On Provable Benefits of Policy Learning from Human Preferences in Contextual Bandit Problems

Anonymous authors
Paper under double-blind review

Abstract

For a real-world decision-making problem, the reward function often needs to be engineered or learned. A popular approach is to utilize human feedback to learn a reward function for training. The most straightforward way to do so is to ask humans to provide ratings for state-action pairs on an absolute scale and take these ratings as reward samples directly. Another popular way is to ask humans to rank a small set of state-action pairs by preference and learn a reward function from these preference data. Recently, preference-based methods have demonstrated substantial success in empirical applications such as InstructGPT. In this work, we develop a theoretical comparison between these human feedback approaches in offline contextual bandits and show how human bias and uncertainty in feedback modelings can affect the theoretical guarantees of these approaches. Through this, our results seek to provide a theoretical explanation for the empirical successes of preference-based methods from a modeling perspective.

1 Introduction

Reward engineering is one of the most crucial aspects in real-world decision-making problems. It is particularly important to bandits and reinforcement learning (RL), since it can be prohibitively expensive to learn the entire environment through random exploration and most existing algorithms rely on a reward function to guide their exploration in a deliberate manner so that they can solve the desired tasks efficiently.

In some cases, it can be straightforward to select reward functions when we have sufficient prior knowledge about the dynamics or rules that govern the problems of interest, such as games and simulated physical systems [32; 33; 47]. However, this is often not the case in practice. Real-world problems can be highly complex, and there may not be a clear choice of reward to use. Therefore, practitioners often have to construct a reward function from scratch for their algorithms to use. Unfortunately, such artificial rewards often end up misaligned with the overall objective of the system and fail to lead to the desired policy. For instance, in the task of teaching a chatbot to converse like a human, assigning a scalar reward to a chatbot's reply is challenging since there is no scale that can objectively evaluate its quality. Therefore, reward engineering poses a significant challenge to policy learning, particularly when it is difficult to quantify policy performance or the system is multi-objective.

To address these challenges, a popular methodology is to learn the reward function from human feedback instead of handcrafting it from scratch. These methods assume a true reward function exists and its corresponding optimal policy is aligned with our goal, but this true reward is not accessible or directly observable. Instead, it needs to be learned with the feedback data from human annotators, who are believed to be able to evaluate an algorithm or agent in a way aligned with the true reward. The most straightforward approach is to ask human annotators to rate subjects on an absolute scale [28; 29; 18]. These ratings can be either used directly as reward samples or incorporated into a pre-designed reward function as a component [50]. We refer to this use of human feedback as "rating". The rating approach has been popular because of its easy implementation and the compatibility of such rating data with most existing algorithms. However, as some empirical studies have shown [55; 22], the reward derived from human ratings is susceptible to bias and uncertainty of the human annotators and can deviate from the true reward. To characterize the ratings given by human

annotators, several models have been proposed and widely adopted in the literature [38; 30]. These existing models, however, fall short on two aspects: (i) they model the uncertainty noise in a simple, isolated form, and (ii) their modeling of the bias is also restrictive, which does not fully capture the bias in practice and can render the problem of policy learning statistically inconsistent.

As an alternative, there has been a growing trend in recent years to use human feedback by comparing subjects rather than rating them individually. These methods are commonly referred to as preference-based methods. In lieu of assigning ratings on an absolute scale, human annotators are given small sets of subjects and tasked with comparing and ranking the subjects within each set. Since some empirical studies have shown that humans are fairly accurate when making choices among a small number of subjects [34; 49; 64], the preference-based approach is believed to be more robust to human bias and uncertainty and learn reward more accurately. The preference feedback of human annotators is commonly modeled with either the Bradley-Terry-Luce model [4] or the Plackett-Luce model [39], both of which have found extensive applications in recommender systems and crowdsourcing research [8; 44; 13]. Recently, preference-based methods have become popular for reward learning in bandit problems and have played a crucial role in the remarkable success of training large language models such as InstructGPT and ChatGPT.

While the preference-based approach has demonstrated notable empirical effectiveness in reward engineering, its theoretical properties remain largely unexplored. Existing results have primarily focused on algorithms for online bandit and RL, whose goal is to maximize a preference metric rather than to learn a reward function [37; 62]. Recently, [77; 75] proved the optimal policy can be learned from preference data in the offline setting with pessimism and maximum likelihood estimation (MLE) and analyzed the suboptimality. [54] showed any robust RL algorithm can be adapted to find the optimal policy with preference data, suggesting preference-based policy learning is no harder than standard robust RL. However, the reason why the preference-based approach outperforms the traditional rating-based approach in practice still remains a question. In this work, we provide a theoretical comparison between these human feedback approaches and propose a theory that aims to explain the advantage of the preference-based approach over rating in policy learning from a modeling perspective. To align with recent applications [11; 36], we focus on tabular contextual bandits in the offline setting.

Specifically, we first consider a new model for human rating data and analyze the suboptimality guarantees of the standard LCB algorithm under it. Our rating model is based on a general class of monotone functions that can account for both human bias and uncertainty with general forms. By incorporating the concept of monotonicity, our model captures the bias observed in real-world human rating and maintains the correct reward ordering. This allows policy learning to be a consistent yet nontrivial statistical problem, which differs from existing rating models that do not preserve the reward ordering or guarantee the consistency of the induced policy learning problem. In addition, our model is able to express a more general form of noise to represent human uncertainty during rating. Through our models, we provide the first known suboptimality analysis for reward engineering with human rating in bandit problems and shed light on how human bias and uncertainty can adversely impact policy learning. Furthermore, we compare our results with the suboptimality result from [77] for the preference-based method pessimistic MLE. The comparison reveals that the preference-based approach enjoys lower suboptimality than the rating-based approach when human bias is extreme in human rating. Lastly, we also consider a new model for human preference with human bias and compare the sample complexity of pessimistic MLE under this new model with the results for human rating. This comparison shows when human bias and uncertainty are equally strong in both types of human feedback, the preference-based approach has no provable advantage over the rating-based one. Altogether, our theory shows the advantage of the preference-based approach can be largely attributed to its modeling with mild human bias and uncertainty, which makes it reasonable to believe the great empirical success of preference-based methods is because human preference data is subject to less bias and uncertainty in practice.

1.1 RELATED WORKS

Preference-based reinforcement learning. Preference-based Reinforcement Learning (PbRL) [11; 46; 6; 56; 57; 12; 1] has been studied under different types of human feedback including action comparison, state comparison and trajectory comparison—see [57; 1] for reviews of the literature. Preference-based feedback has been well-studied in a bandit setting known as dueling bandits [72; 70; 71; 42; 2; 78; 23; 16; 43]. Recently, there is a growing interest in the theoretical guarantees

of PbRL methods, including tabular case [35; 62] and linear and general function approximations [37; 9; 76]. However, these works are focused on the online setting and their methods are not applicable to the offline setting.

Offline policy learning. The vast literature on offline policy learning can be divided by the different assumptions on the data sampling distribution. The strongest assumption is one that requires all state-action pairs can be sampled [67; 14; 7; 58]. A similar assumption requires that the occupancy measure of every possible policy be dominated by the data sampling distribution, which is common in the function approximation setting [3; 48]. One of the weakest assumptions is one that requires the occupancy measure of an optimal policy be dominated by the data sampling distribution. To learn the optimal policy under this setting, the principle of pessimism [24; 5; 21] was introduced and has inspired many algorithms [69; 40; 27; 60; 73; 74; 59; 61]. In particular, the sample complexity of pessimistic algorithms has been extensively studied in the tabular case [45; 63; 26; 65; 41; 60; 66; 40] and linear MDPs [21; 59; 74; 53; 15; 68]. In this spirit, the algorithms we study in this work also use human feedback with pessimism.

Notation Given any vector $x \in \mathbb{R}^{SA}$ that represents a function $x : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, we use $x(s,a)$ to denote the entry corresponding to the state-action pair (s,a). For any random sample X with distribution P and a function $f(\cdot)$ of X, we denote the expectation of $f(X)$ over P with $\mathbb{E}_{X \sim P}[f(X)]$ or $\mathbb{E}_X[f(X)]$. Similarly, we denote the variance of $f(X)$ over P with $\text{Var}_X(f(X))$. Lastly, we denote equality in distribution with $\overset{d}{=}$.

2 PRELIMINARIES

In this section, we make a brief review of the contextual bandit problem and policy learning in the offline setting.

2.1 CONTEXTUAL BANDIT

We consider a contextual bandit represented by $(\mathcal{S}, \mathcal{A}, r, \rho)$. Specifically, we focus on a tabular setting, where $\mathcal{S} := 1, 2, \ldots, S$ denotes the state space of size S, and $\mathcal{A} := 1, 2, \ldots, A$ denotes the action space of size A. The function $r : \mathcal{S} \times \mathcal{A} \to [0, R]$ represents the true reward function, which is assumed to be deterministic and unknown in this paper. Here, $r(s,a)$ is the immediate reward obtained when taking action $a \in \mathcal{A}$ in state $s \in \mathcal{S}$. ρ denotes the initial state distribution of the bandit.

A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ specifies how actions are selected given a state, where $\pi(\cdot|s) \in \Delta(\mathcal{A})$ represents the action selection probability vector at state $s \in \mathcal{S}$. We also use $\pi(s)$ to denote the action selected by policy π at state s. We denote the state-action visitation distribution of π starting from the initial distribution ρ with d_ρ^π. The value function of policy π is defined as follows:

$$V^\pi(s) := \mathbb{E}_{s \sim \rho, a \sim \pi(\cdot|s)} \left[r(s,a) \right].$$

Lastly, an optimal policy $\pi^\star$ maximizes the value function for all states simultaneously.

2.2 OFFLINE POLICY LEARNING

We consider the offline setting of policy learning, in which the learner is given a dataset of i.i.d. samples generated under a sampling distribution. While the sampling distribution can take different forms under different feedback models, the task is always to learn a policy π from the offline dataset that performs as well as the optimal policy as possible. In particular, we evaluate the performance of π by the suboptimality metric defined as follow:

$$\text{SubOpt}(\pi) := \mathbb{E}_{s \sim \rho} \left[V^{\pi^\star}(s) - V^\pi(s) \right]. \tag{1}$$

Here the suboptimality measures the performance difference between the optimal policy $\pi^\star$ and π in the problem bandit. Naturally, one aims to minimize the suboptimality and find an algorithm whose suboptimality converges to zero as the sample size n approaches infinity.

3 HUMAN RATING MODELS

One of the most straightforward ways to use human feedback is to let human annotators evaluate on an absolute scale. Since such data can be readily used in most algorithms in the existing literature, the human rating approach has become very popular and one of the most important to study. As

evidenced in behavioral studies, human ratings are subject to both bias and uncertainty [18; 31]. Specifically, under the influence of their own personalities and past experiences, human annotators can exhibit personal opinion bias during rating, leading to deviations from the true score. Furthermore, due to the tedious and challenging nature of the rating process, the evaluation of the same subject by the same human annotator can fluctuate over time, giving rise to what is known as intra-observer uncertainty. In light of these phenomena, [18] propose a model that aims to characterize the ratings from a human annotator in the real world, which has been widely adopted in the existing literature [38; 30; 28]. In the following, we present this model in the single annotator case under the contextual bandit setting. For any fixed state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ with true reward $r(s, a)$, a rating from human annotator $\widetilde{r}(s, a)$ can be written as

$$\widetilde{r}(s, a) = r(s, a) + \Delta_{(s,a)} + \epsilon. \tag{2}$$

Here, $\Delta_{(s,a)}$ represents the bias of the human annotator for action a at state s. Learners and algorithms have no knowledge of such bias and would take these observed ratings as reward samples directly. In most works [18; 31], $\Delta_{(s,a)}$ is modeled as an unknown constant; in [38], $\Delta_{(s,a)}$ is a Gaussian random variable with an unknown, non-zero mean. ϵ is a random noise representing the intra-observer uncertainty, which is modeled with a zero-mean Gaussian random variable in these works.

Apparently, such a human rating model bears several limitations. In (2), the bias $\Delta_{(s,a)}$ has an unknown non-zero expectation, which makes it impossible to recover the true reward $r(s, a)$ exactly. Furthermore, identifying the optimal action for a given state s becomes infeasible when the bias causes a flip in the expected reward, i.e, $\mathbb{E}[\widetilde{r}(s, a)] > \mathbb{E}[\widetilde{r}(s, \pi^\star(s))]$ for any $a \notin \pi^\star(s)$, in which case the policy learning problem becomes inconsistent under this model. However, in the real world, the bias of a human annotator should not keep him or her from identifying the best action in expectation, but it is likely to take a much more general form. Neither of these is reflected in the current modeling with additive constant bias. On the other hand, the uncertainty is only modeled with an additive sub-gaussian random variable in these works. While this simple noise model is considered in many theoretical works, it cannot capture the setting when reward samples are generated from human rating. In practice, such uncertainty noise can be higher-order and more complex [20]. In addition, the uncertainty might also be affected by the bias and reward at different state-action pairs and have an intricate role in the final observation $\widetilde{r}(s, a)$.

To model human rating more realistically while keeping the policy learning problem consistent, we propose a new model under which the rating $\widetilde{r}(s, a)$ can be expressed with

$$\widetilde{r}(s, a) = h(r(s, a), \epsilon), \tag{3}$$

where $h(\cdot, \cdot)$ is a general, deterministic transformation and ϵ is a random variable sampled from $\mathcal{N}(0, \sigma^2)$ and independent from (s, a). For notational simplicity, we define $\bar{h}(r) := \mathbb{E}_\epsilon[h(r, \epsilon)]$. This can be interpreted as the reward function for this bandit instance in the human annotator's mind, which he or she uses to produce ratings. We can refer to $\bar{h}(r)$ as the biased reward and the function $\bar{h}(\cdot)$ as the expected bias function.

While the h transformation can be general, for the policy learning problem to make sense, we assume it satisfies the following three condition. In this work, we only consider models that satisfy these conditions:

Condition 1. The function $\bar{h}(\cdot)$ satisfies

$$\bar{h}(r_1), \bar{h}(r_2) \in [0, R] \quad \text{and} \quad \bar{h}(r_1) > \bar{h}(r_2)$$

for any $r_1, r_2 \in [0, R]$ such that $r_1 > r_2$. In addition, $\bar{h}(0) = 0$.

This condition assumes that $\bar{h}(\cdot)$ is strictly monotone, implying that the biased reward function should preserve the ranking under the true reward r in expectation. This condition is particularly important, as it ensures that the consistency of the policy learning problem. Therefore, we can identify the optimal policy based on human rating data in expectation.

Remark 1. The monotonicity in Condition 1 also guarantees that any group of samples can be correctly compared and ranked in expectation, which is a necessary condition for the use of preference-based methods. This unifies the admissibility assumption in the rating models and preference models, which is crucial for the validity of our subsequent theoretical comparison between the two approaches.

Remark 2. We require $\bar{h}(0) = 0$ only to rule out arbitrary constant shift in $\bar{h}$ because shifting the reward by a constant is trivial and does not change the policy learning problem or any theoretical guarantee.

Condition 2. For any $r \in [0, R]$, $h(r, \epsilon)$ is a degree-q polynomial in ϵ and symmetric in ϵ about its expectation, i.e.,

$$-(h(r, \epsilon) - \mathbb{E}_\epsilon[h(r, \epsilon)]) \overset{d}{=} h(r, \epsilon') - \mathbb{E}_{\epsilon'}[h(r, \epsilon')],$$

where ϵ' is a random variable identically distributed as ϵ.

Since $h(r, \cdot)$ can be a very general function, the human uncertainty noise in the final observation $\widetilde{r}(s, a)$ is allowed to have a complicated dependency on the bias and the true reward, even though the randomness only comes from an independent Gaussian random variable ϵ. For instance, the original white noise ϵ may end up amplified or reshaped by the true reward and human's internal bias and cause the final ratings to exhibit a complex concentration around $\bar{h}(r)$. This not only provides more realism and flexibility in modeling but also presents a greater theoretically challenge compared to the uncertainty considered in (2), which is modeled with a simple additive Gaussian noise with no interactions with the true reward and human bias.

Remark 3. Condition 2 is only a regulation on the effect of uncertainty—the uncertainty noise should not favor any particular direction (though the bias still can). This is in line with the real world, where the random noise concentrates evenly around the expectation and the effect of uncertainty diminishes in expectation.

Condition 3. For any $r_1, r_2 \in [0, R]$ such that $r_1 \geq r_2$, there are positive constants $C_{h,1}, C_{h,2} > 0$ such that

$$\bar{h}^{-1}(r_1) - \bar{h}^{-1}(r_2) \leq C_{h,1} \cdot \bar{h}^{-1}(r_1 - r_2) \quad \text{and} \quad \bar{h}(r_1) - \bar{h}(r_2) \leq C_{h,2} \cdot \bar{h}(r_1 - r_2).$$

This is a technical condition on the regularity of the expected bias function. It ensures that the bias does not transform the reward too drastically, which eases our theoretical analysis.

Overall, this h transformation can model very general behavior. For example, many human annotators with strong personal opinions tend to exhibit an extreme bias in their evaluations, making them rate subjects with low true reward even lower and rate those with high true reward even higher on average. Our model can capture such bias with a convex $\bar{h}(\cdot)$, with a concrete example and its theoretical guarantees detailed in Appendix B.

4 RESULTS FOR HUMAN RATING

Before the theoretical comparison with the preference-based approach, let us first establish some theoretical results for our more general rating model. In particular, we analyze the suboptimality of the LCB algorithm under our more practical rating model. These results can provide some theoretical explanation for how human bias and uncertainty could adversely affect policy learning.

In the case of human rating, we are given an offline dataset $\mathcal{D} = \{(s_i, a_i, \widetilde{r}_i)\}_{i=1}^n$. The state-action pairs in $\mathcal{D}$ are generated in an i.i.d. fashion according to a sampling distribution over the state-action space. The sampling probability of the state-action pair (s, a) is denoted with $d(s, a)$. For each (s_i, a_i), the human annotator provides a *rating sample* $\widetilde{r}_i$ following the rating model (3) based on the true reward $r(s_i, a_i)$.

Let us also make a brief review of the standard LCB approach for offline policy learning [21; 40; 66]. In the existing literature, it is common to assume the knowledge of a reasonable upper bound on the variance of reward observations. Similarly, we assume there exists an upper bound on the variance $\text{Var}_\epsilon(h(r, \epsilon))$ for all $r \in [0, R]$, which we denote with $V_{R,\sigma}^2$ and can depend on R and σ. Recall that the learner has no knowledge of the transformation h, but let us assume the learner can make a reasonable estimate $\widetilde{V}_{R,\sigma}^2$ for the true variance $V_{R,\sigma}^2$ such that $\widetilde{V}_{R,\sigma}^2 = c_V V_{R,\sigma}^2$, where $c_V > 0$ is an absolute constant. To learn the optimal policy with at least $1 - \delta$ success probability, the standard LCB algorithm (Algorithm 1) uses a penalty in the form of

$$b_m = c_b \sqrt{\frac{\widetilde{V}_{R,\sigma}^2 \log \frac{SA}{\delta}}{m}} \tag{4}$$

with an appropriately chosen constant c_b.

To understand the effects of human bias and uncertainty on policy learning under our more realistic rating model, let us establish the lower bound on the suboptimality of the LCB algorithm. We will consider two scenarios with different coverage assumptions for the offline dataset $\mathcal{D}$.

Algorithm 1: LCB for contextual bandits

---

1 **Input:** Offline rating dataset $\mathcal{D}$, confidence level $\delta \in (0, 1)$.
2 **for** *all* $(s, a) \in \mathcal{S} \times \mathcal{A}$ **do**
3   Set $n_{(s,a)} = \sum_{i=1}^{n} \mathbb{1}\{(s_i, a_i) = (s, a)\}$;
4   Set $\widetilde{r}(s, a) = \frac{1}{n} \sum_{i=1}^{n} \widetilde{r}_i \mathbb{1}\{(s_i, a_i) = (s, a)\}$;
5   Set $\widehat{r}(s, a) = \max\{\widetilde{r}(s, a) - b_{n_{(s,a)}}, 0\}$;
6 **return** $\widehat{\pi}_{\mathrm{LCB}}(\cdot) = \arg\max_{a \in \mathcal{A}} \widehat{r}(\cdot, a)$.

4.1 LOWER BOUND UNDER PARTIAL COVERAGE

As [40; 66] have shown, to learn the optimal policy in the offline setting, it is sufficient for the sampling distribution of the offline dataset to cover the state-action pairs that the optimal policy can reach. Concretely, this assumption can be written as follows:

Assumption 1. *There exists an optimal policy $\pi^{\star}$ such that $d(s, a) > 0$ whenever $d_{\rho}^{\pi^{\star}}(s, a) > 0$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$.*

Under this assumption, it makes sense to define a concentrability coefficient $C^{\star}$ as follows:

$$C^{\star} := \max_{(s,a) \in \mathcal{X}} \frac{d_{\rho}^{\pi^{\star}}(s, a)}{d(s, a)}, \tag{5}$$

where the set $\mathcal{X}$ is the set of all state-action pairs that the sampling distribution of $\mathcal{D}$ can cover, i.e., $\mathcal{X} := \{(s, a) \in \mathcal{S} \times \mathcal{A} : d(s, a) > 0\}$. Under Assumption 1, if the reward can be observed exactly or with only additive sub-gaussian noise, the LCB algorithm (Algorithm 1) with penalty (4) is guaranteed to converge to the optimal policy [40; 66]. However, theory suggests it does not converge in the worst case when the reward function is engineered from human rating. In particular, let us consider the setting beyond the standard additive sub-gaussian noise, which has been well-studied in the existing literature. That is, let us consider a more practical model in the form of (3) with $q \geq 2$. We can prove that even when the rating model preserves the correct reward ordering in expectation and keeps the policy learning problem consistent, it is possible that the LCB algorithm does not converge to the optimal policy and must suffer constant suboptimality.

Theorem 1. *For any fixed constant $0 < \delta < 1$, there exists a contextual bandit instance with initial state distribution ρ such that if one samples a dataset $\mathcal{D}$ of size $n \geq c(\delta, c_b, c_V, q, \sigma, R)$ using a sampling distribution d satisfying Assumption 1 with $C^{\star} = 2$ and runs Algorithm 1 on $\mathcal{D}$, the output policy $\widehat{\pi}_{\mathrm{LCB}}$ must suffer constant suboptimality, i.e.,*

$$\mathbb{E}_{\mathcal{D}}[\mathrm{SubOpt}(\widehat{\pi}_{\mathrm{LCB}})] = c_0 R, \tag{6}$$

where c_0 is a universal constant and $c(\delta, c_b, c_V, q, \sigma, R)$ is a constant depending on $\delta, c_b, c_V, q, \sigma, R$.

This result is reminiscent of Proposition 1 in [40], which constructs a bandit and shows the empirically best policy chooses a suboptimal action with constant probability under Assumption 1. The very same work also shows that by adding a pessimism penalty, the LCB algorithm (Algorithm 1) can converge to the optimal policy under the same data coverage assumption. In contrast, our theorem shows that even when we make pessimistic estimates and penalize less-observed state-action pairs in human rating data, a constant suboptimality can still ensue. This shows a disadvantage of using human rating as reward samples directly: although the estimation problem induced by human rating is still consistent, using LCB with only the knowledge of variance is not sufficient for convergence. Instead, the learner needs to know the shape of the noise distribution, but it is unrealistic to model the human uncertainty accurately in practice. We also provide an upper bound result for the case when the learner has complete knowledge of the uncertainty noise distribution in Appendix C.

Proof sketch In a bandit instance with special reward design, we first find the lower bound for the probability that suboptimal actions are only observed for a very small number of times in the offline dataset. Such state-action pairs can have huge fluctuation in their empirical reward average and mislead the algorithm. Then, we find the lower bound on the probability that a state-action pair (s, a) such that $\widehat{r}(s, a) > \widehat{r}(s, a^{\star})$ exists, which can cause the algorithm to always select the suboptimal action a and suffer suboptimality. Different from Proposition 1 in [40], in which the reward noise for suboptimal actions is defined with two Dirac delta functions, the noise under our

rating model is unbounded and can be viewed as a Gaussian chaos, so we compute this probability using a method from the corresponding literature. Moreover, in the same paper, a bandit instance is sufficient to induce constant suboptimality as long as its action space is designed large. In our case, since the pessimism penalty in Algorithm 1 accounts for the bandit size and larger bandit instances are penalized more, it requires a careful balance in the design of our bandit instance.

4.2 LOWER BOUND UNDER FULL COVERAGE

Uniform coverage is another popular coverage assumption for offline policy learning [67; 51; 17]. It can be written as follows:

Assumption 2. *The sampling distribution satisfies $d(s, a) > 0$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$.*

This coverage assumption is much stronger than Assumption 1 and makes the offline policy learning problem much easier. Under Assumption 2, many algorithms without the pessimism principle can also be shown to provably converge to the optimal policy [7; 58]. Moreover, [21] showed that the suboptimality of algorithms with pessimism can decay faster when the data are well-explored. In this setting, we establish a lower bound on the suboptimality of Algorithm 1 under Assumption 2.

Theorem 2. *For any fixed constant $0 < \delta < 1$, there exists a contextual bandit instance with initial state distribution ρ such that if one samples a dataset $\mathcal{D}$ of size $n \geq \max\{48\sigma^4, 60\}$ using a sampling distribution d satisfying Assumption 2 with $d(s, a) = \frac{1}{SA}$ for every $s \in \mathcal{S}$ and $a \in \mathcal{A}$ and runs Algorithm 1 on $\mathcal{D}$, the output policy $\widehat{\pi}_{\mathrm{LCB}}$ must suffer suboptimality at least*

$$\mathbb{E}_{\mathcal{D}}[\mathrm{SubOpt}(\widehat{\pi}_{\mathrm{LCB}})] = c_0 \cdot \bar{h}^{-1}\left(\sqrt{\frac{V_{R,\sigma}^2}{n}}\right),$$

where c_0 is a constant that depends on q.

In fact, under uniform data coverage as in Theorem 2, pessimism becomes unnecessary and this result holds no matter what penalty b_n is used in the algorithm. This theorem demonstrates another disadvantage of human rating: even when the data covers the entire state-action space and learning is no longer impeded by the lack of knowledge of human uncertainty, the suboptimality is still bottlenecked by human bias.

5 COMPARISON WITH PREFERENCE-BASED APPROACH

In contrast to rating, the preference-based approach relies on models that characterize how a human annotator would rank a group of subjects by reward. In this case, the feedback is simply the most preferred subject to the human annotator within the group. Such feedback actually contains less information than rating. Preference data are also incompatible with standard bandit algorithms and require special adaptation to use [54]. However, the preference-based approach has received much attention recently because some have found it easier and more accurate for human to make preferences than rating [34; 49; 64]. In this section, we compare the human rating approach with the preference-based approach.

5.1 HUMAN PREFERENCE UNDER BTL

Let us consider the most basic case of human preference called pairwise comparison, which involves the ranking between a pair of state-action pairs based on their rewards. This is predominantly modeled with the Bradley-Terry-Luce (BTL) model [4], under which a human annotator gives a binary response $y = \{0, 1\}$ following a Bernoulli distribution when asked to compare two state-action pairs (s, a^0) and (s, a^1) with $a^0 \neq a^1$:

$$P(y|s, a, a') = \frac{\exp(r(s, a^y))}{\exp(r(s, a^0)) + \exp(r(s, a^1))}. \tag{7}$$

Like our rating model in (3), the BTL model admits a consistent statistical problem. The learner is given a dataset $\mathcal{D}' = \{(s_i, a_i^0, a_i^1, y_i)\}_{i=1}^n$, which contains i.i.d. human preference samples from some sampling distribution. y_i is the binary human preference feedback for the comparison between (s_i, a_i^0) and (s_i, a_i^1). We denote the sampling probability of the state-action-action triplet (s, a^0, a^1) with $d(s, a^0, a^1)$.

To find the optimal policy with human preference data, we can use pessimistic MLE [77], which first computes a reward function by MLE and then outputs the optimal policy corresponding to a

---

Algorithm 2: Pessimistic MLE for contextual bandits

---

1 **Input:** Offline preference dataset $\mathcal{D}'$, confidence level $\delta \in (0, 1)$.
2 Construct the reward function set $\mathcal{F} := \{v \in \mathbb{R}^{SA} : \mathbf{1}^\top v = 0, \|v\|_\infty \le R\}$;
3 Set

$$\widetilde{r} = \arg\max_{f \in \mathcal{F}} \sum_{i=1}^{n} \log\left(\frac{\mathbb{1}\{y_i = 1\}\exp(f(s_i, a_i^1))}{\exp(f(s_i, a_i^0)) + \exp(f(s_i, a_i^1))} + \frac{\mathbb{1}\{y_i = 0\}\exp(f(s_i, a_i^0))}{\exp(f(s_i, a_i^0)) + \exp(f(s_i, a_i^1))} \right);$$

4 Construct empirical covariance matrix

$$\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{1}_{(s_i, a_i^0)} - \mathbf{1}_{(s_i, a_i^1)}\right)\left(\mathbf{1}_{(s_i, a_i^0)} - \mathbf{1}_{(s_i, a_i^1)}\right)^\top;$$

5 Construct the pessimistic reward function set

$$\mathcal{F}_{\mathrm{CR}}(\widetilde{r}) = \left\{ f \in \mathcal{F} \ : \ \sqrt{(f - \widetilde{r})^\top \widehat{\Sigma}(f - \widetilde{r})} \le b'_n \right\};$$

6 **return** $\widehat{\pi}_{\mathrm{PMLE}} = \arg\max_\pi \min_{\widehat{r} \in \mathcal{F}_{\mathrm{CR}}(\widetilde{r})} \mathbb{E}_{s \sim \rho}[\widehat{r}(s, \pi(s))]$.

---

pessimistic version of this MLE reward (Algorithm 2). The data coverage assumption is similar to Assumption 1, which essentially requires the sampling distribution to covers the state-actions pairs that optimal policy can reach. In the tabular case, this assumption can be written as follows:

Assumption 3. *There exists an optimal policy $\pi^\star$ such that the pairwise concentrability coefficient*

$$C^\dagger := \sqrt{\sup_{v \in [-1,1]^{SA}:\mathbf{1}^\top v = 0} \frac{\left(\sum_{(s,a)} d_\rho^{\pi^\star}(s,a)v(s,a) \right)^2}{\sum_{(s,a^0,a^1)} d(s,a^0,a^1)\left(v(s,a^0) - v(s,a^1)\right)^2}} \tag{8}$$

is bounded.

[77] proved the convergence of pessimistic MLE in the linear bandit setting. The following theorem is a special case of Theorem 3.2 from [77] with some modification, which expresses everything in the tabular setting. This shows when we assume human preference follows the BTL model, pessimistic MLE can provably converge to the optimal policy under the mild data coverage assumption of Assumption 3 and its suboptimality decays at a fast rate of $O(1/\sqrt{n})$. This result marks a clear distinction from the negative results for human rating.

Theorem 3. *Denote $\gamma = \frac{1}{2 + \exp(R\sqrt{SA}) + \exp(-R\sqrt{SA})}$. Suppose Assumption 3 holds. For any fixed constant $0 < \delta < 1$, if one runs Algorithm 2 with*

$$b'_m = c'_b \sqrt{\frac{SA + \log\frac{1}{\delta}}{\gamma^2 m}},$$

where c'_b is an appropriately chosen universal constant, with probability $1 - \delta$, the suboptimality of the output policy $\widehat{\pi}_{\mathrm{PMLE}}$ satisfies

$$\mathrm{SubOpt}(\widehat{\pi}_{\mathrm{PMLE}}) \le c_0 C^\dagger R \left(\sqrt{\frac{SA + \log\frac{1}{\delta}}{\gamma^2 n}} + \sqrt{\frac{S^2 A^2 \log\frac{n}{\delta}}{n}} \right),$$

where c_0 is a universal constant.

We can compare the suboptimality in this theorem with the results for the rating-based approach. A comparison with Theorem 1 shows the uncertainty in human ratings may require the data to have stronger coverage in order to converge to the optimal policy. A comparison with Theorem 2 shows when the bias in human ratings distorts the reward function and makes it more extreme and drastic (less smooth in the Lipschitz sense), the $\bar{h}^{-1}(\cdot)$ can slow down the suboptimality's decay with respect to the sample size. In fact, we can observe that the preference-based approach enjoys faster suboptimality decay because preference feedback contains no bias and mild uncertainty noise according to the BTL model. While such modeling is justified by empirical evidences, it makes one wonder whether the advantage of preference-based methods mostly comes from the modeling aspect. To delve into this further, let us make another theoretical analysis for the case when preference data are affected by human bias.

5.2 HUMAN PREFERENCE UNDER BIASED BTL

Let us introduce a new model for human preference called the biased BTL model. This model considers the case when human preferences are also subject to bias just like the rating model (3) and the feedback is generated with respect to the biased reward. In particular, the binary feedback $\widetilde{y} = \{0, 1\}$ for (s, a^0) and (s, a^1) follows:

$$P(\widetilde{y}|s, a, a') = \frac{\exp(\bar{h}(r(s, a^{\widetilde{y}})))}{\exp(\bar{h}(r(s, a^0))) + \exp(\bar{h}(r(s, a^1)))}, \tag{9}$$

where $\bar{h}$ is the expected bias function from (3).

We consider the performance of pessimistic MLE (Algorithm 2) again with human preference data generated under this model. While the data are generated under human bias this time, we still run pessimistic MLE on the new data as before. Different from the suboptimality results in the previous section, we focus on the sample complexity for learning the optimal policy. We take a gap-dependent approach in our analysis to consider the case when human bias closes the biased optimality gap $r(s, \pi^\star(s)) - r(s, a)$ and the actual optimality gap $\bar{h}(r(s, \pi^\star(s))) - \bar{h}(r(s, a))$ remains big, where a is the second best action at s. This echoes with the type of undesirable bias we considered in the last comparison, which is true when human annotators have more extreme standards at heart. In a simple bandit instance, we can obtain the following result and notice the samples needed to find the optimal policy with the preference-based approach is no less than the samples needed for the rating-based approach.

Theorem 4. *Consider any single-state bandit instance with $\mathcal{A} = \{a_1, a_2\}$ and $0 \leq \bar{h}(r(a_1)) < \bar{h}(r(a_2)) \leq 1$. For any fixed constant $0 < \delta < 1$, let n_{rate} be the total number of samples needed to learn the optimal action with probability at least $1 - \delta$ in the human rating setting under observation model* (3) *with additive sub-gaussian uncertainty noise and uniform data coverage $n_{a_1} = n_{a_2}$, and let n_{pref} be the number of samples needed to learn the optimal action with probability at least $1 - \delta$ in the human preference setting with observation model* (9). *It always holds that*

$$\frac{n_{rate}}{n_{pref}} < 0.25\sigma^2. \tag{10}$$

We can see that when the variance proxy of the uncertainty noise σ^2 is no larger than 4 in human rating (the expected reward is bounded in $[0, 1]$), the samples needed in the rating-based approach is always fewer than the preference-based approach. This shows if one assumes a similar amount of human bias and uncertainty in both types of human feedback, the preference-based approach is no more sample-efficient. This actually contradicts with the empirical observations in the existing literature, which suggests preference-based methods have superior performance. Hence, our theory shows the bias-free modeling plays a great role in the lower sample complexity of preference-based methods, and our theoretical results can conversely confirm the standard BTL modeling of human preference feedback—it is reasonable to believe human preference data is indeed subject to less bias and uncertainty in practice.

6 CONCLUSION

In this work, we have studied policy learning using human feedback for reward engineering in bandit. Specifically, we have provided a theoretical comparison between human rating methods and preference-based methods, which shows human bias and uncertainty can have considerable adverse effect on policy learning. Our theory also suggests the preference-based approach has no provable advantage over the traditional rating-based approach when the two types of human feedback are modeled with equally strong human bias and uncertainty. This implies the reason for the empirical success of preference-based methods might be that human preference data are subject to milder human bias and uncertainty. Beyond this work, it is still open for future work to investigate the case when the human feedback is generated from a mixture model representing a group of annotators and provide a comparison between rating methods and preference-based methods in this setting.

REFERENCES

[1] Youssef Abdelkareem, Shady Shehata, and Fakhri Karray. Advances in preference-based reinforcement learning: A review. In *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2527–2532. IEEE, 2022.

[2] Nir Ailon, Zohar Shay Karnin, and Thorsten Joachims. Reducing dueling bandits to cardinal bandits. In *ICML*, volume 32, pages 856–864, 2014.

[3] András Antos, Csaba Szepesvári, and Rémi Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. In *Proceedings of the 19th Annual Conference on Learning Theory*, COLT'06, pages 574–588, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 3540352945. doi: 10.1007/11776420_42. URL https://doi.org/10.1007/11776420_42.

[4] Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324, 1952.

[5] Jacob Buckman, Carles Gelada, and Marc G Bellemare. The importance of pessimism in fixed-dataset policy optimization. *arXiv preprint arXiv:2009.06799*, 2020.

[6] Róbert Busa-Fekete, Balázs Szörényi, Paul Weng, Weiwei Cheng, and Eyke Hüllermeier. Preference-based reinforcement learning: evolutionary direct policy search using a preference-based racing algorithm. *Machine Learning*, 97(3):327–351, 2014.

[7] Jinglin Chen and Nan Jiang. Information-theoretic considerations in batch reinforcement learning. In *International Conference on Machine Learning*, pages 1042–1051. PMLR, 2019.

[8] Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 193–202, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450318693. doi: 10.1145/2433396.2433420. URL https://doi.org/10.1145/2433396.2433420.

[9] Xiaoyu Chen, Han Zhong, Zhuoran Yang, Zhaoran Wang, and Liwei Wang. Human-in-the-loop: Provably efficient preference-based reinforcement learning with general function approximation. In *International Conference on Machine Learning*, pages 3773–3793. PMLR, 2022.

[10] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952.

[11] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

[12] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, pages 4299–4307, 2017.

[13] Jialin Dong, Kai Yang, and Yuanming Shi. Ranking from crowdsourced pairwise comparisons via smoothed matrix manifold optimization. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 949–956, 2017. doi: 10.1109/ICDMW.2017.130.

[14] Yaqi Duan, Zeyu Jia, and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*, pages 2701–2709. PMLR, 2020.

[15] Dylan J Foster, Akshay Krishnamurthy, David Simchi-Levi, and Yunzong Xu. Offline reinforcement learning: fundamental barriers for value function approximation. *arXiv preprint arXiv:2111.10919*, 2021.

[16] Pratik Gajane, Tanguy Urvoy, and Fabrice Clérot. A relative exponential weighing algorithm for adversarial utility-based dueling bandits. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 218–227, 2015.

[17] Botao Hao, Xiang Ji, Yaqi Duan, Hao Lu, Csaba Szepesvári, and Mengdi Wang. Bootstrapping fitted q-evaluation for off-policy inference. In *International Conference on Machine Learning*, pages 4074–4084. PMLR, 2021.

[18] Lucjan Janowski and Margaret Pinson. The accuracy of subjects in a quality experiment: A theoretical subject model. *IEEE Transactions on Multimedia*, 17(12):2210–2224, 2015. doi: 10.1109/TMM.2015.2484963.

[19] Xiang Ji, Minshuo Chen, Mengdi Wang, and Tuo Zhao. Sample complexity of nonparametric off-policy evaluation on low-dimensional manifolds using deep networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9x3CO0ZU9LR.

[20] Xin Jin, Le Wu, Xiaodong Li, Siyu Chen, Siwei Peng, Jingying Chi, Shiming Ge, Chenggen Song, and Geng Zhao. Predicting aesthetic score distribution through cumulative jensen-shannon divergence. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[21] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline RL? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

[22] H John Bernardin, Stephanie Thomason, M Ronald Buckley, and Jeffrey S Kane. Rater rating-level bias and accuracy in performance appraisals: The impact of rater personality, performance management competence, and rater accountability. *Human Resource Management*, 55 (2):321–340, 2016.

[23] Junpei Komiyama, Junya Honda, Hisashi Kashima, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in dueling bandit problem. In *COLT*, pages 1141–1154, 2015.

[24] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative Q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33: 1179–1191, 2020.

[25] Rafał Latała. Estimates of moments and tails of gaussian chaoses. 2006.

[26] Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*, 2022.

[27] Gene Li, Cong Ma, and Nathan Srebro. Pessimism for offline linear contextual bandits using ℓ_p confidence sets. *arXiv preprint arXiv:2205.10671*, 2022.

[28] Jing Li, Suiyi Ling, Junle Wang, and Patrick Le Callet. A probabilistic graphical model for analyzing the subjective visual quality assessment data from crowdsourcing. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, pages 3339–3347, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379885. doi: 10.1145/3394171.3413619. URL https://doi.org/10.1145/3394171.3413619.

[29] Jing Li, Suiyi Ling, Junle Wang, Zhi Li, and Patrick Le Callet. Gpm: A generic probabilistic model to recover annotator's behavior and ground truth labeling. *arXiv preprint arXiv:2003.00475*, 2020.

[30] Zhi Li and Christos G Bampis. Recover subjective quality scores from noisy measurements. In *2017 Data compression conference (DCC)*, pages 52–61. IEEE, 2017.

[31] Zhi Li, Christos G Bampis, Lukáš Krasula, Lucjan Janowski, and Ioannis Katsavounidis. A simple model for subject behavior in subjective experiments. *arXiv preprint arXiv:2004.02067*, 2020.

[32] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, February 2015. ISSN 00280836. URL http://dx.doi.org/10.1038/nature14236.

[33] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR, 2016.

[34] Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. Rankme: Reliable human ratings for natural language generation. *arXiv preprint arXiv:1803.05928*, 2018.

[35] Ellen R Novoseller, Yanan Sui, Yisong Yue, and Joel W Burdick. Dueling posterior sampling for preference-based reinforcement learning. *arXiv preprint arXiv:1908.01289*, 2019.

[36] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[37] Aldo Pacchiano, Aadirupa Saha, and Jonathan Lee. Dueling rl: reinforcement learning with trajectory preferences. *arXiv preprint arXiv:2111.04850*, 2021.

[38] María Pérez-Ortiz, Aliaksei Mikhailiuk, Emin Zerman, Vedad Hulusic, Giuseppe Valenzise, and Rafał K. Mantiuk. From pairwise comparisons and rating to a unified quality scale. *IEEE Transactions on Image Processing*, 29:1139–1151, 2020. doi: 10.1109/TIP.2019.2936103.

[39] Robin L. Plackett. The analysis of permutations. *Journal of The Royal Statistical Society Series C-applied Statistics*, 24:193–202, 1975.

[40] Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information Processing Systems*, 34:11702–11716, 2021.

[41] Tongzheng Ren, Jialian Li, Bo Dai, Simon S Du, and Sujay Sanghavi. Nearly horizon-free offline reinforcement learning. *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

[42] Aadirupa Saha and Aditya Gopalan. Battle of bandits. In *Uncertainty in Artificial Intelligence*, 2018.

[43] Ayush Sekhari, Karthik Sridharan, Wen Sun, and Runzhe Wu. Contextual bandits and imitation learning via preference-based active queries. *arXiv preprint arXiv:2307.12926*, 2023.

[44] Nihar Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran, and Martin Wainwright. Estimation from pairwise comparisons: Sharp minimax bounds with topology dependence. In *Artificial intelligence and statistics*, pages 856–865. PMLR, 2015.

[45] Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International Conference on Machine Learning*, pages 19967–20025. PMLR, 2022.

[46] Daniel Shin, Anca D Dragan, and Daniel S Brown. Benchmarks and algorithms for offline preference-based reward learning. *arXiv preprint arXiv:2301.01392*, 2023.

[47] David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489, January 2016. doi: 10.1038/nature16961.

[48] Csaba Szepesvári and Rémi Munos. Finite time bounds for sampling based fitted value iteration. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, pages 880–887, New York, NY, USA, 2005. Association for Computing Machinery. ISBN 1595931805. doi: 10.1145/1102351.1102462. URL https://doi.org/10.1145/1102351.1102462.

[49] Kevin R. Tarlow, Daniel F. Brossart, Alexandra M. McCammon, Alexander J. Giovanetti, M. Camille Belle, and Joshua Philip. Reliable visual analysis of single-case data: A comparison of rating, ranking, and pairwise methods. *Cogent Psychology*, 8(1):1911076, 2021. doi: 10.1080/23311908.2021.1911076. URL https://doi.org/10.1080/23311908.2021.1911076.

[50] Ana C. Tenorio-Gonzalez, Eduardo F. Morales, and Luis Villaseñor Pineda. Dynamic reward shaping: Training a robot by voice. In *Proceedings of the 12th Ibero-American Conference on Advances in Artificial Intelligence*, IBERAMIA'10, pages 483–492, Berlin, Heidelberg, 2010. Springer-Verlag. ISBN 3642169511.

[51] Masatoshi Uehara, Masaaki Imaizumi, Nan Jiang, Nathan Kallus, Wen Sun, and Tengyang Xie. Finite sample analysis of minimax offline reinforcement learning: Completeness, fast rates and first-order efficiency. *arXiv preprint arXiv:2102.02981*, 2021.

[52] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019. doi: 10.1017/9781108627771.

[53] Ruosong Wang, Dean P Foster, and Sham M Kakade. What are the statistical limits of offline rl with linear function approximation? *arXiv preprint arXiv:2010.11895*, 2020.

[54] Yuanhao Wang, Qinghua Liu, and Chi Jin. Is rlhf more difficult than standard rl? *arXiv preprint arXiv:2306.14111*, 2023.

[55] Simon K Warfield, Kelly H Zou, and William M Wells. Validation of image segmentation by estimating rater bias and variance. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 366(1874):2361–2375, 2008.

[56] Christian Wirth, Johannes Furnkranz, Gerhard Neumann, et al. Model-free preference-based reinforcement learning. In *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, pages 2222–2228, 2016.

[57] Christian Wirth, Riad Akrour, Gerhard Neumann, and Johannes Fürnkranz. A survey of preference-based reinforcement learning methods. *The Journal of Machine Learning Research*, 18(1):4945–4990, 2017.

[58] Tengyang Xie and Nan Jiang. Q* approximation schemes for batch reinforcement learning: A theoretical comparison. In *Conference on Uncertainty in Artificial Intelligence*, pages 550–559. PMLR, 2020.

[59] Tengyang Xie, Ching-An Cheng, Nan Jiang, Paul Mineiro, and Alekh Agarwal. Bellman-consistent pessimism for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34:6683–6694, 2021.

[60] Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in Neural Information Processing Systems*, 34:27395–27407, 2021.

[61] Tengyu Xu and Yingbin Liang. Provably efficient offline reinforcement learning with trajectory-wise reward. *arXiv preprint arXiv:2206.06426*, 2022.

[62] Yichong Xu, Ruosong Wang, Lin Yang, Aarti Singh, and Artur Dubrawski. Preference-based reinforcement learning with finite-time guarantees. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 18784–18794. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/d9d3837ee7981e8c064774da6cdd98bf-Paper.pdf.

[63] Yuling Yan, Gen Li, Yuxin Chen, and Jianqing Fan. The efficacy of pessimism in asynchronous q-learning. *arXiv preprint arXiv:2203.07368*, 2022.

[64] Georgios N Yannakakis and Héctor P Martínez. Ratings are overrated! *Frontiers in ICT*, 2:13, 2015.

[65] Ming Yin and Yu-Xiang Wang. Optimal uniform ope and model-based offline reinforcement learning in time-homogeneous, reward-free and task-agnostic settings. *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

[66] Ming Yin and Yu-Xiang Wang. Towards instance-optimal offline reinforcement learning with pessimism. *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, 34, 2021.

[67] Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1567–1575. PMLR, 2021.

[68] Ming Yin, Yaqi Duan, Mengdi Wang, and Yu-Xiang Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. In *International Conference on Learning Representation*, 2022.

[69] Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *Advances in Neural Information Processing Systems*, 33:14129–14142, 2020.

[70] Yisong Yue and Thorsten Joachims. Interactively optimizing information retrieval systems as a dueling bandits problem. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1201–1208. ACM, 2009.

[71] Yisong Yue and Thorsten Joachims. Beat the mean bandit. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 241–248, 2011.

[72] Yisong Yue, Josef Broder, Robert Kleinberg, and Thorsten Joachims. The k-armed dueling bandits problem. *Journal of Computer and System Sciences*, 78(5):1538–1556, 2012.

[73] Andrea Zanette. When is realizability sufficient for off-policy reinforcement learning? *arXiv preprint arXiv:2211.05311*, 2022.

[74] Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 34:13626–13640, 2021.

[75] Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Provable offline reinforcement learning with human feedback. *arXiv preprint arXiv:2305.14816*, 2023.

[76] Wenhao Zhan, Masatoshi Uehara, Wen Sun, and Jason D Lee. How to query human feedback efficiently in rl? *arXiv preprint arXiv:2305.18505*, 2023.

[77] Banghua Zhu, Jiantao Jiao, and Michael I Jordan. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. *arXiv preprint arXiv:2301.11270*, 2023.

[78] Masrour Zoghi, Shimon Whiteson, Remi Munos, Maarten de Rijke, et al. Relative upper confidence bound for the k-armed dueling bandit problem. In *JMLR Workshop and Conference Proceedings*, number 32, pages 10–18. JMLR, 2014.

A   ADDITIONAL NOTATION

Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be two functions. We denote the their composition $f(g(\cdot))$ with $(f \circ g)(\cdot)$. We use $\mathcal{N}(\mu, \sigma^2)$ to denote a Gaussian distribution with mean μ and variance σ^2. For a probability event $\mathcal{E}$, we denote its complement event with $\mathcal{E}^C$. For a vector v, $\|v\|_2$ denotes the ℓ_2-norm of the vector v. For a positive semidefinite matrix A, $\|v\|_A$ denotes a semi-norm of the vector v with respect to the matrix A with $\|v\|_A = \sqrt{v^\top A v}$. We denote the set of all positive integers with $\mathbb{Z}_{>0}$.

B   AN EXAMPLE OF HUMAN RATING MODEL

Consider a setting in which the true reward function satisfies $r(s, a) \in [0, 1]$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. For any r,
$$h(r, \epsilon) = r^2 + r^2 \epsilon |\epsilon| \quad \text{and} \quad \bar{h}(r) = r^2. \tag{11}$$
This specific rating model has an interpretation that aligns with how human annotators give ratings in practice. Many human annotators with strong personal opinions and taste tend to exhibit an extreme bias in their evaluations, making them rate subjects with low true reward even lower and rate those with high true reward even higher on average. This is captured by the quadratic form of $\bar{h}(\cdot)$. On the other hand, when the white noise ϵ has a large magnitude, the impact of uncertainty on the final rating should be even larger and more drastic. Since annotators tend to be more deliberate yet more uncertain about subjects with large true reward, the fluctuation of the final rating should also increase with the magnitude of the true reward of a subject. Note that the uncertainty $r^2 \epsilon |\epsilon|$ is a subexponential random variable and actually not a polynomial of ϵ; however, our theory still applies, since the tail of $r^2 \epsilon |\epsilon|$ is dominated by the quadratic $r^2 \epsilon^2$. More generally, our theory is applicable to any uncertainty noise whose tail is dominated by a degree-q polynomial of ϵ. Furthermore, one can easily check that this h-function satisfies all three conditions in Section 3.

Now, let us also provide some corollaries for the suboptimality of the LCB algorithm (Algorithm 1) under the example rating model in (11). In the partial coverage setting under Assumption 1, we have the following corollary.

Corollary 1. *For any fixed constant $0 < \delta < 1$, there exists a contextual bandit instance with initial state distribution ρ such that if one samples a dataset $\mathcal{D}$ of size $n \geq c(\delta, c_b, c_V, \sigma, R)$ using a sampling distribution d satisfying Assumption 1 with $C^\star = 2$ and runs Algorithm 1 on $\mathcal{D}$, the output policy $\widehat{\pi}_{\text{LCB}}$ must suffer constant suboptimality, i.e.,*
$$\mathbb{E}_{\mathcal{D}}[\text{SubOpt}(\widehat{\pi}_{\text{LCB}})] = c_0, \tag{12}$$
where c_0 is a universal constant and $c(\delta, c_b, c_V, \sigma, R)$ is a constant depending on $\delta, c_b, c_V, \sigma, R$.

The proof of this corollary is deferred to Appendix E.2.

Let us also provide a corollary for the suboptimality of the LCB algorithm under the same example rating model in the full coverage setting of Assumption 2.

Corollary 2. *For any fixed constant $0 < \delta < 1$, there exists a contextual bandit instance with initial state distribution ρ such that if one samples a dataset $\mathcal{D}$ of size $n \geq \max\{48\sigma^4, 60\}$ using a sampling distribution d satisfying Assumption 2 with $d(s, a) = \frac{1}{SA}$ for every $s \in \mathcal{S}$ and $a \in \mathcal{A}$ and runs Algorithm 1 on $\mathcal{D}$, the output policy $\widehat{\pi}_{\text{LCB}}$ must suffer suboptimality at least*
$$\mathbb{E}_{\mathcal{D}}[\text{SubOpt}(\widehat{\pi}_{\text{LCB}})] = \frac{c_0 \sigma}{n^{1/4}},$$
where c_0 is a universal constant.

The proof of this corollary is deferred to Appendix F.2. This corollary shows the suboptimality can decay more slowly under the influence of human annotator bias.

C   UPPER BOUND UNDER ASSUMPTION 1 WITH KNOWLEDGE OF NOISE

To compare with the lower bounds for human rating, we prove an upper bound on the suboptimality of the LCB algorithm (Algorithm 1) in the most benign case that the learner has full knowledge of the uncertainty noise distribution of the rating model and design the LCB penalty b_n accordingly. This assumes the learner is able to find the confidence interval with any δ, which is equivalent to knowing the cumulative density function of the distribution and can be unrealistic for real human feedback data in practice. This upper bound result provides a more direct comparison with the preference-based approach and demonstrates how human bias can affect the suboptimality when the uncertainty noise can be coped with.

Theorem 5. Suppose Assumption 1 holds. For any fixed constant $0 < \delta < 1$, if one runs Algorithm 1 with

$$b_m = c_b \sqrt{\frac{V_{R,\sigma}^2 \log^q \frac{SA}{\delta}}{m}}$$

and an appropriately chosen universal constant c_b, as soon as $n > 8 \log \frac{2SA}{\delta} / \bar{d}$, where $\bar{d} := \min_{(s,a) \in \mathcal{X}} d(s,a)$ and $\mathcal{X} := \{(s,a) \in \mathcal{S} \times \mathcal{A} : d(s,a) > 0\}$, with probability $1 - \delta$, the suboptimality of the output policy $\widehat{\pi}_{\mathrm{LCB}}$ satisfies

$$\mathrm{SubOpt}(\widehat{\pi}_{\mathrm{LCB}}) \leq c_0 \sum_{(s,a) \in \mathcal{X}} d_\rho^{\pi^\star}(s,a) \cdot \bar{h}^{-1} \left(\sqrt{\frac{V_{R,\sigma}^2 \log^q \frac{SA}{\delta}}{n \cdot d(s,a)}} \right),$$

where c_0 is a constant that depends on q.

This theorem shows that even when the algorithm has full knowledge of the human uncertainty in the rating model, human bias can still influence the suboptimality of $\widehat{\pi}_{\mathrm{LCB}}$ negatively. It demonstrates the effect of bias on the suboptimality is truly unavoidable when using human rating directly. This can be further illustrated with our example model (11) as follows, which shows the suboptimality still decays more slowly because of the quadratic human bias.

Corollary 3. Suppose Assumption 1 holds. For any fixed constant $0 < \delta < 1$, if one runs Algorithm 1 with

$$b_m = \sqrt{\frac{64\sigma^4 \log \frac{SA}{\delta}}{m} + \frac{8\sigma^2 \log \frac{SA}{\delta}}{m}},$$

as soon as $n > 8 \log \frac{2SA}{\delta} / \bar{d}$, where $\bar{d} := \min_{(s,a) \in \mathcal{X}} d(s,a)$ and $\mathcal{X} := \{(s,a) \in \mathcal{S} \times \mathcal{A} : d(s,a) > 0\}$, with probability $1 - \delta$, the suboptimality of the output policy $\widehat{\pi}_{\mathrm{LCB}}$ satisfies

$$\mathrm{SubOpt}(\widehat{\pi}_{\mathrm{LCB}}) \leq 2 \sum_{(s,a) \in \mathcal{X}} d_\rho^{\pi^\star}(s,a) \left(\frac{256\sigma^4 \log \frac{SA}{\delta}}{n \cdot d(s,a)} \right)^{1/4} + d_\rho^{\pi^\star}(s,a) \sqrt{\frac{32\sigma^2 \log \frac{SA}{\delta}}{n \cdot d(s,a)}}.$$

D   SUPPORTING LEMMAS

Lemma 1 (Hoeffding's inequality). Given $X_1, \cdots, X_n$ independent sub-gaussian random variables, each X_i with variance proxy σ_i^2. It holds that

$$\mathbb{P} \left[\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \geq t \right] \leq \exp \left(-\frac{t^2}{2 \sum_{i=1}^n \sigma_i^2} \right).$$

Lemma 2 (Bernstein's inequality). Given $X_1, \cdots, X_n$ independent sub-exponential random variables, each X_i with parameters (τ_i^2, α_i). Define

$$\tau_*^2 := \sum_{i=1}^n \tau_i^2 \quad \text{and} \quad \alpha_* := \max_i \alpha_i.$$

It holds that

$$\mathbb{P} \left[\sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \geq t \right] \leq \begin{cases} \exp \left(-\frac{t^2}{2\tau_*^2} \right), & \text{if } 0 < t < \frac{\tau_*^2}{\alpha_*}; \\ \exp \left(-\frac{t}{2\alpha_*} \right), & \text{if } t \geq \frac{\tau_*^2}{\alpha_*}. \end{cases}$$

The following lemma is modified from Lemma B.1 in [66]. Recall $n_{(s,a)}$ denotes the number of samples in the offline dataset that visits (s,a). This lemma is about the concentration of $n_{(s,a)}$, the number of samples in the offline dataset that visits (s,a).

Lemma 3. Given a dataset $\mathcal{D}$ with n i.i.d. samples $\mathcal{D} = \{(s_i, a_i)\}_{i=1}^n$ from a sampling distribution d, let $d(s,a)$ be the probability (s,a) is sampled from the outcome space $\mathcal{S} \times \mathcal{A}$ and $n_{(s,a)}$ be the number of samples in $\mathcal{D}$ for (s,a). For any $\delta \in (0,1)$, the event

$$d(s,a) \frac{n}{2} \leq n_{(s,a)} \leq d(s,a) \frac{3n}{2} \tag{13}$$

holds simultaneously for all $(s,a) \in \mathcal{S} \times \mathcal{X}$ with probability $1 - \delta$, as soon as $n > 8SA \log \frac{2SA}{\delta} / \bar{d}$, where $\bar{d} := \min_{(s,a) \in \mathcal{X}} d(s,a)$ and $\mathcal{X} := \{(s,a) \in \mathcal{S} \times \mathcal{A} : d(s,a) > 0\}$

Proof. First, consider the event $d(s,a)\frac{n}{2} > n_{(s,a)}$ for some fixed $(s,a) \in \mathcal{S} \times \mathcal{A}$.

Note that we can view $n_{(s,a)}$ as a sum of independent Bernoulli variables, i.e., $n_{(s,a)} = \sum_{i=1}^{n} \mathbb{1}\{(s_i, a_i) = (s,a)\}$, so $n_{(s,a)}$ follows a binomial distribution with parameters $(d(s,a), n)$. Recall the multiplicative Chernoff bound on a binomial random variable:

Lemma 4 (Multiplicative Chernoff bound [10]). *Let X be a Binomial random variable with parameters (p, n). Denote its mean with μ. For any $\epsilon \in (0,1]$, we have*

$$\mathbb{P}[X < (1-\epsilon)\mu] < e^{-\frac{\epsilon^2 \mu}{2}} \tag{14}$$

and

$$\mathbb{P}[X \geq (1+\epsilon)\mu] < e^{-\frac{\epsilon^2 \mu}{3}}. \tag{15}$$

Take $\epsilon = \frac{1}{2}$. (14) suggests $d(s,a)\frac{n}{2} > n_{(s,a)}$ holds with probability at most $e^{-nd(s,a)/8}$. Similarly, taking $\epsilon = \frac{1}{2}$ in (15) suggests $n_{(s,a)} > d(s,a)\frac{3n}{2}$ holds with probability at most $e^{-nd(s,a)/12}$. Taking the union bound on these two events, we have $d(s,a)\frac{n}{2} > n_{(s,a)}$ or $n_{(s,a)} > d(s,a)\frac{3n}{2}$ with probability at most $2e^{-n\bar{d}/8}$ for this fixed (s,a).

Then, we take a union bound over all $(s,a) \in \mathcal{S} \times \mathcal{A}$, which can give us the advertised result. $\qquad\square$

Moreover, since the uncertainty in our rating model in 3 can end up with a complex concentration, we need a lemma that can give the upper and lower tail bounds of Gaussian chaos.

Lemma 5 (Corollary 1 in [25]). *Let X be a zero-mean Gaussian random variable, and let $f : \mathbb{R} \to \mathbb{R}$ be a polynomial for degree $q \in \mathbb{Z}_{>0}$. Then*

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \geq c_q \exp\left(-\left(\frac{t^2}{c_1 \operatorname{Var}(f(X))}\right)^{1/q}\right) \tag{16}$$

and

$$\mathbb{P}[|f(X) - \mathbb{E}[f(X)]| \geq t] \leq C_q \exp\left(-\left(\frac{t^2}{C_1 \operatorname{Var}(f(X))}\right)^{1/q}\right), \tag{17}$$

where $c_1, C_1 > 0$ are absolute constants and $c_q, C_q > 0$ are absolute constants depending on q.

Proposition 1. *Let $\epsilon \sim \mathcal{N}(0, \sigma^2)$. $\epsilon|\epsilon|$ is a sub-exponential random variable with parameters $(4\sigma^4, 4\sigma^2)$.*

Proof. Let $X = Z^2$, where $Z \sim \mathcal{N}(0,1)$. For $0 < \lambda < \frac{1}{2}$,

$$\mathbb{E}[e^{\lambda(X-1)/2}] = (1-2\lambda)^{-1/2}e^{-\lambda} = e^{-\frac{1}{2}\log(1-2\lambda)-\lambda} \geq e^{\lambda^2}$$

and by [52], for $0 < \lambda \leq \frac{1}{4}$,

$$\mathbb{E}[e^{\lambda(X-1)/2}] \leq e^{2\lambda^2}.$$

Thus, let $Y = (\sigma Z)^2$. For $0 < \lambda < \frac{1}{2\sigma^2}$,

$$\mathbb{E}[e^{\lambda(Y-\sigma^2)/2}] = e^{-\frac{1}{2}\log(1-2\lambda\sigma^2)-\lambda\sigma^2} \geq e^{\lambda^2\sigma^4}$$

and by [52], for $0 < \lambda \leq \frac{1}{4\sigma^2}$,

$$\mathbb{E}[e^{\lambda(Y-\sigma^2)/2}] \leq e^{2\lambda^2\sigma^4}.$$

$\qquad\square$

E PROOFS FOR HUMAN RATING UNDER ASSUMPTION 1

E.1 PROOF OF THEOREM 1

Proof. To prove this theorem, it suffices to construct a bandit instance with $\mathcal{S} = \{s\}$ and $\mathcal{A} = \{a_1, \cdots, a_A\}$, where $A = n^2$. Let the true reward function r be defined such that two conditions are satisfied: (i) let $r(s, a_1) = r(s, a_i) + c(R, \sigma, q)\bar{h}^{-1}(V_{R,\sigma})$ for any a_i such that $i = 2, \cdots, A$, where $c(R, \sigma, q)$ is a constant that might depend on R, q, σ and makes sure that $c(R, \sigma, q)\bar{h}^{-1}(V_{R,\sigma}) = cR$ for some absolute constant $c > 0$ and $r \in [0, R]^{SA}$; (ii) let $\operatorname{Var}\left(h(r(s, a_i), \epsilon) - \bar{h}(r(s, a_i))\right) = c(q)V_{R,\sigma}^2$ for any a_i such that $i = 2, \cdots, A$, where $c(q)$ is a constant depending on q. We can select

$c(R, \sigma, q)$ and $c(q)$ so that such a reward function r exists. It can be observed that a_1 is the optimal action at s.

For the sampling distribution d, let $d(s, a_1) = \frac{1}{2}$ and $d(s, a) = \frac{1}{2(A-1)}$ for any a_i such that $i = 2, \cdots, A$.

Consider the event that each of the suboptimal actions is observed less than once in the offline dataset, i.e., $n_{(s,a)} = 0$ or $n_{(s,a)} = 1$ for all a_i such that $i = 2, \cdots, A$. For convenience, let us denote this event with $\mathcal{E}_1$. Conditioned on the total number of observations of suboptimal actions in the dataset, we can obtain the following via a simple counting argument

$$\mathbb{P}\left[\mathcal{E}_1 \mid \sum_{i=2}^{A} n_{(s,a_i)}\right] = \frac{(A-1)(A-2) \cdot (A - \sum_{i=2}^{A} n_{(s,a_i)})}{(A-1)^{\sum_{i=2}^{A} n_{(s,a_i)}}} \geq \left(\frac{A-n}{A-1}\right)^n. \tag{18}$$

where the last inequality is because $(\frac{A-x}{A-1})^x$ is monotonically decreasing for $x \geq 1$ and the fact $n \geq \sum_{i=2}^{A} n_{(s,a_i)}$.

In addition, let us plug in $A = n^2$, and we can observe

$$\left(\frac{A-n}{A-1}\right)^n \geq \lim_{n \to \infty} \left(\frac{n^2 - n}{n^2 - 1}\right)^n = \frac{1}{e}.$$

Combining the two inequalities above, we can arrive at $\mathbb{P}[\mathcal{E}_1 \mid \sum_{i=2}^{A} n_{(s,a_i)}] \geq \frac{1}{e}$. Since this bound holds for any value of $\sum_{i=2}^{A} n_{(s,a_i)}$, we have $\mathbb{P}[\mathcal{E}_1] \geq \frac{1}{e}$.

On the other hand, let us consider the event $n_{(s,a_1)} < 0.9n$ and denote it with $\mathcal{E}_2$. Note that this event is equivalent to the event $\sum_{i=2}^{A} n_{(s,a_i)} \geq 0.1n$. Since $n_{(s,a_1)}$ follows the binomial distribution with parameters $(n, \frac{1}{2})$, by (14) in Lemma 4, we have

$$\mathbb{P}\left[n_{(s,a_1)} < 0.9n\right] = 1 - \mathbb{P}\left[n_{(s,a_1)} > 0.9n\right] \geq 1 - e^{-8n/75} \geq 0.99517 \tag{19}$$

as long as $n \geq 50$.

Now, we are ready to give a lower bound for the probability of the event that a suboptimal arm is finally selected by Algorithm 1, i.e., there exists $a \neq a^\star, \widehat{r}(s, a) > \widehat{r}(s, a^\star)$.

To this end, let us define a new event $\mathcal{E} := \mathcal{E}_1 \cap \mathcal{E}_2$. The aforementioned probability can be decomposed as follows:

$$\mathbb{P}\left[\exists a \neq a^\star, \widehat{r}(s, a) > \widehat{r}(s, a^\star)\right]$$

$$= 1 - \mathbb{P}\left[\forall a \neq a^\star, \widehat{r}(s, a) \leq \widehat{r}(s, a^\star)\right]$$

$$= 1 - \left(\mathbb{P}\left[\mathcal{E} \wedge \forall a \neq a^\star, \widehat{r}(s, a) \leq \widehat{r}(s, a^\star)\right] + \mathbb{P}\left[\mathcal{E}^C \wedge \forall a \neq a^\star, \widehat{r}(s, a) \leq \widehat{r}(s, a^\star)\right]\right)$$

$$\geq 1 - \left(\mathbb{P}\left[\mathcal{E} \wedge \forall a \neq a^\star, \widehat{r}(s, a) \leq \widehat{r}(s, a^\star)\right] + \mathbb{P}\left[\mathcal{E}^C\right]\right).$$

Let us first focus on finding an upper bound for the probability of the intersection of $\mathcal{E}$ and the event that the optimal arm is selected by the algorithm correctly, i.e., for all $a \neq a^\star, \widehat{r}(s, a) \leq \widehat{r}(s, a^\star)$.

Consider a fixed suboptimal action $a \neq a^\star$ with only one observation in the dataset, i.e., $n_{(s,a)} = 1$. We have

$$\mathbb{P}\left[\widehat{r}(s, a) \leq \widehat{r}(s, a^\star)\right]$$

$$= \mathbb{P}\left[\left(h(r(s, a), \epsilon) - \bar{h}(r(s, a))\right) - \left(\widetilde{r}(s, a^\star) - \bar{h}(r(s, a^\star))\right)\right.$$

$$\left. \leq \bar{h}(r(s, a^\star)) - \bar{h}(r(s, a)) + b_1 - b_{n_{(s,a^\star)}}\right]$$

$$\leq \mathbb{P}\left[\left(h(r(s, a), \epsilon) - \bar{h}(r(s, a))\right) - \left(\widetilde{r}(s, a^\star) - \bar{h}(r(s, a^\star))\right) \leq \bar{h}(r(s, a^\star)) - \bar{h}(r(s, a)) + b_1\right]$$

$$\overset{(i)}{=} \mathbb{P}\left[\left(h(r(s, a), \epsilon) - \bar{h}(r(s, a))\right) + \left(\widetilde{r}(s, a^\star) - \bar{h}(r(s, a^\star))\right) \leq \bar{h}(r(s, a^\star)) - \bar{h}(r(s, a)) + b_1\right]$$

$$= 1 - \mathbb{P}\left[\left(h(r(s, a), \epsilon) - \bar{h}(r(s, a))\right) + \left(\widetilde{r}(s, a^\star) - \bar{h}(r(s, a^\star))\right)\right.$$

$$\geq \bar{h}(r(s,a^\star)) - \bar{h}(r(s,a)) + b_1 \Big]$$

$$\overset{(ii)}{\leq} 1 - c_q \exp\left(-\left(\frac{\big(\bar{h}(r(s,a^\star)) - \bar{h}(r(s,a)) + b_1\big)^2}{c_1\Big(\mathrm{Var}\Big(h(r(s,a),\epsilon) - \bar{h}(r(s,a))\Big) + \mathrm{Var}\Big(\widetilde{r}(s,a^\star) - \bar{h}(r(s,a^\star))\Big)\Big)}\right)^{1/q}\right)$$

$$\leq 1 - c_q \exp\left(-\left(\frac{\big(\bar{h}(r(s,a^\star)) - \bar{h}(r(s,a)) + b_1\big)^2}{c_1\,\mathrm{Var}\Big(h(r(s,a),\epsilon) - \bar{h}(r(s,a))\Big)}\right)^{1/q}\right).$$

In the series of equalities and inequalities above, (i) is Condition 2 of $\mathcal{H}$ described in Section 3, which the uncertainty has symmetric concentration.

(ii) can be obtained by applying (16) in Lemma 5 to $\widetilde{r}(s,a^\star) - \bar{h}(r(s,a^\star))$, which is a sum of $n_{(s,a^\star)}$ human rating samples at $(s,a^\star)$ and one sample at (s,a).

Recall that the pessimism penalty in Algorithm 1 is

$$b_n(s,a) = c_b \sqrt{\frac{\widetilde{V}_{R,\sigma}^2 \log \frac{SA}{\delta}}{n}}.$$

Bringing this and everything else into the inequality above, we have

$$\mathbb{P}\left[\widehat{r}(s,a) \leq \widehat{r}(s,a^\star)\right]$$

$$\leq 1 - c_q \exp\left(-\left(\frac{\big(\bar{h}(r(s,a^\star)) - \bar{h}(r(s,a)) + b_1\big)^2}{c_1\,\mathrm{Var}\Big(h(r(s,a),\epsilon) - \bar{h}(r(s,a))\Big)}\right)^{1/q}\right)$$

$$\overset{(iii)}{\leq} 1 - c_q \exp\left(-\frac{\big(\bar{h}(r(s,a^\star)) - \bar{h}(r(s,a))\big)^{2/q} + b_1^{2/q}}{c_1^{2/q}\,\mathrm{Var}\Big(h(r(s,a),\epsilon) - \bar{h}(r(s,a))\Big)^{2/q}}\right)$$

$$\overset{(iv)}{\leq} 1 - c_q \exp\left(-\frac{\big(C_{h,2}\bar{h}\big(r(s,a^\star) - r(s,a)\big)\big)^{2/q} + b_1^{2/q}}{c_1^{2/q}\,\mathrm{Var}\Big(h(r(s,a),\epsilon) - \bar{h}(r(s,a))\Big)^{2/q}}\right)$$

$$= 1 - c_q \exp\left(-\frac{\big(C_{h,2}c(R,\sigma,q)V_{R,\sigma}\big)^{2/q} + b_1^{2/q}}{c_1^{2/q}\,\mathrm{Var}\Big(h(r(s,a),\epsilon) - \bar{h}(r(s,a))\Big)^{2/q}}\right)$$

$$= 1 - c_{q,1} \exp\left(-\frac{C_{h,2}^{2/q}(c(R,\sigma,q))^{2/q}V_{R,\sigma}^{2/q} + \big(c_b^2\widetilde{V}_{R,\sigma}^2 \log\frac{SA}{\delta}\big)^{1/q}}{c_1^{2/q}(c(q))^{2/q}V_{R,\sigma}^{2/q}}\right)$$

$$= 1 - c_{q,1} \exp\left(-\frac{C_{h,2}^{2/q}(c(R,\sigma,q))^{2/q}V_{R,\sigma}^{2/q} + \big(c_b^2 c_V V_{R,\sigma}^2 \log\frac{SA}{\delta}\big)^{1/q}}{c_1^{2/q}(c(q))^{2/q}V_{R,\sigma}^{2/q}}\right)$$

$$= 1 - c_{q,1}e^{-\left(\frac{C_{h,2}c(R,\sigma,q)}{c_1 c(q)}\right)^{2/q}} \exp\left(-\left(\frac{c_b\sqrt{c_V}}{c_1 c(q)}\right)^{2/q} \log^{1/q}\frac{n^2}{\delta}\right). \tag{20}$$

In the inequalities above, (iii) can be obtained by Jensen's inequality, because the function $x^{2/q}$ is concave for $x \geq 0$ when $q \geq 2$.

(iv) is obtained by Condition 3 of $\mathcal{H}$ described in Section 3.

We can further obtain the following once $n \geq c(q, c_b, c_V, \delta, \sigma, R)$:

$$\mathbb{P}\left[\mathcal{E} \wedge \forall a \neq a^\star, \widehat{r}(s,a) \leq \widehat{r}(s,a^\star)\right]$$

$$= (\mathbb{P}\left[\widehat{r}(s, a_2) \leq \widehat{r}(s, a^\star)\right])^{0.1n}$$

$$\leq \left(1 - c_{q,1}e^{-\left(\frac{C_{h,2}c(R,\sigma,q)}{c_1 c(q)}\right)^{2/q}} \exp\left(-\left(\frac{c_b\sqrt{c_V}}{c_1 c(q)}\right)^{2/q} \log^{1/q}\frac{n^2}{\delta}\right)\right)^{0.1n}$$

$$\overset{(v)}{\leq} \left(1 - c_{q,1}e^{-\left(\frac{C_{h,2}c(R,\sigma,q)}{c_1 c(q)}\right)^{2/q}}\left(\frac{\delta^{1/q}}{n^{2/q}}\right)^{\left(\frac{c_b\sqrt{c_V}}{c_1 c(q)}\right)^{2/q}}\right)^{0.1n} \tag{21}$$

$$\overset{(vi)}{<} 0.17. \tag{22}$$

Above, (v) becomes true once n surpasses some threshold that depends on $q, c_b, c_V, \delta, c_1, C_{h,2}, \sigma, R$. For the case $q \geq 3$, the quantity in (21) decreases monotonically after a certain point that depends on $q, c_b, c_V, \delta, c_1, C_{h,2}, \sigma, R$ and its limit goes to 0 as n approaches infinity, so (vi) is true once $n \geq c(q, c_b, c_V, \delta, \sigma, R)$, where $c(q, c_b, c_V, \delta, \sigma, R)$ is a constant depending on $q, c_b, c_V, \delta, \sigma, R$. For the case $q = 2$, the quantity in (21) monotonically increases towards its limit, which is strictly less than 1, and we can choose $c(R, \sigma, q)$ and $c(q)$ appropriately so that (vi) holds.

On the other hand, through a union bound argument, we have

$$\mathbb{P}\left[\mathcal{E}^C\right] \leq \mathbb{P}\left[\mathcal{E}_1^C\right] + \mathbb{P}\left[\mathcal{E}_2^C\right] = \left(1 - \frac{1}{e}\right) - (1 - 0.99517) < 0.63. \tag{23}$$

Overall, combining (22) and (23), we can obtain a lower bound on the probability that Algorithm 1 fails to identify the optimal policy:

$$\mathbb{P}\left[\exists a \neq a^\star, \widehat{r}(s, a) > \widehat{r}(s, a^\star)\right]$$
$$\geq 1 - \left(\mathbb{P}\left[\mathcal{E} \wedge \forall a \neq a^\star, \widehat{r}(s, a) \leq \widehat{r}(s, a^\star)\right] + \mathbb{P}\left[\mathcal{E}^C\right]\right)$$
$$> 1 - 0.17 - 0.63$$
$$= 0.2.$$

Finally, using this probability lower bound, we arrive at the desired lower bound on the excepted suboptimality:

$$\mathbb{E}_{\mathcal{D}}[\text{SubOpt}(\widehat{\pi}_{\text{LCB}})] = \mathbb{P}\left[\exists a \neq a^\star, \widehat{r}(s, a) > \widehat{r}(s, a^\star)\right] \cdot (r(s, a^\star) - r(s, a))$$
$$= 0.2c(R, \sigma, q)\bar{h}^{-1}\left(V_{R,\sigma}\right)$$
$$= c_0 R.$$

$\square$

E.2 PROOF OF COROLLARY 1

Proof. To prove this corollary, we can construct a bandit instance with the same state space $\mathcal{S}$ and action space $\mathcal{A}$ as in the proof for Theorem 1. Recall the specific human rating function for this corollary is $h(r, \epsilon) = r^2 + r^2\epsilon|\epsilon|$ with $\bar{h}(r) = r^2$. The variance of $h(r, \epsilon)$ is $3r^4\sigma^4$. Let the true reward function r be defined such that $r(s, a_1) = \frac{1}{2} + \frac{1}{4 \cdot 3^{1/4}\sigma}(3r^4\sigma^4)^{1/4} = \frac{3}{4}$ and $r(s, a_2) = \frac{1}{2}$. It can be checked that this reward function satisfies the two conditions in the proof for Theorem 1. The remaining of the proof is similar to the proof for Theorem 1. We can conclude that Algorithm 1 suffers a suboptimality of $\frac{1}{4}$ with constant probability depending on $q, c_b, c_V, \delta, \sigma$. $\square$

F PROOFS FOR HUMAN RATING UNDER ASSUMPTION 2

F.1 PROOF OF THEOREM 2

Proof. To prove this theorem, it suffices to construct a bandit instance with $\mathcal{S} = \{s\}$ and $\mathcal{A} = \{a_1, a_2\}$. Let the true reward function r be defined such that two conditions are satisfied: (i) let $r(s, a_1) = r(s, a_2) + \bar{h}^{-1}\left(\sqrt{\frac{V_{R,\sigma}^2}{n}}\right)$. Note that when n is sufficiently large, i.e., when $n \geq c(R, \sigma, q)$, we can make sure $r \in [0, R]^{SA}$. For (ii), we let $\text{Var}\left(h(r(s, a_1), \epsilon) - \bar{h}(r(s, a_i))\right) = c_1(q)V_{R,\sigma}^2$ and $\text{Var}\left(h(r(s, a_2), \epsilon) - \bar{h}(r(s, a_i))\right) = c_2(q)V_{R,\sigma}^2$, where $c_1(q)$ and $c_2(q)$ are constants

depending on q. We can select $c(R, \sigma, q)$, $c_1(q)$ and $c_2(q)$ so that such a reward function r exists. It can be observed that a_1 is the optimal action at s. Also recall the sampling distribution d is uniform, i.e., $d(s, a) = \frac{1}{SA}$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Let us consider the regime when $n > 8SA \log(\frac{SA}{0.1})/\bar{d}$, where $\bar{d} := \min_{s,a}\{d(s, a) : d(s, a) > 0\}$. In our setting, this is equivalent to $n \geq 120 > 32 \log(40)$. By Lemma 3, for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ simultaneously,

$$\frac{1}{2}n \cdot d(s, a) \leq n_{(s,a)} \leq \frac{3}{2}n \cdot d(s, a) \tag{24}$$

with probability at least 0.9.

In the event of (24), we have $cn_{(s,a_1)} = n_{(s,a_2)}$ with $\frac{1}{3} \leq c \leq 3$. Since we can design the reward function adversarially so that a_1 is the arm with more samples in the offline dataset, we can assume $1 \leq c_s \leq 3$ without loss of generality.

Similar to the proof of Theorem 1, we want to give a lower bound for the probability of the event that a suboptimal arm is finally selected by Algorithm 1, i.e., $\hat{r}(s, a_2) > \hat{r}(s, a_1)$. We can rewrite this probability as follows:

$$\mathbb{P}\left[\hat{r}(s, a_2) > \hat{r}(s, a_1)\right]$$

$$= \mathbb{P}\left[\tilde{r}(s, a_2) - b_{n_{(s,a_2)}} > \tilde{r}(s, a_1) - b_{n_{(s,a_1)}}\right]$$

$$= \mathbb{P}\left[\tilde{r}(s, a_2) - \tilde{r}(s, a_1) > b_{n_{(s,a_2)}} - b_{n_{(s,a_1)}}\right]$$

$$= \mathbb{P}\left[\left(\tilde{r}(s, a_2) - \bar{h}(r(s, a_2))\right) - \left(\tilde{r}(s, a_1) - \bar{h}(r(s, a_1))\right)\right.$$

$$\left. > \bar{h}(r(s, a_1)) - \bar{h}(r(s, a_2)) + b_{n_{(s,a_2)}} - b_{n_{(s,a_1)}}\right]$$

$$= \mathbb{P}\left[\left(\tilde{r}(s, a_2) - \bar{h}(r(s, a_2))\right) + \left(\tilde{r}(s, a_1) - \bar{h}(r(s, a_1))\right)\right.$$

$$\left. > \bar{h}(r(s, a_1)) - \bar{h}(r(s, a_2)) + b_{n_{(s,a_2)}} - b_{n_{(s,a_1)}}\right],$$

where the last step is due to Condition 2 of $\mathcal{H}$ described in Section 3, which the uncertainty has symmetric concentration.

We can invoke (16) in Lemma 5 on the quantity $\left(\tilde{r}(s, a_2) - \bar{h}(r(s, a_2))\right) - \left(\tilde{r}(s, a_1) - \bar{h}(r(s, a_1))\right)$ above, which is a sum of $n_{(s,a_1)}$ human rating samples at (s, a_1) and $n_{(s,a_2)}$ human rating samples at (s, a_2). This can give us:

$$\mathbb{P}\left[\left(\tilde{r}(s, a_2) - \bar{h}(r(s, a_2))\right) + \left(\tilde{r}(s, a_1) - \bar{h}(r(s, a_1))\right)\right.$$

$$\left. > \bar{h}(r(s, a_1)) - \bar{h}(r(s, a_2)) + b_{n_{(s,a_2)}} - b_{n_{(s,a_1)}}\right]$$

$$\geq c_q \exp\left(-\left(\frac{\left(\bar{h}(r(s, a_1)) - \bar{h}(r(s, a_2))\right)^2}{c_1\left(\frac{c_1(q)}{n_{(s,a_1)}} + \frac{c_2(q)}{n_{(s,a_2)}}\right)V_{R,\sigma}^2}\right)^{1/q}\right)$$

$$\geq c_q \exp\left(-\left(\frac{\left(C_{h,2}\bar{h}\left(r(s, a_1) - r(s, a_2)\right)\right)^2}{c_1\left(\frac{c_1(q)}{n_{(s,a_1)}} + \frac{c_2(q)}{n_{(s,a_2)}}\right)V_{R,\sigma}^2}\right)^{1/q}\right)$$

$$= c_q \exp\left(-\left(\frac{C_{h,2}^2 V_{R,\sigma}^2}{nc_1\left(\frac{c_1(q)}{n_{(s,a_1)}} + \frac{c_2(q)}{n_{(s,a_2)}}\right)V_{R,\sigma}^2}\right)^{1/q}\right)$$

$$\geq c_q \exp\left(-\left(\frac{C_{h,2}^2 V_{R,\sigma}^2}{nc_1\left(\frac{c_1(q)}{\frac{3}{2}n \cdot d(s,a_1)} + \frac{c_2(q)}{\frac{3}{2}n \cdot d(s,a_2)}\right)V_{R,\sigma}^2}\right)^{1/q}\right)$$

$$\geq c_q \exp\left(-\left(\frac{3C_{h,2}^2}{4c_1(c_1(q) + c_2(q))}\right)^{1/q}\right)$$

$$=: c_0.$$

Overall, Algorithm 1 is guaranteed to incur an expected suboptimality at least $c_0\bar{h}^{-1}\left(\sqrt{\frac{V_{R,\sigma}^2}{n}}\right)$ as soon as $n \geq \max\{c(R,\sigma,q), 120\}$. □

F.2 PROOF OF COROLLARY 2

Proof. To prove this corollary, we can construct a bandit instance with the same state space $\mathcal{S}$ and action space $\mathcal{A}$ as in the proof for Theorem 2. Recall the specific human rating function for this corollary is $h(r, \epsilon) = r^2 + r^2\epsilon|\epsilon|$ with $\bar{h}(r) = r^2$. The variance of $h(r, \epsilon)$ is $3r^4\sigma^4$. Let the true reward function r be defined such that $r(s, a_1) = \frac{1}{2} + \sigma\left(\frac{1}{n}\right)^{1/4}$ and $r(s, a_2) = \frac{1}{2}$. It can be checked that this reward function satisfies the two conditions in the proof for Theorem 1. The remaining of the proof is similar to the proof for Theorem 1. We can conclude that Algorithm 1 suffers a suboptimality of $\sigma\left(\frac{1}{n}\right)^{1/4}$ with constant probability depending on q as soon as $n \geq \max\{48\sigma^4, 60\}$. □

G PROOFS FOR HUMAN RATING UPPER BOUNDS

G.1 PROOF OF THEOREM 5

Proof. In this proof, let us denote the output of Algorithm 1 $\hat{\pi}_{\mathrm{LCB}}$ with $\hat{\pi}$ for short. By the definition in (1), we can decompose the suboptimality of the output $\hat{\pi}$ of Algorithm 1 as follows:

$$\mathrm{SubOpt}(\hat{\pi})$$
$$= \mathbb{E}_{s\sim\rho}[r(s, \pi^\star(s)) - r(s, \hat{\pi}(s))]$$
$$= \mathbb{E}_{s\sim\rho}[(\bar{h}^{-1} \circ \bar{h})(r(s, \pi^\star(s))) - (\bar{h}^{-1} \circ \bar{h})(r(s, \hat{\pi}(s)))]$$
$$\overset{(i)}{\leq} C_{h,1}\mathbb{E}_{s\sim\rho}\left[\bar{h}^{-1}\left(\bar{h}(r(s, \pi^\star(s))) - \bar{h}(r(s, \hat{\pi}(s)))\right)\right]$$
$$= C_{h,1}\mathbb{E}_{s\sim\rho}\left[\bar{h}^{-1}\left(\bar{h}(r(s, \pi^\star(s))) - \hat{r}(s, \pi^\star(s)) + \hat{r}(s, \pi^\star(s)) - \hat{r}(s, \hat{\pi}(s)) + \hat{r}(s, \hat{\pi}(s)) - \bar{h}(r(s, \hat{\pi}(s)))\right)\right]$$
$$\overset{(ii)}{\leq} C_{h,1}\mathbb{E}_{s\sim\rho}\left[\bar{h}^{-1}\left(\bar{h}(r(s, \pi^\star(s))) - \hat{r}(s, \pi^\star(s)) + \hat{r}(s, \hat{\pi}(s)) - \bar{h}(r(s, \hat{\pi}(s)))\right)\right]$$
$$\overset{(iii)}{\leq} C_{h,1}\mathbb{E}_{s\sim\rho}\left[\bar{h}^{-1}\left(\bar{h}(r(s, \pi^\star(s))) - \hat{r}(s, \pi^\star(s))\right)\right]$$
$$\overset{(iv)}{\leq} C_{h,1}\mathbb{E}_{s\sim\rho}\left[\bar{h}^{-1}\left(\left|\bar{h}(r(s, \pi^\star(s))) - \tilde{r}(s, \pi^\star(s))\right| + b_{n_{(s, \pi^\star(s))}}\right)\right]$$
$$\overset{(v)}{\leq} C_{h,1}\mathbb{E}_{(s,a)\sim d_\rho^{\pi^\star}}\left[\bar{h}^{-1}\left(C'\sqrt{\frac{V_{R,\sigma}^2 \log^q \frac{SA}{\delta}}{n_{(s,a)}}}\right)\right]$$
$$\overset{(vi)}{\leq} C_{h,1}\mathbb{E}_{(s,a)\sim d_\rho^{\pi^\star}}\left[\bar{h}^{-1}\left(C\sqrt{\frac{V_{R,\sigma}^2 \log^q \frac{SA}{\delta}}{n \cdot d(s,a)}}\right)\right]$$
$$\overset{(vii)}{\leq} c_0\mathbb{E}_{(s,a)\sim d_\rho^{\pi^\star}}\left[\bar{h}^{-1}\left(\sqrt{\frac{V_{R,\sigma}^2 \log^q \frac{SA}{\delta}}{n \cdot d(s,a)}}\right)\right]$$
$$\leq c_0 \sum_{(s,a)\in\mathcal{X}} d_\rho^{\pi^\star}(s,a) \cdot \bar{h}^{-1}\left(\sqrt{\frac{V_{R,\sigma}^2 \log^q \frac{SA}{\delta}}{n \cdot d(s,a)}}\right). \tag{25}$$

In the series of equalities and inequalities above, (i) is due to Condition 3 of $\mathcal{H}$ described in Section 3.

(ii) can be obtained because of the fact that in Algorithm 1, $\widehat{\pi}(s) = \arg\max_a \widehat{r}(s, a)$, so for every $s \in \mathcal{S}$, $\widehat{r}(s, \widehat{\pi}(s)) \geq \widehat{r}(s, \pi^\star(s))$.

(iii) can be obtained because the pessimistic penalty b_n in Theorem 5 guarantees $\widehat{r}(s, \widehat{\pi}(s)) \leq \bar{h}(r(s, \widehat{\pi}(s)))$ with probability $1 - \delta$.

To show this, we can focus on bounding the probability of the event

$$\widetilde{r}(s, a) - \mathbb{E}[\widetilde{r}(s, a)] \leq b_{n_{(s,a)}}$$

for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ simultaneously. For simplicity in the remaining part of this proof, let us denote this event with $\mathcal{E}$. This involves an understanding of the concentration of the rating observation $h(r(s, a), \epsilon)$ under (3).

By Lemma 5 and a union bound argument, for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ simultaneously, we have

$$\mathbb{P}\left[|\widetilde{r}(s, a) - \mathbb{E}[\widetilde{r}(s, a)]| \geq t\right] \leq SA \cdot C_q \exp\left(-\left(\frac{t^2}{C_1 V_{R,\sigma}^2}\right)^{1/q}\right).$$

We can solve for t in the inequality above. Given b_n in Theorem 5 with a sufficiently large c_b, this gives

$$|\widetilde{r}(s, a) - \mathbb{E}[\widetilde{r}(s, a)]| \leq c(q)\sqrt{\frac{V_{R,\sigma}^2 \log^q \frac{SA}{\delta}}{n_{(s,a)}}} \leq b_{n_{(s,a)}} \tag{26}$$

for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ simultaneously with probability at least $1 - \delta$. Here, $c(q)$ is an absolute constant depending on q.

Note that $\bar{h}(r(s, a)) = \mathbb{E}[\widetilde{r}(s, a)]$. Thus, we can establish (iii) on the event $\mathcal{E}$, which has probability at least $1 - \delta$.

(iv) is obtained by the monotoncity of $\bar{h}^{-1}(\cdot)$. Since $\bar{h}(r(s, a)) - \widehat{r}(s, a) \leq |\bar{h}(r(s, a)) - \widetilde{r}(s, a)| + b_{n_{(s,a)}}$, $\bar{h}^{-1}(\bar{h}(r(s, a)) - \widehat{r}(s, a)) \leq \bar{h}^{-1}(|\bar{h}(r(s, a)) - \widetilde{r}(s, a)| + b_{n_{(s,a)}})$, for any $(s, a) \in \mathcal{S} \times \mathcal{A}$.

(v) can be obtained by finding an upper bound for $|\bar{h}(r(s, \pi^\star(s))) - \widetilde{r}(s, \pi^\star(s))|$, which follows from (26). That is, on the event $\mathcal{E}$, we have

$$|\bar{h}(r(s, \pi^\star(s))) - \widetilde{r}(s, \pi^\star(s))| \leq c(q)\sqrt{\frac{V_{R,\sigma}^2 \log^q \frac{SA}{\delta}}{n_{(s,\pi^\star(s))}}}$$

for some absolute constant $c(q)$ depending on q. Then, the addition of the bound above and $b_{n_{(s,\pi^\star(s))}}$ gives (v).

(vi) can be obtained from an invocation of Lemma 3, which guarantees $n_{(s,a)} = cd(s, a)n$ for some constant $\frac{1}{2} \leq c \leq \frac{3}{2}$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$ simultaneously, with probability at least $1 - \delta$. We can denote this event with $\mathcal{E}'$.

Finally, (vii) is obtained because we can pull the constant factor inside $\bar{h}^{-1}$ out due to Condition 3 of $\mathcal{H}$ described in Section 3.

The advertised bound can be obtained after taking a union bound on the probability of the complement of $\mathcal{E}$ and the complement of $\mathcal{E}'$, which only changes (25) by a constant factor. □

G.2 Proof of Corollary 3

Proof. In this proof, let us denote the output of Algorithm 1 $\widehat{\pi}_{\text{LCB}}$ with $\widehat{\pi}$ for short. By the definition in (1), we can decompose the suboptimality of the output $\widehat{\pi}$ of Algorithm 1 as follows:

$$\begin{aligned}
&\text{SubOpt}(\widehat{\pi}) \\
&= \mathbb{E}_{s\sim\rho}[r(s, \pi^\star(s)) - r(s, \widehat{\pi}(s))] \\
&= \mathbb{E}_{s\sim\rho}\left[\sqrt{r^2(s, \pi^\star(s))} - \sqrt{r^2(s, \widehat{\pi}(s))}\right] \\
&\overset{(i)}{\leq} \mathbb{E}_{s\sim\rho}\left[\sqrt{r^2(s, \pi^\star(s)) - r^2(s, \widehat{\pi}(s))}\right] \\
&= \mathbb{E}_{s\sim\rho}\left[\sqrt{r^2(s, \pi^\star(s)) - \widehat{r}(s, \pi^\star(s)) + \widehat{r}(s, \pi^\star(s)) - \widehat{r}(s, \widehat{\pi}(s)) + \widehat{r}(s, \widehat{\pi}(s)) - r^2(s, \widehat{\pi}(s))}\right]
\end{aligned}$$

$$\overset{(ii)}{\le} \mathbb{E}_{s\sim\rho}\left[\sqrt{r^2(s,\pi^\star(s)) - \widehat{r}(s,\pi^\star(s)) + \widehat{r}(s,\widehat{\pi}(s)) - r^2(s,\widehat{\pi}(s))}\right]$$

$$\overset{(iii)}{\le} \mathbb{E}_{s\sim\rho}\left[\sqrt{r^2(s,\pi^\star(s)) - \widehat{r}(s,\pi^\star(s))}\right]$$

$$\overset{(iv)}{\le} \mathbb{E}_{s\sim\rho}\left[\sqrt{\left|\bar{h}(r(s,\pi^\star(s))) - \widetilde{r}(s,\pi^\star(s))\right| + b_{n_{(s,\pi^\star(s))}}}\right]$$

$$\overset{(v)}{\le} \mathbb{E}_{(s,a)\sim d_\rho^{\pi^\star}}\left[\left(2\sqrt{\frac{64\sigma^4 \log\frac{SA}{\delta}}{n_{(s,a)}} + \frac{16\sigma^2 \log\frac{SA}{\delta}}{n_{(s,a)}}}\right)^{1/2}\right]$$

$$\le 2 \sum_{(s,a)\in\mathcal{X}} d_\rho^{\pi^\star}(s,a)\left(\frac{256\sigma^4 \log\frac{SA}{\delta}}{n\cdot d(s,a)}\right)^{1/4} + d_\rho^{\pi^\star}(s,a)\sqrt{\frac{32\sigma^2 \log\frac{SA}{\delta}}{n\cdot d(s,a)}}. \tag{27}$$

In the series of equalities and inequalities above, (i) is due to Condition 3 of $\mathcal{H}$ described in Section 3, which can be proved for $\bar{h}(\cdot) = (\cdot)^2$ with Jensen's inequality.

(ii) can be obtained because of the fact that in Algorithm 1, $\widehat{\pi}(s) = \arg\max_a \widehat{r}(s,a)$, so for every $s \in \mathcal{S}$, $\widehat{r}(s,\widehat{\pi}(s)) \ge \widehat{r}(s,\pi^\star(s))$.

(iii) can be obtained because the pessimistic penalty b_n in Corollary 3 guarantees $\widehat{r}(s,\widehat{\pi}(s)) \le r^2(s,\widehat{\pi}(s))$ with probability $1 - \delta/2$.

To show this, we can focus on bounding the probability of the event
$$\widetilde{r}(s,a) - \mathbb{E}[\widetilde{r}(s,a)] \le b_{n_{(s,a)}}$$
for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ simultaneously. For simplicity in the remaining part of this proof, let us denote this event with $\mathcal{E}$. This involves an understanding of the concentration of the rating observation $h(r(s,a),\epsilon)$ under (3).

Note that $\epsilon|\epsilon|$ is a sub-exponential random variable with parameters $(4\sigma^4, 4\sigma^2)$ by Proposition 1.

Now, by Lemma 2 and a union bound argument, for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ simultaneously, we have

$$|\widetilde{r}(s,a) - \mathbb{E}[\widetilde{r}(s,a)]| \le \sqrt{\frac{64\sigma^4 \log\frac{SA}{\delta}}{n_{(s,a)}}} + \frac{8\sigma^2 \log\frac{SA}{\delta}}{n_{(s,a)}} \le b_{n_{(s,a)}}$$

with probability at least $1 - \delta/2$.

Note that $\bar{h}(r(s,a)) = \mathbb{E}[\widetilde{r}(s,a)]$. Thus, we can establish (iii) on the event $\mathcal{E}$, which has probability at least $1 - \delta/2$.

(iv) is obtained by the monotoncity of $\bar{h}^{-1}(\cdot) = \sqrt{\cdot}$. Since $\bar{h}(r(s,a)) - \widehat{r}(s,a) \le |\bar{h}(r(s,a)) - \widetilde{r}(s,a)| + b_{n_{(s,a)}}$, $\bar{h}^{-1}(\bar{h}(r(s,a)) - \widehat{r}(s,a)) \le \bar{h}^{-1}(|\bar{h}(r(s,a)) - \widetilde{r}(s,a)| + b_{n_{(s,a)}})$, for any $(s,a) \in \mathcal{S}\times\mathcal{A}$.

(v) can be obtained by finding an upper bound for $|\bar{h}(r(s,\pi^\star(s))) - \widetilde{r}(s,\pi^\star(s))|$, which follows from (26). That is, on the event $\mathcal{E}$, we have

$$|\bar{h}(r(s,\pi^\star(s))) - \widetilde{r}(s,\pi^\star(s))| \le \sqrt{\frac{64\sigma^4 \log\frac{SA}{\delta}}{n_{(s,a)}}} + \frac{8\sigma^2 \log\frac{SA}{\delta}}{n_{(s,\pi^\star(s))}}.$$

Then, the addition of the bound above and $b_{n_{(s,\pi^\star(s))}}$ gives (v).

(vi) can be obtained from an invocation of Lemma 3, which guarantees $n_{(s,a)} = cd(s,a)n$ for some constant $\frac{1}{2} \le c \le \frac{3}{2}$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$ simultaneously, with probability at least $1 - \delta/2$. We can denote this event with $\mathcal{E}'$.

The advertised bound can be obtained after taking a union bound on the probability of the complement of $\mathcal{E}$ and the complement of $\mathcal{E}'$, which changes (27) by a factor of 2. $\qquad\square$

H PROOF OF THEOREM 3

[77] considers the setting in which the reward function $r_{\theta^\star}(s,a)$ is parameterized by $\theta^\star \in \mathbb{R}^d$ in a linear fashion, i.e., $r_{\theta^\star}(s,a) = (\theta^\star)^\top \phi(s,a)$. Here, $\phi(s,a)$ is a known feature in $\mathbb{R}^d$. The

tabular setting we consider is a special case of the linear setting. To write the tabular setting in this notation, we can let $\phi(s,a) = e_{(s,a)}$, where $e_{(s,a)} \in \mathbb{R}^{SA}$ is the canonical basis vector for the (s,a)-th dimension, and let $\theta^\star \in \mathbb{R}^{SA}$ be a shifted version of the tabular reward function/vector $r \in [0,R]^{SA}$, i.e., $\theta^\star(s,a) = r(s,a) - \frac{1}{SA}\sum_{(s,a)} r(s,a) \in [-R,R]^{SA}$. This ensures $\mathbf{1}^\top \theta^\star = 0$, which is required by [77] for the identifiability of $\theta^\star$. As [77] shows, Algorithm 2 converges to $\theta^\star$.

In the analysis below, we omit the subscript of $\widehat{\pi}_{\text{PMLE}}$ and write the output policy of Algorithm 2 as $\widehat{\pi}$ for simplicity. We also denote the shifted true reward with $\bar{r}(s,a) := r(s,a) - \frac{1}{SA}\sum_{(s,a)} r(s,a)$ for any $s \in \mathcal{S}$ and $a \in \mathcal{A}$ and let $\widehat{r}$ be the one from Line 6 of Algorithm 2. The following analysis largely overlaps with Theorem 3.2 in [77], but we handle the distributional mismatch term with more care and provide a characterization specific to the tabular setting.

The suboptimality of Algorithm 2 can be decomposed as follows:

$$
\begin{aligned}
&\text{SubOpt}(\widehat{\pi}_{\text{PMLE}}) \\
&= \mathbb{E}_{s\sim\rho}[r(s,\pi^\star(s)) - r(s,\widehat{\pi}(s))] \\
&= \mathbb{E}_{s\sim\rho}[\bar{r}(s,\pi^\star(s)) - \bar{r}(s,\widehat{\pi}(s))] \\
&= \mathbb{E}_{s\sim\rho}\Big[\bar{r}(s,\pi^\star(s)) - \widehat{r}(s,\pi^\star(s)) + \widehat{r}(s,\pi^\star(s)) - \widehat{r}(s,\widehat{\pi}(s)) + \widehat{r}(s,\widehat{\pi}(s)) - \bar{r}(s,\widehat{\pi}(s))\Big] \\
&\overset{(i)}{\leq} \mathbb{E}_{s\sim\rho}\Big[\bar{r}(s,\pi^\star(s)) - \widehat{r}(s,\pi^\star(s)) + \widehat{r}(s,\widehat{\pi}(s)) - \bar{r}(s,\widehat{\pi}(s))\Big] \\
&\overset{(ii)}{\leq} \mathbb{E}_{s\sim\rho}\Big[\bar{r}(s,\pi^\star(s)) - \widehat{r}(s,\pi^\star(s))\Big] \\
&\overset{(iii)}{\leq} \sqrt{\sup_{v:\|v\|_\infty \leq 2R, \mathbf{1}^\top v = 0} \frac{\left(\sum_{(s,a)} d_\rho^{\pi^\star}(s,a) v(s,a)\right)^2}{\sum_{(s,a^0,a^1)} d(s,a^0,a^1)\big(v(s,a^0) - v(s,a^1)\big)^2}} \; \|\bar{r} - \widehat{r}\|_\Sigma \\
&\overset{(iv)}{=} \sqrt{\sup_{v:\|v\|_\infty \leq 1, \mathbf{1}^\top v = 0} \frac{\left(\sum_{(s,a)} d_\rho^{\pi^\star}(s,a) v(s,a)\right)^2}{\sum_{(s,a^0,a^1)} d(s,a^0,a^1)\big(v(s,a^0) - v(s,a^1)\big)^2}} \; \|\bar{r} - \widehat{r}\|_\Sigma \\
&\leq C^\dagger \left(\|\bar{r} - \widetilde{r}\|_\Sigma + \|\widetilde{r} - \widehat{r}\|_\Sigma\right) \\
&\overset{(v)}{\leq} 2C^\dagger \left(CR\sqrt{\frac{SA + \log\frac{1}{\delta}}{\gamma^2 n}} + cR\sqrt{\frac{S^2 A^2 \log\frac{n}{\delta}}{n}}\right) \\
&\leq c_0 C^\dagger R \left(\sqrt{\frac{SA + \log\frac{1}{\delta}}{\gamma^2 n}} + \sqrt{\frac{S^2 A^2 \log\frac{n}{\delta}}{n}}\right).
\end{aligned}
\tag{28}
$$

In the analysis above, (i) can be obtained because of Line 6 in Algorithm 2, $\widehat{\pi}(s) = \arg\max_\pi \mathbb{E}_{s\sim\rho}[\widehat{r}(s,\pi(s))]$, so $\mathbb{E}_{s\sim\rho}[\widehat{r}(s,\widehat{\pi}(s))] \geq \mathbb{E}_{s\sim\rho}[\widehat{r}(s,\pi^\star(s))]$.

(ii) is because $\widehat{r}$ is selected pessimistically from $\mathcal{F}_{\text{CR}}(\widetilde{r})$ in Line 6 of Algorithm 2 and $\bar{r} \in \mathcal{F}_{\text{CR}}(\widetilde{r})$ with probability at least $1 - \delta/2$. This is guaranteed by Lemma 3.1 of [77], which proves with probability at least $1 - \delta$,

$$
\|\widetilde{r} - \bar{r}\|_{\widehat{\Sigma}} \leq CR\sqrt{\frac{SA + \log\frac{1}{\delta}}{\gamma^2 n}},
\tag{29}
$$

where $C > 0$ is an absolute constant. Hence, we have $\mathbb{E}_{s\sim\rho}[\widehat{r}(s,\widehat{\pi}(s))] \leq \mathbb{E}_{s\sim\rho}[\bar{r}(s,\widehat{\pi}(s))]$.

In (iii), we change the measure by invoking the following lemma:

Lemma 6 (modified from Lemma 1 in [19]). Given a function class $\mathcal{F}$ that contains functions mapping from $\mathcal{X}$ to $\mathbb{R}$ and a probability distribution p_1 supported on $\mathcal{X}$ and a joint quasiprobability probability distribution p_2 supported on $\mathcal{X} \times \mathcal{X}$, for any $g \in \mathcal{F}$,

$$
\mathbb{E}_{x\sim p_1}[g(x)] \leq \sqrt{\mathbb{E}_{(x,x')\sim p_2}[g(x)g(x')] \sup_{g\in\mathcal{F}} \frac{\mathbb{E}_{p_1}[g(x)]^2}{\mathbb{E}_{p_2}[g(x)g(x')]}}.
$$

Moreover, note that Σ is a scaled Laplacian matrix. Since $\bar{r}, \widehat{r} \in \mathcal{F}$ ($\mathcal{F}$ is defined in Algorithm 2) and $v^\top \Sigma v = \sum_{i,j} \Sigma_{i,j} v_i v_j$ for any vector v, this allows us to write (the denominator of) the distributional mismatch term in the form of $C^\dagger$ defined in (8).

(iv) can be obtained because the distributional mismatch term $C^\dagger$ is constant with respect to any nonzero scaling of v.

(v) can be obtained by invoking the following lemma, which provides an upper bound on the difference between the ground truth $\bar{r}$ and the empirical estimation $\widetilde{r}$ with respect to Σ, the covariance matrix for the population sampling distribution.

Lemma 7. With probability at least $1 - \delta$, Algorithm 2 satisfies

$$\|\widetilde{r} - \bar{r}\|_\Sigma \leq CR\sqrt{\frac{SA + \log\frac{1}{\delta}}{\gamma^2 n}} + cR\sqrt{\frac{S^2 A^2 \log\frac{n}{\delta}}{n}}, \tag{30}$$

where $\gamma = \frac{1}{2 + \exp(R\sqrt{SA}) + \exp(-R\sqrt{SA})}$ and $C, c > 0$ are absolute constants.

The proof of Lemma 7 is deferred to Section H.1.

Finally, we conclude with (28), which leads to the advertised result.

H.1 PROOF OF LEMMA 7

We can start with

$$\|\widetilde{r} - \bar{r}\|_\Sigma = \sqrt{\|\widetilde{r} - \bar{r}\|_\Sigma^2}$$

$$\leq 2\|\widetilde{r} - \bar{r}\|_{\widehat{\Sigma}}^2 + c\frac{SA\log\frac{n}{\delta}}{n}\|\widetilde{r} - \bar{r}\|_2^2, \tag{31}$$

in which the inequality is due to the following lemma paraphrased from [37]:

Lemma 8 (Lemma 7 in [37]). For any $\delta \in (0, 1)$, Algorithm 2 satisfies

$$\|\widetilde{r} - \bar{r}\|_\Sigma^2 \leq 2\|\widetilde{r} - \bar{r}\|_{\widehat{\Sigma}}^2 + c^2\frac{SA\log\frac{n}{\delta}}{n}\|\widetilde{r} - \bar{r}\|_2^2 \tag{32}$$

with probability at least $1 - \delta$. $\widehat{\Sigma}$ is the empirical covariance matrix defined in Algorithm 2. Σ is the population covariance matrix, i.e., $\Sigma = \mathbb{E}[\widehat{\Sigma}]$. $c > 0$ is an absolute constant.

To proceed with (31), we invoke Lemma 3.1 in [77] for an upper bound on $\|\widetilde{r} - \bar{r}\|_{\widehat{\Sigma}}$, which gives

$$\|\widetilde{r} - \bar{r}\|_\Sigma \leq \sqrt{C^2 R^2 \frac{SA + \log\frac{1}{\delta}}{\gamma^2 n} + c^2 \frac{SA\log\frac{n}{\delta}}{n}\|\widetilde{r} - \bar{r}\|_2^2}$$

$$\leq CR\sqrt{\frac{SA + \log\frac{1}{\delta}}{\gamma^2 n}} + c\sqrt{\frac{SA\log\frac{n}{\delta}}{n}}\|\widetilde{r} - \bar{r}\|_2$$

$$\leq CR\sqrt{\frac{SA + \log\frac{1}{\delta}}{\gamma^2 n}} + cR\sqrt{\frac{S^2 A^2 \log\frac{n}{\delta}}{n}},$$

where the last step is because $\|\widetilde{r} - \bar{r}\|_2 \leq 2R\sqrt{SA}$. $C > 0$ is an absolute constant.

I PROOF OF THEOREM 4

We consider a simple bandit with one state s and two actions $\{a_1, a_2\}$. Since there is only one state in this bandit, we omit the state notation and use the shorthand (a) for (s, a). The true reward of this bandit satisfies $0 \leq r(a_1) < r(a_2) \leq 1$. By Condition 1 of our human rating model, the biased reward also satisfies $0 \leq h(r(a_1)) < h(r(a_2)) \leq 1$.

Without loss of generality, let a human preference feedback $\widetilde{y} = \{0, 1\}$ for this bandit instance be in the following form: $\widetilde{y}_i = 1$ if the human annotator prefers a_2, and $\widetilde{y}_i = 0$ if the human annotator prefers a_1.

For the preference-based method, the MLE objective is to obtain a reward estimate f, or two scalar estimates $f(a_1)$ and $f(a_2)$ that maximizes

$$\sum_{i=1}^n \log\left(\frac{\mathbb{1}\{\widetilde{y}_i = 1\}\exp(f(a_2))}{\exp(f(a_1)) + \exp(f(a_2))} + \frac{\mathbb{1}\{\widetilde{y}_i = 0\}\exp(f(a_1))}{\exp(f(a_1)) + \exp(f(a_2))}\right).$$

We can see that the problem is equivalent to finding a scalar $r(a_2) - r(a_1)$ that maximizes

$$\sum_{i=1}^{n} \log \left(\frac{\mathbb{1}\{\widetilde{y}_i = 1\}}{1 + \exp(-(f(a_2) - f(a_1)))} + \frac{\mathbb{1}\{\widetilde{y}_i = 0\}}{1 + \exp(f(a_2) - f(a_1))} \right).$$

This is also equivalent to the MLE for the parameter $p := \frac{1}{1+\exp(-(\bar{h}(r(a_2))-\bar{h}(r(a_1))))}$ of a Bernoulli distribution with samples $\{\widetilde{y}_i\}_{i=1}^{n}$. We know that the solution of this Bernoulli MLE is $\frac{1}{n} \sum_{i=1}^{n} \widetilde{y}_i$.

To identify the optimal arm a_2 correctly, the estimated parameter needs to be greater than $\frac{1}{2}$ to show a preference for a_2. In the following, we can compute a threshold for the number of samples n such that the failure probability (the algorithm mistakenly identifies a_1 as the optimal arm) is at most δ.

$$\mathbb{P}\left[\frac{1}{n} \sum_{i=1}^{n} \widetilde{y}_i \le \frac{1}{2} \right] \le \delta$$

This can be done by invoking the Chernoff bound (Lemma 4), which is tight in most cases. Specifically, we can invoke (14) in Lemma 4 with $\epsilon = 1 - \frac{1}{2p}$ and solve for n. Thus, the minimum number of samples needed for the preference-based approach to identify the optimal policy n_{pref} is:

$$n_{\text{pref}} = \frac{2 \log \frac{1}{\delta}}{p \epsilon^2} = \frac{2 \log \frac{1}{\delta}}{p(1 - \frac{1}{2p})^2}, \tag{33}$$

which implies we can identify the optimal arm correctly with probability at least $1 - \delta$ once the number of samples exceeds this threshold.

For the human rating case, let us denote the human rating data with two sets: $\{\widetilde{r}_i^{(1)}\}_{i=1}^{n_1}$ is the reward ratings for a_1, and $\{\widetilde{r}_i^{(2)}\}_{i=1}^{n_2}$ is the reward ratings for a_2. Recall for any a, $\widetilde{r}(a)$ denotes the empirical average for the reward of a based on the rating data, and $\widehat{r}(a)$ denotes the pessimistic estimate for the reward of a based on $\widetilde{r}(a)$. Notice when the data coverage is uniform, pessimism no longer plays any role in the algorithm.

In this setting, the estimated reward for a_1 and a_2 needs to satisfy $\widehat{r}(a_1) < \widehat{r}(a_2)$ in order to identify the optimal arm a_2 correctly. We want to find the number of samples needed so that the failure probability is controlled below probability δ.

$$\begin{aligned}
&\mathbb{P}\left[\widehat{r}(a_1) \ge \widehat{r}(a_2) \right] \\
=\ &\mathbb{P}\left[\widetilde{r}(a_1) \ge \widetilde{r}(a_2) \right] \\
=\ &\mathbb{P}\left[\widetilde{r}(a_1) - \bar{h}(r(a_1)) \ge \widetilde{r}(a_2) - \bar{h}(r(a_1)) \right] \\
=\ &\mathbb{P}\left[\widetilde{r}(a_1) - \bar{h}(r(a_1)) \ge \widetilde{r}(a_2) - \bar{h}(r(a_2)) + \bar{h}(r(a_2)) - \bar{h}(r(a_1)) \right] \\
=\ &\mathbb{P}\left[\frac{1}{n_1} \sum_{i=1}^{n_1} \widetilde{r}_i^{(1)} - \bar{h}(r(a_1)) \ge \frac{1}{n_2} \sum_{j=1}^{n_2} \widetilde{r}_j^{(2)} - \bar{h}(r(a_2)) + \bar{h}(r(a_2)) - \bar{h}(r(a_1)) \right] \\
\overset{(i)}{=}\ &\mathbb{P}\left[\frac{1}{n_1} \sum_{i=1}^{n_1} \widetilde{r}_i^{(1)} - \bar{h}(r(a_1)) + \frac{1}{n_2} \sum_{j=1}^{n_2} \widetilde{r}_j^{(2)} - \bar{h}(r(a_2)) \ge \bar{h}(r(a_2)) - \bar{h}(r(a_1)) \right] \tag{34}
\end{aligned}$$

To bound the probability in (34) with δ, we can invoke Hoeffding's inequality (Lemma 1) and solve for n_1 and n_2, which gives

$$\frac{1}{\frac{1}{n_1} + \frac{1}{n_2}} \ge \frac{\sigma^2 \log \frac{1}{\delta}}{2 \left(\bar{h}(r(a_2)) - \bar{h}(r(a_1)) \right)^2}. \tag{35}$$

Recall $n_1 = n_2$. Thus, the minimum number of samples needed for the rating-based approach to identify the optimal policy n_{rate} is:

$$n_{\text{rate}} = \frac{\sigma^2 \log \frac{1}{\delta}}{\left(\bar{h}(r(a_2)) - \bar{h}(r(a_1)) \right)^2}.$$

Finally, we have

$$\frac{n_{\text{rate}}}{n_{\text{pref}}} = \frac{\sigma^2 p (1 - \frac{1}{2p})^2}{2 \left(\bar{h}(r(a_2)) - \bar{h}(r(a_1)) \right)^2},$$

where $p = \frac{1}{1+\exp(-(\bar{h}(r(a_2))-\bar{h}(r(a_1))))}$.

Note $\frac{n_{\text{rate}}}{n_{\text{pref}}}$ is monotonically decreasing for $\bar{h}(r(a_2)) - \bar{h}(r(a_1)) \in [0, 1]$ and $\frac{n_{\text{rate}}}{n_{\text{pref}}}$ approaches $0.25\sigma^2$ as $\bar{h}(r(a_2)) - \bar{h}(r(a_1))$ approaches 0, so we can obtain the advertised result.