Reinforcement Learning from Human Feedback without Reward Inference: Model-Free Algorithm and Instance-Dependent Analysis

Qining Zhang
qiningz@umich.edu
University of Michigan, Ann Arbor

Honghao Wei
honghao.wei@wsu.edu
Washington State University

Lei Ying
leiying@umich.edu
University of Michigan, Ann Arbor

Abstract

In this paper, we study reinforcement learning from human feedback (RLHF) under an episodic Markov decision process with a general trajectory-wise reward model. We developed a model-free RLHF best policy identification algorithm, called BSAD, without explicit reward model inference, which is a critical intermediate step in the contemporary RLHF paradigms for training large language models (LLM). The algorithm identifies the optimal policy directly from human preference information in a backward manner, employing a dueling bandit sub-routine that constantly duels actions to identify the superior one. BSAD adopts a reward-free exploration and best-arm-identification-like adaptive stopping criteria to equalize the visitation among all states in the same decision step while moving to the previous step as soon as the optimal action is identifiable, leading to a provable, instance-dependent sample complexity $\tilde{\mathcal{O}}(c_{\mathcal{M}} SA^3 H^3 M \log \frac{1}{\delta})$[1] which resembles the result in classic RL, where $c_{\mathcal{M}}$ is the instance-dependent constant and M is the batch size. Moreover, BSAD can be transformed into an explore-then-commit algorithm with logarithmic regret and generalized to discounted MDPs using a frame-based approach. Our results show: (i) sample-complexity-wise, RLHF is not significantly harder than classic RL and (ii) end-to-end RLHF may deliver improved performance by avoiding pitfalls in reward inferring such as overfit and distribution shift.

1 Introduction

Reinforcement learning (RL), with a wide range of applications in gaming AIs (Bradley Knox & Stone, 2008; MacGlashan et al., 2017; Warnell et al., 2018), recommendation systems (Yang et al., 2024; Zeng et al., 2016; Kohli et al., 2013), autonomous driving (Wei et al., 2024; Schwarting et al., 2018; Kiran et al., 2022) , and large language model (LLM) training (Wu et al., 2021; Nakano et al., 2021; Ouyang et al., 2022; Ziegler et al., 2019; Stiennon et al., 2020), has achieved tremendous success in the past decade. A typical reinforcement learning problem involves an agent and an environment, where at each step, the agent observes the state, takes a certain action, and then receives a reward signal. The state of the environment then transits to another state, and this process continues. However, most RL advances remain in the simulator environment where the data acquisition process heavily depends on the crafted reward signal, which limits RL from more realistic applications such as LLM, as defining a universal reward is generally difficult. In recent years, using human feedback as reward signals to train and fine-tune LLMs has delivered significant empirical successes for AI

[1] we use $\mathcal{O}(\cdot)$ to hide instance-independent constants and use $\tilde{\mathcal{O}}(\cdot)$ to further hide logarithmic terms except $\log \frac{1}{\delta}$.

Setting	Algorithm	Sample Complexity	Space	Instance	Policy
RL	MOCA	$\mathcal{O}\left(\frac{H^3SA\log\frac{1}{\delta}}{\Delta_{\min}^2 p_{\max}^{\pi}}\right)$	model-based	dependent	Opt
	Q-Learning	$\tilde{\mathcal{O}}\left(\frac{H^4SA\log\frac{1}{\delta}}{\varepsilon^2}\right)$	model-free	independent	ε-Opt
RLHF	P2R-Q	$\tilde{\mathcal{O}}\left(\frac{H^4SA\log\frac{1}{\delta}}{\varepsilon^2}\right)$	model-free	independent	ε-Opt
	PEPS	$\tilde{\mathcal{O}}\left(\frac{H^2S^2A\log\frac{1}{\delta}}{\varepsilon^2}+\frac{S^4H^3\log^3\frac{1}{\delta}}{\varepsilon}\right)$	model-based	independent	ε-Opt
	BSAD(Ours)	$\mathcal{O}\left(\frac{H^3MSA^3\log\frac{1}{\delta}}{(\overline{\Delta}_{\min}^M p_{\max}^{\pi})^2}\right)$	model-free	dependent	Opt

Table 1: Comparison of RL and RLHF algorithms with MOCA (Wagenmaker et al., 2022), Q-Learning (Jin et al., 2018), PEPS (Xu et al., 2020), and P2R (Wang et al., 2023) with Q-learning. S, A, and H are the number of states, actions, and planing steps. δ is confidence level, M is the batch size. $\Delta_{\min}$ is the minimum value function gap, $\overline{\Delta}_{\min}$ characterizes the preference probability gap (Def. 1), and $p_{\max}^{\pi}$ characterizes the maximum state visitation probability (Def. 2).

alignment problems and produced dialog AIs such as the ChatGPT (Ouyang et al., 2022). This paradigm where the reward of the state and actions is inferred from real human preferences, instead of being handcrafted, is referred to as *Reinforcement Learning from Human Feedback* (RLHF). A typical RLHF algorithm on LLMs involves three steps: (i) pre-train a network with supervised learning, (ii) infer a reward model from human feedback, in the form of comparisons or rankings among trajectories (responses), and (iii) use classic RL algorithm to fine-tune the pre-trained model. An accurate reward model that aligns with human preferences is the key to the superiority of RLHF.

Pitfalls of Reward Inference: However, most reward models are trained on a maximum likelihood estimator (MLE) (Christiano et al., 2017; Wang et al., 2023; Saha et al., 2023) under Bradley-Terry model (Bradley & Terry, 1952). This paradigm exhibits pitfalls: (i) the reward models easily overfit the dataset which produces in-distribution errors, and (ii) the reward models fail to measure out-of-distribution state-action pairs during fine-tuning. Even though attempts such as pessimistic estimations (Zhu et al., 2023; Zhan et al., 2023b;a) and regularity conditions are made to improve the accuracy and consistency of reward models, it remains a question of whether reward inference is indeed required. Can we develop a model-free RLHF algorithm without reward inference, which has provable instance-dependent sample complexity?

Contributions: We study an episodic RLHF problem with general trajectory rewards and propose a model-free algorithm called *Batched Sequential Action Dueling* (BSAD) which identifies the optimal action for each state backwardly using action dueling with batched trajectories to obtain human preferences. To equalize the state visitation of the same planning step, we adopt a reward-free exploration strategy and adaptive stopping criteria, which enables learning the exact optimal policy with an instance-dependent sample complexity (Theorem. 1) similar to classic RL with reward (Wagenmaker et al., 2022), as long as the batch size is chosen carefully. Moreover, our results only assume the existence of a uniformly optimal stationary policy and do not require the existence of a Condorcet winner, as we will show the optimal policy is the Condorcet winner when human preferences are obtained with large batch sizes. To the best of our knowledge, BSAD is the first RLHF algorithm with instance-dependent sample complexity, and a transformation of BSAD will provide the first model-free explore-then-commit RLHF algorithm with logarithmic regret.

Comparison to (Xu et al., 2020): From the best of our knowledge, the only algorithm with no reward inference (explicit/implicit) is PEPS (Xu et al., 2020). Our paper is different in (i) BSAD is model-free and takes $\mathcal{O}(SA^2)$ space complexity, while PEPS is model-based and takes $\mathcal{O}(S^2A^2)$ space complexity, (ii) BSAD employs adaptive stopping criteria which leads to an instance-dependent sample complexity with improved dependence in S and δ, while PEPS uses fixed exploration horizon and only has worst-case bounds, (iii) we assume the trajectory reward and require the existence of uniformly deterministic optimal policy which slightly generalizes the classic reward, while PEPS requires the existence of Condorcet winner and stochastic triangle inequality, and (iv) we also generalize to discounted MDPs. The complete comparison of BSAD and related algorithms is summarized in Tab. 1, and a thorough review of related work is deferred to the appendix.

2 Preliminaries

Episodic MDP: An episodic Markov decision process (MDP) is a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, H, P, \mu_0)$, where $\mathcal{S}$ is the state space with $|\mathcal{S}| = S$, $\mathcal{A}$ is the action space with $|\mathcal{A}| = A$, H is the planning horizon, $P = \{P_h\}_{h=1}^H$ is the transition kernels, and μ_0 is the initial distribution. At each episode k, the agent chooses a policy π^k, which is a collection of H functions $\{\pi_h^k : \mathcal{S} \to \mathcal{A}\}_{h=1}^H$, and nature samples an initial state s_1^k from the initial distribution μ_0. Then, at step h, the agent takes an action $a_h^k = \pi_h^k(s_h^k)$ after observing state s_h^k. The environment then moves to a new state s_{h+1}^k sampled from the distribution $P_h(\cdot|s_h^k, a_h^k)$ without revealing any feedback. After each episode, the trajectory of all state-action pairs is collected, which we use τ^k to denote, i.e., $\tau^k = \tau_{1:H}^k = \{(s_h^k, a_h^k)\}_{h=1}^H$.

Trajectory Reward Model: In this paper, we assume the expected reward of each trajectory τ is a general function $f(\tau)$ which maps trajectory to real values, a slight generalization of the cumulative reward structure. Let Ψ be the set of all partial or complete trajectories. Then, we assume there exists a function $f : \Psi \to [0, D]$ which is the expected reward of the MDP $\mathcal{M}$, where D is a positive constant. The reward of a certain trajectory may be random, but humans will evaluate trajectories based on the expected reward. The cumulative reward model is $f(\tau) = \sum_{h=1}^H r(s_h, a_h)$. Under the trajectory reward, we can formulate the Q-function as follows:

$$\begin{aligned} V_h^\pi(s) =& \mathbb{E}^\pi \left[f(\tau_{h:H}) \middle| s_h = s \right] = \mathbb{E}\left[f(\tau_{h:H}) \middle| s_h = s, a_h = \pi(s), \tau_{h+1:H} \sim \pi \right], \\ Q_h^\pi(s,a) =& \mathbb{E}^\pi \left[f(\tau_{h:H}) \middle| s_h = s, a_h = a \right] = \mathbb{E}\left[f(\tau_{h:H}) \middle| s_h = s, a_h = a, \tau_{h+1:H} \sim \pi \right]. \end{aligned}$$

The optimal policy π^* is defined as $\pi^* = \arg\max_\pi \mathbb{E}_{\mu_0}[V_1^\pi(x_1)]$. Without regularity on f, learning the π^* may fundamentally take $\Omega(A^H)$ samples. Therefore, we impose the following assumption:

Assumption 1 *There exists a uniformly optimal deterministic stationary policy π^* for the MDP, i.e., $\pi^* = \arg\max_\pi V_h^\pi(s), \forall(h, s)$.*

Under the assumption, we define the value function gap for sub-optimal actions similar to classic MDPs as $\Delta_h(s,a) = V_h^*(s) - Q_h^*(s,a) = \max_{a'} Q_h^*(s,a') - Q_h^*(s,a)$. Let $\Delta_{\min} = \min_{h,s,a\neq\pi^*(s)} \Delta_h(s,a)$. For simplicity, we assume the optimal action $\pi_h^*(s)$ is unique for each (h, s). Otherwise, we can incorporate $\Delta_{\min}$ into the algorithm so that the duel between the two optimal actions will terminate in a finite time. As a special case, Convex MDPs (Zahavy et al., 2021), e.g., pure exploration (Hazan et al., 2019), apprenticeship learning (Abbeel & Ng, 2004), and adversarial RL (Rosenberg & Mansour, 2019), satisfy Assumption 1 when the optimal policy is deterministic.

Human Feedback: The agent has access to an oracle (a human expert) that evaluates the average quality (reward) of two trajectory batches. At the end of each episode, the agent has the opportunity to choose two sets of (partial) trajectories, denoted by $\mathcal{D}_0$ and $\mathcal{D}_1$ with cardinality M_0 and M_1, to query the human for which has the higher average reward. We slightly abuse the notation τ to let τ_0^i and τ_1^i be the i-th (partial) trace in $\mathcal{D}_0$ and $\mathcal{D}_1$ respectively, i.e., $\mathcal{D}_0 = \{\tau_0^1, \tau_0^2, \cdots, \tau_0^{M_0}\}$, and $\mathcal{D}_1 = \{\tau_1^1, \tau_1^2, \cdots, \tau_1^{M_1}\}$. Each of them may contain only certain steps. After observing the two sets of trajectories, the oracle will give a one-bit feedback $\sigma \in \{0, 1\}$ to the agent to indicate the dataset he/she favors. For simplicity, let $\overline{f}(\mathcal{D}_1)$ and $\overline{f}(\mathcal{D}_0)$ denote the average trajectory reward of $\mathcal{D}_1$ and $\mathcal{D}_0$. Existing works mostly assume the Bradley-Terry model for preference generalization, i.e., the preference probability is a logistic function of the reward difference, i.e.,

$$\mathbb{P}\left(\mathcal{D}_1 \succ \mathcal{D}_0\right) = u\left(\overline{f}(\mathcal{D}_1) - \overline{f}(\mathcal{D}_0)\right) = \frac{1}{1 + \exp\left(\overline{f}(\mathcal{D}_1) - \overline{f}(\mathcal{D}_0)\right)},$$

where $u : \mathbb{R} \to [0, 1]$ is referred to as the link function (Bengs et al., 2021) which characterizes the structure of preference models. Other link functions, such as linear function, probit function, cloglog function, and cauchit function, have also been well-studied in dueling bandits (Ailon et al., 2014) and generalized linear models (Razzaghi, 2013; McCulloch, 2000), but not RLHF. In this paper, we use a 0-1 link function that indicates the favored set with higher reward, i.e.,

$$\sigma = \mathsf{HumanFeedback}(\mathcal{D}_0, \mathcal{D}_1) = \underset{i\in\{0,1\}}{\arg\max}\, \overline{f}(\mathcal{D}_i) = \underset{i\in\{0,1\}}{\arg\max}\, \frac{1}{M_i}\sum_{m=1}^{M_i} f(\tau_i^m).$$

Generalization to other link functions can be achieved through revising the probability gap definition below (Def. 1). Furthermore, we show in Fig. 1 that single trajectory preference may not align with the expected reward and thus batched comparison is necessary, and it may be easier for humans to identify a better response if the trajectory batches resemble each other with the same initial state, which motivates the comparison between partial and batched trajectories. Typically, in an LLM training setting, for each candidate policy, the human evaluator will look at multiple responses generated respectively and then assess which policy has a better average quality. Similarly for UAV training, humans will watch multiple UAV trajectories for each policy and declare which policy is better based on the average quality of the movement, i.e., success rate, stability, etc. When the batch sizes are not unbearably large, batched preference assessment of trajectories should not be essentially harder than single trajectory preference assessment.

Problem Formulation: Our goal is to design a learning algorithm to interact with the MDP and learn the optimal policy π^* from the human feedback as quickly as possible. A learning algorithm Alg consists of (i) a sampling rule which decides which policy to choose at each episode and whether to query the human agent, (ii) a stopping rule which decides a stopping time when the learner wishes to output an learned policy, and (iii) a decision rule which decides which policy $\hat{\pi}$ to output. We call an algorithm δ-PAC if it outputs an optimal policy with probability at least $1-\delta$. Our goal is to design such an algorithm to minimize sample complexity K:

$$\min \mathbb{E}[K], \text{ such that } \mathbb{P}(\hat{\pi} = \pi^*) \geq 1 - \delta.$$

3 Main Results for Episodic MDPs

In this paper, we focus on the instance-dependent performance. To characterize the structure of the MDPs under human feedback, we introduce the notion of probability gaps in Def. 1 for each state and sub-optimal action, which is a generalization of the calibrated pairwise preference probability considered in the dueling bandits literature (Yue et al., 2012; Yue & Joachims, 2011). We also define the state visitation probability $p_h^\pi(s)$ of a given policy π in Def. 2.

Definition 1 (Probability Gap) *Given (h,s) and a sub-optimal action a, the probability gap $\overline{\Delta}_h^M(s,a)$ for comparison of two trajectory sets with cardinality both being M is defined as:*

$$\overline{\Delta}_h^M(s,a) = \underbrace{\mathbb{P}\left(\sum_{m=1}^{M} f(\tau_0^m) > \sum_{m=1}^{M} f(\tau_1^m) \,\middle|\, \tau_0^m \sim \pi^*,\ \tau_1^m \sim \{a_h = a, \pi^*\}\right)}_{\overline{p}_h^M(s,a)} - \frac{1}{2},$$

where the traces $\{\tau_0^1, \cdots, \tau_0^M\}$ are independently sampled starting from state (h,s) using the optimal policy $\{\pi_k^*\}_{k=h}^H$, while $\{\tau_1^1, \cdots, \tau_1^M\}$ are independently sampled starting from state (h,s) using immediate action $a_h = a$ and the optimal policy $\{\pi_k^*\}_{k=h+1}^H$ afterwards. Let $\overline{\Delta}_{\min}^M = \min_{h,s,a} \overline{\Delta}_h^M(s,a)$.

Definition 2 (State Visitation Probability) *Given $(h,s) \in [H] \times \mathcal{S}$, the visitation probability (occupancy measure) of policy π is defined as follows:*

$$p_h^\pi(s) = \mathbb{P}\left(s_h = s | s_0 \sim \mu_0,\ a_{h'} \sim \pi(s_{h'}),\ \forall h' < h\right).$$

Let $p_{\max}^\pi = \min_{h,s} \max_\pi p_h^\pi(s)$, and we assume it is positive. We will use both the probability gap and the state visitation probability to characterize our instance-dependent performance.

Algorithm 1: BASD for Episodic MDPs

initialize for all (h, s, a), $J_h(s,a) \leftarrow 1$, $L_h(s,a) \leftarrow 0$, $M_h(s) \leftarrow 0$, and $l \leftarrow H$, $k \leftarrow 0$;
initialize for all (h, s, a, a'), $w_h(s,a,a') \leftarrow 0$, $N_h(s,a,a') \leftarrow 0$, $\hat{\pi}_h(s) = \mathcal{D}_h^0(s) = \mathcal{D}_h^1(s) = \emptyset$;
define $\iota \equiv c\log(\frac{SAHk}{\delta})$, $\beta_t = \sqrt{\frac{H\iota}{\max\{t,1\}}}$, and $\alpha_t = \frac{H+1}{H+t}$;
$\hat{\sigma}_h(s,a,a') \equiv \frac{w_h(s,a,a')}{N_h(s,a,a')}$ or $\frac{1}{2}$ if $N_h(s,a,a') = 0$, $b_h(s,a,a') \equiv \sqrt{\frac{\iota}{\max\{N_h(s,a,a'),1\}}}$, $\forall(h,s,a,a')$;
while $l \geq 1$ **do**
  receive s_1, $k = k+1$;
  for *step* $h = 1 : l-1$ **do** `// reward-free exploration`
    take action $a_h \leftarrow \arg\max_a J_h(s_h, a)$ and observe s_{h+1}, $L_h(s_h,a_h) \leftarrow L_h(s_h,a_h) + 1$;
    $W_{h+1}(s_{h+1}) \leftarrow \min\{1, \max_a J_{h+1}(s_{h+1}, a)\}$;
    $J_h(s_h,a_h) \leftarrow (1-\alpha_t)J_h(s_h,a_h) + \alpha_t[W_{h+1}(s_{h+1}) + 2\beta_t]$ where $t = L_h(s_h,a_h)$;
  $M_l(s_l) \leftarrow M_l(s_l) + 1$. $W_l(s_l) \leftarrow \min\{1, b_{M_l(s_l)}\}$;
  call action dueling sub-routine B-RUCB$(l, s_l, M_l(s_l))$; `// action dueling`
  if $\forall s$, $\exists a$, *such that* $\forall a'$, $\hat{\sigma}_l(s,a,a') - b_l(s,a,a') \geq 0.5$ **then**
    $\forall s$, $\hat{\pi}_l(s) \in \{a | \forall a', \hat{\sigma}_l(s,a,a') - b_l(s,a,a') \geq 0.5\}$;
    $l \leftarrow l-1$. $J_h(s,a) \leftarrow 1$, $L_h(s,a) \leftarrow 0, \forall(h,s,a)$, $k \leftarrow 0$; `// backward search`
return $\hat{\pi}$

Algorithm 2: B-RUCB: a batched dueling bandits sub-routine

Input: step h, state s, candidate policy $\hat{\pi}$, past visits $M_h(s)$.
if $M_h(s) \pmod{2M} \leq M$ **then**
  if $M_h(s) \equiv 1 \pmod{M}$ **then** `// select relative optimal arm`
    $\mathcal{C}_h(s) = \{a | \forall a' : \hat{\sigma}_h(s,a,a') + b_h(s,a,a') \geq 0.5\}$, sample $\hat{a}_s$ uniformly from $\mathcal{C}_h(s)$;
    $\mathcal{D}_h^0(s) \leftarrow \emptyset$, $\mathcal{D}_h^1(s) \leftarrow \emptyset$;
  take action $a_h \leftarrow \hat{a}_s$ and observe s_{h+1}, and use policy $\hat{\pi}$ for steps afterwards;
  $\mathcal{D}_h^0(s) = \mathcal{D}_h^0(s) \cup \{(s_h,a_h), \cdots, (s_H,a_H)\}$;
else
  if $M_h(s) \equiv 1 \pmod{M}$ **then** `// select combating arm based on UCB`
    $\tilde{a}_s = \arg\max_{a \neq \hat{a}_s}\{\hat{\sigma}_h(s,a,\hat{a}_s) + b_h(s,a,\hat{a}_s)\}$;
  take action $a_h \leftarrow \tilde{a}_s$ and observe s_{h+1}, and use policy $\hat{\pi}$ for steps afterwards ;
  $\mathcal{D}_h^1(s) = \mathcal{D}_h^1(s) \cup \{(s_h,a_h), \cdots, (s_H,a_H)\}$;
if $M_h(s) \equiv 0 \pmod{2M}$ **then** `// query human every 2M episodes`
  query feedback $\sigma = \mathsf{HumanFeedback}\left(\mathcal{D}_h^0(s), \mathcal{D}_h^1(s)\right)$;
  $w_h(s,\tilde{a}_s,\hat{a}_s) \leftarrow w_h(s,\tilde{a}_s,\hat{a}_s) + \sigma$, $w_h(s,\hat{a}_s,\tilde{a}_s) \leftarrow w_h(s,\hat{a}_s,\tilde{a}_s) + 1 - \sigma$;
  $N_h(s,\tilde{a}_s,\hat{a}_s) = N_h(s,\tilde{a}_s,\hat{a}_s) + 1$;
return

3.1 Algorithm for Episodic RLHF

In this section, we propose an algorithm called BASD (Alg. 1) to solve the RLHF for episodic MDPs. The algorithm can be divided into two major modules: (i) an action dueling sub-routine generalizing the RUCB algorithm from the dueling bandits (Zoghi et al., 2014), and (ii) a reward-free exploration strategy to equalize the visitation probability of each state to minimize the overall sample complexity.

Backward Action Dueling: BSAD identifies the optimal policy for each state using a backward search. The backbone is to employ a batched version of the RUCB algorithm (Zoghi et al., 2014), called B-RUCB in Alg. 2, which is called in step l and controls the action selection policy from step l to H. Namely, it chooses the action a_l at step l using the RUCB dueling bandits principle and then uses the candidate optimal policy $\hat{\pi}$ for steps afterward. If the policy $\hat{\pi}$ is indeed the optimal

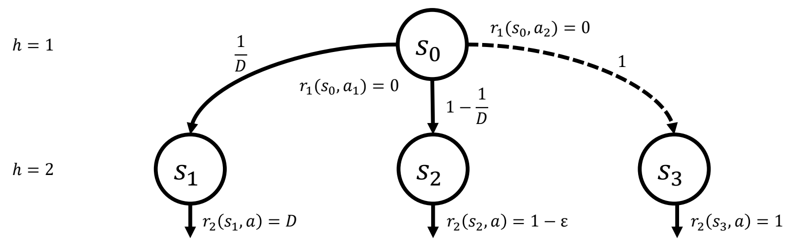


Figure 1: MDP where $\pi_h^*(s)$ is not the Condorcet winner: there are 3 states ($\{s_1, s_2, s_3\}$) at step 2 with 1 action, and 1 state s_0 in step 1 with 2 actions. With action a_1, the state transits w.p. $1/D$ to state s_1 with reward D, and w.p. $1 - 1/D$ to state s_2 which gives reward $1 - \varepsilon$, where $0 < \epsilon < 1$. With action 2, the state transits deterministically to state s_3 with reward 1.

policy π^*, the average reward from step l to H constitutes an unbiased estimator of $Q_h^*(s_l, a_l)$, which resembles dueling bandits. Different from classic RUCB, we query human feedback every $2M$ episode with batches and we will show later that it allows the optimal action $\pi_h^*(s)$ to be the action favored by the human oracle (Condorcet winner). Moreover, we adopt a stopping rule for each (h, s) that if there exists one action a whose lower confidence bound of the preference probability estimation $\hat{\sigma}_h(s, a, a')$ is larger than half for all other actions, the optimal action is found. Specifically, we use $\mathcal{T}_h(s)$ to denote the stopping rule for state (h, s), i.e., $\mathcal{T}_h(s) = \{\exists a, \ \forall a', \ \hat{\sigma}_l(s, a, a') - b_l(s, a, a') \geq 0.5\}$. Then, the criteria for l to move from h to $h-1$ is equivalent to $\cap_{s=1}^{S} \mathcal{T}_h(s)$. Running B-RUCB with the stopping rule identifies the optimal action $\pi_l^*(s)$ for all states at step l with high probability.

Reward-free Exploration: To minimize the sample complexity, it is ideal that every state has a similar visitation probability so that action identification can be performed simultaneously for all the states. Our chosen model-free reward-free exploration between step 1 to step $l-1$ contributes towards this goal. We slightly adapted the UCBZero algorithm originally proposed in (Zhang et al., 2020) in our BSAD algorithm so that the overall algorithm is model-free. This strategic exploration policy will guarantee that we visit each state on step l proportional to the maximum visitation probability over all possible policy π starting from the initial distribution.

3.2 Theoretical Results

It is well-known from dueling bandits literature (Zoghi et al., 2014) that the RUCB algorithm only requires the existence of the Condorcet winner to identify the optimal action, where the Condorcet winner refers to an action that is preferred with probability larger than half when compared to any other action. Similar to the definition in dueling bandits, for any state (h, s) and any size M, we say the optimal action $\pi_h^*(s)$ is the Condorcet winner if the preference probability $\overline{p}_h^M(s, a)$ is larger than half for all other actions a. For any comparison-based algorithm to identify the optimal policy, the optimal policy must be the Condorcet winner. We will first characterize the existence of the Condorcet winner when human experts are queried with batch size M large enough.

Lemma 1 *Given an MDP $\mathcal{M}$ and for any (h, s), the action $\pi_h^*(s)$ associated with the optimal policy π^* is the Condorcet winner in the* HumanFeedback *comparison as long as $M = \Omega(D^2 \Delta_{\min}^{-2})$.*

Existence of Condorcet Winner: In general, the optimal action $\pi_h^*(s)$, although it maximizes the expected reward, is not necessarily the Condorcet winner with arbitrary M. To see this, consider a two-step MDP with traditional cumulative reward as shown in Fig. 1. For state s_0 and $D > 2$ in step 1, the optimal action is a_1 which gives expected reward $1 + (1 - D^{-1})(1 - \varepsilon)$ larger than 1 given by action a_2. However, if we choose $M = 1$ and query human feedback of the duel between action a_1 and a_2, the human expert will only prefer action a_1 if the state transits to s_1, which only occurs with probability $1/D$ and could be much less than half. Therefore, the optimal action a_1 for state s_0 is not the Condorcet winner. Similarly, it is also not hard to construct counter-examples with more than three actions to show that the Condorcet winner does not exist. However, Lemma. 1 shows that the optimal action $\pi_h^*(s)$ is indeed the Condorcet winner at every state (h, s) as long as the batch size M is large enough. The bound is proportional to D^2 which characterizes the variance of reward for a single trajectory and inversely proportional to the square of the minimum value function gap $\Delta_{\min}$, which characterizes the distinguishability among actions. The proof of Lemma. 1 is deferred to the appendix, where we apply concentration inequalities to lower bound the preference probability $\overline{p}_h^M(s, a)$. Next, we characterize the sample complexity of BSAD.

Theorem 1 *Given an MDP $\mathcal{M}$, fix δ and suppose M is chosen large enough such that the optimal policy π^* is the Condorcet winner for all states (h, s). Then with probability at least $1 - \mathcal{O}(\delta)$, the BSAD algorithm terminates within K episodes and returns the optimal policy $\hat{\pi} = \pi^*$ with:*

$$K = \tilde{\mathcal{O}}\left(\sum_{h=1}^{H} \frac{SA^3h^2M\log(\frac{1}{\delta})}{\min_{s,a}\max_{\pi}[\overline{\Delta}_h^M(s,a)p_h^{\pi}(s)]^2}\right).$$

Proof Roadmap: Our main Theorem. 1 conveys two messages: (i) BSAD is δ-PAC, and (ii) BSAD has provable instance-dependent sample complexity bound under general reward model. The proof of Theorem. 1 is deferred to appendix. To obtain the correctness guarantee, we decompose the probability of making a mistake into the sum of probabilities where the mistake is made on a certain step h. Then, using a backward induction argument, we show that the total mistake probability is small. To obtain the sample complexity bound, we fix (h, s) and then bound the number of comparisons between two actions. Next, we bound the total number of comparisons and the total number of episodes needed to identify the optimal action for this (h, s). This can be achieved by summing up the number of comparisons between all pairs of arms before the stopping criteria $\mathcal{T}_h(s)$ for that state is satisfied. Lemma. 2 characterizes the sample complexity for any state (h, s):

Lemma 2 *Given an MDP $\mathcal{M}$, fix δ and suppose M is large enough. For fixed (h, s), the number of episodes with $l = h$ and $s_h = s$ until the criteria $\mathcal{T}_h(s)$ is bounded with high probability by:*

$$M_h(s) = \tilde{\mathcal{O}}\left(\sum_{i=2}^{A} \frac{i}{\overline{\Delta}_h^M(s,a_i)^2} M\log\left(\frac{1}{\delta}\right)\right) = \tilde{\mathcal{O}}\left(\frac{A^2M\log\left(\frac{1}{\delta}\right)}{\min_a \overline{\Delta}_h^M(s,a)^2}\right),$$

where $\{a_1, a_2, \cdots, a_A\}$ is a permutation of the action set $\mathcal{A}$ such that a_1 is the optimal action and $\overline{\Delta}_h^M(s,a_2) \le \overline{\Delta}_h^M(s,a_2) \le \cdots, \overline{\Delta}_h^M(s,a_A)$.

Notice that our bound in Lemma. 2 is different from the original RUCB algorithm provided in (Zoghi et al., 2014, Theorem 4) due to (i) we study a PAC setting while the vanilla RUCB focuses on regret minimization and (ii) we chose a larger confidence bonus so that our bound only have logarithmic dependence on δ. After bounding the sample complexity to identify the optimal action for each state, we need to relate $M_h(s)$ to the total number of episodes through reward-free exploration. We show in Lemma. 3 that the number of episodes spent for a step $l = h$ is bounded by the number of visitations $M_h(s)$, which is analog to (Zhang et al., 2020, Theorem 3).

Lemma 3 *Given an MDP $\mathcal{M}$, fix δ and suppose M is large enough. For a fixed (h, s), suppose we have $l = h$ and $k = K_h$ in the current episode, we have:*

$$\forall s,\ K_h \le \mathcal{O}\left(\frac{SAh^2M_h(s)}{\max_{\pi} p_h^{\pi}(s)^2}\right).$$

Combining both Lemma. 2 and Lemma. 3, we will be able to prove Theorem. 1:

$$K = \sum_{h=1}^{H} \max_s \mathcal{O}\left(\frac{SAh^2M_h(s)}{\max_{\pi} p_h^{\pi}(s)^2}\right) = \tilde{\mathcal{O}}\left(\sum_{h=1}^{H} \frac{SA^3h^2M\log(\frac{1}{\delta})}{\min_{s,a}\max_{\pi}[\overline{\Delta}_h^M(s,a)p_h^{\pi}(s)]^2}\right).$$

RLHF Algorithm with Logarithm Regret: It is very simple to adapt the BSAD algorithm to an explore-then-commit type algorithm for regret minimization by choosing $\delta = T^{-1}$. Then, the sample complexity bound will convert into a regret bound in the order of $\mathcal{O}(\log T)$. To the best of our knowledge, this is the first RLHF algorithm with logarithmic regret performance.

Instance Dependence and Connection to Classical RL: Our sample complexity bound in Theorem. 1 has a linear dependence on the number of states S, a polynomial on the number of actions A and the planning horizon H, and a logarithmic dependence on the inverse of confidence δ. Moreover, it characterizes how the sample complexity depends on fine-grained structures of the MDP $\mathcal{M}$ itself. It is also inversely proportional to the square of the probability gap $\overline{\Delta}_h^M(s,a)$ which resembles the sample complexity or regret bounds in the dueling bandit literature, and also resembles the dependence of the value function gap $\Delta_h(s,a)$ in the sample complexity bounds for traditional tabular RL, e.g., (Wagenmaker et al., 2022, Theorem 2). Moreover, the inverse proportional dependence of the maximum state visitation probability over all policies also resembles the traditional RL. In fact, with M chosen in the same order as in Lemma. 1 and using concentration inequalities, the sample complexity bound can be converted depending on the value function gap as follows:

$$K = \tilde{\mathcal{O}}\left(\frac{SA^3H^3D^2\log(\frac{1}{\delta})}{\min_{h,s,a}\Delta_h(s,a)^2\max_\pi p_h^\pi(s)^2}\right).$$

This shows that RLHF is almost no harder than classic RL given the appropriate parameter, except for a polynomial factor on the number of actions A and the planning horizon H. This finding coincides with (Wang et al., 2023) and sheds light on the similarity between RLHF and classic RL. Notice that our result is derived from a general reward model where the Bellman equations do not hold. Therefore, our result also seemingly implies that the fundamental backbone of RL is the existence of uniformly optimal stationary policy instead of the Bellman equations.

4 Generalization to Discounted MDPs

In this section, we generalize the BSAD algorithm to discounted MDPs with the traditional state-action reward function $r(s,a) \in [0,1]$ and discount factor γ. Our approach is to segment the time horizon into frames with length $H = \Theta(\frac{1}{1-\gamma}\log\frac{1}{\varepsilon(1-\gamma)^2})$. Then, we run BSAD (Alg. 1) with horizon H on the discounted MDP, as if it is episodic. This frame-based adaptation delivers provable instance-dependent sample complexity shown in Theorem. 2. Discussions are deferred to the appendix.

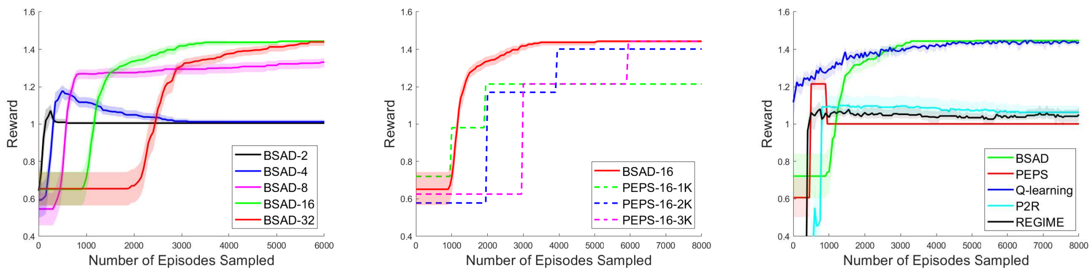
Theorem 2 *suppose M is chosen large enough. Then with probability $1-\mathcal{O}(\delta)$,* BSAD *terminates within K episodes and returns an ε-optimal policy with:*

$$K = \tilde{\mathcal{O}}\left(\frac{SA^3M\log(\frac{1}{\delta})\log^3(\frac{1}{\varepsilon})}{(1-\gamma)^3\min_{h,s,a}\overline{\Delta}_h^M(s,a)^2\max_\pi\min_{s'}p_h^\pi(s|s')^2}\right),$$

where $\overline{\Delta}_h^M(s,a)$ to be the probability gap for action a and trajectories of length $H-h$ compared to the Condorcet winner of that state s, and $p_h^\pi(s|s')$ is the visitation probability of s after h steps starting from state s' with policy π. Both definitions are analog to the definitions in episodic MDPs.

5 Numerical Results

In this section, we study the empirical performance of BSAD on an MDP based on Fig. 1 with $D = 10$. The only difference is we replicate two copies of s_0 in the first step with different initial distributions. For these states, the optimal policy is not the Condorcet winner under a single trajectory comparison but will become the Condorcet winner when the batch size increases. We compare BSAD to existing value-based model-free RLHF algorithms, with and without reward inference, where the performance is measured by the value function $\mathbb{E}_{\mu_0}[V_1^{\hat{\pi}}(s)]$ of the candidate policy evaluated on the true MDP. The baselines that we chose are (i) a model-free and batched adaptation of PEPS (Xu et al., 2020) (no



(a) batch size (b) adaptive stopping (c) performance

Figure 2: numerical experiment on a three-state two-step MDP: (a) shows the proposed BSAD algorithm with different batch sizes. (b) compares BSAD with adaptive stopping to batched version of PEPS with fixed exploration horizon. (c) compares BSAD to model-free RLHF and RL algorithms. Results are averaged over 100 trajectories and shaded areas represent bootstrap confidence intervals.

reward inference) which uses UCBZero (Zhang et al., 2020), (ii) Q-learning with P2R (Wang et al., 2023) (reward inference) where the candidate policy $\hat{\pi}$ is the greedy policy, and (iii) REGIME (Zhan et al., 2023b) (reward inference) with UCBZero and pessimistic Q-learning (Shi et al., 2022) as offline RL oracle, where each point is obtained through a $1k$-episode offline RL algorithm. We also compare to classic RL algorithms, i.e., Q-learning (Jin et al., 2018).

Algorithm	BSAD(ours)	PEPS	Q-learning	P2R	REGIME
Running Time (ms)	171.21	179.23	1090.12	5898.30	4613.73

Table 2: running time comparisons on 1 CPU averaged over 50 trajectories.

Fig. 2a shows the effect of batch size. When using a small batch size, i.e., $M = 2, 4$, the Condorcet winner at $h = 1$ is not optimal, and BSAD converges to a sub-optimal policy. When M is large, BSAD identifies the optimal policy, and the sample complexity displays a decrease-then-increase trend, which coincides with Theorem. 1. Specifically, when M increases, the probability gap in the denominator increases sharply, leading to reduced sample complexity, and as M continues to increase, M in the numerator starts to dominate. This justifies BSAD is adaptive to MDP instances. Fig. 2b shows the comparison of BSAD to a batched version of PEPS with different exploration horizons. The observation that the curve of BSAD lies uniformly above all PEPS curves shows the necessity of adaptive algorithm design. Specifically, our design of adaptive stopping criteria identifies the optimal policy earlier and adapts to the different distinguishability in different states, which results in improved regret performance. In Fig. 2c, we compare BSAD to Q-learning and RLHF algorithms with reward inference. First, we observe that BSAD has almost the same performance as Q-learning which uses the reward information, which shows RLHF is almost no harder than classic RL. However, our algorithm applies to the general trajectory reward function while Q-learning cannot be used anymore. BSAD exhibits superior performance than other RLHF algorithms also in running time as shown in Table. 2, because training reward models with MLE is difficult and takes much larger sample and computational complexity, let alone the best policy can only be obtained when the reward model is accurate enough. This observation somewhat justifies the reward model is unnecessary given it suffers from pitfalls like over-fitting and distribution shift.

6 Conclusion

We studied RLHF under both episodic MDPs with trajectory reward structure, a generalization of the classic cumulative reward. We propose an algorithm called BSAD which enjoys a provable instance-dependent sample complexity that resembles the result in classic RL with reward. We also generalize our results to discounted MDPs. Our results show RLHF is almost no harder than classic RL, and the current dominating reward model training module in RLHF may be unnecessary.

Acknowledgments

The work of Qining Zhang and Lei Ying is supported in part by NSF under grants 2112471, 2134081, 2207548, 2240981, and 2331780.

References

Pieter Abbeel and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning*, ICML '04, pp. 1, New York, NY, USA, 2004. Association for Computing Machinery. ISBN 1581138385. doi: 10.1145/1015330.1015430. URL https://doi.org/10.1145/1015330.1015430.

Nir Ailon, Zohar Karnin, and Thorsten Joachims. Reducing dueling bandits to cardinal bandits. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 856–864, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/ailon14.html.

Robert E. Bechhofer. A sequential multiple-decision procedure for selecting the best one of several normal populations with a common unknown variance, and its use with various experimental designs. *Biometrics*, 14(3):408–429, 1958. ISSN 0006341X, 15410420. URL http://www.jstor.org/stable/2527883.

Viktor Bengs, Róbert Busa-Fekete, Adil El Mesaoudi-Paul, and Eyke Hüllermeier. Preference-based online learning with dueling bandits: A survey. *Journal of Machine Learning Research*, 22(7): 1–108, 2021. URL http://jmlr.org/papers/v22/18-546.html.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444. URL http://www.jstor.org/stable/2334029.

W. Bradley Knox and Peter Stone. Tamer: Training an agent manually via evaluative reinforcement. In *2008 7th IEEE International Conference on Development and Learning*, pp. 292–297, 2008. doi: 10.1109/DEVLRN.2008.4640845.

Niladri Chatterji, Aldo Pacchiano, Peter Bartlett, and Michael Jordan. On the theory of reinforcement learning with once-per-episode feedback. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 3401–3412. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/1bf2efbbe0c49b9f567c2e40f645279a-Paper.pdf.

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 Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 3773–3793. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/chen22ag.html.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/309fee4e541e51de2e41f21bebb342aa-Paper.pdf.

Christoph Dann, Tor Lattimore, and Emma Brunskill. Unifying pac and regret: Uniform pac bounds for episodic reinforcement learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/17d8da815fa21c57af9829fb0a869602-Paper.pdf.

Christoph Dann, Teodor Vanislavov Marinov, Mehryar Mohri, and Julian Zimmert. Beyond value-function gaps: Improved instance-dependent regret bounds for episodic reinforcement learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 1–12. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/000c076c390a4c357313fca29e390ece-Paper.pdf.

Rémy Degenne, Thomas Nedelec, Clement Calauzenes, and Vianney Perchet. Bridging the gap between regret minimization and best arm identification, with application to a/b tests. In Kamalika Chaudhuri and Masashi Sugiyama (eds.), *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 1988–1996. PMLR, 16–18 Apr 2019. URL https://proceedings.mlr.press/v89/degenne19a.html.

Yihan Du, Anna Winnicki, Gal Dalal, Shie Mannor, and R Srikant. Exploration-driven policy optimization in rlhf: Theoretical insights on efficient data utilization. *arXiv preprint arXiv:2402.10342*, 2024.

Miroslav Dudík, Katja Hofmann, Robert E. Schapire, Aleksandrs Slivkins, and Masrour Zoghi. Contextual dueling bandits. In Peter Grünwald, Elad Hazan, and Satyen Kale (eds.), *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pp. 563–587, Paris, France, 03–06 Jul 2015. PMLR. URL https://proceedings.mlr.press/v40/Dudik15.html.

Yonathan Efroni, Nadav Merlis, and Shie Mannor. Reinforcement learning with trajectory feedback. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8):7288–7295, May 2021. doi: 10.1609/aaai.v35i8.16895. URL https://ojs.aaai.org/index.php/AAAI/article/view/16895.

Dylan J Foster, Alexander Rakhlin, David Simchi-Levi, and Yunzong Xu. Instance-dependent complexity of contextual bandits and reinforcement learning: A disagreement-based perspective. *arXiv preprint arXiv:2010.03104*, 2020.

Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/8b0d268963dd0cfb808aac48a549829f-Paper.pdf.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In Vitaly Feldman, Alexander Rakhlin, and Ohad Shamir (eds.), *29th Annual Conference on Learning Theory*, volume 49 of *Proceedings of Machine Learning Research*, pp. 998–1027, Columbia University, New York, New York, USA, 23–26 Jun 2016. PMLR. URL https://proceedings.mlr.press/v49/garivier16a.html.

Aurélien Garivier and Emilie Kaufmann. Nonasymptotic sequential tests for overlapping hypotheses applied to near-optimal arm identification in bandit models. *Sequential Analysis*, 40(1):61–96, 2021. doi: 10.1080/07474946.2021.1847965. URL https://doi.org/10.1080/07474946.2021.1847965.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 4447–4455. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

Elad Hazan, Sham Kakade, Karan Singh, and Abby Van Soest. Provably efficient maximum entropy exploration. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2681–2691. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/hazan19a.html.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/d3b1fb02964aa64e257f9f26a31f72cf-Paper.pdf.

Marc Jourdan, Degenne Rémy, and Kaufmann Emilie. Dealing with unknown variances in best-arm identification. In Shipra Agrawal and Francesco Orabona (eds.), *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pp. 776–849. PMLR, 20 Feb–23 Feb 2023. URL https://proceedings.mlr.press/v201/jourdan23a.html.

Shivaram Kalyanakrishnan, Ambuj Tewari, Peter Auer, and Peter Stone. Pac subset selection in stochastic multi-armed bandits. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, pp. 227–234, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851. URL https://icml.cc/2012/papers/359.pdf.

Emilie Kaufmann and Shivaram Kalyanakrishnan. Information complexity in bandit subset selection. In Shai Shalev-Shwartz and Ingo Steinwart (eds.), *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pp. 228–251, Princeton, NJ, USA, 12–14 Jun 2013. PMLR. URL https://proceedings.mlr.press/v30/Kaufmann13.html.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *J. Mach. Learn. Res.*, 17(1):1–42, jan 2016. ISSN 1532-4435. URL https://jmlr.org/papers/volume17/kaufman16a/kaufman16a.pdf.

Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 2023.

Chinmaya Kausik, Mirco Mutti, Aldo Pacchiano, and Ambuj Tewari. A framework for partially observed reward-states in rlhf. *arXiv preprint arXiv:2402.03282*, 2024.

B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A. Al Sallab, Senthil Yogamani, and Patrick Pérez. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):4909–4926, 2022. doi: 10.1109/TITS.2021.3054625.

Pushmeet Kohli, Mahyar Salek, and Greg Stoddard. A fast bandit algorithm for recommendation to users with heterogenous tastes. *Proceedings of the AAAI Conference on Artificial Intelligence*, 27(1):1135–1141, Jun. 2013. doi: 10.1609/aaai.v27i1.8463. URL https://ojs.aaai.org/index.php/AAAI/article/view/8463.

Dingwen Kong and Lin Yang. Provably feedback-efficient reinforcement learning via active reward learning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 11063–11078. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/476c289f685e27936aa089e9d53a4213-Paper-Conference.pdf.

James MacGlashan, Mark K. Ho, Robert Loftin, Bei Peng, Guan Wang, David L. Roberts, Matthew E. Taylor, and Michael L. Littman. Interactive learning from policy-dependent human feedback. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2285–2294. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/macglashan17a.html.

Charles E. McCulloch. Generalized linear models. *Journal of the American Statistical Association*, 95(452):1320–1324, 2000. ISSN 01621459. URL http://www.jstor.org/stable/2669780.

Mirco Mutti, Riccardo De Santi, Piersilvio De Bartolomeis, and Marcello Restelli. Convex reinforcement learning in finite trials. *Journal of Machine Learning Research*, 24(250):1–42, 2023. URL http://jmlr.org/papers/v24/22-1514.html.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

Ellen Novoseller, Yibing Wei, Yanan Sui, Yisong Yue, and Joel Burdick. Dueling posterior sampling for preference-based reinforcement learning. In Jonas Peters and David Sontag (eds.), *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pp. 1029–1038. PMLR, 03–06 Aug 2020. URL https://proceedings.mlr.press/v124/novoseller20a.html.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Manish Prajapat, Mojmír Mutnỳ, Melanie N Zeilinger, and Andreas Krause. Submodular reinforcement learning. *arXiv preprint arXiv:2307.13372*, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 53728–53741. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

Mehdi Razzaghi. The probit link function in generalized linear models for data mining applications. *Journal of Modern Applied Statistical Methods*, 12:164–169, 2013. URL https://jmasm.com/index.php/jmasm/article/view/650/651.

Aviv Rosenberg and Yishay Mansour. Online convex optimization in adversarial Markov decision processes. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th*

International Conference on Machine Learning, volume 97 of *Proceedings of Machine Learning Research*, pp. 5478–5486. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/rosenberg19a.html.

Aadirupa Saha and Akshay Krishnamurthy. Efficient and optimal algorithms for contextual dueling bandits under realizability. In Sanjoy Dasgupta and Nika Haghtalab (eds.), *Proceedings of The 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pp. 968–994. PMLR, 29 Mar–01 Apr 2022. URL https://proceedings.mlr.press/v167/saha22a.html.

Aadirupa Saha, Aldo Pacchiano, and Jonathan Lee. Dueling rl: Reinforcement learning with trajectory preferences. In Francisco Ruiz, Jennifer Dy, and Jan-Willem van de Meent (eds.), *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 6263–6289. PMLR, 25–27 Apr 2023. URL https://proceedings.mlr.press/v206/saha23a.html.

Wilko Schwarting, Javier Alonso-Mora, and Daniela Rus. Planning and decision-making for autonomous vehicles. *Annual Review of Control, Robotics, and Autonomous Systems*, 1(Volume 1, 2018):187–210, 2018. ISSN 2573-5144. doi: https://doi.org/10.1146/annurev-control-060117-105157. URL https://www.annualreviews.org/content/journals/10.1146/annurev-control-060117-105157.

Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 19967–20025. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/shi22c.html.

Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/10a5ab2db37feedfdeaab192ead4ac0e-Paper.pdf.

Satinder P Singh and Richard C Yee. An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, 16:227–233, 1994. URL https://doi.org/10.1007/BF00993308.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 3008–3021. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

Andrea Tirinzoni, Aymen Al Marjani, and Emilie Kaufmann. Near instance-optimal pac reinforcement learning for deterministic mdps. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 8785–8798. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/39c60dda48ebf0a2e5dda52ce08eb5c8-Paper-Conference.pdf.

Andrea Tirinzoni, Aymen Al-Marjani, and Emilie Kaufmann. Optimistic pac reinforcement learning: the instance-dependent view. In Shipra Agrawal and Francesco Orabona (eds.), *Proceedings of The 34th International Conference on Algorithmic Learning Theory*, volume 201 of *Proceedings of Machine Learning Research*, pp. 1460–1480. PMLR, 20 Feb–23 Feb 2023. URL https://proceedings.mlr.press/v201/tirinzoni23a.html.

Andrew J Wagenmaker, Max Simchowitz, and Kevin Jamieson. Beyond no regret: Instance-dependent pac reinforcement learning. In Po-Ling Loh and Maxim Raginsky (eds.), *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 358–418. PMLR, 02–05 Jul 2022. URL https://proceedings.mlr.press/v178/wagenmaker22a.html.

Yuanhao Wang, Qinghua Liu, and Chi Jin. Is RLHF more difficult than standard RL? a theoretical perspective. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=sxZLrBqg50.

Garrett Warnell, Nicholas Waytowich, Vernon Lawhern, and Peter Stone. Deep tamer: Interactive agent shaping in high-dimensional state spaces. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr. 2018. doi: 10.1609/aaai.v32i1.11485. URL https://ojs.aaai.org/index.php/AAAI/article/view/11485.

Honghao Wei, Zixian Yang, Xin Liu, Zhiwei Qin, Xiaocheng Tang, and Lei Ying. A reinforcement learning and prediction-based lookahead policy for vehicle repositioning in online ride-hailing systems. *IEEE Transactions on Intelligent Transportation Systems*, 25(2):1846–1856, 2024. doi: 10.1109/TITS.2023.3312048.

Jeff Wu, Long Ouyang, Daniel M Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. Recursively summarizing books with human feedback. *arXiv preprint arXiv:2109.10862*, 2021.

Haike Xu, Tengyu Ma, and Simon Du. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. In Mikhail Belkin and Samory Kpotufe (eds.), *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pp. 4438–4472. PMLR, 15–19 Aug 2021. URL https://proceedings.mlr.press/v134/xu21a.html.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 18784–18794. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/d9d3837ee7981e8c064774da6cdd98bf-Paper.pdf.

Kunhe Yang, Lin Yang, and Simon Du. Q-learning with logarithmic regret. In Arindam Banerjee and Kenji Fukumizu (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1576–1584. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/yang21b.html.

Zixian Yang, Xin Liu, and Lei Ying. Exploration, exploitation, and engagement in multi-armed bandits with abandonment. *Journal of Machine Learning Research*, 25(9):1–55, 2024. URL http://jmlr.org/papers/v25/22-1251.html.

Yisong Yue and Thorsten Joachims. Beat the mean bandit. In *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, pp. 241–248, Madison, WI, USA, 2011. Omnipress. ISBN 9781450306195. URL http://www.icml-2011.org/papers/200_icmlpaper.pdf.

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. ISSN 0022-0000. doi: https://doi.org/10.1016/j.jcss.2011.12.028. URL https://www.sciencedirect.com/science/article/pii/S0022000012000281. JCSS Special Issue: Cloud Computing 2011.

Tom Zahavy, Brendan O' Donoghue, Guillaume Desjardins, and Satinder Singh. Reward is enough for convex mdps. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 25746–25759.

Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/d7e4cdde82a894b8f633e6d61a01ef15-Paper.pdf.

Chunqiu Zeng, Qing Wang, Shekoofeh Mokhtari, and Tao Li. Online context-aware recommendation with time varying multi-armed bandit. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 2025–2034, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450342322. doi: 10.1145/2939672.2939878. URL https://doi.org/10.1145/2939672.2939878.

Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D. Lee, and Wen Sun. Provable offline reinforcement learning with human feedback. In *ICML 2023 Workshop The Many Facets of Preference-Based Learning*, 2023a. URL https://openreview.net/forum?id=AY1dsKpNTu.

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

Qining Zhang and Lei Ying. Fast and regret optimal best arm identification: Fundamental limits and low-complexity algorithms. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 16729–16769. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/35fdecdf8861bc15110d48fbec3193cf-Paper-Conference.pdf.

Xuezhou Zhang, Yuzhe Ma, and Adish Singla. Task-agnostic exploration in reinforcement learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 11734–11743. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/8763d72bba4a7ade23f9ae1f09f4efc7-Paper.pdf.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 43037–43067. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/zhu23f.html.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

Masrour Zoghi, Shimon Whiteson, Remi Munos, and Maarten Rijke. Relative upper confidence bound for the k-armed dueling bandit problem. In Eric P. Xing and Tony Jebara (eds.), *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pp. 10–18, Bejing, China, 22–24 Jun 2014. PMLR. URL https://proceedings.mlr.press/v32/zoghi14.html.