUAN, N. and CHEN, W. (2024). ToRA: A tool-integrated reasoning agent for mathematical problem solving. In *The Twelfth International Conference on Learning Representations*.
URL https://openreview.net/forum?id=Ep0TtjVoap

GUI, L., GÂRBACEA, C. and VEITCH, V. (2024). Bonbon alignment for large language models and the sweetness of best-of-n sampling. *arXiv preprint arXiv:2406.00832* .

GULCEHRE, C., PAINE, T. L., SRINIVASAN, S., KONYUSHKOVA, K., WEERTS, L., SHARMA, A., SIDDHANT, A., AHERN, A., WANG, M., GU, C. ET AL. (2023). Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998* .

GUO, S., ZHANG, B., LIU, T., LIU, T., KHALMAN, M., LLINARES, F., RAME, A., MESNARD, T., ZHAO, Y., PIOT, B. ET AL. (2024). Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792* .

HAARNOJA, T., TANG, H., ABBEEL, P. and LEVINE, S. (2017). Reinforcement learning with deep energy-based policies. In *International conference on machine learning*. PMLR.

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

HENDRYCKS, D., BURNS, C., KADAVATH, S., ARORA, A., BASART, S., TANG, E., SONG, D. and STEINHARDT, J. (2021). Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874* .

LAN, G. (2023). Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming* **198** 1059–1106.

LANGFORD, J. and ZHANG, T. (2007). The epoch-greedy algorithm for multi-armed bandits with side information. *Advances in neural information processing systems* **20**.

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

LIU, T., ZHAO, Y., JOSHI, R., KHALMAN, M., SALEH, M., LIU, P. J. and LIU, J. (2023). Statistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657* .

MEI, J., XIAO, C., SZEPESVARI, C. and SCHUURMANS, D. (2020). On the global convergence rates of softmax policy gradient methods. In *International conference on machine learning*. PMLR.

META, A. (2024). Introducing meta llama 3: The most capable openly available llm to date. *Meta AI* .

NAKANO, R., HILTON, J., BALAJI, S., WU, J., OUYANG, L., KIM, C., HESSE, C., JAIN, S., KOSARAJU, V., SAUNDERS, W. ET AL. (2021). Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332* .

NEU, G., JONSSON, A. and GÓMEZ, V. (2017). A unified view of entropy-regularized markov decision processes. *arXiv preprint arXiv:1705.07798* .

OPENAI (2023). Gpt-4 technical report. *ArXiv* **abs/2303.08774**.

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

RAFAILOV, R., SHARMA, A., MITCHELL, E., MANNING, C. D., ERMON, S. and FINN, C. (2024). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems* **36**.

RUSSO, D. and VAN ROY, B. (2013). Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems* **26**.

SCHULMAN, J., CHEN, X. and ABBEEL, P. (2017a). Equivalence between policy gradients and soft q-learning. *arXiv preprint arXiv:1704.06440* .

SCHULMAN, J., LEVINE, S., ABBEEL, P., JORDAN, M. and MORITZ, P. (2015). Trust region policy optimization. In *International Conference on Machine Learning*. PMLR.

SCHULMAN, J., WOLSKI, F., DHARIWAL, P., RADFORD, A. and KLIMOV, O. (2017b). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* .

SHANI, L., EFRONI, Y. and MANNOR, S. (2020). Adaptive trust region policy optimization: Global convergence and faster rates for regularized mdps. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34.

SHAO, Z., WANG, P., ZHU, Q., XU, R., SONG, J., ZHANG, M., LI, Y., WU, Y. and GUO, D. (2024). Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300* .

SONG, Y., SWAMY, G., SINGH, A., BAGNELL, D. and SUN, W. (2024). The importance of online data: Understanding preference fine-tuning via coverage. In *ICML 2024 Workshop: Aligning Reinforcement Learning Experimentalists and Theorists*.

SUTTON, R. S. (2018). Reinforcement learning: An introduction. *A Bradford Book* .

SZEPESVÁRI, C. (2022). *Algorithms for reinforcement learning*. Springer nature.

TEAM, G., ANIL, R., BORGEAUD, S., WU, Y., ALAYRAC, J.-B., YU, J., SORICUT, R., SCHALKWYK, J., DAI, A. M., HAUTH, A. ET AL. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* .

TONG, Y., ZHANG, X., WANG, R., WU, R. and HE, J. (2024). Dart-math: Difficulty-aware rejection tuning for mathematical problem-solving. *arXiv preprint arXiv:2407.13690* .

TOUVRON, H., MARTIN, L., STONE, K., ALBERT, P., ALMAHAIRI, A., BABAEI, Y., BASHLYKOV, N., BATRA, S., BHARGAVA, P., BHOSALE, S. ET AL. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* .

WANG, H., LIN, Y., XIONG, W., YANG, R., DIAO, S., QIU, S., ZHAO, H. and ZHANG, T. (2024a). Arithmetic control of llms for diverse user preferences: Directional preference alignment with multi-objective rewards. *arXiv preprint arXiv:2402.18571* .

WANG, H., XIONG, W., XIE, T., ZHAO, H. and ZHANG, T. (2024b). Interpretable preferences via multi-objective reward modeling and mixture-of-experts. *arXiv preprint arXiv:2406.12845* .

WANG, R., SALAKHUTDINOV, R. R. and YANG, L. (2020). Reinforcement learning with general value function approximation: Provably efficient approach via bounded eluder dimension. *Advances in Neural Information Processing Systems* **33** 6123–6135.

WANG, Y., LIU, Q. and JIN, C. (2023a). Is rlhf more difficult than standard rl? a theoretical perspective. *Advances in Neural Information Processing Systems* **36** 76006–76032.

WANG, Z., DONG, Y., DELALLEAU, O., ZENG, J., SHEN, G., EGERT, D., ZHANG, J. J., SREEDHAR, M. N. and KUCHAIEV, O. (2024c). Helpsteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673* .

WANG, Z., DONG, Y., ZENG, J., ADAMS, V., SREEDHAR, M. N., EGERT, D., DELALLEAU, O., SCOWCROFT, J. P., KANT, N., SWOPE, A. ET AL. (2023b). Helpsteer: Multi-attribute helpfulness dataset for steerlm. *arXiv preprint arXiv:2311.09528* .

WU, Y., SHARIFF, R., LATTIMORE, T. and SZEPESVÁRI, C. (2016). Conservative bandits. In *International Conference on Machine Learning*. PMLR.

WU, Y., SUN, Z., YUAN, H., JI, K., YANG, Y. and GU, Q. (2024). Self-play preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675* .

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

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

XIONG, W., SHI, C., SHEN, J., ROSENBERG, A., QIN, Z., CALANDRIELLO, D., KHALMAN, M., JOSHI, R., PIOT, B., SALEH, M. ET AL. (2024b). Building math agents with multi-turn iterative preference learning. *arXiv preprint arXiv:2409.02392* .

YE, C., XIONG, W., GU, Q. and ZHANG, T. (2023). Corruption-robust algorithms with uncertainty weighting for nonlinear contextual bandits and markov decision processes. In *International Conference on Machine Learning*. PMLR.

YE, C., XIONG, W., ZHANG, Y., JIANG, N. and ZHANG, T. (2024a). A theoretical analysis of nash learning from human feedback under general kl-regularized preference. *arXiv preprint arXiv:2402.07314* .

YE, C., YANG, R., GU, Q. and ZHANG, T. (2024b). Corruption-robust offline reinforcement learning with general function approximation. *Advances in Neural Information Processing Systems* **36**.

YUE, Y., BRODER, J., KLEINBERG, R. and JOACHIMS, T. (2012). The k-armed dueling bandits problem. *Journal of Computer and System Sciences* **78** 1538–1556.

ZHANG, T. (2023). *Mathematical analysis of machine learning algorithms*. Cambridge University Press.

ZHANG, X., XIONG, W., CHEN, L., ZHOU, T., HUANG, H. and ZHANG, T. (2024). From lists to emojis: How format bias affects model alignment. *arXiv preprint arXiv:2409.11704* .

ZHANG, Z., ZHOU, Y. and JI, X. (2020). Almost optimal model-free reinforcement learningvia reference-advantage decomposition. *Advances in Neural Information Processing Systems* **33** 15198–15207.

ZHAO, H., HE, J. and GU, Q. (2023a). A nearly optimal and low-switching algorithm for reinforcement learning with general function approximation. *arXiv preprint arXiv:2311.15238* .

ZHAO, Y., JOSHI, R., LIU, T., KHALMAN, M., SALEH, M. and LIU, P. J. (2023b). Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425* .

ZHONG, H., FENG, G., XIONG, W., ZHAO, L., HE, D., BIAN, J. and WANG, L. (2024). Dpo meets ppo: Reinforced token optimization for rlhf. *arXiv preprint arXiv:2404.18922* .

ZHU, B., JIAO, J. and JORDAN, M. I. (2023). Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. *arXiv preprint arXiv:2301.11270* .

ZIEGLER, D. M., STIENNON, N., WU, J., BROWN, T. B., RADFORD, A., AMODEI, D., CHRISTIANO, P. and IRVING, G. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593* .

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We clearly state the contributions of the paper in the abstract and introduction. The main claims are that we provide a theoretical analysis of the role of reverse-KL regularization in decision-making models including contextual bandits and reinforcement learning from preference feedback, highlighting its significance in terms of sample complexity. We also show that with sufficient coverage from the reference policy, a mixed sampling strategy can achieve a sample complexity that exhibits only an additive dependence on the coverage coefficient without the need for explicit exploration or additional structural assumptions.

Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We detail in our conclusion the assumptions that are crucial for our analysis but may be relaxed in future work.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

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

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.

- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

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

   Answer: [NA]

   Justification: This paper is mostly theoretical and provides necessary details to reproduce the synthetic experiments. The paper does not include real-world experiments.

   Guidelines:

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

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

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

   Answer: [NA]

   Justification: This paper is purely theoretical with only synthetic experiments.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [NA]

   Justification: This paper is purely theoretical with only synthetic experiments.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [NA]

   Justification: This paper is purely theoretical.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [NA]

   Justification: This paper is purely theoretical.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

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

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [No]

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

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

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

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [NA]

    Justification: We do not use existing assets.

    Guidelines:

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

13. **New assets**

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

    Answer: [NA]

    Justification: We do not release new assets.

Guidelines:

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

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

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

    Answer: [NA]

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

    Guidelines:

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

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

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

    Answer: [NA]

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

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [NA]

    Justification: All the proof techniques in this theoretical research are proposed by human.

    Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.

A Additional Related Work

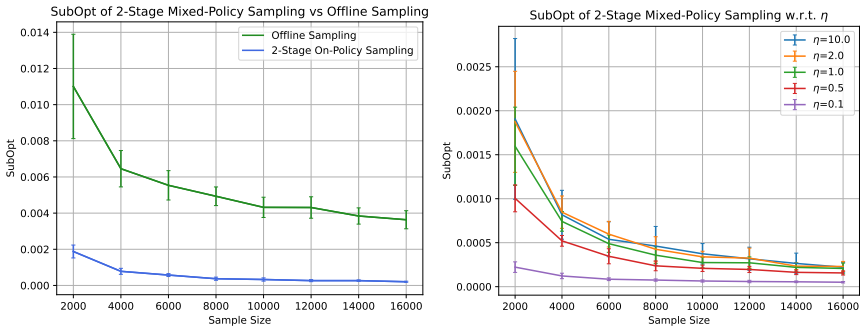
Analyses of Policy Optimization with Regularization While it is previously unknown whether regularization can improve the sample complexity of policy optimization without additional assumptions, there are some works that provided a sharp convergence rate in the presence of regularization (Mei et al., 2020; Shani et al., 2020; Agarwal et al., 2020, 2021; Lan, 2023). However, these works either assumed the access of exact or unbiased policy gradient or required uniform value function approximation error, which are not the standard case in general sample-based RL setting. For instance, Lan (2023) provided a sharp convergence rate for policy optimization with KL-regularization, assuming the access to an unbiased value function estimator (Condition 4.1) and the bounded infinity norm on the error (Conditions 4.2, 4.3), which is standard in the literature of optimization. However, RL algorithms usually make biased estimation to balance exploration and exploitation. Instead of focusing on the influence of regularization on the optimization, our work aims to understand how the KL-regularization affects the exploration and exploitation trade-off in the bandit and RLHF settings through a novel analysis on the optimal sample complexity.

RLHF Algorithms There are mainly three types of RLHF algorithms: offline, online and hyrbid. The most well-known offline algorithms are Slic (Zhao et al., 2023b), Direct Preference Optimization (DPO) (Rafailov et al., 2024), Identity-PO (IPO) (Azar et al., 2024) and (SPIN) (Chen et al., 2024). They aim to approximate the closed-form solution of the optimization problem on a fixed offline dataset. For the online algorithms, the most representative one is Proximal Policy Optimization (PPO) (Schulman et al., 2017b). PPO has been used in the Chat-GPT (OpenAI, 2023), Gemini (Team et al., 2023), and Claude (Bai et al., 2022). However, the deep RL method PPO is known to be sample inefficient and unstable, making its success hard to reproduce for the open-source community. In response to this, there have been many efforts to propose alternative algorithms to the PPO algorithm. The Reward ranked fine-tuning (RAFT) (also known as rejection sampling finetuning) (Dong et al., 2023; Touvron et al., 2023; Gulcehre et al., 2023; Gui et al., 2024) is a stable framework requiring minimal hyper-parameter tuning, which iteratively learns from the best-of-n policy (Nakano et al., 2021). This framework proves to be particularly effective in the reasoning task such as (Gou et al., 2024; Tong et al., 2024). However, the RAFT-like algorithms only use the positive signal by imitating the best-of-n sampling. To further improve the efficiency, there is an emerging body of literature that proposes online direct preference optimization by extending DPO or IPO to an online iterative framework (Xiong et al., 2024a; Guo et al., 2024; Wu et al., 2024; Calandriello et al., 2024; Xiong et al., 2024b). Finally, for the third type, the common point of hybrid and online algorithms is that they both require further interaction with the preference oracle and on-policy data collection. The difference is that hybrid algorithms start with a pre-collected dataset (Xiong et al., 2024a; Song et al., 2024; Touvron et al., 2023), while the online algorithms learn from scratch.

RLHF Theory The theoretical study of RLHF can date back to the dueling bandits (Yue et al., 2012) and follow-up work on MDPs (Wang et al., 2023a; Zhu et al., 2023). However, these works deviate from the practice because they do not realize the significance of KL-regularization and still choose the greedy policy that simply maximizes the reward. After this line of work, Xiong et al. (2024a); Ye et al. (2024a); Song et al. (2024) highlight the KL-regularization theoretically and incorporate the KL term into the learning objective. However, they circumvent the special advantages of KL-regularization and still follow the techniques in bandit analysis, thus obtaining loose bounds. In our paper, we establish a new lower bound and a sharper upper bound for the KL-regularized framework, thus validating the empirical advantage of KL-regularization. There are also some works extending KL-regularized RLHF from bandit problems to the Markov decision process (MDP) problems (Zhong et al., 2024; Xiong et al., 2024b). We expect that our techniques can also be extended to the MDP setting, which we leave for future work.

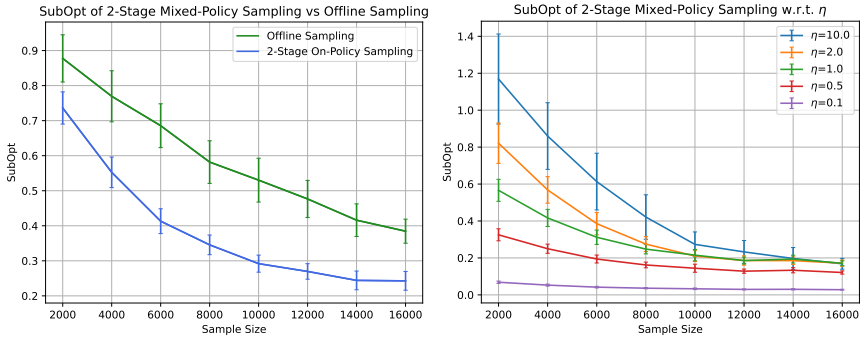
B Experimental Results

In this section, we conduct experiments with synthetic data to investigate the benefit of mixed-policy sampling and the effect of KL-regularization coefficient on the sample complexity of the problem. We plot the experimental results for RL from preference feedback in Figure 2 and the results for KL-regularized contextual bandits in Figure 1. All the trials are repeated for 10 times and plotted with the standard variation.



(a) Benefit of mixed-policy sampling (b) Results under different KL-penalty

Figure 1: Suboptimality gap for KL-regularized contextual bandits.



(a) Benefit of mixed-policy sampling (b) Results under different KL-penalty

Figure 2: Suboptimality gap for reinforcement learning from preference feedback.

Algorithm 2 Active Querying for KL-Regularized Bandits

Input: η, ϵ, Θ.
for $i = 1, \ldots, m$ **do**
  Sample context $x \sim d_0$.
  if there exists $a \in \mathcal{A}$ such that $\mathcal{D}_{\mathcal{R}}(x, a; D_0) \geq u$ **then**
    Query the reward signal r for (x, a) and update the dataset $D_0 \leftarrow D_0 \cup (x, a, r)$.
  end if
end for
Compute the least square estimate $\widehat{\theta}_0$ of the reward function based on D_0.

We consider the case where context distribution d_0 is a projected Gaussian distribution over the unit sphere and $\mathcal{A}$ is a discrete set with $|\mathcal{A}| = 5$. We construct the reward functions as $R(\phi, x, a) = \langle x, \phi(a) \rangle$, parameterized by a mapping ϕ from $\mathcal{A}$ to $\mathbb{R}^{10}$, and set the reference policy π_0 to be the uniform random policy. To generate ϕ_*, we sample $\phi_*(a)$ independently for each $a \in \mathcal{A}$ according to another projected gaussian distribution over the sphere with radius equal to 5. In Figure 2(a), we compare the suboptimality gaps of mixed-policy sampling with $m = n$ to those of offline sampling using π_0 under the same sample sizes. The result indicates that the usage of mixed-policy sampling reduces the suboptimality gap by a large margin. In Figure 2(b), it is shown that the sample complexity is remarkably affected by the KL-regularization term, corroborating our sharp analysis for regularized RLHF.

C Alternative Approach with Active Query

In this section, we remove the converage assumption and show that if we could actively query the reward signal, we can still achieve the sharper sample complexity bound. The key idea is to actively query the context-action pairs that are most informative for the current reward model in stage 1, and then follow the same steps as in stage 2 of Algorithm 1.

In the following definition, we introduce the uncertainty measure $\mathcal{D}_{\mathcal{R}}(x, a; D)$ which characterizes the informativeness of the context-action pair (x, a) with respect to a dataset D.

Definition C.1 (Uncertainty Measure)**.** For a given dataset $D = \{x_i, a_i\}_{i=1}^m$, we define the uncertainty measure $\mathcal{D}_{\mathcal{R}}(x, a; D)$ as follows:

$$\mathcal{D}_{\mathcal{R}}(x, a; D) := \sup_{R_1, R_2 \in \mathcal{R}} \frac{\big|R_1(x, a) - R_2(x, a)\big|}{\sqrt{1 + \sum_{i=1}^m \big|R_1(x_i, a_i) - R_2(x_i, a_i)\big|^2}}.$$

Definition C.2 (Definition of eluder dimension, Russo and Van Roy 2013)**.** The eluder dimension of a function class $\mathcal{G}$ with domain $\mathcal{Z}$ is defined as follows:

- A point $z \in \mathcal{Z}$ is ϵ-dependent on $z_1, z_2, \cdots, z_k \in \mathcal{Z}$ with respect to $\mathcal{G}$, if for all $g_1, g_2 \in \mathcal{G}$ such that $\sum_{i=1}^k [g_1(z_i) - g_2(z_i)]^2 \le \epsilon$, it holds that $|g_1(z) - g_2(z)| \le \epsilon$.
- Further z is said to be ϵ-independent of $z_1, z_2, \cdots, z_k$ with respect to $\mathcal{G}$, if z is not dependent on $z_1, z_2, \cdots, z_k$.
- The eluder dimension of $\mathcal{G}$, denoted by $\dim_E(\mathcal{G}, \epsilon)$, is the length of the longest sequence of elements in $\mathcal{Z}$ such that every element is ϵ'-independent of its predecessors for some $\epsilon' \ge \epsilon$.

Lemma C.3. Let $\dim_E(\mathcal{R})$ be the eluder dimension of the reward function class $\mathcal{R}$ as defined in Definition C.2. In Algorithm 2, if we set $m = \widetilde{O}(\dim_E(\mathcal{R})\eta/\epsilon u^2)$, then with probability at least $1 - \delta$, $\mathbb{P}_{x \sim d_0}\{\max_{a \in \mathcal{A}} \mathcal{D}_{\mathcal{R}}(x, a; D_0) \ge u\} \le \epsilon/\eta$.

Proof. We denote the m sampled context by $x_1, \ldots, x_m$ numbered in the order of operations, and denote by $D_{0,i}$ the dataset D_0 at the beginning of the i-th iteration. Also, we define $\mu_i := \mathbb{P}_{x \sim d_0}\{\max_{a \in \mathcal{A}} \mathcal{D}_{\mathcal{R}}(x, a; D_{0,i}) \ge u\}$.

Applying freedman's inequality (Lemma G.6), we have with probability at least $1 - \delta$:

$$\begin{aligned}
\sum_{i=1}^m \big[\mu_i - \mathbb{1}\{\max_{a \in \mathcal{A}} \mathcal{D}_{\mathcal{R}}(x_i, a; D_{0,i}) \ge u\}\big] &\le 2\sqrt{\sum_{i=1}^m \mu_i \cdot \log\big((2m+2)/\delta\big)} + 2\sqrt{\log\big((2m+2)/\delta\big)} \\
&\quad + 2\log\big((2m+2)/\delta\big) \\
&\le \frac{1}{2}\sum_{i=1}^m \mu_i + 4\log\big((2m+2)/\delta\big) + 2\sqrt{\log\big((2m+2)/\delta\big)}
\end{aligned}$$

where the last inequality follows from Young's inequality.

Rearranging the terms, we obtain:

$$\begin{aligned}
\sum_{i=1}^m \mu_i &\le 2|D_0| + 8\log\big((2m+2)/\delta\big) + 4\sqrt{\log\big((2m+2)/\delta\big)} \\
&\le 2\dim_E(\mathcal{R})/u^2 + 8\log\big((2m+2)/\delta\big) + 4\sqrt{\log\big((2m+2)/\delta\big)}
\end{aligned}$$

where the last inequality holds due to the definition of eluder dimension.

Hence, by setting $m = \widetilde{O}(\dim_E(\mathcal{R})\eta/\epsilon u^2)$, we can ensure that $\frac{1}{m}\sum_{i=1}^m \mu_i \le \frac{\epsilon}{\eta}$. □

Theorem C.4. Suppose we alternatively execute Algorithm 2 for stage 1 in Algorithm 1, and suppose that the reward signals lies in [0, 1]. If we set $u = \frac{1}{\eta\sqrt{16\log\left(2N_{\mathcal{R}}(\epsilon_c)/\delta\right)+5}}$, set m as in Lemma C.3, and set $n = \Theta(\eta \cdot \log(N_{\mathcal{R}}(\epsilon_c)/\delta)/\epsilon)$, then with probability at least $1 - 6\delta$, the output policy $\pi_{\widehat{\theta}}^{\eta}$ is $O(\epsilon)$ optimal.

Proof. By Lemma C.3, we have with probability at least $1-\delta$, $\mathbb{P}_{x\sim d_0}\{\max_{a\in\mathcal{A}}\mathcal{D}_{\mathcal{R}}(x,a;D_0)\geq u\}\leq \epsilon/\eta$.

Let $\mathcal{X}_0:=\{x\in\mathcal{X}:\max_{a\in\mathcal{A}}\mathcal{D}_{\mathcal{R}}(x,a;D_0)<u\}$. Similar to the proof of Lemma E.4, by Lemma E.2 and the definition of u, we have with probability at least $1-4\delta$, for all $x\in\mathcal{X}_0$,

$$\eta|R(\widehat{\theta}_0,x,a)-R(\theta_*,x,a)|\leq 1, \eta|R(\widehat{\theta},x,a)-R(\theta_*,x,a)|\leq 1. \tag{C.1}$$

By Lemma 2.9, we have

$$Q(\pi^*)-Q(\pi^\eta_{\widehat{\theta}})\leq \eta\mathbb{E}_{x\sim d_0}\Big[\sum_{a\in\mathcal{A}}\pi^\eta_f(a|x)\big(R(\widehat{\theta},x,a)-R(\theta_* x,a)\big)^2\Big].$$

From (C.1), we have for any $x\in\mathcal{X}_0$, $a\in\mathcal{A}$,

$$\frac{\pi^\eta_f(a|x)}{\pi^\eta_{\widehat{\theta}_0}(a|x)}\leq e^4.$$

Therefore, we can proceed as in the proof of Theorem 2.8 to obtain

$$\begin{aligned}Q(\pi^*)-Q(\pi^\eta_{\widehat{\theta}})\leq \eta e^4\mathbb{E}_{x\sim d_0}\Big[\mathbb{1}(x\in\mathcal{X}_0)\sum_{a\in\mathcal{A}}\pi^\eta_{\widehat{\theta}_0}(a|x)\big(R(\widehat{\theta},x,a)-R(\theta_* x,a)\big)^2\Big]\\ +\eta\cdot\mathbb{P}_{x\sim d_0}\{\max_{a\in\mathcal{A}}\mathcal{D}_{\mathcal{R}}(x,a;D_0)\geq u\}.\end{aligned}$$

Then we can complete the proof by Lemma E.3 and the union bound. □

D KL-Regularized Algorithm for RLHF

In this section, we present a two-stage mixed-policy sampling algorithm for RLHF in Algorithm 3, which can be seen as an extention of Algorithm 1. There are two stages in the algorithm.

In the first stage, we sample m context-action pairs $\{(\widetilde{x}_i,\widetilde{a}^1_i,\widetilde{a}^2_i,\widetilde{y}_i)\}_{i=1}^m$ from the BT model and call the preference oracle to get the preference labels. We then compute the MLE estimator of the reward function $\widehat{\theta}_0$ based on the preference feedback in line 6. Afterwards, we apply the planning oracle to compute the optimal policy $\pi^\eta_{\widehat{\theta}_0}$ based on the reward function $\widehat{\theta}_0$ in line 7. Line 6 and line 7 correspond to the practical implementation of RLHF (Ouyang et al., 2022; Bai et al., 2022; Touvron et al., 2023) given a dataset of preference feedback.

In the second stage, we sample n context-action pairs $\{(x_i,a^1_i,a^2_i,y_i)\}_{i=1}^n$ using the intermediate policy $\pi^\eta_{\widehat{\theta}_0}$ and call the preference oracle to get the preference labels. We then compute the MLE estimator of the reward function $\widehat{\theta}$ based on the preference feedback from both stages. Finally, we apply the planning oracle to compute the optimal policy $\pi^\eta_{\widehat{\theta}}$ based on the reward function $\widehat{\theta}$.

E Proofs from Section 2

E.1 Proof of Theorem 2.6

Proof. Consider a simple case when $|\mathcal{X}|=M$ and $|\mathcal{A}|=2$. We suppose that the context x is drawn uniformly from $\mathcal{X}$ at the beginning of each round. Let Θ be the set consisting of mappings from $\mathcal{X}$ to $\mathcal{A}=\{0,1\}$. For each $\theta\in\Theta$, we have $R(\theta,x,a)=\begin{cases}1/2+c & \text{if } a=\theta(x),\\ 1/2 & \text{if } a\neq\theta(x),\end{cases}$ where $c\in(0,1/4)$ is a constant, and $\theta(x)$ is the optimal action under context x when the model is θ.

For any $(\theta,x,a)\in\Theta\times\mathcal{X}\times\mathcal{A}$, we assume the reward feedback $r\sim Bernoulli(R(\theta,x,a))$ when the model is θ and a is chosen under context x.

We pick a pair of model θ_1,θ_2 in Θ, such that $\theta_1(x)=\begin{cases}\theta_2(x) & \text{if } x\neq x_0,\\ 1-\theta_2(x) & \text{if } x=x_0.\end{cases}$

Algorithm 3 Two-stage Mixed-Policy Sampling from Preference Feedback (TMPS-PF)

Input: η, ϵ, π_0, Θ.

▷ Use policy π_0 to achieve sufficient data coverage

for $i = 1, \ldots, m$ **do**

  Sample context $\widetilde{x}_i \sim d_0$ and 2 actions $\widetilde{a}_i^1, \widetilde{a}_i^2 \sim \pi_0(\cdot|\widetilde{x}_i)$.

  Observe preference label $\widetilde{y}_i \in \{0, 1\}$ from the preference oracle defined in Definition 3.1.

end for

Compute the MLE estimator of the reward function based on $\{(\widetilde{x}_i, \widetilde{a}_i^1, \widetilde{a}_i^2, \widetilde{y}_i)\}_{i=1}^m$:

$$\widehat{\theta}_0 \leftarrow \operatorname*{argmax}_{\theta} \sum_{i=1}^{m} \widetilde{y}_i \cdot \log \sigma(R(\theta, \widetilde{x}_i, \widetilde{a}_i^1) - R(\theta, \widetilde{x}_i, \widetilde{a}_i^2)) + (1 - \widetilde{y}_i) \cdot \log \sigma(R(\theta, \widetilde{x}_i, \widetilde{a}_i^2) - R(\theta, \widetilde{x}_i, \widetilde{a}_i^1)).$$

Apply the planning oracle to compute $\pi_{\widehat{\theta}_0}^{\eta}(\cdot|\cdot) \propto \pi_0(\cdot|\cdot) \exp\big(\eta R(\widehat{\theta}_0, \cdot, \cdot)\big)$.

▷ Use policy $\pi_{\widehat{\theta}_0}^{\eta}$ to sample new responses

for $i = 1, \ldots, n$ **do**

  Sample context $x_i \sim d_0$ and 2 actions $a_i^1, a_i^2 \sim \pi_{\widehat{\theta}_0}^{\eta}(\cdot|x_i)$.

  Observe preference label $y_i \in \{0, 1\}$ from the preference oracle defined in Definition 3.1.

end for

Compute the MLE estimator of the reward function using $\{(x_i, a_i^1, a_i^2, y_i)\}_{i=1}^n$ together with $\{(\widetilde{x}_i, \widetilde{a}_i^1, \widetilde{a}_i^2, \widetilde{y}_i)\}_{i=1}^m$:

$$\begin{aligned}\widehat{\theta} \leftarrow \operatorname*{argmax}_{\theta} &\sum_{i=1}^{m} \widetilde{y}_i \cdot \log \sigma(R(\theta, \widetilde{x}_i, \widetilde{a}_i^1) - R(\theta, \widetilde{x}_i, \widetilde{a}_i^2)) + (1 - \widetilde{y}_i) \cdot \log \sigma(R(\theta, \widetilde{x}_i, \widetilde{a}_i^2) - R(\theta, \widetilde{x}_i, \widetilde{a}_i^1)) \\ &+ \sum_{i=1}^{n} y_i \cdot \log \sigma(R(\theta, x_i, a_i^1) - R(\theta, x_i, a_i^2)) + (1 - y_i) \cdot \log \sigma(R(\theta, x_i, a_i^2) - R(\theta, x_i, a_i^1))\end{aligned}$$

Output $\pi_{\widehat{\theta}}^{\eta}(\cdot|\cdot) \propto \pi_0(\cdot|\cdot) \exp\big(\eta R(\widehat{\theta}, \cdot, \cdot)\big)$.

We denote by $\mathbb{P}_\theta$, $\mathbb{E}_\theta$ the probability measure and expectation under the model θ. Let $N(x)$ be the number of times the context x is observed in the first T rounds for an $x \in \mathcal{X}$.

For two Bernoulli random variables X and Y with parameters $1/2 - c$ and $1/2 + c$, we have

$$\begin{aligned}\mathrm{KL}(X\|Y) &= (1/2 - c) \log \frac{1/2 - c}{1/2 + c} + (1/2 + c) \log \frac{1/2 + c}{1/2 - c} \\ &= 2c \cdot \log \frac{1 + 2c}{1 - 2c} \leq 16c^2\end{aligned}$$

where the inequality follows from the fact that $\log(1 + x) \leq x$ for $x \geq 0$ and $c \in (0, 1/4)$.

Applying Pinsker's inequality (Lemma G.3), we have for all event A measurable with respect to the filtration generated by the observations,

$$|\mathbb{P}_{\theta_1}(A) - \mathbb{P}_{\theta_2}(A)| \leq \sqrt{8c^2 \mathbb{E}_{\theta_1}[N(x_0)]} = \sqrt{8c^2 T/M},$$

where the first inequality follows from the chain rule of KL divergence, and the fact that $\mathbb{E}_{\theta_1}[N(x_0)] = T/M$.

Set A to be the event that $\pi_{\text{out}}(\theta_1(x_0)|x_0) > 1/2$. Then we have

$$\mathbb{P}_{\theta_1}(\pi_{\text{out}}(\theta_1(x_0)|x_0) \leq 1/2) + \mathbb{P}_{\theta_2}(\pi_{\text{out}}(\theta_2(x_0)|x_0) \leq 1/2) \geq 1 - |\mathbb{P}_{\theta_1}(A) - \mathbb{P}_{\theta_2}(A)| \geq 1 - \sqrt{8c^2 T/M}.$$

If the model θ is uniformly drawn from Θ, then we have

$$\mathbb{E}_{\theta \sim \text{Unif}(\Theta)} \mathbb{P}_\theta(\pi_{\text{out}}(\theta(x_0)) \leq 1/2) \geq \frac{1}{2} - \sqrt{2c^2 T/M}$$

for an arbitrary x_0.

Then we consider the following suboptimality, where we set π_0 to be the uniform random policy:

$$\mathbb{E}_{\pi_{\theta_*}^{\eta}}\left[R(\theta_*, x, a) - \frac{1}{\eta} \log \frac{\pi_{\theta_*}^{\eta}(a|x)}{\pi_0(a|x)}\right] - \mathbb{E}_{\pi_{\text{out}}}\left[R(\theta_*, x, a) - \frac{1}{\eta} \log \frac{\pi_{\text{out}}(a|x)}{\pi_0(a|x)}\right]$$

$$= \frac{1}{\eta}\mathbb{E}_{\pi^{\eta}_{\theta_*}}\left[\log\frac{\pi_0(a|x)\cdot\exp\big(\eta R(\theta_*,x,a)\big)}{\pi^{\eta}_{\theta_*}(a|x)}\right] - \frac{1}{\eta}\mathbb{E}_{\pi_{\text{out}}}\left[\log\frac{\pi_0(a|x)\cdot\exp\big(\eta R(\theta_*,x,a)\big)}{\pi_{\text{out}}(a|x)}\right]$$
$$= \frac{1}{\eta}\mathbb{E}_{\pi_{\text{out}}}\left[\log\frac{\pi_{\text{out}}(a|x)}{\pi^*(a|x)}\right],$$

where the last equality follows from the fact that $\pi^{\eta}_{\theta_*} \propto \pi_0(a|x)\cdot\exp(\eta R(\theta_*,x,a))$. Note that we handle the difference in reward and the KL-divergence term together, which is distinct from the standard analysis of the lower bound for contextual bandits.

To bound the suboptimality gap, we further have

$$\begin{aligned}
&\mathbb{E}_{\theta\sim\text{Unif}(\Theta)}\mathbb{E}_{\pi_{\text{out}}}\left[\log\frac{\pi_{\text{out}}(a|x)}{\pi^*(a|x)}\right]\\
&= \mathbb{E}_{\theta\sim\text{Unif}(\Theta)}\frac{1}{M}\sum_{x\in\mathcal{X}}\mathbb{E}_{a\sim\pi_{\text{out}}(\cdot|x)}\left[\log\frac{\pi_{\text{out}}(a|x)}{\pi^*(a|x)}\right]\\
&\geq \mathbb{E}_{\theta\sim\text{Unif}(\Theta)}\frac{1}{M}\sum_{x\in\mathcal{X}}\mathbb{P}_\theta(\pi_{\text{out}}(\theta(x))\leq 1/2)\cdot\left[\frac{1}{2}\log\frac{1+\exp(-\eta c)}{2}+\frac{1}{2}\log\frac{1+\exp(\eta c)}{2}\right]\\
&\geq \left(\frac{1}{2}-\sqrt{2c^2T/M}\right)\left[\frac{1}{2}\log\frac{1+\exp(-\eta c)}{2}+\frac{1}{2}\log\frac{1+\exp(\eta c)}{2}\right],
\end{aligned}\tag{E.1}$$

where the first inequality follows from the fact that $\text{KL}(\pi_{\text{out}}(\cdot|x)\|\pi^*(\cdot|x)) \geq \text{KL}(\pi_{\text{unif}}(\cdot|x)\|\pi^*(\cdot|x))$ if $\pi_{\text{out}}(\theta(x)) \leq 1/2$. Here π_{unif} is the uniform distribution over $\mathcal{A}$. Note that

$$\frac{\mathrm{d}}{\mathrm{d}u}\left[\frac{1}{2}\log\frac{1+e^{-u}}{2}+\frac{1}{2}\log\frac{1+e^{u}}{2}\right]\Big|_{u=0} = \frac{1}{2}\left[\frac{1}{1+\exp(-u)}-\frac{1}{1+\exp(u)}\right]\Big|_{u=0} = 0,$$
$$\frac{\mathrm{d}^2}{\mathrm{d}u^2}\left[\frac{1}{2}\log\frac{1+e^{-u}}{2}+\frac{1}{2}\log\frac{1+e^{u}}{2}\right] = \frac{\exp(u)}{[1+\exp(u)]^2}.$$

Thus, applying Taylor's expansion on the right-hand side of (E.1), we have

$$\mathbb{E}_{\theta\sim\text{Unif}(\Theta)}\mathbb{E}_{\pi_{\text{out}}}\left[\log\frac{\pi_{\text{out}}(a|x)}{\pi^*(a|x)}\right] \geq \frac{1}{2}\cdot\left(\frac{1}{2}-\sqrt{2c^2T/M}\right)\eta^2c^2\cdot\frac{1}{3+\exp(\eta c)},$$

which follows from the Taylor's expansion where $f(x) = f(0)+f'(0)x+\frac{1}{2}f''(z)x^2$ where $z\in[0,\eta c]$, and the fact that $\frac{1}{2}\frac{e^z}{(1+e^z)^2} = \frac{1}{2}\frac{1}{e^{-z}+e^z+2} \leq \frac{1}{2}\frac{1}{3+e^{\eta c}}$.

When $\epsilon < 1/64\eta$, we can set $c = 8\sqrt{\epsilon/\eta}$. To achieve a suboptimality gap of ϵ, we need to satisfy:

$$\frac{1}{2}\cdot\left(\frac{1}{2}-\sqrt{2c^2T/M}\right)\eta^2c^2\cdot\frac{1}{3+\exp(\eta c)} \leq \eta\epsilon,$$

indicating that $T \geq \frac{\eta M}{2048\epsilon} = \Omega(\frac{\eta M}{\epsilon})$.

When $\epsilon \geq 1/64\eta$, or equivalently, $\eta \geq 1/64\epsilon$, we employ a different lower bound for (E.1) as follows:

$$\begin{aligned}
\frac{1}{2}\log\frac{1+\exp(-\eta c)}{2}+\frac{1}{2}\log\frac{1+\exp(\eta c)}{2} &= \frac{1}{2}\log\frac{2+\exp(\eta c)+\exp(-\eta c)}{4}\\
&\geq \frac{1}{2}\cdot\frac{1}{2}\left(\log\frac{\exp(\eta c)+\exp(-\eta c)}{2}\right)\\
&\geq \frac{1}{4}(\eta c-\log 2),
\end{aligned}\tag{E.2}$$

where the first inequality follows from Jensen's inequality.

Substituting (E.2) into (E.1), we have

$$\epsilon \geq \frac{1}{\eta}\mathbb{E}_{\theta\sim Unif(\Theta)}\mathbb{E}_{\pi_{\text{out}}}\left[\log\frac{\pi_{\text{out}}(a|x)}{\pi^*(a|x)}\right] \geq \frac{1}{4}\cdot\left(\frac{1}{2}-\sqrt{c^2T/2M}\right)(\eta c-\log 2)\cdot\frac{1}{\eta}.$$

Set $c = 64\epsilon$. Then we have $T = \Omega(M/\epsilon^2)$.

Note that for the given ϵ, $N_{\mathcal{R}}(\epsilon) = M$, and since d_0 is the uniform distribution over $\mathcal{X}$, we have $D^2 = O(M)$, which completes the proof.

□

E.2 Proof of Theorem 2.8

We start with the following lemma, which provides an on-policy generalization bound for the reward function. Due to the on-policy nature of the algorithm (i.e., the usage of intermediate $\pi^{\eta}_{\widehat{\theta}_0}$), we can leverage the covering number of the reward function class $\mathcal{R}$ to derive the generalization error. Since we are using a fixed policy $\pi^{\eta}_{\widehat{\theta}_0}$ to sample in the second stage, we can derive the generalization error of the reward function as follows:

Lemma E.1 (Generalization error of reward function)**.** For an arbitrary policy π, a set of context-action pairs $\{(x_i, a_i)\}_{i=1}^n$ generated i.i.d. from π, and a distance threshold $0 < \epsilon_c \leq B$, we have with probability at least $1-\delta$, for any pair of parameters θ_1 and θ_2,

$$\begin{aligned}
&\mathbb{E}_\pi |R(\theta_1, x, a) - R(\theta_2, x, a)|^2 \\
&\leq \frac{2}{n}\sum_{i=1}^n |R(\theta_1, x_i, a_i) - R(\theta_2, x_i, a_i)|^2 + \frac{32B^2}{3n}\log(2N_{\mathcal{R}}(\epsilon_c)/\delta) + 10\epsilon_c B.
\end{aligned}$$

Proof. We first consider an ϵ_c-net $\mathcal{R}^c$ of the reward function class $\mathcal{R}$ where $\mathcal{R}^c = \{R(\theta, \cdot, \cdot) | \theta \in \Theta^c\}$ with size $N_{\mathcal{R}}(\epsilon_c)$. For any $R(\theta, \cdot, \cdot) \in \mathcal{R}$, there exists θ^c such that $\|R(\theta, \cdot, \cdot) - R(\theta^c, \cdot, \cdot)\|_\infty \leq \epsilon_c$.

By Lemma G.1, for each pair of $\theta_1^c, \theta_2^c \in \Theta^c$ (corresponding to θ_1, θ_2), we have with probability at least $1-\delta$,

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^n (R(\theta_1^c, x_i, a_i) - R(\theta_2^c, x_i, a_i))^2 - \mathbb{E}_\pi |R(\theta_1^c, x, a) - R(\theta_2^c, x, a)|^2\right| \\
&\leq \sqrt{\frac{2\mathrm{Var}_\pi |R(\theta_1^c, x, a) - R(\theta_2^c, x, a)|^2}{n}\log(2/\delta)} + \frac{2}{3n}B^2\log(2/\delta) \\
&\leq \sqrt{\frac{2B^2\mathbb{E}_\pi |R(\theta_1^c, x, a) - R(\theta_2^c, x, a)|^2}{n}\log(2/\delta)} + \frac{2}{3n}B^2\log(2/\delta)
\end{aligned}$$

where the second inequality follows from the fact that $R(\theta_1^c, x, a), R(\theta_2^c, x, a) \leq B$.

Using union bound over all $\theta_1^c, \theta_2^c \in \Theta^c$, we have with probability at least $1-\delta$, for all $\theta_1^c, \theta_2^c \in \Theta^c$,

$$\begin{aligned}
&\mathbb{E}_\pi |R(\theta_1^c, x, a) - R(\theta_2^c, x, a)|^2 - \frac{1}{n}\sum_{i=1}^n (R(\theta_1^c, x_i, a_i) - R(\theta_2^c, x_i, a_i))^2 \\
&\leq \sqrt{\frac{4B^2\mathbb{E}_\pi |R(\theta_1^c, x, a) - R(\theta_2^c, x, a)|^2}{n}\log(2N_{\mathcal{R}}(\epsilon_c)/\delta)} + \frac{4B^2}{3n}\log(2N_{\mathcal{R}}(\epsilon_c)/\delta),
\end{aligned}$$

from which we further obtain the following inequality by Lemma G.2,

$$\mathbb{E}_\pi |R(\theta_1^c, x, a) - R(\theta_2^c, x, a)|^2 \leq \frac{2}{n}\sum_{i=1}^n (R(\theta_1^c, x_i, a_i) - R(\theta_2^c, x_i, a_i))^2 + \frac{32B^2}{3n}\log(2N_{\mathcal{R}}(\epsilon_c)/\delta). \tag{E.3}$$

Then we can complete the proof by the definition of ϵ-net. □

Next, we provide the following lemma, which gives an upper bound on the cumulative square error of the learned reward function.

Lemma E.2 (Confidence bound for reward function)**.** For an arbitrary policy π, and a set of data $\{(x_i, a_i, r_i)\}_{i=1}^n$ generated i.i.d. from π, suppose that $\widehat{\theta}$ is the least squares estimator of θ_*, i.e., $\widehat{\theta} = \arg\min_{\theta\in\Theta} \sum_{i=1}^n (R(\theta, x_i, a_i) - r_i)^2$. Then for any threshold $\epsilon_c > 0$, with probability at least $1-\delta$, it holds that

$$\sum_{i=1}^n (R(\widehat{\theta}, x_i, a_i) - R(\theta_*, x_i, a_i))^2 \leq 16B^2\log(2N_{\mathcal{R}}(\epsilon_c)/\delta) + 4\epsilon_c nB.$$

Proof. We have the following inequality for $\sum_{i=1}^n (R(\widehat{\theta}, x_i, a_i) - R(\theta_*, x_i, a_i))^2$,

$$\begin{aligned}
&\sum_{i=1}^n (R(\widehat{\theta}, x_i, a_i) - R(\theta_*, x_i, a_i))^2 \\
&= \sum_{i=1}^n (R(\widehat{\theta}, x_i, a_i) - r_i)^2 - \sum_{i=1}^n (R(\theta_*, x_i, a_i) - r_i)^2 \\
&\quad + 2\sum_{i=1}^n (R(\widehat{\theta}, x_i, a_i) - R(\theta_*, x_i, a_i)(r_i - R(\theta_*, x_i, a_i)) \\
&\leq 2\sum_{i=1}^n (R(\widehat{\theta}, x_i, a_i) - R(\theta_*, x_i, a_i))(r_i - R(\theta_*, x_i, a_i)),
\end{aligned}$$

where the last inequality follows from the fact that $\sum_{i=1}^n (R(\widehat{\theta}, x_i, a_i) - r_i)^2 \leq \sum_{i=1}^n (R(\theta_*, x_i, a_i) - r_i)^2$.

We then consider an ϵ_c-net $\mathcal{R}^c$ of the reward function class $\mathcal{R}$ where $\mathcal{R}^c = \{R(\theta, \cdot, \cdot) | \theta \in \Theta^c\}$ with size $N_{\mathcal{R}}(\epsilon_c)$. For any $R(\theta, \cdot, \cdot) \in \mathcal{R}$, there exists θ^c such that $\|R(\theta, x, a) - R(\theta^c, x, a)\|_\infty \leq \epsilon_c$. From Azuma-Hoeffding inequality, with probability at least $1 - \delta$, it holds for all $\theta \in \Theta^c$ that

$$\begin{aligned}
&\sum_{i=1}^n (R(\theta, x_i, a_i) - R(\theta_*, x_i, a_i))(r_i - R(\theta_*, x_i, a_i)) \\
&\leq \sqrt{2B^2 \sum_{i=1}^n (R(\theta, x_i, a_i) - R(\theta_*, x_i, a_i))^2 \log(2N_{\mathcal{R}}(\epsilon_c)/\delta)}.
\end{aligned}$$

Then we further have with probability at least $1 - \delta$, there exists $\|R(\theta^c, \cdot, \cdot) - R(\widehat{\theta}, \cdot, \cdot)\| \leq \epsilon_c$ such that

$$\begin{aligned}
&\sum_{i=1}^n (R(\widehat{\theta}, x_i, a_i) - R(\theta_*, x_i, a_i))(r_i - R(\theta_*, x_i, a_i)) \\
&\leq \sqrt{2B^2 \sum_{i=1}^n (R(\theta, x_i, a_i) - R(\theta_*, x_i, a_i))^2 \log(2N_{\mathcal{R}}(\epsilon_c)/\delta) + 2\epsilon_c nB},
\end{aligned}$$

which implies that

$$\sum_{i=1}^n (R(\widehat{\theta}, x_i, a_i) - R(\theta_*, x_i, a_i))^2 \leq 16B^2 \log(2N_{\mathcal{R}}(\epsilon_c)/\delta) + 4\epsilon_c nB \tag{E.4}$$

by Lemma G.2. □

With the above lemmas, we are now ready to prove the following lemma that bounds the estimation error of the reward function $R(\widehat{\theta}, \cdot, \cdot)$ under the sampled policy $\pi^\eta_{\widehat{\theta}_0}$.

Lemma E.3. Let $\widehat{\theta}_0$ be the least squares estimator of the reward function based on the data $\{(x_i^0, a_i^0, r_i^0)\}_{i=1}^m$ generated from π_0 as defined in Algorithm 1. Then for any threshold $\epsilon_c > 0$, with probability at least $1 - 2\delta$, we have

$$\mathbb{E}_{\pi^\eta_{\widehat{\theta}_0}} |R(\widehat{\theta}, x, a) - R(\theta_*, x, a)|^2 \leq \frac{43B^2}{n} \log(2N_{\mathcal{R}}(\epsilon_c)/\delta) + 10\epsilon_c(1 + m/n)B.$$

Proof. By Lemma E.1, we have with probability at least $1 - \delta$, the following upper bound holds for $\mathbb{E}_{\pi^\eta_{\widehat{\theta}_0}} |R(\theta_1, x, a) - R(\theta_2, x, a)|^2$,

$$\mathbb{E}_{\pi^\eta_{\widehat{\theta}_0}} |R(\theta_1, x, a) - R(\theta_2, x, a)|^2$$

$$\leq \frac{2}{n}\sum_{i=1}^{n}|R(\theta_1,x_i,a_i)-R(\theta_2,x_i,a_i)|^2+\frac{32B^2}{3n}\log(2N_{\mathcal{R}}(\epsilon_c)/\delta)+10\epsilon_c B. \tag{E.5}$$

By Lemma E.2, with probability at least $1-\delta$

$$\sum_{i=1}^{n}|R(\theta_*,x_i,a_i)-R(\widehat{\theta},x_i,a_i)|^2\leq 16B^2\log(2N_{\mathcal{R}}(\epsilon_c)/\delta)+4\epsilon_c(n+m)B. \tag{E.6}$$

Then we can complete the proof using a union bound and substituting (E.6) into (E.5). □

Lemma E.4. If $m\geq 128\eta^2D^2B^2\cdot\log(2N_{\mathcal{R}}(\epsilon_c)/\delta))$, and there exists a positive constant $c_{m,n}>0$ such that $n=c_{m,n}m$ in Algorithm 1 and Assumption 2.5 holds, then by taking $\epsilon_c\leq\min\{B,(8(1+c_{m,n})B\eta^2D^2)^{-1}\}$, with probability at least $1-3\delta$, we have

$$\eta|R(\widehat{\theta}_0,x,a)-R(\theta_*,x,a)|\leq 1,\quad \eta|R(\widehat{\theta},x,a)-R(\theta_*,x,a)|\leq 1$$

for any pair $(x,a)\in\mathcal{X}\times\mathcal{A}$ such that $\pi_0(a|x)>0$.

Proof. By Lemma E.1, with probability at least $1-\delta$, for all $\theta_1,\theta_2\in\Theta$, we have

$$\mathbb{E}_{\pi_0}|R(\theta_1,x,a)-R(\theta_2,x,a)|^2\leq\frac{2}{m}\sum_{i=1}^{m}|R(\theta_1,x_i^0,a_i^0)-R(\theta_2,x_i^0,a_i^0)|^2+\frac{32B^2}{3m}\log(2N_{\mathcal{R}}(\epsilon_c)/\delta).$$

By Lemma E.2, with probability at least $1-\delta$, we have

$$\sum_{i=1}^{m}|R(\widehat{\theta}_0,x_i^0,a_i^0)-R(\theta_*,x_i^0,a_i^0)|^2\leq 16B^2\log(2N_{\mathcal{R}}(\epsilon_c)/\delta)+4\epsilon_c m.$$

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

$$\sum_{i=1}^{m}|R(\theta_*,x_i^0,a_i^0)-R(\widehat{\theta},x_i^0,a_i^0)|^2\leq 16B^2\log(2N_{\mathcal{R}}(\epsilon_c)/\delta)+4\epsilon_c(m+n)B.$$

Similar to the proof of Lemma E.3, we have if $m\geq 128\eta^2D^2B^2\cdot\log(2N_{\mathcal{R}}(\epsilon_c)/\delta)$, $n=c_{m,n}n$, then with probability at least $1-3\delta$,

$$\mathbb{E}_{\pi_0}|R(\theta_*,x,a)-R(\widehat{\theta}_0,x,a)|^2\leq 1/\eta^2D^2,\quad \mathbb{E}_{\pi_0}|R(\theta_*,x,a)-R(\widehat{\theta},x,a)|^2\leq 1/\eta^2D^2.$$

which implies that $\eta|R(\widehat{\theta}_0,x,a)-R(\theta_*,x,a)|\leq 1$ and $\eta|R(\widehat{\theta},x,a)-R(\theta_*,x,a)|\leq 1$ for all $(x,a)\in\mathcal{X}\times\mathcal{A}$ such that $\pi_0(a|x)>0$. □

Lemma E.5 (Restatement of Lemma 2.9)**.** For any estimator $\widehat{\theta}\in\Theta$, and the policy $\pi_{\widehat{\theta}}^{\eta}$ satisfying Definiton 2.4, we have

$$\begin{aligned}Q(\pi^*)-Q(\pi_{\widehat{\theta}}^{\eta})&=\eta\mathbb{E}_{x\sim d_0}\Big[\sum_{a\in\mathcal{A}}\pi_f^{\eta}(a|x)\Delta^2(x,a)-\sum_{a_1,a_2\in\mathcal{A}}\pi_f^{\eta}(a_1|x)\pi_f^{\eta}(a_2|x)\Delta(x,a_1)\Delta(x,a_2)\Big]\\&\leq\eta\mathbb{E}_{x\sim d_0}\Big[\sum_{a\in\mathcal{A}}\pi_f^{\eta}(a|x)\Delta^2(x,a)\Big],\end{aligned}$$

where $\Delta(x,a)=R(\widehat{\theta},x,a)-R(\theta_*,x,a)$, $f(\cdot,\cdot)=\gamma R(\widehat{\theta},\cdot,\cdot)+(1-\gamma)R(\theta_*,\cdot,\cdot)$ ($\gamma\in(0,1)$) the inequality uses the fact that second term on the right-hand side of the equality is $(\sum_{a\in\mathcal{A}}\pi_f^{\eta}(a|x)\Delta(x,a))^2\geq 0$.

Proof of Lemma 2.9. We have

$$\begin{aligned}&\mathbb{E}_{\pi_{\theta_*}^{\eta}}\left[R(\theta_*,x,a)-\frac{1}{\eta}\log\frac{\pi_{\theta_*}^{\eta}(a|x)}{\pi_0(a|x)}\right]-\mathbb{E}_{\pi_{\widehat{\theta}}^{\eta}}\left[R(\theta_*,x,a)-\frac{1}{\eta}\log\frac{\pi_{\widehat{\theta}}^{\eta}(a|x)}{\pi_0(a|x)}\right]\\&=\frac{1}{\eta}\mathbb{E}_{\pi_{\theta_*}^{\eta}}\left[\log\frac{\pi_0(a|x)\cdot\exp\big(\eta R(\theta_*,x,a)\big)}{\pi_{\theta_*}^{\eta}(a|x)}\right]-\frac{1}{\eta}\mathbb{E}_{\pi_{\widehat{\theta}}^{\eta}}\left[\log\frac{\pi_0(a|x)\cdot\exp\big(\eta R(\theta_*,x,a)\big)}{\pi_{\widehat{\theta}}^{\eta}(a|x)}\right]\end{aligned}$$

$$= \frac{1}{\eta}\mathbb{E}_{x\sim d_0}\big[\log Z^\eta_{\theta_*}(x)\big] - \frac{1}{\eta}\mathbb{E}_{x\sim d_0}\big[\log Z^\eta_{\widehat{\theta}}(x)\big] - \mathbb{E}_{x\sim d_0}\bigg[\sum_{a\in\mathcal{A}}\pi^\eta_{\widehat{\theta}}(a|x)\cdot\big(R(\theta_*,x,a) - R(\widehat{\theta},x,a)\big)\bigg],$$

where the first equality follows from the definition of the KL-divergence, the second equality follows from Lemma G.5.

For an arbitrary reward function $f : \mathcal{X}\times\mathcal{A}\to\mathbb{R}$, let $\Delta_f(x,a) = f(x,a) - R(\theta_*,x,a)$. Consider the following first derivative of $J(f) = \log Z^\eta_f(x) - \eta\sum_{a\in\mathcal{A}}\pi^\eta_f(a|x)\cdot\Delta_f(x,a)$, where $Z^\eta_f(x) = \sum_{a\in\mathcal{A}}\pi_0(a|x)\cdot\exp(\eta\cdot f(x,a))$ and $\pi^\eta_f(a|x)\propto\pi_0(a|x)\cdot\exp(\eta\cdot f(x,a))$.

$$\begin{aligned}
&\frac{\partial}{\partial\Delta_f(x,a)}\bigg[\log Z^\eta_f(x) - \eta\sum_{a\in\mathcal{A}}\pi^\eta_f(a|x)\cdot\Delta_f(x,a)\bigg]\\
&= \frac{1}{Z^\eta_f(x)}\cdot\pi_0(a|x)\exp\big(\eta\cdot f(x,a)\big)\cdot\eta - \eta\cdot\pi^\eta_f(a|x)\\
&\quad - \eta\cdot\Delta_f(x,a)\cdot\frac{\pi_0(a|x)\cdot\exp\big(\eta\cdot f(x,a)\big)}{Z^\eta_f(x)}\cdot\eta + \eta\cdot\Delta_f(x,a)\cdot\frac{\big[\pi_0(a|x)\cdot\exp\big(\eta\cdot f(x,a)\big)\big]^2}{[Z^\eta_f(x)]^2}\cdot\eta\\
&\quad + \eta\sum_{a'\in\mathcal{A}\setminus\{a\}}\frac{\pi_0(a'|x)\cdot\exp\big(\eta\cdot f(x,a')\big)}{Z^\eta_f(x)}\cdot\eta\cdot\Delta_f(x,a')\cdot\frac{\pi_0(a|x)\cdot\exp\big(\eta\cdot f(x,a)\big)}{Z^\eta_f(x)}\\
&= -\eta^2\pi^\eta_f(a|x)\Delta_f(x,a) + \eta^2[\pi^\eta_f(a|x)]^2\cdot\Delta_f(x,a) + \eta^2\sum_{a'\in\mathcal{A}\setminus\{a\}}\pi^\eta_f(a'|x)\pi^\eta_f(a|x)\Delta_f(x,a')
\end{aligned}$$

where the first equality is derived by taking the derivative of $\log Z^\eta_f(x)$ and the second term with respect to Δ_f. Therefore, by the Mean Value Theorem, there exists an $f(\cdot,\cdot) = \gamma R(\widehat{\theta},\cdot,\cdot) + (1-\gamma)R(\theta_*,\cdot,\cdot)$ for some $\gamma\in[0,1]$ such that

$$\begin{aligned}
&\mathbb{E}_{x\sim d_0}[J(R(\widehat{\theta},\cdot,\cdot)) - J(R(\theta_*,\cdot,\cdot))] = \frac{1}{\eta}\mathbb{E}_{x\sim d_0}\bigg[-\eta^2\sum_{a\in\mathcal{A}}\pi^\eta_f(a|x)\cdot\gamma\cdot\big(R(\widehat{\theta},x,a) - R(\theta_*,x,a)\big)^2\bigg]\\
&\quad + \frac{1}{\eta}\mathbb{E}_{x\sim d_0}\bigg[\gamma\eta^2\sum_{a_1\in\mathcal{A}}\sum_{a_2\in\mathcal{A}}\pi^\eta_f(a_1|x)\pi^\eta_f(a_2|x)\big(R(\widehat{\theta},x,a_1) - R(\theta_*,x,a_1)\big)\big(R(\widehat{\theta},x,a_2) - R(\theta_*,x,a_2)\big)\bigg]\\
&\geq -\eta\cdot\mathbb{E}_{\pi^\eta_f}\big[\big(R(\widehat{\theta},x,a) - R(\theta_*,x,a)\big)^2\big]
\end{aligned}$$

where the last inequality holds since

$$\begin{aligned}
&\sum_{a_1\in\mathcal{A}}\sum_{a_2\in\mathcal{A}}\pi^\eta_f(a_1|x)\pi^\eta_f(a_2|x)\big(R(\widehat{\theta},x,a_1) - R(\theta_*,x,a_1)\big)\big(R(\widehat{\theta},x,a_2) - R(\theta_*,x,a_2)\big)\\
&= \big[\mathbb{E}_{a\sim\pi^\eta_f(\cdot|x)}[R(\widehat{\theta},x,a) - R(\theta_*,x,a)]\big]^2 \geq 0.
\end{aligned}$$

□

Now, we are ready to prove the theorem.

Proof of Theorem 2.8. By Lemma 2.9, we have

$$Q(\pi^*) - Q(\pi^\eta_{\widehat{\theta}}) \leq \eta\mathbb{E}_{x\sim d_0}\bigg[\sum_{a\in\mathcal{A}}\pi^\eta_f(a|x)\big(R(\widehat{\theta},x,a) - R(\theta_* x,a)\big)^2\bigg].$$

By Lemma E.4, if $m\geq 128\eta^2D^2B^2\cdot\log(2N_{\mathcal{R}}(\epsilon_c)/\delta)$, for any $(x,a)\in\mathcal{X}\times\mathcal{A}$ such that $\pi_0(a|x) > 0$, it holds that

$$\eta|R(\widehat{\theta}_0,x,a) - R(\theta_*,x,a)|\leq 1,\quad \eta|R(\widehat{\theta},x,a) - R(\theta_*,x,a)|\leq 1,$$

which means that for any $(x,a)\in\mathcal{X}\times\mathcal{A}$

$$\frac{\pi^\eta_f(a|x)}{\pi^\eta_{\widehat{\theta}_0}(a|x)}\leq e^4.$$

Let $\epsilon_c = \min\{\frac{\epsilon}{(1+c^{-1}_{m,n})B}, \frac{1}{8(1+c_{m,n})B\eta^2D^2}, B\}$. By Lemma E.3, if $m\geq 128\eta^2D^2B^2\cdot\log(2N_{\mathcal{R}}(\epsilon_c)/\delta)$ and $n\geq\eta/\epsilon\cdot B^2\log(N_{\mathcal{R}}(\epsilon_c)/\delta)$ and $n = c_{m,n}m$ then with high probability the output policy $\pi^\eta_{\widehat{\theta}}$ is $O(\epsilon)$ optimal. □

E.3 Proof of Corollary 2.13

Proof of Corollary 2.13. The proof follows the same lines as Theorem 3.7 by replacing the data coverage condition with the local-coverage condition. It still holds that

$$Q(\pi^*) - Q(\pi^{\eta}_{\widehat{\theta}_0}) \le \eta \cdot \mathbb{E}_{\pi^{\eta}_f}\left[\left(R(\widehat{\theta}_0, x, a) - R(\theta_*, x, a)\right)^2\right]$$

where $\pi^{\eta}_f(a|x) \propto \pi_0(a|x) \cdot \exp(\eta \cdot f(x,a))$ and $f(\cdot,\cdot) = \gamma R(\widehat{\theta}_0, \cdot, \cdot) + (1-\gamma)R(\theta_*, \cdot, \cdot)$ for some $\gamma \in (0,1)$. Thus, We have $\mathrm{KL}(\pi^{\eta}_f(a|x)\|\pi_0) \le 2\eta B$, which further implies that

$$Q(\pi^*) - Q(\pi^{\eta}_{\widehat{\theta}}) \le \eta \cdot C_{\rho_{\mathrm{KL}}} \cdot O\left(\frac{1}{n} B \log(N_{\mathcal{R}}(\epsilon_c)/\delta) + B(1 + c^{-1}_{m,n})\epsilon_c\right)$$

by Lemma F.4. Then we can conclude by substituting the value of m into the suboptimality gap. □

F Proofs from Section 3

F.1 Proof of Theorem 3.6

Proof of Theorem 3.6. The proof follows a similar construction as the one for Theorem 2.6. Consider a simple case when $|\mathcal{X}| = M$ and $|\mathcal{A}| = 2$. We suppose that the context x is drawn uniformly from $\mathcal{X}$ at the beginning of each round. Let Θ be the set consisting of mappings from $\mathcal{X}$ to $\mathcal{A} = \{0,1\}$. For each $\theta \in \Theta$, we have $R(\theta, x, a) = \begin{cases} c & \text{if } a = \theta(x), \\ 0 & \text{if } a \neq \theta(x), \end{cases}$ where $c \in (0, 1/4)$ is a constant, and $\theta(x)$ is the optimal action under context x when the model is θ.

We pick a pair of model θ_1, θ_2 in Θ, such that $\theta_1(x) = \begin{cases} \theta_2(x) & \text{if } x \neq x_0, \\ 1 - \theta_2(x) & \text{if } x = x_0. \end{cases}$

We denote by $\mathbb{P}_\theta$, $\mathbb{E}_\theta$ the probability measure and expectation under the model θ.

We have the following upper bound for two Bernoulli distribution $y_1 \sim Bernoulli(\sigma(c))$ and $y_2 \sim Bernoulli(\sigma(-c))$ with $\sigma(x) = 1/(1+\exp(-x))$:

$$\begin{aligned}
\sigma(c)\log\frac{\sigma(c)}{\sigma(-c)} + \sigma(-c)\log\frac{\sigma(-c)}{\sigma(c)} &= 2\cdot(\frac{1}{2} - \sigma(-c))\log\frac{\sigma(c)}{\sigma(-c)} \\
&= \frac{1-e^{-c}}{1+e^{-c}} \cdot \log\frac{1+e^{c}}{1+e^{-c}} \\
&\le \frac{1-e^{-c}}{1+e^{-c}} \cdot \frac{e^{c}-e^{-c}}{1+e^{-c}} \\
&\le \frac{c\cdot(c+e^{\frac{1}{4}}c)}{[(1+e^{-\frac{1}{4}})]^2} \le c^2,
\end{aligned}$$

where the first equality follows from the fact that $\sigma(-c) = 1 - \sigma(c)$, the first inequality holds since $\log x \le x - 1$, and the second inequality holds since $c \le 1/4$.

Applying Pinsker's inequality (Lemma G.3), we have for all event A measurable with respect to the filtration generated by the observations,

$$|\mathbb{P}_{\theta_1}(A) - \mathbb{P}_{\theta_2}(A)| \le \sqrt{c^2 \mathbb{E}_{\theta_1}[N(x_0)]} = \sqrt{c^2 T/M},$$

where the first inequality follows from the chain rule of KL divergence, and the fact that $\mathbb{E}_{\theta_1}[N(x_0)] = T/M$.

Set A to be the event that $\pi_{\text{out}}(\theta_1(x_0)|x_0) > 1/2$. Then we have

$$\mathbb{P}_{\theta_1}(\pi_{\text{out}}(\theta_1(x_0)|x_0) \le 1/2) + \mathbb{P}_{\theta_2}(\pi_{\text{out}}(\theta_2(x_0)|x_0) \le 1/2) \ge 1 - |\mathbb{P}_{\theta_1}(A) - \mathbb{P}_{\theta_2}(A)| \ge 1 - \sqrt{c^2 T/M}.$$

If the model θ is uniformly drawn from Θ, then we have

$$\mathbb{E}_{\theta\sim\mathrm{Unif}(\Theta)}\mathbb{P}_\theta(\pi_{\text{out}}(\theta(x_0)) \le 1/2) \ge \frac{1}{2} - \sqrt{c^2 T/4M}$$

for an arbitrary x_0.

Then we consider the following suboptimality gap:

$$\begin{aligned}
&\mathbb{E}_{\pi^{\eta}_{\theta_*}}\left[R(\theta_*,x,a)-\frac{1}{\eta}\log\frac{\pi^{\eta}_{\theta_*}(a|x)}{\pi_0(a|x)}\right]-\mathbb{E}_{\pi_{\text{out}}}\left[R(\theta_*,x,a)-\frac{1}{\eta}\log\frac{\pi_{\text{out}}(a|x)}{\pi_0(a|x)}\right]\\
&=\frac{1}{\eta}\mathbb{E}_{\pi^{\eta}_{\theta_*}}\left[\log\frac{\pi_0(a|x)\cdot\exp\big(\eta R(\theta_*,x,a)\big)}{\pi^{\eta}_{\theta_*}(a|x)}\right]-\frac{1}{\eta}\mathbb{E}_{\pi_{\text{out}}}\left[\log\frac{\pi_0(a|x)\cdot\exp\big(\eta R(\theta_*,x,a)\big)}{\pi_{\text{out}}(a|x)}\right]\\
&=\frac{1}{\eta}\mathbb{E}_{\pi_{\text{out}}}\left[\log\frac{\pi_{\text{out}}(a|x)}{\pi^*(a|x)}\right],
\end{aligned}$$

where the last equality follows from the fact that $\pi^{\eta}_{\theta_*}\propto\pi_0(a|x)\cdot\exp(\eta R(\theta_*,x,a))$. To bound the suboptimality gap, we further have

$$\begin{aligned}
&\mathbb{E}_{\theta\sim\text{Unif}(\Theta)}\mathbb{E}_{\pi_{\text{out}}}\left[\log\frac{\pi_{\text{out}}(a|x)}{\pi^*(a|x)}\right]\\
&=\mathbb{E}_{\theta\sim\text{Unif}(\Theta)}\frac{1}{M}\sum_{x\in\mathcal{X}}\mathbb{E}_{a\sim\pi_{\text{out}}(\cdot|x)}\left[\log\frac{\pi_{\text{out}}(a|x)}{\pi^*(a|x)}\right]\\
&\geq\mathbb{E}_{\theta\sim\text{Unif}(\Theta)}\frac{1}{M}\sum_{x\in\mathcal{X}}\mathbb{P}_{\theta}(\pi_{\text{out}}(\theta(x))\leq 1/2)\cdot\left[\frac{1}{2}\log\frac{1+\exp(-\eta c)}{2}+\frac{1}{2}\log\frac{1+\exp(\eta c)}{2}\right]\\
&\geq\Big(\frac{1}{2}-\sqrt{c^2T/4M}\Big)\left[\frac{1}{2}\log\frac{1+\exp(-\eta c)}{2}+\frac{1}{2}\log\frac{1+\exp(\eta c)}{2}\right]
\end{aligned}\tag{F.1}$$

Note that

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}u}\left[\frac{1}{2}\log\frac{1+e^{-u}}{2}+\frac{1}{2}\log\frac{1+e^{u}}{2}\right]\Big|_{u=0}=\frac{1}{2}\left[\frac{1}{1+\exp(-u)}-\frac{1}{1+\exp(u)}\right]\Big|_{u=0}=0,\\
&\frac{\mathrm{d}^2}{\mathrm{d}u^2}\left[\frac{1}{2}\log\frac{1+e^{-u}}{2}+\frac{1}{2}\log\frac{1+e^{u}}{2}\right]=\frac{\exp(u)}{[1+\exp(u)]^2}.
\end{aligned}$$

Thus, applying Taylor's expansion on the right-hand side of (F.1), we have

$$\mathbb{E}_{\theta\sim\text{Unif}(\Theta)}\mathbb{E}_{\pi_{\text{out}}}\left[\log\frac{\pi_{\text{out}}(a|x)}{\pi^*(a|x)}\right]\geq\frac{1}{2}\cdot\Big(\frac{1}{2}-\sqrt{c^2T/4M}\Big)\eta^2c^2\cdot\frac{1}{3+\exp(\eta c)}$$

When $\epsilon<1/64\eta$, we can set $c=8\sqrt{\epsilon/\eta}$. To achieve a suboptimality gap of ϵ, we need to satisfy:

$$\frac{1}{2}\cdot\Big(\frac{1}{2}-\sqrt{c^2T/4M}\Big)\eta^2c^2\cdot\frac{1}{3+\exp(\eta c)}\leq\eta\epsilon,$$

indicating that $T\geq\frac{\eta M}{512\epsilon}=\Omega(\frac{\eta M}{\epsilon})$.

When $\epsilon\geq 1/64\eta$, or equivalently, $\eta\geq 1/64\epsilon$, we employ a different lower bound for (E.1) as follows:

$$\begin{aligned}
\frac{1}{2}\log\frac{1+\exp(-\eta c)}{2}+\frac{1}{2}\log\frac{1+\exp(\eta c)}{2}&=\frac{1}{2}\log\frac{2+\exp(\eta c)+\exp(-\eta c)}{4}\\
&\geq\frac{1}{2}\cdot\frac{1}{2}\Big(\log\frac{\exp(\eta c)+\exp(-\eta c)}{2}\Big)\\
&\geq\frac{1}{4}(\eta c-\log 2),
\end{aligned}\tag{F.2}$$

where the first inequality follows from Jensen's inequality. Substituting (F.2) into (F.1), we have

$$\epsilon\geq\frac{1}{\eta}\mathbb{E}_{\theta\sim\text{Unif}(\Theta)}\mathbb{E}_{\pi_{\text{out}}}\left[\log\frac{\pi_{\text{out}}(a|x)}{\pi^*(a|x)}\right]\geq\frac{1}{4}\cdot\Big(\frac{1}{2}-\sqrt{c^2T/4M}\Big)(\eta c-\log 2)\cdot\frac{1}{\eta}.$$

Set $c=64\epsilon$. Then we have $T=\Omega(M/\epsilon^2)$.

□

F.2 Proof of Theorem 3.7

First, we provide the following lemma for the connection between the likelihood loss and the reward difference, which is a key step to upper bound the reward difference between $\widehat{\theta}$ and θ_*.

Lemma F.1. For an arbitrary policy π, and a set of context-action pairs $\{(x_i, a_i^1, a_i^2, y_i)\}_{i=1}^n$ generated i.i.d. from the BT model and π, we have with probability at least $1-\delta$, for any $s \le n$,

$$\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{s} \mathcal{L}(\theta|x_i, a_i^1, a_i^2, y_i) - \mathcal{L}(\theta_*|x_i, a_i^1, a_i^2, y_i) \\
&\le \log(1/\delta) - \frac{1}{8}e^{-B}\sum_{i=1}^{s}\big([R(\theta, x_i, a_i^2) - R(\theta, x_i, a_i^1)] - [R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)]\big)^2
\end{aligned}$$

Proof. Applying Lemma G.4 to the sequence

$$\left\{-\frac{1}{2}y_i\cdot\log\frac{\sigma(R(\theta_*, x_i, a_i^1) - R(\theta_*, x_i, a_i^2))}{\sigma(R(\theta, x_i, a_i^1) - R(\theta, x_i, a_i^2))} - \frac{1}{2}(1-y_i)\cdot\log\frac{\sigma(R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1))}{\sigma(R(\theta, x_i, a_i^2) - R(\theta, x_i, a_i^1))}\right\}_{i=1}^{n},$$

We have with probability at least $1-\delta$, for all $s \le n$,

$$\begin{aligned}
&\frac{1}{2}\sum_{i=1}^{s} \mathcal{L}(\theta|x_i, a_i^1, a_i^2, y_i) - \mathcal{L}(\theta_*|x_i, a_i^1, a_i^2, y_i) \\
&\le \log(1/\delta) + \sum_{i=1}^{s}\log\Bigg(\sqrt{\sigma(R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1))\cdot\sigma(R(\theta, x_i, a_i^2) - R(\theta, x_i, a_i^1))} \\
&\qquad + \sqrt{\sigma(R(\theta_*, x_i, a_i^1) - R(\theta_*, x_i, a_i^2))\cdot\sigma(R(\theta, x_i, a_i^1) - R(\theta, x_i, a_i^2))}\Bigg) \\
&\le \log(1/\delta) + \sum_{i=1}^{s}\Bigg(\sqrt{\sigma(R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1))\cdot\sigma(R(\theta, x_i, a_i^2) - R(\theta, x_i, a_i^1))} \\
&\qquad + \sqrt{\sigma(R(\theta_*, x_i, a_i^1) - R(\theta_*, x_i, a_i^2))\cdot\sigma(R(\theta, x_i, a_i^1) - R(\theta, x_i, a_i^2))} - 1\Bigg) \\
&= \log(1/\delta) - \frac{1}{2}\sum_{i=1}^{s}\left(\sqrt{\sigma(R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1))} - \sqrt{\sigma(R(\theta, x_i, a_i^2) - R(\theta, x_i, a_i^1))}\right)^2 \\
&\quad - \frac{1}{2}\sum_{i=1}^{s}\left(\sqrt{\sigma(R(\theta_*, x_i, a_i^1) - R(\theta_*, x_i, a_i^2))} - \sqrt{\sigma(R(\theta, x_i, a_i^1) - R(\theta, x_i, a_i^2))}\right)^2 \\
&\le \log(1/\delta) - \frac{1}{8}\sum_{i=1}^{s}\big(\sigma(R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)) - \sigma(R(\theta, x_i, a_i^2) - R(\theta, x_i, a_i^1))\big)^2 \\
&\le \log(1/\delta) - \frac{1}{8}e^{-B}\sum_{i=1}^{s}\big([R(\theta, x_i, a_i^2) - R(\theta, x_i, a_i^1)] - [R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)]\big)^2,
\end{aligned}$$

where the second inequality holds due to $\log(1+r) \le r$ for $r > -1$, the equality follows from the fact that $\sigma(r) + \sigma(-r) = 1$ and the last inequality holds since $\sigma'(r) = \sigma(r)\cdot(1-\sigma(r)) \ge e^{-B}$ for all $r \in [-B, B]$. □

To further control the error bound for the reward function with the help of Lemma F.1, we introduce the following lemma to show that the likelihood function class $\mathcal{L}$ can be well-covered by the ϵ-net of the reward function class $\mathcal{R}$.

Lemma F.2 (Covering number of $\mathcal{L}$)**.** For any $\epsilon_c > 0$, consider an ϵ_c-net $\mathcal{R}^c = \{R(\theta, \cdot, \cdot)|\theta \in \Theta^c\}$ for the reward function class $\mathcal{R}$ with size $N_{\mathcal{R}}(\epsilon_c)$. Then for any $\theta \in \Theta$, there exists $\theta^c \in \Theta^c$ such that for any $s \in [n]$,

$$\sum_{i=1}^{s}\mathcal{L}(\theta|x_i, a_i^1, a_i^2, y_i) \le \sum_{i=1}^{s}\mathcal{L}(\theta^c|x_i, a_i^1, a_i^2, y_i) + 2s\epsilon_c.$$

Proof. For any $r \in \mathbb{R}$, we have

$$\frac{\mathrm{d}\log(\sigma(r))}{\mathrm{d}r} = \frac{1}{\sigma(r)} \cdot \sigma(r) \cdot (1 - \sigma(r)) = 1 - \sigma(r) \in (0, 1).$$

Thus, for any $\theta \in \Theta$, $x \in \mathcal{X}$, $a^1, a^2 \in \mathcal{A}$ and $y \in \{0, 1\}$, there exists $\theta^c \in \Theta^c$ such that

$$\begin{aligned}&\left|\mathcal{L}(\theta|x, a^1, a^2, y) - \mathcal{L}(\theta^c|x, a^1, a^2, y)\right| \\ &\leq \left|[R(\theta, x, a^1) - R(\theta, x, a^2)] - [R(\theta^c, x, a^1) - R(\theta^c, x, a^2)]\right| = 2\epsilon_c.\end{aligned}$$

□

With the above two lemmas, we are now ready to provide the confidence bound for the MLE estimator of the reward function.

Lemma F.3. Consider a set of context-action pairs $\{(x_i, a_i^1, a_i^2, y_i)\}_{i=1}^n$ where labels $\{y_i\}_{i=1}^n$ are generated independently from the BT model. Suppose that $\widehat{\theta}$ is the MLE estimator as defined in Definition 3.3. We have with probability at least $1 - \delta$,

$$\sum_{i=1}^n \left([R(\widehat{\theta}, x_i, a_i^2) - R(\widehat{\theta}, x_i, a_i^1)] - [R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)]\right)^2 \leq O(e^B \log(N_{\mathcal{R}}(\epsilon_c)/\delta) + e^B n\epsilon_c).$$

Proof. By Lemmas F.1 and F.2, we have with probability at least $1 - \delta$, for any $\theta \in \Theta$,

$$\begin{aligned}&\frac{1}{2}\sum_{i=1}^n \mathcal{L}(\theta|x_i, a_i^1, a_i^2, y_i) - \mathcal{L}(\theta_*|x_i, a_i^1, a_i^2, y_i) \\ &\leq \log(N_{\mathcal{R}}(\epsilon_c)/\delta) - \frac{1}{8}e^{-B}\sum_{i=1}^n \left([R(\theta, x_i, a_i^2) - R(\theta, x_i, a_i^1)] - [R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)]\right)^2 + O(n\epsilon_c).\end{aligned}$$

Since $\widehat{\theta}$ is the MLE estimator, we have $\sum_{i=1}^n \mathcal{L}(\theta|x_i, a_i^1, a_i^2, y_i) - \mathcal{L}(\theta_*|x_i, a_i^1, a_i^2, y_i) \geq 0$, which further implies

$$\begin{aligned}0 \leq \log(N_{\mathcal{R}}(\epsilon_c)/\delta) - \frac{1}{8}e^{-B}\sum_{i=1}^n &\left([R(\theta, x_i, a_i^2) - R(\theta, x_i, a_i^1)] - [R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)]\right)^2 \\ &+ O(n\epsilon_c).\end{aligned}$$

Then we have

$$\sum_{i=1}^n \left([R(\widehat{\theta}, x_i, a_i^2) - R(\widehat{\theta}, x_i, a_i^1)] - [R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)]\right)^2 \leq O(e^B \log(N_{\mathcal{R}}(\epsilon_c)/\delta) + e^B n\epsilon_c).$$

□

Finally, we provide the on-policy confidence bound for the squared reward difference between the MLE estimator $\widehat{\theta}$ and the optimal reward function θ_*.

Lemma F.4. Consider an arbitrary policy π, and a set of context-action pairs $\{(x_i, a_i^1, a_i^2, y_i)\}_{i=1}^n$ generated i.i.d. from the BT model and π. Suppose that $\widehat{\theta}$ is the MLE estimator. We have with probability at least $1 - 2\delta$, there exists a mapping $b : \mathcal{X} \to \mathbb{R}$ such that

$$\mathbb{E}_\pi\left[\left(R(\widehat{\theta}, x, a) - R(\theta_*, x, a) - b(x)\right)^2\right] \leq O\left(\frac{1}{n}e^B \log(N_{\mathcal{R}}(\epsilon_c)/\delta) + e^B\epsilon_c\right).$$

Proof. By Lemma F.3, we have with probability at least $1 - \delta$,

$$\sum_{i=1}^n \left([R(\widehat{\theta}, x_i, a_i^2) - R(\widehat{\theta}, x_i, a_i^1)] - [R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)]\right)^2 \leq O(e^B \log(N_{\mathcal{R}}(\epsilon_c)/\delta) + e^B n\epsilon_c).$$

We consider an ϵ_c-net $\mathcal{R}^c = \{R(\theta, \cdot, \cdot)|\theta \in \Theta^c\}$ for the reward function class $\mathcal{R}$ with size $N_{\mathcal{R}}(\epsilon_c)$. For any $R(\theta, \cdot, \cdot)$, there exists $R(\theta^c, \cdot, \cdot)$ such that

$$\left|R(\theta, x, a) - R(\theta^c, x, a)\right| \leq O(\epsilon_c)$$

for all $x \in \mathcal{X}, a \in \mathcal{A}$.

Applying Lemma G.1, with probability at least $1-\delta$, we have

$$\begin{aligned}
&\sum_{i=1}^{n} \left([R(\theta^c, x_i, a_i^2) - R(\theta^c, x_i, a_i^1)] - [R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)]\right)^2 \\
&\quad - n\mathbb{E}_{x\sim d_0}\mathbb{E}_{a^1,a^2\sim\pi}\left[\left(R(\theta^c, x, a^1) - R(\theta_*, x, a^1) - R(\theta^c, x, a^2) + R(\theta_*, x, a^2)\right)^2\right] \\
&\leq \sqrt{\sum_{i=1}^{n} 4B^2\mathbb{E}_{x\sim d_0}\mathbb{E}_{a^1,a^2\sim\pi}\left[\left(R(\theta^c, x, a^1) - R(\theta_*, x, a^1) - R(\theta^c, x, a^2) + R(\theta_*, x, a^2)\right)^2\right]\log(N_{\mathcal{R}}(\epsilon_c)/\delta)} \\
&\quad + \frac{8}{3}B^2\log(N_{\mathcal{R}}(\epsilon_c)/\delta)
\end{aligned}$$

for all $\theta^c \in \Theta^c$. By Lemma G.2 and the definition of Θ^c, we further have

$$\begin{aligned}
&\mathbb{E}_{x\sim d_0}\mathbb{E}_{a^1,a^2\sim\pi}\left[\left(R(\widehat{\theta}, x, a^1) - R(\theta_*, x, a^1) - R(\widehat{\theta}, x, a^2) + R(\theta_*, x, a^2)\right)^2\right] \\
&\leq O(\frac{1}{n}B^2\log(N_{\mathcal{R}}(\epsilon_c)/\delta) + \frac{1}{n}\sum_{i=1}^{n}\left([R(\widehat{\theta}, x_i, a_i^2) - R(\widehat{\theta}, x_i, a_i^1)] - [R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)]\right)^2 + B\epsilon_c),
\end{aligned} \tag{F.3}$$

from which we can further derive that

$$\begin{aligned}
&\mathbb{E}_{x\sim d_0}\mathbb{E}_{a^1,a^2\sim\pi}\left[\left(R(\widehat{\theta}, x, a^1) - R(\theta_*, x, a^1) - R(\widehat{\theta}, x, a^2) + R(\theta_*, x, a^2)\right)^2\right] \\
&\leq O(\frac{1}{n}e^B\log(N_{\mathcal{R}}(\epsilon_c)/\delta) + e^B\epsilon_c)
\end{aligned}$$

with probability at least $1-2\delta$ by Lemma F.3 and the union bound.

We can then complete the proof by setting

$$b(x) = \mathbb{E}_{a^2\sim\pi(\cdot|x)}\left[R(\widehat{\theta}, x, a^2) - R(\theta_*, x, a^2)\right].$$

□

Lemma F.5 (Coverage of π_* and $\pi_{\widehat{\theta}}$ by $\pi_{\widehat{\theta}_0}$)**.** If $m \geq 32\eta^2D^2e^B\log(N_{\mathcal{R}}(\epsilon_c))$, $n = c_{m,n}m$ and $\epsilon_c \leq \frac{1}{(1+c_{m,n})e^B\eta^2D^2}$ in Algorithm 3 and Assumption 2.5 holds, then with probability at least $1-4\delta$, there exists $b_1 : \mathcal{X} \to \mathbb{R}$ and $b_2 : \mathcal{X} \to \mathbb{R}$ such that

$$\eta|R(\widehat{\theta}_0, x, a) - R(\theta_*, x, a) - b_1(x)| \leq 1, \quad \eta|R(\widehat{\theta}, x, a) - R(\theta_*, x, a) - b_2(x)| \leq 1$$

for all $x \in \mathcal{X}, a \in \mathcal{A}$ such that $\pi_0(a|x) > 0$.

Proof. By Lemma F.3 and the union bound, we have with probability at least $1-\delta$, it holds that

$$\begin{aligned}
&\sum_{i=1}^{m}\left([R(\widehat{\theta}, \widetilde{x}_i, \widetilde{a}_i^2) - R(\widehat{\theta}, \widetilde{x}_i, \widetilde{a}_i^1)] - [R(\theta_*, \widetilde{x}_i, \widetilde{a}_i^2) - R(\theta_*, \widetilde{x}_i, \widetilde{a}_i^1)]\right)^2 \\
&\quad + \sum_{i=1}^{n}\left([R(\widehat{\theta}, x_i, a_i^2) - R(\widehat{\theta}, x_i, a_i^1)] - [R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)]\right)^2 \\
&\leq O(e^B\log(N_{\mathcal{R}}(\epsilon_c)/\delta) + e^B(n+m)\epsilon_c).
\end{aligned} \tag{F.4}$$

Consider an ϵ_c-net $\mathcal{R}^c = \{R(\theta, \cdot, \cdot)|\theta \in \Theta^c\}$ for the reward function class $\mathcal{R}$ with size $N_{\mathcal{R}}(\epsilon_c)$. For any $R(\theta, \cdot, \cdot)$, there exists $R(\theta^c, \cdot, \cdot)$ such that

$$\left|R(\theta, x, a) - R(\theta^c, x, a)\right| \leq O(\epsilon_c)$$

for all $x \in \mathcal{X}, a \in \mathcal{A}$.

Applying Lemma G.1, with probability at least $1-\delta$, we have

$$\sum_{i=1}^{m}\left([R(\theta^c, \widetilde{x}_i, \widetilde{a}_i^2) - R(\theta^c, \widetilde{x}_i, \widetilde{a}_i^1)] - [R(\theta_*, x_i, a_i^2) - R(\theta_*, x_i, a_i^1)]\right)^2$$

$$
\begin{aligned}
&- m\mathbb{E}_{x\sim d_0}\mathbb{E}_{a^1,a^2\sim\pi_0}\big[\big(R(\theta^c,x,a^1)-R(\theta_*,x,a^1)-R(\theta^c,x,a^2)+R(\theta_*,x,a^2)\big)^2\big] \\
&\le \sqrt{\sum_{i=1}^{m} 4B^2\mathbb{E}_{x\sim d_0}\mathbb{E}_{a^1,a^2\sim\pi_0}\big[\big(R(\theta^c,x,a^1)-R(\theta_*,x,a^1)-R(\theta^c,x,a^2)+R(\theta_*,x,a^2)\big)^2\big]\log(N_{\mathcal{R}}(\epsilon_c)/\delta)} \\
&\quad + \frac{8}{3}B^2\log(N_{\mathcal{R}}(\epsilon_c)/\delta)
\end{aligned}
$$

for all $\theta^c \in \Theta^c$. By Lemma G.2 and the definition of Θ^c, we further have

$$
\begin{aligned}
&\mathbb{E}_{x\sim d_0}\mathbb{E}_{a^1,a^2\sim\pi}\big[\big(R(\widehat{\theta},x,a^1)-R(\theta_*,x,a^1)-R(\widehat{\theta},x,a^2)+R(\theta_*,x,a^2)\big)^2\big] \\
&\le O\bigg(\frac{1}{m}B^2\log(N_{\mathcal{R}}(\epsilon_c)/\delta\bigg) \\
&\quad + \frac{1}{m}\sum_{i=1}^{n}\big([R(\widehat{\theta},\widetilde{x}_i,\widetilde{a}_i^2)-R(\widehat{\theta},\widetilde{x}_i,\widetilde{a}_i^1)]-[R(\theta_*,\widetilde{x}_i,\widetilde{a}_i^2)-R(\theta_*,\widetilde{x}_i,\widetilde{a}_i^1)]\big)^2 + B\epsilon_c).
\end{aligned}
\tag{F.5}
$$

Substituting (F.4) into (F.5), we have with probability at least $1-2\delta$,

$$
\begin{aligned}
&\mathbb{E}_{x\sim d_0}\mathbb{E}_{a^1,a^2\sim\pi_0}\big[\big(R(\widehat{\theta},x,a^1)-R(\theta_*,x,a^1)-R(\widehat{\theta},x,a^2)+R(\theta_*,x,a^2)\big)^2\big] \\
&\le O\big(\frac{1}{m}e^B\log(N_{\mathcal{R}}(\epsilon_c)/\delta)+e^B\cdot\frac{n+m}{m}\cdot\epsilon_c\big).
\end{aligned}
$$

Therefore, there exists a mapping $b_2:\mathcal{X}\to\mathbb{R}$ such that

$$
\mathbb{E}_{\pi_0}\big[\big(R(\widehat{\theta},x,a)-R(\theta_*,x,a)-b_2(x)\big)^2\big] \le O\big(\frac{1}{m}e^B\log(N_{\mathcal{R}}(\epsilon_c)/\delta)+e^B\cdot\frac{n+m}{m}\cdot\epsilon_c\big).
$$

By Lemma F.4, we have with probability at least $1-2\delta$, there exists a mapping $b_1:\mathcal{X}\to\mathbb{R}$ such that

$$
\mathbb{E}_{\pi_0}\big[\big(R(\widehat{\theta}_0,x,a)-R(\theta_*,x,a)-b_1(x)\big)^2\big] \le O\big(\frac{1}{m}e^B\log(N_{\mathcal{R}}(\epsilon_c)/\delta)+e^B(1+c_{m,n})\epsilon_c\big).
$$

Hence, we can complete the proof by a union bound over the two events and Assumption 3.5. □

Now we are ready to prove Theorem 3.7.

Proof of Theorem 3.7. Let b be the mapping defined in Lemma F.4 for $\widehat{\theta}$ We have

$$
\begin{aligned}
&\mathbb{E}_{\pi^\eta_{\theta_*}}\bigg[R(\theta_*,x,a)-\frac{1}{\eta}\log\frac{\pi^\eta_{\theta_*}(a|x)}{\pi_0(a|x)}\bigg]-\mathbb{E}_{\pi^\eta_{\widehat{\theta}}}\bigg[R(\theta_*,x,a)-\frac{1}{\eta}\log\frac{\pi^\eta_{\widehat{\theta}}(a|x)}{\pi_0(a|x)}\bigg] \\
&=\frac{1}{\eta}\mathbb{E}_{\pi^\eta_{\theta_*}}\bigg[\log\frac{\pi_0(a|x)\cdot\exp\big(\eta R(\theta_*,x,a)\big)}{\pi^\eta_{\theta_*}(a|x)}\bigg]-\frac{1}{\eta}\mathbb{E}_{\pi^\eta_{\widehat{\theta}}}\bigg[\log\frac{\pi_0(a|x)\cdot\exp\big(\eta R(\theta_*,x,a)\big)}{\pi^\eta_{\widehat{\theta}}(a|x)}\bigg] \\
&=\frac{1}{\eta}\mathbb{E}_{x\sim d_0}\big[\log Z^\eta_{\theta_*}(x)\big]-\frac{1}{\eta}\mathbb{E}_{x\sim d_0}\big[\log Z^\eta_{\widehat{\theta}}(x)\big]-\mathbb{E}_{x\sim d_0}\bigg[\sum_{a\in\mathcal{A}}\pi^\eta_{\widehat{\theta}}(a|x)\cdot\big(R(\theta_*,x,a)-R(\widehat{\theta},x,a)\big)\bigg].
\end{aligned}
$$

For an arbitrary reward function $f:\mathcal{X}\times\mathcal{A}\to\mathbb{R}$, let $\Delta(x,a)=f(x,a)-R(\theta_*,x,a)$. Consider the following first derivative of $J(f)=\log Z^\eta_f(x)-\eta\sum_{a\in\mathcal{A}}\pi^\eta_f(a|x)\cdot\Delta(x,a)$, where $Z^\eta_f(x)=\sum_{a\in\mathcal{A}}\pi_0(a|x)\cdot\exp(\eta\cdot f(x,a))$ and $\pi^\eta_f(a|x)\propto\pi_0(a|x)\cdot\exp(\eta\cdot f(x,a))$.

Similar to the proof of Theorem 2.8, we still have

$$
\begin{aligned}
&\frac{\partial}{\partial\Delta(x,a)}\bigg[\log Z^\eta_f(x)-\eta\sum_{a\in\mathcal{A}}\pi^\eta_f(a|x)\cdot\Delta(x,a)\bigg] \\
&=\frac{1}{Z^\eta_f(x)}\cdot\pi_0(a|x)\exp\big(\eta\cdot f(x,a)\big)\cdot\eta-\eta\cdot\pi^\eta_f(a|x) \\
&\quad-\eta\cdot\Delta(x,a)\cdot\frac{\pi_0(a|x)\cdot\exp\big(\eta\cdot f(x,a)\big)}{Z^\eta_f(x)}\cdot\eta+\eta\cdot\Delta(x,a)\cdot\frac{\big[\pi_0(a|x)\cdot\exp\big(\eta\cdot f(x,a)\big)\big]^2}{[Z^\eta_f(x)]^2}\cdot\eta
\end{aligned}
$$

$$+\eta \sum_{a'\in\mathcal{A}\setminus\{a\}} \frac{\pi_0(a'|x)\cdot\exp\big(\eta\cdot f(x,a')\big)}{Z_f^\eta(x)}\cdot\eta\cdot\Delta(x,a')\cdot\frac{\pi_0(a|x)\cdot\exp\big(\eta\cdot f(x,a)\big)}{Z_f^\eta(x)}$$
$$=-\eta^2\pi_f^\eta(a|x)\Delta(x,a)+\eta^2[\pi_f^\eta(a|x)]^2\cdot\Delta(x,a)+\eta^2\sum_{a'\in\mathcal{A}\setminus\{a\}}\pi_f^\eta(a'|x)\pi_f^\eta(a|x)\Delta(x,a').$$

Note that

$$J(R(\widehat{\theta},x,\cdot))=\log Z_{\widehat{\theta}}^\eta(x)-\eta\sum_{a\in\mathcal{A}}\pi_{\widehat{\theta}}^\eta(a|x)\cdot\big(R(\widehat{\theta},x,a)-R(\theta_*,x,a)\big)$$
$$=\log\sum_{a\in\mathcal{A}}\pi_0(a|x)\cdot\exp(\eta(R(\widehat{\theta},x,a)-b(x)))-\eta\sum_{a\in\mathcal{A}}\pi_{\widehat{\theta}}^\eta(a|x)\cdot\big(R(\widehat{\theta},x,a)-R(\theta_*,x,a)-b(x)\big)$$
$$=J(R(\widehat{\theta},x,\cdot)-b(x)).$$

Therefore, there exists $f(\cdot,\cdot)=\gamma[R(\widehat{\theta},\cdot,\cdot)-b(\cdot)]+(1-\gamma)R(\theta_*,\cdot,\cdot)$ such that $\quad(\gamma\in(0,1))$

$$\begin{aligned}
&\mathbb{E}_{x\sim d_0}[J(R(\widehat{\theta},\cdot,\cdot))-J(R(\theta_*,\cdot,\cdot))]\\
&=\frac{1}{\eta}\mathbb{E}_{x\sim d_0}\Big[-\eta^2\sum_{a\in\mathcal{A}}\pi_f^\eta(a|x)\cdot\gamma\cdot\big(R(\widehat{\theta},x,a)-R(\theta_*,x,a)-b(x)\big)^2\Big]\\
&\quad+\frac{1}{\eta}\mathbb{E}_{x\sim d_0}\Big[\gamma\eta^2\sum_{a_1\in\mathcal{A}}\sum_{a_2\in\mathcal{A}}\pi_f^\eta(a_1|x)\pi_f^\eta(a_2|x)\big(R(\widehat{\theta},x,a_1)-R(\theta_*,x,a_1)-b(x)\big)\\
&\quad\big(R(\widehat{\theta},x,a_2)-R(\theta_*,x,a_2)-b(x)\big)\Big]\\
&\geq-\eta\cdot\mathbb{E}_{\pi_f^\eta}\big[\big(R(\widehat{\theta},x,a)-R(\theta_*,x,a)-b(x)\big)^2\big]
\end{aligned}$$

By Lemma F.2, if $m\geq 32\eta^2D^2e^B\cdot\log(2N_{\mathcal{R}}(\epsilon_c)/\delta)$, for any $(x,a)\in\mathcal{X}\times\mathcal{A}$ such that $\pi_0(a|x)>0$, it holds that

$$\eta|R(\widehat{\theta}_0,x,a)-R(\theta_*,x,a)-b_1(x)|\leq 1,\quad \eta|R(\widehat{\theta},x,a)-R(\theta_*,x,a)-b_2(x)|\leq 1,$$

which means that

$$\frac{\pi_f^\eta}{\pi_{\widehat{\theta}_0}^\eta}\leq e^4.$$

Let $\epsilon_c=\min\{\frac{\epsilon}{2(1+c_{m,n}^{-1})e^B},\frac{1}{(1+c_{m,n})e^B\eta^2D^2}\}$. By Lemma F.4, under the condition of the theorem, with high probability the output policy $\pi_{\widehat{\theta}}^\eta$ is $O(\epsilon)$ optimal. □

F.3 Proof of Corollary 3.10

In this subsection, we also discuss our result under the local-coverage condition (Definition 2.11).

Proof of Corollary 3.10. The proof follows the same lines as Theorem 3.7 by replacing the data coverage condition with the local-coverage condition. It still holds that

$$Q(\pi^*)-Q(\pi_{\widehat{\theta}_0}^\eta)\leq\eta\cdot\mathbb{E}_{\pi_f^\eta}\big[\big(R(\widehat{\theta}_0,x,a)-R(\theta_*,x,a)-b(x)\big)^2\big],$$

where $\pi_f^\eta(a|x)\propto\pi_0(a|x)\cdot\exp(\eta\cdot f(x,a))$ and $f(\cdot,\cdot)=\gamma[R(\widehat{\theta}_0,\cdot,\cdot)-b(\cdot)]+(1-\gamma)R(\theta_*,\cdot,\cdot)$ for some $\gamma\in(0,1)$. Thus, We have $\mathrm{KL}(\pi_f^\eta(a|x)\|\pi_0)\leq 2\eta B$, which further implies that

$$Q(\pi^*)-Q(\pi_{\widehat{\theta}}^\eta)\leq\eta\cdot C_{\rho_{\mathrm{KL}}}\cdot O\big(\frac{1}{n}e^B\log(N_{\mathcal{R}}(\epsilon_c)/\delta)+e^B(1+c_{m,n}^{-1})\epsilon_c\big)$$

by Lemma F.4. Then we can conclude by substituting the value of m into the suboptimality gap. □

G Auxiliary Lemmas

Lemma G.1 (Freedman's Inequality)**.** Let $M, v > 0$ be fixed constants. Let $\{X_i\}_{i=1}^n$ be a stochastic process, $\{\mathcal{G}_i\}_i$ be a sequence of σ-fields, and X_i be $\mathcal{G}_i$-measurable, while almost surely

$$\mathbb{E}[X_i|\mathcal{G}_i] = 0, |X_i| \le M, \text{ and } \sum_{i=1}^n \mathbb{E}[X_i^2|\mathcal{G}_{i-1}] \le v.$$

Then for any $\delta > 0$, with probability at least $1-\delta$, it holds that

$$\sum_{i=1}^n X_i \le \sqrt{2v\log(1/\delta)} + \frac{2}{3}M\log(1/\delta).$$

Lemma G.2. Suppose $a, b \ge 0$. If $x^2 \le a + b\cdot x$, then $x^2 \le 2b^2 + 2a$.

Proof. By solving the root of quadratic polynomial $q(x) := x^2 - b\cdot x - a$, we obtain $\max\{x_1, x_2\} = (b+\sqrt{b^2+4a})/2$. Hence, we have $x \le (b+\sqrt{b^2+4a})/2$ provided that $q(x) \le 0$. Then we further have

$$x^2 \le \frac{1}{4}\left(b + \sqrt{b^2+4a}\right)^2 \le \frac{1}{4}\cdot 2\left(b^2+b^2+4a\right) \le 2b^2 + 2a. \tag{G.1}$$

□

Lemma G.3 (Pinsker's Inequality)**.** If $\mathbb{P}_1$, $\mathbb{P}_2$ are two probability measures on a common measurable space $(\Omega, \mathcal{F})$, then it holds that

$$\delta(\mathbb{P}_1, \mathbb{P}_2) \le \sqrt{\frac{1}{2}\mathrm{KL}(\mathbb{P}_1\|\mathbb{P}_2)},$$

where $\delta(\cdot,\cdot)$ is the total variation distance and $\mathrm{KL}(\cdot\|\cdot)$ is the Kullback-Leibler divergence.

Lemma G.4 (Lemma A.4, Foster et al. 2021)**.** For any sequence of real-valued random variables $(X_t)_{t\le T}$ adapted to a filtration $(\mathcal{F}_t)_{t\le T}$, it holds that with probability at least $1-\delta$, for all $T' \le T$,

$$\sum_{t=1}^{T'} X_t \le \sum_{t=1}^{T'} \log\big(\mathbb{E}_{t-1}[e^{X_t}]\big) + \log(1/\delta).$$

Lemma G.5 (Solution of KL-regularized Optimization (Proposition 7.16 of Zhang 2023))**.** For any fixed $x \in \mathcal{X}$ and reward function R, we have

$$\begin{aligned}&\max_\pi \mathbb{E}_{a\sim\pi(\cdot|x)}\Big[R(x,a) - \eta^{-1}\mathrm{KL}\big(\pi(\cdot|x)\|\pi_0(\cdot|x)\big)\Big]\\ &= \frac{1}{\eta}\cdot\log\mathbb{E}_{a\sim\pi_0(\cdot|x)}\exp\big(\eta R(x,a)\big),\end{aligned}$$

where $Z_R(x)$ is the normalization constant and the minimizer of the loss functional is

$$\pi_R^\eta(a|x) = \frac{1}{Z_R(x)}\pi_0(a|x)\exp\Big(\eta R(x,a)\Big).$$

Lemma G.6 (Lemma 10, Zhang et al. 2020)**.** Let $(M_n)_{n\ge0}$ be a martingale such that $M_0 = 0$ and $|M_n - M_{n-1}| \le c$ for some $c > 0$ and any $n \ge 1$. Let $\mathrm{Var}_n = \sum_{k=1}^n E[(M_k - M_{k-1})^2|\mathcal{F}_{k-1}]$ for $n \ge 0$, where $\mathcal{F}_k = \sigma(M_1, M_2, \ldots, M_k)$. Then for any positive integer n and any $\varepsilon, p > 0$, we have that

$$\mathbb{P}\left(|M_n| \ge 2\sqrt{\mathrm{Var}_n \log\left(\frac{1}{p}\right)} + 2\sqrt{\varepsilon\log\left(\frac{1}{p}\right)} + 2c\log\left(\frac{1}{p}\right)\right) \le \left(2nc^2/\varepsilon + 2\right)p$$