
Efficient Preference-Based Reinforcement Learning: Randomized Exploration Meets Experimental Design

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 We study reinforcement learning from human feedback in general Markov decision
2 processes, where agents learn from trajectory-level preference comparisons. A
3 central challenge in this setting is to design algorithms that select informative
4 preference queries to identify the underlying reward while ensuring theoretical
5 guarantees. We propose a meta-algorithm based on randomized exploration, which
6 avoids the computational challenges associated with optimistic approaches and
7 remains tractable. We establish both regret and last-iterate guarantees under mild
8 reinforcement learning oracle assumptions. To improve query complexity, we
9 introduce and analyze an improved algorithm that collects batches of trajectory
10 pairs and applies optimal experimental design to select informative comparison
11 queries. The batch structure also enables parallelization of preference queries,
12 which is relevant in practical deployment as feedback can be gathered concurrently.
13 Empirical evaluation confirms that the proposed method is competitive with reward-
14 based reinforcement learning while requiring a small number of preference queries.

15 1 Introduction

16 Reinforcement learning (RL) is a fundamental paradigm in machine learning, where agents learn
17 to make sequential decisions by interacting with an environment to maximize cumulative rewards
18 [Barto, 2021]. RL has enabled advances in domains such as game play [Silver et al., 2017], robotics
19 [Todorov et al., 2012], or autonomous driving [Lu et al., 2023]. However, the practicality of RL is
20 hindered by the challenge of designing rewards: crafting a reward function that aligns with human
21 objectives is often difficult, and a misspecified reward function can lead to suboptimal or unsafe
22 behavior [Amodei et al., 2016, Hadfield-Menell et al., 2017]. This motivates the development of
23 principled alternatives to manual reward design.

24 Rather than relying on manually specified reward functions, reinforcement learning from human
25 feedback (RLHF) guides learning through preference feedback: at each step, a human oracle compares
26 trajectories and indicates which is preferable [Christiano et al., 2017]. This preference signal is often
27 much easier to provide than engineering a reward function [Pereira et al., 2019, Lee et al., 2023].
28 RLHF has proven to be effective in robotics [Jain et al., 2013] and, more recently, fine-tuning of large
29 language models [Stiennon et al., 2020, Ziegler et al., 2019, Rafailov et al., 2023]. This highlights
30 the practical relevance of RLHF compared to reward-based learning.

31 Despite its empirical success, the theoretical foundations of RLHF are still in development. Existing
32 works first studied the simpler setting of dueling bandits. In this context, the learner selects pairs
33 of actions and observes noisy preference feedback [Yue et al., 2012, Komiyama et al., 2015]. Clas-
34 sical algorithms for regret minimization in this setting include approaches based on zeroth-order
35 optimization [Yue and Joachims, 2009] or the principle of optimism [Ailon et al., 2014]. A key
36 challenge in this setting is reducing the number of preference queries. For this purpose, several recent

works propose strategic query selection strategies for dueling bandits [Das et al., 2024, Liu et al., 2024, Scheid et al., 2024, Mukherjee et al., 2024], often hinging on optimal experimental design mechanisms [Pukelsheim, 2006]. However, such approaches are usually limited to finite-armed bandits, where the resulting optimization problems can be solved efficiently.

In the online RL setting, the theory of RLHF has received increasing attention, with several works establishing either regret or probably approximately correct (PAC) guarantees. PAC-RL methods aim to identify a near-optimal policy with high probability [Xu et al., 2020, Novoseller et al., 2020, Zhu et al., 2023], while regret-based approaches provide bounds on the cumulative reward during learning [Pacchiano et al., 2021, Chen et al., 2022, Wu and Sun, 2023]. Similar to dueling bandits, a central challenge in RLHF is to actively select informative trajectory comparisons to drive learning. In RL, however, this active-learning problem presents additional difficulties: First, the learner cannot freely choose arbitrary state-action pairs or trajectories, but must reach them through exploration [Wagenmaker et al., 2022]. Second, many existing approaches with guarantees [Pacchiano et al., 2021, Zhan et al., 2024] rely on maximizing an exploration bonus involving a norm of state distributions – a problem which is known to be computationally intractable even in tabular settings [Efroni et al., 2021].

To sidestep these challenges, another line of work focuses on RLHF with offline data. In this setting, learning proceeds over a fixed pre-collected dataset of trajectory preferences [Zhu et al., 2023, Zhan et al., 2023]. Although this offline paradigm avoids the need for online exploration and active query selection, it depends critically on having access to sufficiently diverse and informative preference data a priori [Rashidinejad et al., 2021, Xie et al., 2021, Zanette et al., 2021, Zanette, 2023] — a requirement that can be difficult to meet in practice. Hence, this merely shifts the exploration and active learning challenges to the data collection phase.

Despite progress on statistical guarantees in RLHF, two key challenges remain open: the design of tractable algorithms for active preference query selection and reducing the workload of human preference annotators. In existing approaches, a human must provide feedback at every round, which is impractical in real-world applications. Our goal in this work is to develop RLHF algorithms that are computationally efficient, reduce the demand for human feedback, and actively select informative queries. Some recent work has made progress on tractability. For instance, Wu and Sun [2023] proposes a randomized exploration algorithm with regret guarantees, but their method is limited to linear dynamics. In parallel, Wang et al. [2023] introduces a general reduction from RLHF to standard RL and establishes PAC-style guarantees under RL oracle access. However, neither approach addresses the open challenges of reducing feedback requirements or enabling active query selection.

Contributions In this work, we focus on reinforcement learning from human feedback (RLHF) and develop meta-algorithms that reduce the RLHF problem to standard RL by leveraging existing RL algorithms as subroutines. Our approach combines randomized exploration for tractability, lazy updates to reduce human workload, and experimental design to actively select informative preference queries. We provide both regret and PAC-style guarantees under RL oracle assumptions. Our contributions are as follows:

- We propose general meta-algorithms for RLHF using RL oracles, and provide provable regret and PAC guarantees in general MDPs.
- We present a second meta-algorithm with better scalability and query efficiency thanks to: Lazy updates, inspired by linear bandits [Abbasi-Yadkori et al., 2011], which enables parallelization of the preference oracle calls; Greedy optimal design, which selects high-quality preference queries and improves sample efficiency.
- We provide empirical results showing that: Our algorithm is implementable and competitive with reward-based RL; The improved algorithm achieves comparable performance while significantly reducing the query complexity.

2 Preliminaries

Notation Let \mathbb{N} and \mathbb{R} denote the sets of natural and real numbers, respectively. We write $\|\cdot\|$ for the Euclidean norm and $\langle \cdot, \cdot \rangle$ for the standard inner product in \mathbb{R}^d . Moreover, for a positive definite matrix $A \in \mathbb{R}^{d \times d}$, we denote $\|x\|_A := \sqrt{\langle x, Ax \rangle}$ for the Mahalanobis norm. Furthermore, we denote the closed Euclidean ball of radius $a > 0$ by $\mathcal{B}^d(a) \subset \mathbb{R}^d$, and for a compact subset $\mathcal{X} \subset \mathbb{R}^n$,

we denote the set of all probability measures supported on \mathcal{X} by $\Delta_{\mathcal{X}}$. Finally, we use the standard notation $\mathcal{O}(n)$ and $\Omega(n)$ for asymptotic upper and lower bounds, as well as $\tilde{\mathcal{O}}(n) = \mathcal{O}(n \text{ polylog}(n))$ for suppressing polylogarithmic terms.

Setting We consider an infinite-horizon¹ Markov decision process (MDP) $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \nu_0, P, r, \gamma\}$ with compact state and action spaces $\mathcal{S} \subset \mathbb{R}^n$ and $\mathcal{A} \subset \mathbb{R}^m$, respectively, initial state distribution $s_0 \sim \nu_0$, transition law $s_{h+1} \sim P(\cdot | s_h, a_h)$, and discount rate $\gamma \in (0, 1)$. We assume a linear reward model $r_{\theta^*}(s, a) := \langle \theta^*, \phi(s, a) \rangle$, where $\|\theta^*\| \leq B$ and $\phi : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}^d$ is a feature mapping such that $\max_{s,a} \|\phi(s, a)\| \leq L$. We denote the set of all trajectories as $\mathcal{T} := (\mathcal{S} \times \mathcal{A})^\infty$, and the distribution over \mathcal{T} induced by a stationary Markov policy $\pi : \mathcal{S} \rightarrow \Delta_{\mathcal{A}}$ as \mathbb{P}_π . For a trajectory $\tau = (s_0, a_0, s_1, \dots) \in \mathcal{T}$, we denote the discounted sum of features by $\phi(\tau) := \sum_{h=0}^{\infty} \gamma^h \phi(s_h, a_h)$ and the feature expectation of a policy π by $\phi(\pi) := \mathbb{E}_{\tau \sim \mathbb{P}_\pi} [\phi(\tau)]$. Furthermore, given a reward parameter θ , we denote the value of a policy π by $V_\theta^\pi := \mathbb{E}_{\tau \sim \mathbb{P}_\pi} [\sum_{h=0}^{\infty} \gamma^h r_\theta(s_h, a_h)] = \langle \theta, \phi(\pi) \rangle$ and the optimal value by $V_\theta^* := \max_\pi V_\theta^\pi$.

Interaction protocol For each round $t = 1, \dots, T$ of RLHF, a learner, the MDP, and a preference oracle interact as follows. The learner selects two policies π_t and π'_t , and executes them to obtain two trajectories $\tau_t \sim \mathbb{P}_{\pi_t}$ and $\tau'_t \sim \mathbb{P}_{\pi'_t}$. Subsequently, the learner may query the preference oracle, which returns a binary label $y_t = \mathbb{1}(\tau_t \succ \tau'_t) \in \{0, 1\}$. The label equals one if the trajectory τ_t is preferred over τ'_t , denoted as $\tau_t \succ \tau'_t$, and zero otherwise. Each such interaction is one RLHF round.

Preference model We consider a stochastic preference model characterized by a preference function $\mathcal{P} : \mathcal{T} \times \mathcal{T} \rightarrow [0, 1]$, assigning to each pair of trajectories $\tau, \tau' \in \mathcal{T}$ the probability $\mathcal{P}(\tau \succ \tau')$ of preferring τ to τ' . We make the following assumption about the preference model.

Assumption 2.1 (Bradley–Terry model). The preference function \mathcal{P} satisfies for all $\tau, \tau' \in \mathcal{T}$

$$\mathcal{P}(\tau \succ \tau') = \sigma(\langle \theta^*, \phi(\tau) - \phi(\tau') \rangle), \quad (1)$$

where $\sigma(x) = 1/(1 + e^{-x})$ denotes the sigmoid function.

This preference model is a special case of the Plackett–Luce model [Plackett, 1975, Luce et al., 1959], and is commonly used in the dueling bandit setting as well as the RLHF framework [Yue et al., 2012, Christiano et al., 2017, Ouyang et al., 2022]. In particular, this model satisfies the strong stochastic transitivity property [Strzalecki, 2025]. That is, for any three trajectories $\tau_1 \succ \tau_2 \succ \tau_3$ we have $\mathcal{P}(\tau_1 \succ \tau_3) > \max\{\mathcal{P}(\tau_1 \succ \tau_2), \mathcal{P}(\tau_2 \succ \tau_3)\}$, capturing the monotonicity of preferences.

Remark 2.2. In practice, we cannot compare trajectories of infinite length. Fortunately, many environments terminate in finite time, and otherwise one may truncate each trajectory at horizon $H = \mathcal{O}(\log_\gamma(\varepsilon))$, introducing at most an ε error in value estimates (see e.g. [Schlaginhaufen and Kamgarpour, 2024]). For simplicity, however, we omit this truncation step in our presentation.

Regret To assess the learner’s online performance, we consider the cumulative regret

$$R(T) = \sum_{t=1}^T \frac{(V_{\theta^*}^* - V_{\theta^*}^{\pi_t}) + (V_{\theta^*}^* - V_{\theta^*}^{\pi'_t})}{2} = \frac{1}{2} \sum_{t=1}^T (2V_{\theta^*}^* - V_{\theta^*}^{\pi_t} - V_{\theta^*}^{\pi'_t}).$$

Cumulative regret has been widely adopted in the RL and RLHF literature [Abbasi-Yadkori et al., 2011, Zanette et al., 2020, Wang et al., 2023, Zhan et al., 2024]. However, cumulative regret doesn’t provide us with a guarantee of the last iterate’s suboptimality. As a second metric, we therefore also consider the suboptimality of an output policy.

Suboptimality The suboptimality of an output policy $\hat{\pi}$ is defined by

$$\text{SubOpt}(\hat{\pi}) := V_{\theta^*}^* - V_{\theta^*}^{\hat{\pi}}.$$

Suboptimality has previously been considered as a performance metric for offline RLHF Zhu et al. [2023], contextual bandits Das et al. [2024], and online RLHF [Wang et al., 2023]. In the following, we propose a meta-algorithm that features two variants: one with theoretical guarantees on cumulative regret, and another specifically ensuring a bound on last-iterate suboptimality.

¹Our results extend directly to the finite-horizon setting as well. However, we focus on the infinite-horizon discounted setting, as it is more commonly encountered in deep reinforcement learning.

3 Randomized Preference Optimization

3.1 Algorithm

Our algorithm proceeds in three essential steps. First, the preference feedback data is used to estimate the reward parameter using maximum likelihood estimation. Then, a reward parameter is sampled from a Gaussian distribution, which is reminiscent of linear Thompson sampling [Abeille and Lazaric, 2017]. Finally, an RL oracle is used to find an approximately optimal policy for the sampled reward parameter, and we query the preference oracle by comparing one trajectory from this new policy against one from the previous policy.

Maximum likelihood estimation Considering our preference model (1), a standard approach for estimating the reward parameter θ^* is via maximum likelihood estimation. Given a pair of trajectories $\tau_k = (s_{h,k}, a_{h,k})_{h=0}^\infty$ and $\tau'_k = (s'_{h,k}, a'_{h,k})_{h=0}^\infty$ we consider the design points $x_k := \phi(\tau_k) - \phi(\tau'_k) = \sum_{h=0}^\infty \gamma^h (\phi(s_{h,k}, a_{h,k}) - \phi(s'_{h,k}, a'_{h,k}))$ and the preference labels $y_k = \mathbb{1}(\tau_k \succ \tau'_k)$. In round t , the preference dataset is $\mathcal{D}_t = \{(x_k, y_k)\}_{k=1}^{t-1}$ and the corresponding (constrained) maximum likelihood estimator (MLE) is given by $\hat{\theta}_t = \arg \min_{\|\theta\| \leq B} \mathcal{L}_{\mathcal{D}_t}(\theta)$, where

$$\mathcal{L}_{\mathcal{D}_t}(\theta) := - \sum_{(x,y) \in \mathcal{D}_t} [y \log \sigma(\langle \theta, x \rangle) + (1-y) \log \sigma(-\langle \theta, x \rangle)], \quad (2)$$

is the negative log-likelihood of the Bradley-Terry model (1). The loss function (2) is the familiar logistic loss from logistic regression [Shalev-Shwartz and Ben-David, 2014]. In particular, it is a convex problem that can be solved efficiently using standard methods such as LBFGS [Liu and Nocedal, 1989]. Moreover, we have the following time-uniform confidence result.

Lemma 3.1. *Let $\lambda > 0$ and define the design matrix at time t given by $V_t = \sum_{s=1}^{t-1} x_s x_s^\top + \lambda I$. Then, with probability $1 - \delta$, for all $t \geq 1$, the true reward parameter θ^* is contained in the ellipsoid*

$$\mathcal{E}_t(\delta) := \left\{ \theta : \left\| \theta - \hat{\theta}_t \right\|_{V_t}^2 \leq \beta_t(\delta)^2 := \mathcal{O} \left(\kappa \left[\log \left(\frac{1}{\delta} \right) + d \log \left(\frac{t}{d} \right) \right] + \lambda \right) \right\}.$$

Here, $\kappa := \max_{\theta \in \mathcal{B}^d(B), x \in \mathcal{B}^d(2LH_\gamma)} 1/\sigma(\langle \theta, x \rangle)$ denotes the Lipschitz constant of the inverse sigmoid function, and $H_\gamma = (1 - \gamma)^{-1}$ the effective horizon of the MDP.

The above lemma hinges on a likelihood-ratio confidence set from Lee et al. [2024]. The proof and the precise constants are deferred to Appendix A.

Remark 3.2. Compared to the standard analysis of stochastic linear bandits Abbasi-Yadkori et al. [2011], our parameter β_t includes an additional factor of $\sqrt{\kappa}$, which arises naturally due to preference-based feedback. This result improves upon the bound provided by Zhu et al. [2023], which incurs a larger factor of κ instead of $\sqrt{\kappa}$. While the $\sqrt{\kappa}$ factor can theoretically be avoided by constructing confidence sets using the Hessian of the negative log-likelihood $\mathcal{L}_{\mathcal{D}_t}$ [Lee et al., 2024, Das et al., 2024], it reappears in the regret bounds as shown in [Das et al., 2024]. Thus, we adopt confidence sets based on V_t , as both V_t and its inverse can be efficiently updated via rank-one operations, unlike Hessian-based approaches.

Randomized exploration Many approaches to regret minimization and pure exploration in RLHF often rely on maximizing an exploration bonus $\|\phi(\pi_t) - \phi(\pi'_t)\|_{V_t^{-1}}$. Although such methods yield provable guarantees for regret [Pacchiano et al., 2021] or last-iterate suboptimality [Das et al., 2024], they are computationally intractable in RL settings (see Appendix E.2 for a discussion). To address this, we adopt a randomized exploration scheme inspired by Thompson sampling algorithms for linear bandits. In line with Abeille and Lazaric [2017], we sample the reward parameter from an inflated version of the confidence set defined in Lemma 3.1, which produces a computationally efficient alternative to optimism-based approaches. Furthermore, we also show that this randomized strategy extends to the pure exploration setting.

RL oracle With the objective of a meta-algorithm, we assume access to the following RL oracle.

Assumption 3.3 (PAC-RL oracle). We assume access to an (ε, δ) -PAC oracle, $A_{\text{RL}}^{\text{PAC}}$, for the RL problem. That is, a polynomial-time algorithm that produces for every $\varepsilon > 0$, $\delta > 0$, and $\theta \in \mathbb{R}^d$ a policy $\pi = A_{\text{RL}}^{\text{PAC}}(\theta, \varepsilon, \delta)$ such that with probability at least $1 - \delta$ we have $V_\theta^* - V_\theta^\pi \leq \varepsilon$.

This assumption of a PAC-RL oracle is satisfied in several settings, including tabular and linear MDPs [Dann et al., 2019, Ménard et al., 2021, Al Marjani et al., 2021, Wagenmaker et al., 2022, He et al., 2021]. Moreover, it is relatively mild compared to stronger oracles considered in the RLHF literature [Zhan et al., 2024], such as reward-free algorithms [Wang et al., 2020, Kaufmann et al., 2021, Ménard et al., 2021]. In practice, common choices for $\mathcal{A}_{\text{RL}}^{\text{PAC}}$ are policy optimization methods such as proximal policy optimization (PPO) [Schulman et al., 2017] or soft actor critic [Haarnoja et al., 2018], which have shown strong empirical performance in continuous control and large-scale applications such as training large language models.

Algorithm statement Our algorithm RP0 presented below comes in two variants: a variant that balances exploration and exploitation for regret minimization ($\alpha = 1$), and a variant for pure exploration ($\alpha = 0$). In line 4, we sample a reward parameter from a confidence set inflated by \sqrt{d} . Then, in line 5, we compute the policy π_t using an RL oracle. Finally, in line 10, we use the new preference data to update the reward parameter using maximum likelihood.

Algorithm 1: Randomized Preference Optimization (RP0)

Input: Number iterations T , confidence $\delta > 0$, regularization $\lambda > 0$, algorithm type $\alpha \in \{0, 1\}$.

```

1 Set  $\varepsilon = 1/\sqrt{T}$  and  $\delta' = \delta/5$ .
2 Initialize design matrix  $V_1 = \lambda I$ , preference data set  $\mathcal{D}_1 = \emptyset$ , parameter  $\hat{\theta}_1$ , and policy  $\pi_0$ .
3 for  $t = 1, 2, \dots, T$  do
4    $\tilde{\theta}_t \sim \mathcal{N}(\alpha \hat{\theta}_t, \beta_t (\delta')^2 V_t^{-1})$ ; // Reward sampling
5    $\pi_t = \mathcal{A}_{\text{RL}}^{\text{PAC}}(\tilde{\theta}_t, \varepsilon, \delta'/T)$ ,  $\pi'_t = \pi_{t-1}$ ; // Update policy with RL
6    $x_t = \phi(\tau_t) - \phi(\tau'_t)$  with  $\tau_t \sim \mathbb{P}_{\pi_t}$ ,  $\tau'_t \sim \mathbb{P}_{\pi'_t}$ ;
7    $y_t = \mathbb{1}(\tau_t \succ \tau'_t)$ ; // Preference feedback
8    $\mathcal{D}_{t+1} = \mathcal{D}_t \cup (y_t, x_t)$ ;
9    $V_{t+1} = V_t + x_t x_t^\top$ ;
10   $\hat{\theta}_{t+1} \in \arg \min_{\|\theta\| \leq B} \mathcal{L}_{\mathcal{D}_{t+1}}(\theta)$ ; // Reward estimation
11 end
Output: Policy  $\hat{\pi} = \mathcal{A}_{\text{RL}}^{\text{PAC}}(\bar{\theta}_T, \varepsilon, \delta')$  with  $\bar{\theta}_T = \frac{1}{T} \sum_{t=1}^T \hat{\theta}_t$ .
```

3.2 Theoretical results

We analyze the regret of Algorithm 1 and the suboptimality of its output policy $\hat{\pi}$.

3.2.1 Regret analysis

We show that Algorithm 1 with $\alpha = 1$ has sublinear regret with high probability.

Theorem 3.4. *Using Algorithm 1 with $\alpha = 1$, it holds for any $\delta > 0$ and $T \in \mathbb{N}$, with probability at least $1 - \delta$ that*

$$R(T) \leq \mathcal{O}\left(\sqrt{\kappa d^3 T \log(dT/\delta)^3}\right).$$

The regret bound of Theorem 3.4 matches the best existing bounds for algorithms with randomized exploration in reinforcement learning, see [Efroni et al., 2021, Ouhamma et al., 2023]. In addition, due to learning from preferences, we have an extra $\sqrt{\kappa}$ factor (see Remark 3.2), which is in line with other recent work on RLHF [Wu and Sun, 2023, Das et al., 2024].

Comparison with prior work For episodic tabular MDPs Pacchiano et al. [2021] prove a regret bound of $\tilde{\mathcal{O}}(\kappa d \sqrt{T})$. Similarly, Chen et al. [2022] considers episodic linear MDPs and derives a regret bound of $\tilde{\mathcal{O}}(d \sqrt{HT})$, avoiding dependence on κ by assuming a linear preference model. However, these approaches rely on a type of optimism which is computationally intractable (see Appendix E.2). Similar to us, Wu and Sun [2023] avoids this challenge by resorting to randomized exploration and proves a $\tilde{\mathcal{O}}(d^3 \sqrt{\kappa T})$ regret bound for linear MDPs. Compared to Wu and Sun [2023], our analysis improves the dependence on dimensionality from d^3 to $d^{3/2}$ and avoids the need for truncation techniques on the value function. Furthermore, our settings differ in two key points: First, we assume access to an RL oracle without restricting the class of MDPs, whereas Wu and Sun [2023]

209 considers linear MDPs. Second, their approach is model-based, while Algorithm 1 is oracle-based
210 and can accommodate both model-based and model-free implementations.

211 **Proof idea** The full proof of this Theorem is provided in Appendix B. Analogous to the analysis
212 of linear Thompson sampling [Abeille and Lazaric, 2017], the main idea is to control the regret
213 by showing that randomized exploration ensures a constant probability of optimism. However,
214 compared to the linear bandit analysis, our setting comes with additional challenges: First, due to
215 preference-based learning we require a different regret decomposition accounting for the reference
216 policy. Second, as we observe preference feedback on trajectories rather than policies, we need
217 to apply Freedman’s inequality (see Lemma A.9) to control the deviation between expected and
218 observed features. Lastly, as we assume a PAC RL oracle – in place of an exact maximization oracle –
219 we need to carefully track the resulting approximation error.

220 3.2.2 Suboptimality gap

221 The pure exploration version of RPO with $\alpha = 0$, in contrast to $\alpha = 1$, may incur linear regret during
222 learning as the policy π_t can be highly suboptimal. However, Theorem 3.5 below shows that it outputs
223 a policy that is $\tilde{O}(1/\sqrt{T})$ -optimal.

224 **Theorem 3.5.** *Using Algorithm 1 with $\alpha = 0$, it holds for any $\delta > 0$ and $T \in \mathbb{N}$, with probability at*
225 *least $1 - \delta$ that*

$$\text{SubOpt}(\hat{\pi}) = \mathcal{O} \left(\sqrt{\frac{\kappa d^3}{T} \log \left(\frac{dT}{\delta} \right)^3} \right).$$

226 *In other words, we need $\tilde{O}(\kappa d^3/\varepsilon^2)$ iterations to output an ε -optimal policy with high probability.*

227 Except for the extra \sqrt{d} dependency, which is inherent to approaches based on Thompson sampling,²
228 we match the last iterate guarantee proposed by Das et al. [2024] for contextual linear bandits, but
229 with an algorithm that is tractable in the full RL setting.

230 **Comparison with prior work** Few works provide suboptimality guarantees in preference-based
231 RL. Wang et al. [2023] propose a meta-algorithm interfacing with a PAC-RL oracle that outputs an
232 ε -optimal policy after $\tilde{O}(\kappa^2 d^3/\varepsilon^2)$ queries. A different approach by Zhan et al. [2024] leverages
233 optimal design to prove a bound of $\tilde{O}((|S|^2|A|d + \kappa^2 d^2)/\varepsilon^2)$, but their method relies on an intractable
234 maximization oracle. In comparison, our algorithm achieves a bound of $\tilde{O}(\kappa d^3/\varepsilon^2)$, improving the
235 dependence on κ over both prior results. The additional factor of d compared to Zhan et al. [2024]
236 is expected, as our method is randomized rather than optimistic, see [Abeille and Lazaric, 2017].
237 Notably, to the best of our knowledge, this is the first algorithm, whether in preference-based or
238 reward-based RL, that provides last-iterate guarantees using randomized exploration.

239 **Proof idea** The proof of Theorem 3.5, presented in Appendix D, builds on Das et al. [2024]’s
240 suboptimality analysis for the contextual bandits setting. However, to sidestep the intractability of
241 maximizing an exploration bonus over policies, we leverage randomized exploration [Abeille and
242 Lazaric, 2017] to ensure a constant probability of optimism. This allows us to derive a bound on the
243 output policy’s suboptimality that mirrors the regret bound, without needing additional assumptions.

244 3.3 Practical limitations

245 While Algorithm 1 is tractable and statistically efficient, it presents certain limitations. First, invoking
246 an RL oracle at every round (line 5) can cause high latency, especially in continuous state spaces.
247 Second, issuing preference queries at each round (line 7) is impractical when the oracle is a human,
248 due to the need for continuous feedback. Finally, requesting a label for all trajectory pairs can be
249 expensive and inefficient, as many comparisons are uninformative.

250 The next section presents a refined algorithm addressing these limitations. The new approach decou-
251 ples trajectory collection from query selection, and queries only the most informative comparisons.

²Recently, Abeille et al. [2025] show that for certain classes of linear bandits the additional \sqrt{d} factor can be avoided. However, their assumptions are not directly applicable to our preference-based RL setting.

252 4 A practical algorithm with efficient query selection

253 We present Algorithm 2, an improved method for preference collection and active query selection.

254 4.1 Algorithm

255 As discussed earlier, we design Algorithm 2 by using lazy updates to collect a batch of trajectory
256 pairs, then applying optimal design to select the informative queries from the batch.

Algorithm 2: Lazy Randomized Preference Optimization with Optimal Design (LRPO-OD)

Input: Number iterations T , confidence $\delta > 0$, regularization $\lambda > 1$, constant $C > 0$.

```

1 Set  $\varepsilon = 1/\sqrt{T}$ ,  $\delta' = \delta/5$ ,  $t_{\text{stop}} = 1$ .
2 Initialize design matrices  $V_0 = W_0 = \lambda I$ , preference datasets  $\mathcal{D} = \mathcal{D}_0 = \emptyset$ , parameter  $\hat{\theta}_0$ .
3 for  $t = 1, 2, \dots, T$  do
4   if  $\det(W_t) > (1 + C) \det(W_{t_{\text{stop}}})$  then
5      $\mathcal{D}_{\text{opt}}, V_t = \text{D-OptDes}(\mathcal{D}, V_{t_{\text{stop}}}, \det(W_t))$ ;           // Greedy D-Optimal design
6      $\mathcal{D}_t = \mathcal{D}_{t_{\text{stop}}}$ ;
7     for  $(\tau, \tau') \in \mathcal{D}_{\text{opt}}$  do
8        $x = \phi(\tau) - \phi(\tau'), y = \mathbb{1}(\tau \succ \tau')$ ;           // Preference feedback
9        $\mathcal{D}_t = \mathcal{D}_t \cup \{(x, y)\}$ ;
10    end
11     $\hat{\theta}_t \in \arg \min_{\|\theta\| \leq B} \mathcal{L}_{\mathcal{D}_t}(\theta)$ ;           // Reward estimation
12     $t_{\text{stop}} = t, \mathcal{D} = \emptyset, \pi' = \mathbf{A}_{\text{RL}}^{\text{PAC}}(\hat{\theta}_t, \varepsilon, \delta'/T)$ ;
13  end
14   $\tilde{\theta}_t \sim \mathcal{N}(\hat{\theta}_{t_{\text{stop}}}, \beta_{t_{\text{stop}}}(\delta')^2 V_{t_{\text{stop}}}^{-1})$ ;           // Reward sampling
15   $\pi_t = \mathbf{A}_{\text{RL}}^{\text{PAC}}(\tilde{\theta}_t, \varepsilon, \delta'/T)$ ;           // Update policy with RL
16   $x_t = \phi(\tau_t) - \phi(\tau'_t)$  with  $\tau_t \sim \mathbb{P}_{\pi_t}, \tau'_t \sim \mathbb{P}_{\pi'}$ ;
17  If  $t = t_{\text{stop}}$ , then  $W_{t+1} = V_t + x_t x_t^\top$ , else:  $W_{t+1} = W_t + x_t x_t^\top$ ;
18   $\mathcal{D} = \mathcal{D} \cup \{(\tau_t, \tau'_t)\}$ ;
19 end

```

258 **Lazy updates** We use an idea from Abbasi-Yadkori et al. [2011] to collect many trajectory pairs
259 without querying the preference oracle. The modification compared to Algorithm 1 is collecting
260 trajectories without updating the MLE $\hat{\theta}_{t_{\text{stop}}}$ until the information gain, represented by $\det(V_t)$,
261 increases by a multiplicative constant; see line 4 of Algorithm 2. We show that this procedure limits
262 the number of batches to $\mathcal{O}(\log(T))$. In other words, the average (over batches) size of a given
263 batch is of order $\mathcal{O}(T/\log(T))$. A key advantage of this lazy update structure is that the preference
264 queries (line 8) can be collected in parallel across all trajectory pairs within a batch. This significantly
265 reduces the workload of the preference oracle, *e.g.* a human annotator, by eliminating the need for
266 round-by-round feedback. In particular, the algorithm no longer pauses at each timestep to wait for
267 preference labels. To our knowledge, this is the first approach to support parallelization of queries in
268 online preference-based reinforcement learning, even in settings as simple as dueling bandits.

269 **D-Optimal design** To select informative preference queries from the collected trajectories above,
270 we leverage tools from optimal experimental design. Specifically, we apply an approximate D-optimal
271 design criterion to each collected batch of trajectory pairs; see Appendix E.1 for background on
272 D-optimal design. Given the current matrix $V_{t_{\text{stop}}}$ and the set of candidate trajectory pairs \mathcal{D} , we use
273 a greedy algorithm to solve the following maximization problem:

$$\max_{\{n_x\}} \log \det \left(V_{t_{\text{stop}}} + \sum_{x \in \mathcal{D}} n_x x x^\top \right) \quad \text{subject to} \quad \sum_{x \in \mathcal{D}} n_x = |\mathcal{D}|, \quad n_x \in \mathbb{N}.$$

274 Due to the submodularity of the log det function for λ greater than one, the greedy procedure of
275 Algorithm 3 achieves an $(1 - 1/e)$ -approximation to the optimal solution; see [Nemhauser et al.,
276 1978, Krause et al., 2008] and Appendix E.1.

Another key feature of Algorithm 3 is that the while loop is stopped early if $\det(V)$ exceeds the threshold value. This threshold is set as the determinant of the naive design, where every trajectory pair is queried once. When this early termination is satisfied, Algorithm 2 requires fewer samples than Algorithm 1. In particular, since the optimal design maximizes the information gain (measured by $\det(V)$), this termination condition is expected to be satisfied frequently in practice.

Algorithm 3: Greedy D-Optimal Design

Input: Dataset \mathcal{D} , current design matrix V , threshold for the determinant value α .

```

1 Initialize and dataset  $\mathcal{D}_{\text{opt}} = \emptyset$ 
2 while  $\det(V) < \alpha$  and  $|\mathcal{D}_{\text{opt}}| \leq |\mathcal{D}|$  do
282 3    $(\tau, \tau') = \arg \max_{(\tau, \tau') \in \mathcal{D}} \det(V + (\phi(\tau) - \phi(\tau'))(\phi(\tau) - \phi(\tau'))^\top)$ 
4    $V = V + xx^\top$ , where  $x = \phi(\tau) - \phi(\tau')$ ;  $\mathcal{D}_{\text{opt}} = \mathcal{D}_{\text{opt}} \cup \{(\tau, \tau')\}$ 
5 end
Output: Curated dataset  $\mathcal{D}_{\text{opt}}$ , design matrix  $V$ .

```

283 4.2 Theoretical result

284 We now provide our high probability regret bound for Algorithm 2.

285 **Theorem 4.1.** *Instantiating Algorithm 2 with $C > 0$ and $\lambda > 1$, it holds for any $\delta > 0$ and $T \in \mathbb{N}$,*
286 *with probability at least $1 - \delta$ that*

$$R(T) \leq \mathcal{O}\left(\sqrt{(1+C)\kappa d^3 T \log(dT/\delta)^3}\right).$$

287 *In addition, the number of times the condition of line 4 holds is at most $\frac{d}{\log(1+C)} \log\left(1 + \frac{T(LH_\gamma)^2}{\lambda d}\right)$.*

288 *Therefore, the size of the batches is on average of order $\tilde{\mathcal{O}}(T/(d \log(T)))$.*

289 Compared to Theorem 3.4, the above regret bound increases only by constant factors. Regarding the
290 number of preference queries, Sekhari et al. [2023] shows that at least $\Omega(T)$ queries are required to
291 achieve $\mathcal{O}(\sqrt{T})$ regret in the worst case. However, optimal design may lead to significantly fewer
292 queries in favorable instances, as demonstrated in our experiments. Importantly, approximate optimal
293 design does not compromise our guarantees, as it ensures a $1/(1 - 1/e)$ approximation to the optimal.

294 **Proof idea** We present here the key ideas of the proof of Theorem 4.1, the full proof is in Appendix
295 C. The lazy update mechanism used in our algorithm does not degrade the regret bound beyond a
296 constant factor. This follows from the analysis of Abbasi-Yadkori et al. [2011], which shows that
297 the elliptical potentials (related to the determinant of the design matrix) grow by at most a factor
298 of $(1 + C)$ compared to the standard design matrix of Algorithm 1. For the experimental design
299 component, we adapt the proof of the standard elliptical lemma [Lattimore and Szepesvári, 2020,
300 Lemma 19.4] to bring out the information gain, measured as $\log \det W_t$. This information gain is
301 then related to $\log \det V_t$ thanks to standard results from submodular optimization. Specifically, the
302 greedy strategy in Algorithm 3 achieves a $(1 - 1/e)$ approximation of the optimal information gain.
303 Together, these results allow us to derive a regret bound for this improved algorithm, which matches
304 the original up to constant factors.

305 5 Experiments

306 We validate our theoretical results on regret minimization (Theorems 3.4 and 4.1) on the
307 Isaac-Cartpole-v0 environment from Nvidia Isaac Lab [Mittal et al., 2023]. In this task the
308 goal is to balance a pole on a cart by applying left or right forces, preventing the pole from falling.
309 We compare RPO (for $\alpha = 1$) with its lazy variants (LRPO without optimal design and LRPO-OD with
310 optimal design). To simulate human preferences, we generate synthetic preferences using the built-in
311 task-specific reward function. Furthermore, we adopt PPO [Schulman et al., 2017] as our RL oracle
312 using 30 PPO steps per iteration of Algorithm 1 and 2. Moreover, to reduce variance in the reward
313 estimate, we sample five trajectories from π_t and π'_t in each RLHF round.

314 Figure 1 shows that all three algorithms achieve performance competitive with RL using ground
315 truth rewards. However, LRPO-OD reaches the highest performance while requiring significantly

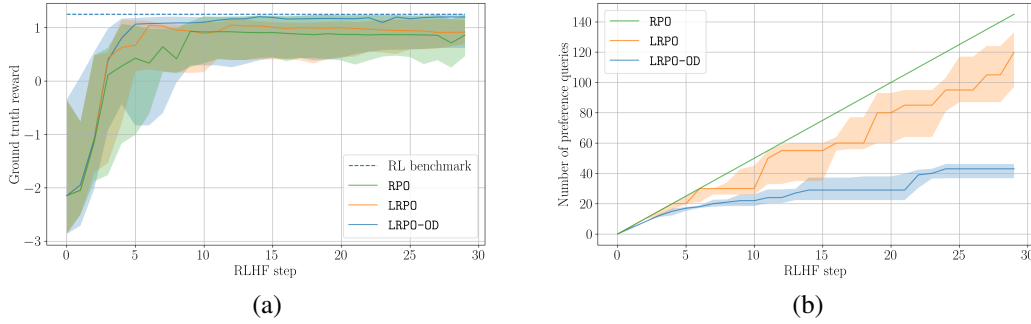


Figure 1: Comparison of RLHF algorithms in terms of (a) the ground truth reward $V_{\theta^*}^{\pi^t}$ (estimated from samples) and (b) number of preference queries performed. In particular, for the choice $\alpha = 1$, we compare RPO (green, Algorithm 1) with its lazy versions LRPO and LRPO-OD (orange & blue). Here, LRPO and LRPO-OD refer to Algorithm 2 without and with optimal design subroutine. The solid lines indicate the median and the shaded areas the 0.2 and 0.8 quantiles, across 10 independent runs. The dashed blue line indicates the mean reward achieved by PPO with the ground truth parameter θ^* .

fewer preference queries. This highlights that, despite theoretical worst-case lower bounds, the number of preference queries can be considerably reduced in practice by selecting informative queries with optimal design. Additional experimental results and further implementation and evaluation details are provided in Appendix F. Furthermore, our code to run these experiments on Isaac Lab’s manager-based environments is included in the supplementary materials and will be publicly released.

6 Conclusion

We introduced a simple meta-algorithm for reinforcement learning from human feedback (RLHF) that leverages randomized exploration to achieve both regret and PAC-style guarantees. Our framework applies to general MDPs and utilizes an RL oracle as a subroutine. Notably, to our knowledge, our PAC-RL guarantee is the first for a randomized algorithm in both RLHF and RL settings. Building on this foundation, we developed an improved variant with better scalability and query efficiency, for which we also established regret guarantees. This variant employs lazy updates to enable parallelization of preference oracle calls, significantly reducing the demand on human annotators. Additionally, we apply greedy optimal design to select informative comparisons. Empirically, our approach is competitive with reward-based RL while requiring significantly fewer preference queries. Overall, our contributions advance the state of RLHF by combining strong theoretical guarantees with practical algorithm design, improving efficiency and broadening applicability to real-world scenarios.

Our work opens several directions for future research. First, we provide separate algorithms for regret minimization and PAC guarantees; it remains unclear whether this separation is necessary, and developing a unified algorithm that simultaneously achieves both objectives is an open challenge. Second, we adopt the Bradley-Terry model for preference generation, which may not fully capture the complexity of real-world human feedback. Extending the framework to richer preference models is an important direction. Third, our approach relies on an RL oracle at every RLHF step, which may be computationally demanding. While using a reward-free algorithm as an RL oracle is theoretically efficient, practical RL implementations are typically based on policy optimization, which is not a reward-free algorithm, and often entails high sample complexity. Thus, it remains open whether we could require fewer RL oracle calls or whether reward-free oracles can be successfully implemented. Finally, our experimental evaluation is limited to simple robotic control tasks with synthetic feedback. Assessing performance on more complex tasks and with real human or LLM-generated feedback would offer a stronger test of the method’s practical applicability.

References

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

349 Marc Abeille and Alessandro Lazaric. Linear Thompson Sampling Revisited. In Aarti Singh and
350 Jerry Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence
351 and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pages 176–184. PMLR,
352 20–22 Apr 2017. URL <https://proceedings.mlr.press/v54/abeille17a.html>.

353 Marc Abeille, David Janz, and Ciara Pike-Burke. When and why randomised exploration works (in
354 linear bandits). *arXiv preprint arXiv:2502.08870*, 2025.

355 Nir Ailon, Zohar Karnin, and Thorsten Joachims. Reducing dueling bandits to cardinal bandits. In
356 *International Conference on Machine Learning*, pages 856–864. PMLR, 2014.

357 Aymen Al Marjani, Aurélien Garivier, and Alexandre Proutiere. Navigating to the best policy in
358 markov decision processes. *Advances in Neural Information Processing Systems*, 34:25852–25864,
359 2021.

360 Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané.
361 Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

362 Alper Atamtürk and Andrés Gómez. Maximizing a class of utility functions over the vertices of a
363 polytope. *Operations Research*, 65(2):433–445, 2017.

364 Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine
365 Learning Research*, 3(Nov):397–422, 2002.

366 Andrew G Barto. Reinforcement learning: An introduction. by richard’s sutton. *SIAM Rev*, 6(2):423,
367 2021.

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

371 Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep
372 reinforcement learning from human preferences. *Advances in neural information processing
373 systems*, 30, 2017.

374 Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable
375 reinforcement learning. In *International Conference on Machine Learning*, pages 1507–1516.
376 PMLR, 2019.

377 Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. Active preference
378 optimization for sample efficient rlhf. *arXiv preprint arXiv:2402.10500*, 2024.

379 Yonathan Efroni, Nadav Merlis, and Shie Mannor. Reinforcement learning with trajectory feedback.
380 In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 7288–7295,
381 2021.

382 David A Freedman. On tail probabilities for martingales. *the Annals of Probability*, pages 100–118,
383 1975.

384 Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy
385 maximum entropy deep reinforcement learning with a stochastic actor. In *International conference
386 on machine learning*, pages 1861–1870. Pmlr, 2018.

387 Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. Inverse
388 reward design. *Advances in neural information processing systems*, 30, 2017.

389 Jiafan He, Dongruo Zhou, and Quanquan Gu. Uniform-pac bounds for reinforcement learning
390 with linear function approximation. *Advances in Neural Information Processing Systems*, 34:
391 14188–14199, 2021.

392 Ashesh Jain, Brian Wojcik, Thorsten Joachims, and Ashutosh Saxena. Learning trajectory preferences
393 for manipulators via iterative improvement. *Advances in neural information processing systems*,
394 26, 2013.

- 395 Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent,
396 and Michal Valko. Adaptive reward-free exploration. In *Algorithmic Learning Theory*, pages
397 865–891. PMLR, 2021.
- 398 Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of*
399 *Mathematics*, 12:363–366, 1960.
- 400 Junpei Komiyama, Junya Honda, Hisashi Kashima, and Hiroshi Nakagawa. Regret lower bound and
401 optimal algorithm in dueling bandit problem. In *Conference on learning theory*, pages 1141–1154.
402 PMLR, 2015.
- 403 Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian
404 processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning*
405 *Research*, 9(2), 2008.
- 406 Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.
- 407 Junghyun Lee, Se-Young Yun, and Kwang-Sung Jun. A unified confidence sequence for generalized
408 linear models, with applications to bandits. *Advances in Neural Information Processing Systems*,
409 37:124640–124685, 2024.
- 410 Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel,
411 Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human
412 feedback. *arXiv preprint arXiv:2302.12192*, 2023.
- 413 Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization.
414 *Mathematical programming*, 45(1):503–528, 1989.
- 415 Pangpang Liu, Chengchun Shi, and Will Wei Sun. Dual active learning for reinforcement learning
416 from human feedback. *arXiv preprint arXiv:2410.02504*, 2024.
- 417 Yiren Lu, Justin Fu, George Tucker, Xinlei Pan, Eli Bronstein, Rebecca Roelofs, Benjamin Sapp,
418 Brandyn White, Aleksandra Faust, Shimon Whiteson, et al. Imitation is not enough: Robustifying
419 imitation with reinforcement learning for challenging driving scenarios. In *2023 IEEE/RSJ*
420 *International Conference on Intelligent Robots and Systems (IROS)*, pages 7553–7560. IEEE,
421 2023.
- 422 R Duncan Luce et al. *Individual choice behavior*, volume 4. Wiley New York, 1959.
- 423 Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent,
424 and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *International*
425 *Conference on Machine Learning*, pages 7599–7608. PMLR, 2021.
- 426 Mayank Mittal, Calvin Yu, Qinxu Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan,
427 Ritvik Singh, Yunrong Guo, Hammad Mazhar, Ajay Mandlekar, Buck Babich, Gavriel State,
428 Marco Hutter, and Animesh Garg. Orbit: A unified simulation framework for interactive robot
429 learning environments. *IEEE Robotics and Automation Letters*, 8(6):3740–3747, 2023. doi:
430 10.1109/LRA.2023.3270034.
- 431 Subhojyoti Mukherjee, Anusha Lalitha, Kousha Kalantari, Aniket Deshmukh, Ge Liu, Yifei Ma, and
432 Branislav Kveton. Optimal design for human feedback. *arXiv preprint arXiv:2404.13895*, 2024.
- 433 George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for
434 maximizing submodular set functions—i. *Mathematical programming*, 14:265–294, 1978.
- 435 Ellen Novoseller, Yibing Wei, Yanan Sui, Yisong Yue, and Joel Burdick. Dueling posterior sampling
436 for preference-based reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*,
437 pages 1029–1038. PMLR, 2020.
- 438 Reda Ouhamma, Debabrota Basu, and Odalric Maillard. Bilinear exponential family of MDPs:
439 frequentist regret bound with tractable exploration & planning. In *Proceedings of the AAAI*
440 *Conference on Artificial Intelligence*, volume 37, pages 9336–9344, 2023.

441 Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong
442 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow
443 instructions with human feedback. *Advances in neural information processing systems*, 35:27730–
444 27744, 2022.

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

447 Bruno L Pereira, Alberto Ueda, Gustavo Penha, Rodrygo LT Santos, and Nivio Ziviani. Online
448 learning to rank for sequential music recommendation. In *Proceedings of the 13th ACM Conference*
449 *on Recommender Systems*, pages 237–245, 2019.

450 Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C:*
451 *Applied Statistics*, 24(2):193–202, 1975.

452 Friedrich Pukelsheim. *Optimal design of experiments*. SIAM, 2006.

453 Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John
454 Wiley & Sons, 2014.

455 Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea
456 Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances*
457 *in Neural Information Processing Systems*, 36:53728–53741, 2023.

458 Paria Rashidinejad, Banghua Zhu, Cong Ma, Jiantao Jiao, and Stuart Russell. Bridging offline rein-
459 forcement learning and imitation learning: A tale of pessimism. *Advances in Neural Information*
460 *Processing Systems*, 34:11702–11716, 2021.

461 R Tyrrell Rockafellar. *Convex analysis*, volume 28. Princeton university press, 1997.

462 Antoine Scheid, Etienne Boursier, Alain Durmus, Michael I Jordan, Pierre Ménard, Eric Moulines,
463 and Michal Valko. Optimal design for reward modeling in rlhf. *arXiv preprint arXiv:2410.17055*,
464 2024.

465 Andreas Schlaginhaufen and Maryam Kamgarpour. Towards the transferability of rewards recovered
466 via regularized inverse reinforcement learning. *arXiv preprint arXiv:2406.01793*, 2024.

467 John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy
468 optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

469 Ayush Sekhari, Karthik Sridharan, Wen Sun, and Runzhe Wu. Contextual bandits and imitation
470 learning with preference-based active queries. *Advances in Neural Information Processing Systems*,
471 36:11261–11295, 2023.

472 Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to*
473 *algorithms*. Cambridge university press, 2014.

474 David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez,
475 Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without
476 human knowledge. *nature*, 550(7676):354–359, 2017.

477 Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford,
478 Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in*
479 *neural information processing systems*, 33:3008–3021, 2020.

480 Tomasz Strzalecki. *Stochastic choice theory*. Cambridge Books, 2025.

481 Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control.
482 In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 5026–5033.
483 IEEE, 2012.

484 Andrew J Wagenmaker, Max Simchowitz, and Kevin Jamieson. Beyond no regret: Instance-dependent
485 pac reinforcement learning. In *Conference on Learning Theory*, pages 358–418. PMLR, 2022.

486 Ruosong Wang, Simon S Du, Lin Yang, and Russ R Salakhutdinov. On reward-free reinforcement
487 learning with linear function approximation. *Advances in neural information processing systems*,
488 33:17816–17826, 2020.

489 Yuanhao Wang, Qinghua Liu, and Chi Jin. Is rlhf more difficult than standard rl? a theoretical
490 perspective. *Advances in Neural Information Processing Systems*, 36:76006–76032, 2023.

491 Runzhe Wu and Wen Sun. Making rl with preference-based feedback efficient via randomization.
492 *arXiv preprint arXiv:2310.14554*, 2023.

493 Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridg-
494 ing sample-efficient offline and online reinforcement learning. *Advances in neural information*
495 *processing systems*, 34:27395–27407, 2021.

496 Yichong Xu, Ruosong Wang, Lin Yang, Aarti Singh, and Artur Dubrawski. Preference-based
497 reinforcement learning with finite-time guarantees. *Advances in Neural Information Processing*
498 *Systems*, 33:18784–18794, 2020.

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

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

504 Andrea Zanette. When is realizability sufficient for off-policy reinforcement learning? In *Interna-*
505 *tional Conference on Machine Learning*, pages 40637–40668. PMLR, 2023.

506 Andrea Zanette, David Brandfonbrener, Emma Brunskill, Matteo Pirota, and Alessandro Lazaric.
507 Frequentist regret bounds for randomized least-squares value iteration. In *International Conference*
508 *on Artificial Intelligence and Statistics*, pages 1954–1964. PMLR, 2020.

509 Andrea Zanette, Martin J Wainwright, and Emma Brunskill. Provable benefits of actor-critic methods
510 for offline reinforcement learning. *Advances in neural information processing systems*, 34:13626–
511 13640, 2021.

512 Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Provable offline
513 reinforcement learning with human feedback. In *ICML 2023 Workshop The Many Facets of*
514 *Preference-Based Learning*, 2023.

515 Wenhao Zhan, Masatoshi Uehara, Wen Sun, and Jason D Lee. Provable reward-agnostic preference-
516 based reinforcement learning. *International Conference on Learning Representations 2024*, 2024.

517 Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human
518 feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*,
519 pages 43037–43067. PMLR, 2023.

520 Yinglun Zhu and Robert Nowak. Efficient active learning with abstention. *Advances in Neural*
521 *Information Processing Systems*, 35:35379–35391, 2022.

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

NeurIPS Paper Checklist

1. Claims

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

Answer: [\[Yes\]](#)

Justification: We provide our theoretical claims in Sections 3 and 4, we also provide our empirical results in Section 5.

Guidelines:

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

2. Limitations

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

Answer: [\[Yes\]](#)

Justification: We address the limitations of our work in Section 6.

Guidelines:

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

3. Theory assumptions and proofs

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

Answer: [\[Yes\]](#)

Justification: We provide our assumptions in Section 2 and our proofs in the appendix.

Guidelines:

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

4. Experimental result reproducibility

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

Answer: [\[Yes\]](#)

Justification: Our experimental environment will be publicly available, and all experimental details can be found in Appendix F.

Guidelines:

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

5. Open access to data and code

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

Answer: [Yes]

Justification: We provide the code used for our experiment with the supplementary files.

Guidelines:

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

6. Experimental setting/details

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

Answer: [Yes]

Justification: The details are provided in Appendix F.

Guidelines:

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

7. Experiment statistical significance

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

Answer: [Yes]

Justification: Our experiment section provides shaded areas for the standard deviation of the performance curves over 10 independent training runs. Additional details can be found therein.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

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

8. Experiments compute resources

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

Answer: [Yes]

Justification: See appendix F.

Guidelines:

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

9. Code of ethics

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

Answer: [Yes]

Justification: We don't see any conflict with the NeurIPS Code of Ethics.

Guidelines:

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

10. Broader impacts

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

Answer: [NA]

Justification: We do not foresee any negative societal impact for our algorithms.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

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

11. Safeguards

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

Answer: [NA]

Justification: We do not release models or datasets in this paper.

Guidelines:

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

12. Licenses for existing assets

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

Answer: [Yes]

Justification: We are citing and crediting the Isaac Lab creators for their RL environment.

Guidelines:

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

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. New assets

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

Answer: [Yes]

Justification: Instructions for reproducing the experiments are provided in the readme.

Guidelines:

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

14. Crowdsourcing and research with human subjects

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

Answer: [NA]

Justification: Our research is not involving any experiments with humans.

Guidelines:

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

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

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

Answer: [NA]

Justification: Our research is not involving any experiments with humans.

Guidelines:

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

16. Declaration of LLM usage

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

840 Answer: [No]

841 Justification: We only use LLMs to improve the writing quality.

842 Guidelines:

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

847 A Technical results

848 This section presents the technical results necessary for our theorems' proofs.

849 A.1 Confidence set

850 The first result is a confidence set for the maximum likelihood estimation. It is an elliptical relaxation
851 of the confidence set provided in Theorem 3.1 of Lee et al. [2024].

852 **Lemma A.1** (Confidence set for the reward estimation). *Define for a given $\lambda > 0$ the following set:*

$$\mathcal{E}_t(\delta) := \left\{ \theta : \left\| \theta - \hat{\theta}_t \right\|_{V_t}^2 \leq \beta_t(\delta)^2 := \kappa \left[\log \left(\frac{1}{\delta} \right) + d \log \left(\max \left\{ e, \frac{4eBLH_\gamma(t-1)}{d} \right\} \right) \right] + 4\lambda B^2 \right\},$$

853 *with $V_t = \sum_{s=1}^{t-1} x_t x_t^\top + \lambda I$, $\kappa = \max_{\|\theta\| \leq B, x \in \mathcal{X}} \frac{1}{\dot{\sigma}(\langle \theta, x \rangle)} = \mathcal{O}(e^{2BLH_\gamma})$, $H_\gamma = (1 - \gamma)^{-1}$, and $\dot{\sigma}$*
854 *is the first derivative of σ .*

855 *Then, it holds that:*

$$\Pr [\exists t \geq 1 : \theta^* \notin \mathcal{E}_t(\delta)] \leq \delta,$$

856 *that is, $\mathcal{E}_t(\delta)$ is a high-probability confidence set for θ^* uniformly in time.*

857 *Proof.* By a first-order Taylor approximation with integral remainder, we have

$$\mathcal{L}_{\mathcal{D}_t}(\theta^*) = \mathcal{L}_{\mathcal{D}_t}(\hat{\theta}_t) + \langle \nabla \mathcal{L}_{\mathcal{D}_t}(\hat{\theta}_t), \theta^* - \hat{\theta}_t \rangle + (\theta^* - \hat{\theta}_t)^\top G_t(\hat{\theta}_t, \theta^*)(\theta^* - \hat{\theta}_t),$$

858 where $\mathcal{L}_{\mathcal{D}_t}$ was defined in Equation (2) and

$$\begin{aligned} G_t(\hat{\theta}_t, \theta^*) &= \int_0^1 (1 - \tau) \left(\sum_{s=1}^{t-1} \dot{\sigma}(\langle \hat{\theta}_t + \tau(\theta^* - \hat{\theta}_t), x_s \rangle) x_s x_s^\top \right) d\tau \\ &= \sum_{s=1}^{t-1} \left[\int_0^1 (1 - \tau) \dot{\sigma}(\langle \hat{\theta}_t + \tau(\theta^* - \hat{\theta}_t), x_s \rangle) d\tau \right] x_s x_s^\top \\ &\succeq \kappa^{-1} \sum_{s=1}^{t-1} x_s x_s^\top. \end{aligned} \tag{3}$$

859 Rearranging terms gives

$$\begin{aligned} \mathcal{L}_{\mathcal{D}_t}(\theta^*) - \mathcal{L}_{\mathcal{D}_t}(\hat{\theta}_t) &\stackrel{(i)}{\geq} (\theta^* - \hat{\theta}_t)^\top G_t(\hat{\theta}_t, \theta^*)(\theta^* - \hat{\theta}_t) \\ &\stackrel{(ii)}{\geq} (\theta^* - \hat{\theta}_t)^\top \left(\kappa^{-1} \sum_{s=1}^{t-1} x_s x_s^\top \right) (\theta^* - \hat{\theta}_t) \\ &= \kappa^{-1} \left\| \theta^* - \hat{\theta}_t \right\|_{V_t}^2 - \kappa^{-1} \lambda \left\| \theta^* - \hat{\theta}_t \right\|^2, \end{aligned}$$

860 where (i) follows from the first order optimality condition for $\hat{\theta}_t$ and (ii) from the lower bound in (3).
861 Rearranging again and applying [Lee et al., 2024, Theorem 3.1] with the Lipschitz constant of $\mathcal{L}_{\mathcal{D}_t}$
862 equal to $L_t = 2LH_\gamma(t-1)$, we get

$$\left\| \theta^* - \hat{\theta}_t \right\|_{V_t}^2 \leq \kappa \left[\log \left(\frac{1}{\delta} \right) + d \log \left(\max \left\{ e, \frac{4eBLH_\gamma(t-1)}{d} \right\} \right) \right] + 4\lambda B^2.$$

863 □

864 A.2 Optimism with constant probability

865 We first recall the following standard concentration and anti-concentration property of the Gaussian
866 distribution.

867 **Lemma A.2** (Appendix A of [Abeille and Lazaric, 2017]). *Let $z \sim \mathcal{N}(0, I)$ be a d -dimensional*
868 *Gaussian random vector. Then, we have:*

869 1. *Anti-concentration:* For any $u \in \mathcal{B}^d(1)$, we have $\Pr[\langle u, z \rangle \geq 1] \geq \frac{1}{4\sqrt{e\pi}}$.

870 2. *Concentration:* $\Pr[\|z\| \leq \sqrt{2d \log(2d/\delta)}] \geq 1 - \delta$.

871 The anti-concentration property yields the following key result, which is required to prove a constant
872 probability of optimism and subsequently control pessimism terms in the regret. Although it has
873 been proven by Abeille and Lazaric [2017] in their linear Thompson sampling analysis, we provide a
874 concise proof based on convex analysis for completeness.

875 **Lemma A.3.** *Let $f : \mathbb{R}^d \rightarrow \mathbb{R}$ be a continuous and convex function, and consider the ellipsoid*
876 $\mathcal{E} := \{x \in \mathbb{R}^d : \|x - x_0\|_A \leq b\}$ *for a positive definite matrix A and $b > 0$. If $\tilde{x} \sim \mathcal{N}(x_0, b^2 A^{-1})$,*
877 *then $\Pr[f(\tilde{x}) \geq \max_{x \in \mathcal{E}} f(x)] \geq 1/(4\sqrt{e\pi})$.*

878 *Proof.* Note that by definition of \tilde{x} we have $\tilde{x} \stackrel{d}{=} x_0 + bA^{-1/2}\tilde{z}$, where $\tilde{z} \sim \mathcal{N}(0, I)$. Hence,
879 considering $g(z) := f(x_0 + bA^{-1/2}z)$, we have

$$p := \Pr\left[f(\tilde{x}) \geq \max_{x \in \mathcal{E}} f(x)\right] = \Pr\left[g(\tilde{z}) \geq \max_{z \in \mathcal{B}^d(1)} g(z)\right],$$

880 where we used that $x_0 + bA^{-1/2}z \in \mathcal{E}$ if and only if $z \in \mathcal{B}^d(1)$. Since g is a continuous convex
881 function, we can choose $\bar{z} \in \arg \max_{z \in \mathcal{B}^d(1)} g(z)$ such that $\|\bar{z}\| = 1$. By optimality of \bar{z} it holds that
882 [Rockafellar, 1997, Theorem 32.4]

$$\partial f(\bar{z}) \subseteq N_{\mathcal{B}^d(1)}(\bar{z}),$$

883 where $N_{\mathcal{B}^d(1)}(\bar{z}) = \{h \in \mathbb{R}^d : h = \lambda \bar{z}, \lambda \geq 0\}$ denotes the normal cone to $\mathcal{B}^d(1)$ at \bar{z} . Hence, by
884 convexity of g we have

$$g(\tilde{z}) \geq g(\bar{z}) + \langle \lambda \bar{z}, \tilde{z} - \bar{z} \rangle = g(\bar{z}) + \lambda(\langle \bar{z}, \tilde{z} \rangle - 1), \quad \text{for some } \lambda \geq 0.$$

885 Therefore, $\langle \bar{z}, \tilde{z} \rangle \geq 1$ implies that $g(\tilde{z}) \geq g(\bar{z})$, which yields the lower bound

$$p = \Pr[g(\tilde{z}) \geq g(\bar{z})] \geq \Pr[\langle \bar{z}, \tilde{z} \rangle \geq 1].$$

886 In light of Lemma A.2, this implies that $p \geq 1/(4\sqrt{e\pi})$. □

887 A.3 Elliptical potential bounds

888 The following lemma provides an upper bound on the sum of norms of sequentially observed vectors
889 in the norm induced by their design matrix. These norms of vectors are commonly called elliptical
890 potentials.

891 **Lemma A.4** (Lemma 19.4 of Lattimore and Szepesvári [2020]). *Let $\{X_t\}_{t \geq 0} \in \mathbb{R}^d$ and for all $t \geq 0$,*
892 *$\|X_t\| \leq L$, let $V_t = \lambda I_d + \sum_{s=0}^{t-1} X_s X_s^\top$ for some $\lambda > 0$. Then,*

$$\sum_{t=1}^T \min\{1, \|X_t\|_{V_t^{-1}}^2\} \leq 2d \log \left(1 + \frac{TL^2}{d\lambda}\right).$$

893 In the analysis of our algorithms, a central challenge is to control the norms of policy feature
894 differences $\phi(\pi_t) - \phi(\pi_t)'$. However, the learner only observes the trajectory features $\phi(\tau_t)$ and
895 $\phi(\tau_t')$, which are random realizations of the policy features. To overcome this, we build on Lemma
896 A.4 and introduce new tools to bound the sum of norms of policy feature differences.

897 **Lemma A.5** (Elliptical lemma). *1) Let $\{X_t\}_{t \geq 0} \in \mathbb{R}^d$ and for all $t \geq 0$, $\|X_t\| \leq L$, let $V_t =$
898 $\lambda I_d + \sum_{s=0}^{t-1} X_s X_s^\top$ for some $\lambda > 0$. Then,*

$$\sum_{t=1}^T \|X_t\|_{V_t^{-1}}^2 \leq 2d \log \left(1 + \frac{TL^2}{d\lambda}\right) + \frac{3dL^2}{\log(2)\lambda} \log \left(1 + \frac{L^2}{\log(2)\lambda}\right).$$

899 2) Let $\{X_t\}_{t \geq 0} \in \mathbb{R}^d$ be a sequence of random vectors adapted to a filtration $\{\mathcal{F}_t\}_{t \geq 0}$. Assume that
 900 for all $t \geq 0$, $\|X_t\| \leq L$ almost surely. Then, for all $\delta > 0$, it holds with probability at least $1 - \delta$
 901 that:

$$\begin{aligned} \forall T \in \mathbb{N}, \quad \sum_{t=1}^T \mathbb{E}[\|X_t\|_{V_t^{-1}}^2 | \mathcal{F}_t] \leq & 2\sqrt{T \left(2d \log \left(1 + \frac{TL^2}{d\lambda} \right) + \frac{3dL^2}{\log(2)\lambda} \log \left(1 + \frac{L^2}{\log(2)\lambda} \right) \right)} \\ & + \frac{8L}{\sqrt{\lambda}} \log(1/\delta), \end{aligned}$$

902 where $V_t = \lambda I_d + \sum_{s=0}^{t-1} X_s X_s^\top$ for some $\lambda > 0$.

903 The first statement is a small improvement over Lemma A.4 because it involves $\|X_t\|_{V_t^{-1}}^2$ instead of
 904 $\min\{1, \|X_t\|_{V_t^{-1}}^2\}$ and maintains a similar upper bound. In the second statement, $\{X_t\}_{t \geq 0}$ represent
 905 trajectory features and their expected values are policy features. Hence, the second statement of the
 906 lemma above allows us to control the elliptical potentials of the policy features while only observing
 907 trajectory features.

908 **Proof. First statement:** The proof of this result is based on the observation in [Lattimore and
 909 Szepesvári, 2020, Exercise 19.3]. Namely, the number of times the term $\|X_t\|_{V_t^{-1}}^2$ can be larger than
 910 one is at most $\frac{3d}{\log(2)} \log(1 + \frac{L^2}{\lambda \log(2)})$.

911 Let's define the rounds $\mathcal{T}_T = \{t \leq T, \|X_t\|_{V_t^{-1}}^2 \leq 1\}$, we have:

$$\begin{aligned} \sum_{t=1}^T \|X_t\|_{V_t^{-1}}^2 &= \sum_{t \in \mathcal{T}_T} \|X_t\|_{V_t^{-1}}^2 + \sum_{t \notin \mathcal{T}_T} \|X_t\|_{V_t^{-1}}^2 \\ &\leq \sum_{t \in \mathcal{T}_T} \min\{1, \|X_t\|_{V_t^{-1}}^2\} + \frac{3dL^2}{\lambda \log(2)} \log(1 + \frac{L^2}{\lambda \log(2)}), \end{aligned}$$

912 where the first term of the last inequality follows by definition of \mathcal{T}_T . The second term follows
 913 because the number of times $1 \leq t \leq T$ not in \mathcal{T}_T is at most $\frac{3d}{\log(2)} \log(1 + \frac{L^2}{\lambda \log(2)})$ as previously
 914 discussed, and because $\|X_t\|_{V_t^{-1}}^2 \leq L^2/\lambda$. Then, the proof is concluded by bounding the first sum
 915 on the right-hand side using Lemma A.4.

916 **Second statement:** The proof proceeds by using Lemma A.9. We have for any $\delta > 0$ that with
 917 probability $1 - \delta$:

$$\begin{aligned} \sum_{t=1}^T \mathbb{E}[\|X_t\|_{V_t^{-1}}^2 | \mathcal{F}_t] &\leq 2 \sum_{t=1}^T \|X_t\|_{V_t^{-1}} + \frac{8L}{\sqrt{\lambda}} \log(1/\delta) \\ &\leq 2\sqrt{T \sum_{t=1}^T \|X_t\|_{V_t^{-1}}^2} + \frac{8L}{\sqrt{\lambda}} \log(1/\delta) \\ &\leq 2\sqrt{T \left(2d \log \left(1 + \frac{TL^2}{d\lambda} \right) + \frac{3dL^2}{\log(2)\lambda} \log \left(1 + \frac{L^2}{\log(2)\lambda} \right) \right)} + \frac{8L}{\sqrt{\lambda}} \log(1/\delta), \end{aligned}$$

918 where the first inequality uses Lemma A.9, the second uses the Cauchy-Schwarz inequality, and the
 919 last follows from the first result of the lemma. \square

920 We now present a variant of the elliptical potentials lemma above, adapted for the case where the
 921 design matrix is updated with optimal design and lazy updates; see Algorithm 2 for more details.

922 **Lemma A.6 (Lazy elliptical lemma).** Let $\{X_t\}_{t \geq 0} \in \mathbb{R}^d$ be a sequence of random vectors adapted
 923 to a filtration $\{\mathcal{F}_t\}_{t \geq 0}$. Assume that for all $t \geq 0$, $\|X_t\| \leq L$. Then, we have the following results:

924 1) For all $\lambda > 0$ and $C > 0$, it holds that

$$\sum_{t=1}^T \|X_t\|_{V_t^{-1}}^2 \leq \frac{2(1+C)d}{1-1/e} \log \left(1 + \frac{TL^2}{d\lambda} \right) + \frac{3dL^2}{\log(2)\lambda} \log \left(1 + \frac{L^2}{\log(2)\lambda} \right).$$

925 where V_t is the matrix in Algorithm 2. Recall that V_t is defined with the following lazy update
926 procedure:

- 927 • Initialize $V_0 = \lambda I$, $W_0 = \lambda I$, $\mathcal{D} = \emptyset$,
- 928 • At a given round $t \in \mathbb{N}$:
 - 929 – Update W_t :
 - 930 * If $\mathcal{D} = \emptyset$ update $W_t = V_t + x_{t-1}x_{t-1}^\top$,
 - 931 * else $W_t = W_{t-1} + x_{t-1}x_{t-1}^\top$, $V_t = V_{t-1}$.
 - 932 – Update set $\mathcal{D} = \mathcal{D} \cup \{x_{t-1}\}$
 - 933 – if $\det W_t > (1+C)\det V_t$:
 - 934 * update $V_t = V_{t-1} + \sum_{x \in \text{D-OptDes}(\mathcal{D}, \det(W_t))} xx^\top$
 - 935 * reset $\mathcal{D} = \emptyset$.

936 the operator D-OptDes is the procedure described in Algorithm 3.

937 2) For all $\delta > 0$, it holds with probability at least $1 - \delta$ that:

$$\forall T \in \mathbb{N}, \quad \sum_{t=1}^T \mathbb{E}[\|X_t\|_{V_t^{-1}} | \mathcal{F}_t] \leq \sqrt{\frac{8(1+C)dT}{(1-1/e)^2} \log \left(1 + \frac{TL^2}{d\lambda} \right) + \frac{12dL^2T}{\log(2)\lambda} \log \left(1 + \frac{L^2}{\log(2)\lambda} \right)} + \frac{8L}{\sqrt{\lambda}} \log(1/\delta).$$

938 Compared to Lemma A.5, we lose a factor of $\frac{1+C}{1-1/2}$ in the first statement and of $\frac{1+C}{(1-1/e)^2}$ in the
939 second statement. The factor $(1+C)$ arises because we only update the matrix V_t once $\det W_t$
940 exceeds $\det V_t$ by a constant $(1+C)$. The $1/(1-1/e)$ factors arise from the use of optimal design.

941 **Proof. First statement:** We proceed in two steps: First, we relate V_t to W_t , similar to the standard
942 proofs for lazy updates (see proof of Theorem 4 of Abbasi-Yadkori et al. [2011] for example); The
943 second step consists of using arguments of approximate optimality of greedy optimal design from
944 [Nemhauser et al., 1978].

945 **Step 1:** Here, we relate the matrix norm in V_t^{-1} to the matrix norm in W_t .

946 We have for all $t \in \mathbb{N}$ that $W_t \geq V_t$, which implies that $V_t^{-1} \geq W_t^{-1}$. Then, using Lemma A.8 we
947 have:

$$\forall x \in \mathbb{R}^d, \quad \|x\|_{V_t^{-1}}^2 \leq \|x\|_{W_t^{-1}}^2 \frac{\det(V_t^{-1})}{\det(W_t^{-1})}. \quad (4)$$

948 By definition of V_t , we know that for all $t \geq 1$, $\det(W_t) \leq (1+C)\det(V_t)$, which implies that
949 $\det(V_t^{-1}) \leq (1+C)\det(W_t^{-1})$. Then, from (4) we obtain that:

$$\forall x \in \mathbb{R}^d, \quad \|x\|_{V_t^{-1}}^2 \leq (1+C)\|x\|_{W_t^{-1}}^2.$$

950 Now, for rounds $t_{\text{stop}} \in \mathbb{N}$ such that $\det V_{t_{\text{stop}}} > (1+C)\det V_{t_{\text{stop}}-1}$, and all rounds $t > t_{\text{stop}}$ such
951 that $\det W_t \leq (1+C)\det V_{t_{\text{stop}}}$ we have:

$$\sum_{q=t_{\text{stop}}}^t \|x_q\|_{V_t^{-1}}^2 = \sum_{q=t_{\text{stop}}}^t \|x_q\|_{V_{t_{\text{stop}}}^{-1}}^2 \leq (1+C) \sum_{q=t_{\text{stop}}}^t \|x_q\|_{W_q^{-1}}^2.$$

952 In addition, we have:

$$\sum_{q=t_{\text{stop}}}^{t-1} \min\{1, \|x_q\|_{V_t^{-1}}^2\} \leq 2(1+C) \sum_{q=t_{\text{stop}}}^{t-1} \log(1 + \|x_q\|_{W_q^{-1}}^2) = 2(1+C) \log(\det(W_t)/\det(W_{t_{\text{stop}}})) ,$$

953 where the last equality follows because $\frac{\det(W_{q+1})}{\det(W_q)} = 1 + \|x_q\|_{W_q^{-1}}^2$ (see the proof of [Abbasi-Yadkori
954 et al., 2011, Lemma 11] for example).

955 *Step 2:* We now relate the terms $(\log \det(W_t))_{t \leq t'_{\text{stop}}}$ to the term $\log \det(V_{t'_{\text{stop}}})$ where t'_{stop} is the
956 first time greater than t_{stop} such that $\det(V_{t'_{\text{stop}}}) > (1 + C) \det(V_{t'_{\text{stop}}-1})$.

957 Based on Algorithm 3, there are two possibilities:

- 958 1. either the new design matrix $V_{t'_{\text{stop}}}$ is such that $\det V_{t'_{\text{stop}}} > \det W_{t'_{\text{stop}}}$,
- 959 2. or the procedure $\text{D-OptDes}(\mathcal{D}, \det(W_{t'_{\text{stop}}}))$ terminates after adding $|\mathcal{D}|$ elements to the
960 matrix $V_{t'_{\text{stop}}}$.

961 In the first case, we have by definition that $\det V_{t'_{\text{stop}}} \geq \det W_{t'_{\text{stop}}}$. We now show that a similar result
962 also holds in the second case.

963 Consider the set \mathcal{D} at time t'_{stop} . The set function $S \subset \mathcal{D} \rightarrow \log \det(V_{t_{\text{stop}}} + \sum_{x \in S} xx^T)$ is
964 submodular [Krause et al., 2008]. Therefore, it follows from [Nemhauser et al., 1978] that the greedy
965 algorithm maximizing the above function is $(1 - 1/e)$ optimal. Namely, consider the set \mathcal{D}_{opt} defined
966 by greedily adding the features $x \in \mathcal{D}$ until $|\mathcal{D}_{\text{opt}}| = |\mathcal{D}|$. Then, it holds that:

$$\log \det \left(V_{t_{\text{stop}}} + \sum_{x \in \mathcal{D}_{\text{opt}}} xx^T \right) \geq (1 - 1/e) \max_{S \subset \mathcal{D}, |S| \leq |\mathcal{D}|} \log \det \left(V_{t_{\text{stop}}} + \sum_{x \in S} xx^T \right).$$

967 From the above, we deduce that for all $t \leq t'_{\text{stop}}$:

$$\begin{aligned} \log \det(W_t) &= \log \det(V_{t_{\text{stop}}} + \sum_{s=t_{\text{stop}}}^t x_s x_s^T) \leq \max_{S \subset \mathcal{D}, |S| \leq |\mathcal{D}|} \log \det \left(V_{t_{\text{stop}}} + \sum_{x \in S} xx^T \right) \\ &\leq \frac{1}{1 - 1/e} \log \det \left(V_{t_{\text{stop}}} + \sum_{x \in \mathcal{D}_{\text{opt}}} xx^T \right). \end{aligned}$$

968 We conclude that in both cases of termination of Algorithm 3, it holds for all $t \leq t'_{\text{stop}}$ that:

969 $\log \det(W_t) \leq \frac{1}{1 - 1/e} \log \det(V_{t'_{\text{stop}}}).$

970 *Combination of steps 1 and 2:* From step 1, we know that:

$$\sum_{q=1}^{t-1} \min\{1, \|x_q\|_{V_t^{-1}}^2\} \leq 2(1 + C) \sum_{q=1}^{t-1} \log(1 + \|x_q\|_{W_q^{-1}}^2) = 2(1 + C) \log(\det(W_t)/\det(W_0)),$$

971 In addition, exercise 19.3 of [Lattimore and Szepesvári, 2020] proves that the number of times where
972 $\|x_q\|_{W_t^{-1}}$ is greater than one is at most $\frac{3d}{\log(2)} \log \left(1 + \frac{L^2}{\lambda \log(2)} \right)$. Therefore, we deduce that for all
973 $T \in \mathbb{N}$:

$$\begin{aligned} \sum_{q=0}^T \|x_q\|_{V_t^{-1}}^2 &\leq \sum_{q=0}^T \min\{1, \|x_q\|_{V_t^{-1}}^2\} + \frac{3dL^2}{\lambda \log(2)} \log \left(1 + \frac{L^2}{\lambda \log(2)} \right) \\ &\leq 2(1 + C) \log(\det(W_t)/\det(W_0)) + \frac{3dL^2}{\lambda \log(2)} \log \left(1 + \frac{L^2}{\lambda \log(2)} \right). \end{aligned}$$

974 Finally, plugging the result of step 2 above yields that:

$$\begin{aligned} \sum_{q=0}^T \|x_q\|_{V_t^{-1}}^2 &\leq \frac{2(1 + C)}{1 - 1/e} \log(\det(V_T)/\det(V_0)) + \frac{3dL^2}{\lambda \log(2)} \log \left(1 + \frac{L^2}{\lambda \log(2)} \right) \\ &\leq \frac{2(1 + C)d}{1 - 1/e} \log \left(1 + \frac{TL^2}{\lambda} \right) + \frac{3dL^2}{\lambda \log(2)} \log \left(1 + \frac{L^2}{\lambda \log(2)} \right). \end{aligned}$$

975 This concludes the proof of the first result.

976 **Second statement:** We know from inequality (4) that for all $t \in [\tau, \tau' - 1]$:

$$\mathbb{E}[\|X_t\|_{V_t^{-1}} | \mathcal{F}_t] \leq \mathbb{E}[\|X_t\|_{W_t^{-1}} | \mathcal{F}_t].$$

977 Then, using Lemma A.9 we obtain:

$$\sum_{t=\tau}^{\tau'-1} \mathbb{E}[\|X_t\|_{V_t^{-1}} | \mathcal{F}_t] \leq 2 \sum_{t=\tau}^{\tau'-1} \|X_t\|_{W_t^{-1}} + \frac{8L}{\sqrt{\lambda}} \log(1/\delta).$$

978 Then, we can conclude the proof using the same arguments as for the first result. \square

979 A.4 Miscellaneous

980 **Proposition A.7.** *The function $f(\theta) = \sup_{\pi \in (\Delta_{\mathcal{A}})^S} \langle \theta, \phi(\pi) \rangle$ is convex and continuous over \mathbb{R}^d .*

981 *Proof.* The function f is the support function of the set $\mathcal{Z} = \{\phi(\pi) : \pi : \mathcal{S} \rightarrow \Delta_{\mathcal{A}}\}$. To prove the
982 convexity, note that for any $\eta \in (0, 1)$, we have

$$f(\eta\theta + (1-\eta)\theta') \leq \eta \sup_{z \in \mathcal{Z}} \langle \theta, z \rangle + (1-\eta) \sup_{z \in \mathcal{Z}} \langle \theta', z \rangle \leq \eta f(\theta) + (1-\eta) f(\theta').$$

983 Furthermore, any convex function is continuous over the relative interior of its effective domain
984 $\text{dom } f = \{x : f(x) < \infty\}$ (see *e.g.* [Rockafellar, 1997, Theorem 10.1]). Since $|f(\theta)| \leq \|\theta\| LH_{\gamma}$,
985 this implies that f is continuous over \mathbb{R}^d . \square

986 **Lemma A.8** (Lemma 12 of [Abbasi-Yadkori et al., 2011]). *Let A, B , and C be positive semi-definite
987 matrices such that $A = B + C$. Then, we have that:*

$$\sup_{x \neq 0} \frac{x^T A x}{x^T B x} \leq \frac{\det(A)}{\det(B)}.$$

988 This next lemma is one of the versions of Freedman's inequality [Freedman, 1975].

989 **Lemma A.9** (Lemma 2 of Zhu and Nowak [2022]). *Let $(Z_t)_{t \leq T}$ be a real-valued sequence of
990 random variables adapted to a filtration \mathcal{F}_t . If $|Z_t| \leq B$ almost surely, then with probability at least
991 $1 - \delta$,*

$$\sum_{t=1}^T Z_t \leq \frac{3}{2} \sum_{t=1}^T \mathbb{E}_t[Z_t] + 4B \log(2\delta^{-1})$$

992 and

$$\sum_{t=1}^T \mathbb{E}_t[Z_t] \leq 2 \sum_{t=1}^T Z_t + 8B \log(2\delta^{-1}).$$

993 B Proof of Theorem 3.4

994 *Proof.* The proof proceeds in four steps. We first define a set of good events of concentration of
995 parameters and sums of trajectory features, we show that they are satisfied with high probability.
996 Second, we decompose the regret into two terms: a pessimism term and an estimation error term. We
997 then show that the pessimism term is controlled by establishing a constant probability of optimism,
998 and the estimation error term is controlled by the estimation error of the maximum likelihood
999 estimator.

1000 Throughout the proof, we work with the two filtrations $\mathcal{F}_t = \sigma(x_1, y_1, \dots, x_{t-1}, y_{t-1})$ and $\mathcal{F}_t^\theta =$
1001 $\sigma(\mathcal{F}_t, \tilde{\theta}_t)$. In particular, both $\hat{\theta}_t$ and V_t are \mathcal{F}_t measurable. Therefore, $\tilde{\theta}_t$ follows a Gaussian
1002 distribution given \mathcal{F}_t , *i.e.* $\tilde{\theta}_t | \mathcal{F}_t \sim \mathcal{N}(\hat{\theta}_t, \beta_t^2 V_t^{-1})$, while it is fully determined by \mathcal{F}_t^θ .

1003 *Step 1 (good events):* Recall that in Algorithm 1 we set $\varepsilon = 1/\sqrt{T}$ and $\delta' = \delta/5$. We define the
1004 following high probability events.

1005 1. Let $\delta'' = \delta'/T$ and $c(\delta'') := \sqrt{2d \log(2d/\delta'')} = \sqrt{2d \log(10dT/\delta)}$. Consider the inflated
 1006 ellipsoid

$$\mathcal{E}_t^{\text{TS}} := \left\{ \theta \in \mathbb{R}^d : \left\| \theta - \hat{\theta}_t \right\|_{V_t} \leq \beta_t(\delta') c(\delta'') \right\} = c(\delta'') \mathcal{E}_t(\delta').$$

1007 We define the events

$$\hat{E}_t := \{\theta^* \in \mathcal{E}_t\}, \quad \tilde{E}_t := \left\{ (1 - \alpha)\hat{\theta}_t + \tilde{\theta}_t \in \mathcal{E}_t^{\text{TS}} \right\}, \quad \text{and } E_t := \hat{E}_t \cap \tilde{E}_t$$

1008 and let $G_1 := \bigcap_{t=1}^T E_t$. This event implies that θ^* lies in the confidence set uniformly over
 1009 times $t \leq T$, and the sampled parameter is close to $\alpha\hat{\theta}_t$.

1010 2. Let G_2 denote the event that for

$$C_T := 2\sqrt{T \left(2d \log \left(1 + \frac{4TL^2 H_\gamma^2}{d\lambda} \right) + \frac{12dL^2 H_\gamma^2}{\log(2)\lambda} \log \left(1 + \frac{4L^2 H_\gamma^2}{\log(2)\lambda} \right) \right)} + \frac{16LH_\gamma}{\sqrt{\lambda}} \log \left(\frac{2}{\delta'} \right),$$

1011 it holds that

$$\sum_{t=1}^T \mathbb{E} \left[\left\| \phi(\tau_t) - \phi(\tau'_t) \right\|_{V_t^{-1}} \mid \mathcal{F}_t \right] \leq C_T,$$

1012 and that

$$\sum_{t=1}^T \mathbb{E} \left[\left\| \phi(\tau_t) - \phi(\tau'_t) \right\|_{V_t^{-1}} \mid \mathcal{F}_t^\theta \right] \leq C_T.$$

1013 3. Let $G_3 = \left\{ \forall t \in [1, T] : V_{\hat{\theta}_t}^* - V_{\hat{\theta}_t}^{\pi_t} \leq \varepsilon \right\}$, where $\pi_t = \mathbf{A}_{\text{RL}}^{\text{PAC}}(\tilde{\theta}_t, \varepsilon, \delta'/T)$.

1014 4. Let G_4 denote the event that $V_{\hat{\theta}_T}^* - V_{\hat{\theta}_T}^{\hat{\pi}} \leq \varepsilon$ where $\hat{\pi} = \mathbf{A}_{\text{RL}}^{\text{PAC}}(\bar{\theta}_T, \varepsilon, \delta')$.

1015 As shown below, with probability at least $1 - \delta$, all the above good events happen at the same time.

1016 **Proposition B.1.** Let $G := \bigcap_{i=1}^4 G_i$. It holds that $\Pr[G] \geq 1 - \delta$.

1017 *Proof.* For the event G_1 , Lemma 3.1 and A.2 imply that $\Pr \left[\bigcup_{t=1}^T \hat{E}_t^{\mathcal{C}} \right] \leq \delta'$ and $\Pr[\tilde{E}_t^{\mathcal{C}}] \leq \delta'/T$ for
 1018 any $t \in [1, T]$. Hence, a union bound yields $\Pr[G_1^{\mathcal{C}}] \leq 2\delta'$. Furthermore, by Lemma A.5 2), we
 1019 have $\Pr[G_2^{\mathcal{C}}] \leq \delta'$. Moreover, by union bound and the definition of $\mathbf{A}_{\text{RL}}^{\text{PAC}}$, we have $\Pr[G_3^{\mathcal{C}}] \leq \delta'$ and
 1020 $\Pr[G_4^{\mathcal{C}}] \leq \delta'$. Hence, we have $\Pr[G] \geq 1 - \sum_{i=1}^4 \Pr[G_i^{\mathcal{C}}] \geq 1 - \delta$. \square

1021 *Remark B.2.* Note that throughout this proof, we consider $\alpha = 1$. However, we defined the good
 1022 events for general $\alpha \in \{0, 1\}$ so we can reuse Proposition B.1 for the proof of Theorem 3.5, which
 1023 requires $\alpha = 0$ instead. In particular, the event G_4 is relevant only for the setting with $\alpha = 0$.

1024 *Step 2 (regret decomposition):* Recall that in algorithm 1, we choose $\pi'_t = \pi_{t-1}$. Hence, we can upper
 1025 bound the cumulative regret as follows

$$\begin{aligned} R(T) &= \frac{1}{2} \sum_{t=1}^T \left(2V_{\theta^*}^* - V_{\theta^*}^{\pi_t} - V_{\theta^*}^{\pi'_t} \right) \\ &\leq \sum_{t=1}^T (V_{\theta^*}^* - V_{\theta^*}^{\pi_t}) + V_{\theta^*}^* - V_{\theta^*}^{\pi'_0} \\ &\leq \sum_{t=1}^T \underbrace{(V_{\theta^*}^* - V_{\theta^*}^{\pi_t})}_{r_t} + BLH_\gamma. \end{aligned}$$

Let $\Delta_t(\theta) := \max_{\pi} \langle \theta, \phi(\pi) - \phi(\pi'_t) \rangle$. On the good event G , we can decompose the instantaneous regret as follows

$$\begin{aligned}
r_t &:= V_{\theta^*}^* - V_{\theta^*}^{\pi_t} \\
&= \left(V_{\theta^*}^* - V_{\theta^*}^{\pi'_t} \right) - \left(V_{\tilde{\theta}_t}^{\pi_t} - V_{\tilde{\theta}_t}^{\pi'_t} \right) + \left(V_{\tilde{\theta}_t}^{\pi_t} - V_{\tilde{\theta}_t}^{\pi'_t} \right) - \left(V_{\theta^*}^{\pi_t} - V_{\theta^*}^{\pi'_t} \right) \\
&\leq \left(V_{\theta^*}^* - V_{\theta^*}^{\pi'_t} \right) - \left(V_{\tilde{\theta}_t}^* - V_{\tilde{\theta}_t}^{\pi'_t} \right) + \varepsilon + \left(V_{\tilde{\theta}_t}^{\pi_t} - V_{\tilde{\theta}_t}^{\pi'_t} \right) - \left(V_{\theta^*}^{\pi_t} - V_{\theta^*}^{\pi'_t} \right) \\
&= \underbrace{\Delta_t(\theta^*) - \Delta_t(\tilde{\theta}_t)}_{r_t^{\text{TS}}} + \underbrace{\langle \tilde{\theta}_t - \theta^*, \phi(\pi_t) - \phi(\pi'_t) \rangle}_{r_t^{\text{MLE}}} + \varepsilon.
\end{aligned}$$

Here, r_t^{TS} is a pessimism term that is negative by construction for optimistic algorithms, and r_t^{MLE} is related to the estimation error of the reward parameter.

Step 3 (bounding r_t^{TS}): This part of the analysis highlights the distinctiveness of randomized exploration. While optimistic algorithms ensure negativity of r_t^{TS} through their intractable optimization procedures, randomized exploration controls it using probability arguments. Specifically, following the proof of Abeille and Lazaric [2017], we begin by bounding r_t^{TS} on the good event via a conditional expectation given the optimism event. The bound on r_t^{TS} then follows by a careful application of the anti-concentration property established in Lemma A.3.

Conditioned on \tilde{E}_t , we can lower bound

$$\Delta_t(\tilde{\theta}_t) \geq \min_{\theta \in \mathcal{E}_t^{\text{TS}}} \Delta_t(\theta) = \max_{\pi} \langle \underline{\theta}_t, \phi(\pi) - \phi(\pi'_t) \rangle =: \underline{\Delta}_t,$$

for some $\underline{\theta}_t \in \mathcal{E}_t^{\text{TS}}$. Moreover, if $O_t := \left\{ \Delta_t(\tilde{\theta}_t) \geq \Delta_t(\theta^*) \right\}$ denotes the event of $\tilde{\theta}_t$ being optimistic at time t , we can upper bound $\Delta_t(\theta^*)$ as follows

$$\Delta_t(\theta^*) \mathbb{1}(E_t) \leq \mathbb{E} \left[\Delta_t(\tilde{\theta}_t) \mathbb{1}(E_t) \mid \mathcal{F}_t, O_t \right].$$

Putting this together, we get

$$\begin{aligned}
r_t^{\text{TS}} \mathbb{1}(E_t) &\leq \mathbb{E} \left[\left(\Delta_t(\tilde{\theta}_t) - \underline{\Delta}_t \right) \mathbb{1}(E_t) \mid \mathcal{F}_t, O_t \right] \\
&\stackrel{(i)}{\leq} \mathbb{E} \left[\left(\langle \tilde{\theta}_t, \phi(\pi_t) - \phi(\pi'_t) \rangle + \varepsilon - \max_{\pi} \langle \underline{\theta}_t, \phi(\pi) - \phi(\pi'_t) \rangle \right) \mathbb{1}(E_t) \mid \mathcal{F}_t, O_t \right] \\
&\stackrel{(ii)}{=} \mathbb{E} \left[\left(\langle \tilde{\theta}_t, \phi(\pi_t) - \phi(\pi'_t) \rangle - \langle \underline{\theta}_t, \phi(\pi_t) - \phi(\pi'_t) \rangle \right) \mathbb{1}(E_t) \mid \mathcal{F}_t, O_t \right] + \varepsilon \\
&\stackrel{(iii)}{\leq} 2\beta_t(\delta')c(\delta'')\mathbb{E} \left[\|\phi(\pi_t) - \phi(\pi'_t)\|_{V_t^{-1}} \mid \mathcal{F}_t, \hat{E}_t, O_t \right] \Pr \left[\hat{E}_t \mid \mathcal{F}_t \right] + \varepsilon,
\end{aligned}$$

where (i) follows from ε -optimality of π_t , in (ii) we used that $\max_{\pi} \langle \underline{\theta}_t, \phi(\pi) - \phi(\pi_t) \rangle \geq 0$, and (iii) follows from the Cauchy-Schwarz inequality. By the law of total probability, we have that

$$\mathbb{E} \left[\|\phi(\pi_t) - \phi(\pi'_t)\|_{V_t^{-1}} \mid \mathcal{F}_t, \hat{E}_t, O_t \right] \leq \mathbb{E} \left[\|\phi(\pi_t) - \phi(\pi'_t)\|_{V_t^{-1}} \mid \mathcal{F}_t, \hat{E}_t \right] / \Pr \left[O_t \mid \mathcal{F}_t, \hat{E}_t \right].$$

Next, as $\theta^* \in \mathcal{E}_t$ on \hat{E}_t , we have

$$\Pr \left[O_t \mid \mathcal{F}_t, \hat{E}_t \right] \geq \Pr \left[\Delta_t(\tilde{\theta}_t) \geq \max_{\theta \in \mathcal{E}_t} \Delta_t(\theta) \mid \mathcal{F}_t \right].$$

Since $\theta \mapsto \Delta_t(\theta)$ is the sum of a linear function and the function $\theta \mapsto \max_{\pi} \langle \theta, \phi(\pi) \rangle$ which is convex and continuous by Proposition A.7. Then, applying Lemma A.3 with $f(\theta) = \Delta_t(\theta)$, $\tilde{x} = \tilde{\theta}_t \mid \mathcal{F}_t$, and $\mathcal{E} = \mathcal{E}_t$, we have

$$\Pr \left[O_t \mid \mathcal{F}_t, \hat{E}_t \right] \geq 1 / (4\sqrt{e\pi}) =: p.$$

As a result, we can upper bound the instantaneous regret as

$$\begin{aligned}
r_t^{\text{TS}} \mathbb{1}(E_t) &\leq \frac{2\beta_t(\delta')c(\delta'')\mathbb{E} \left[\|\phi(\pi_t) - \phi(\pi'_t)\|_{V_t^{-1}} \mid \mathcal{F}_t \right]}{p} + \varepsilon \\
&\leq \frac{2\beta_t(\delta')c(\delta'')\mathbb{E} \left[\|\phi(\tau_t) - \phi(\tau'_t)\|_{V_t^{-1}} \mid \mathcal{F}_t \right]}{p} + \varepsilon,
\end{aligned}$$

1047 where the second inequality follows from the convexity of norms. Applying Lemma A.5 2), we have
 1048 on the good event G that

$$\begin{aligned} \sum_{t=1}^T r_t^{\text{TS}} &\leq \frac{2\beta_T(\delta')c(\delta'')}{p} \sum_{t=1}^T \mathbb{E} \left[\|\phi(\tau_t) - \phi(\tau'_t)\|_{V_t^{-1}} \mid \mathcal{F}_t \right] + T\varepsilon \\ &\leq \frac{2\beta_T(\delta')c(\delta'')}{p} C_T + T\varepsilon, \end{aligned} \quad (5)$$

1049 for the constant

$$C_T = 2\sqrt{T \left(2d \log \left(1 + \frac{4TL^2H_\gamma^2}{d\lambda} \right) + \frac{12dL^2H_\gamma^2}{\log(2)\lambda} \log \left(1 + \frac{4L^2H_\gamma^2}{\log(2)\lambda} \right) \right)} + \frac{16LH_\gamma}{\sqrt{\lambda}} \log \left(\frac{2}{\delta'} \right).$$

1050 *Step 4 (bounding r_t^{MLE}):* Conditioned on event E_t we have

$$\begin{aligned} r_t^{\text{MLE}} &= \langle \tilde{\theta}_t - \theta^*, \phi(\pi_t) - \phi(\pi'_t) \rangle \\ &= \langle \tilde{\theta}_t - \hat{\theta}_t, \phi(\pi_t) - \phi(\pi'_t) \rangle + \langle \hat{\theta}_t - \theta^*, \phi(\pi_t) - \phi(\pi'_t) \rangle \\ &\leq \beta_t(\delta')(1 + c(\delta'')) \|\phi(\pi_t) - \phi(\pi'_t)\|_{V_t^{-1}} \\ &\leq \beta_t(\delta')(1 + c(\delta'')) \mathbb{E} \left[\|\phi(\tau_t) - \phi(\tau'_t)\|_{V_t^{-1}} \mid \mathcal{F}_t^\theta \right], \end{aligned}$$

1051 where the last inequality follows from the convexity of norms. From here, we can again apply the
 1052 second result of Lemma A.5. We deduce that on the good event G , we have

$$\begin{aligned} \sum_{t=1}^T r_t^{\text{MLE}} &\leq \beta_T(\delta')(1 + c(\delta'')) \sum_{t=1}^T \mathbb{E} \left[\|\phi(\tau_t) - \phi(\tau'_t)\|_{V_t^{-1}} \mid \mathcal{F}_t^\theta \right] \\ &\leq \beta_T(\delta')(1 + c(\delta'')) C_T. \end{aligned}$$

1053 In summary, we can conclude that with probability at least $1 - \delta$, the regret can be bounded as follows

$$\begin{aligned} R(T) &\leq \sum_{t=1}^T (r_t^{\text{TS}} + r_t^{\text{MLE}}) + T\varepsilon + BLH_\gamma \\ &\leq \left(\frac{2\beta_T(\delta')c(\delta'')}{p} + \beta_T(\delta')(1 + c(\delta'')) \right) C_T + 2T\varepsilon + BLH_\gamma \\ &\leq \frac{3\beta_T(\delta')c(\delta'')}{p} C_T + 2T\varepsilon + BLH_\gamma \\ &= \frac{3 \left[\sqrt{\kappa \left[\log \left(\frac{5}{\delta} \right) + d \log \left(\max \left\{ e, \frac{4eBLH_\gamma(T-1)}{d} \right\} \right) \right]} + 2\sqrt{\lambda}B \right] \sqrt{2d \log(10dT/\delta)}}{p} \\ &\quad \cdot \left[2\sqrt{T \left(2d \log \left(1 + \frac{4TL^2H_\gamma^2}{d\lambda} \right) + \frac{12dL^2H_\gamma^2}{\log(2)\lambda} \log \left(1 + \frac{4L^2H_\gamma^2}{\log(2)\lambda} \right) \right)} + \frac{16LH_\gamma}{\sqrt{\lambda}} \log \left(\frac{10}{\delta} \right) \right] \\ &\quad + 2\sqrt{T} + BLH_\gamma \\ &= \mathcal{O} \left(\sqrt{\kappa d^3 T \log \left(\frac{dT}{\delta} \right)^3} \right). \end{aligned}$$

1054 □

1055 C Proof of Theorem 4.1

1056 We start the proof by showing that the number of rounds where the design matrix is updated is small.

1057 **Lemma C.1** (Number of design matrix updates). *Using Algorithm 2 with a parameter $C > 0$, it*
 1058 *holds that:*

$$\sum_{t=1}^T \mathbb{1}\{V_t \neq V_{t-1}\} \leq \frac{d}{\log\left(\frac{1+C}{1-1/e}\right)} \log\left(1 + \frac{T(LH_\gamma)^2}{\lambda d}\right).$$

1059 *That is, the number of updates of the matrix V_t is at most logarithmic in the number of interactions T .*

1060 *Proof.* Denote $N_T = \sum_{t=1}^T \mathbb{1}[V_t \neq V_{t-1}]$, and let $T' < T$ be the last time the matrix V_t was updated.
 1061 Then,

$$\begin{aligned} \det(V_T) &\geq (1+C)/(1-1/e) \det(V_{T'}) \\ &\geq ((1+C)/(1-1/e))^{N_T} \det(\lambda I). \end{aligned}$$

1062 The first inequality follows because $\det W_T \geq (1+C) \det V_{T'}$ and because greedy D-Optimal
 1063 design ensures that $\det V_T \geq \frac{1}{1-1/e} \det W_T$, where W_t is defined in Algorithm 2.

1064 Then, using the trace-determinant inequality, we have:

$$\left(\frac{1+C}{1-1/e}\right)^{N_T} \lambda^d \leq \left(\frac{\lambda d + T(LH_\gamma)^2}{d}\right)^d$$

1065 and then,

$$N_T \leq \frac{d}{\log\left(\frac{1+C}{1-1/e}\right)} \log\left(1 + \frac{T(LH_\gamma)^2}{\lambda d}\right).$$

1066 □

1067 We now present the proof for the regret bound, which proceeds similarly to that of Theorem 3.4 up
 1068 to some modifications. The first change is in the regret decomposition, which needs to be adapted
 1069 because we no longer compare to the past policy but rather to $\pi'_t = \mathbf{A}_{\text{RL}}^{\text{PAC}}(\hat{\theta}_{t_{\text{stop}}(t)}, \varepsilon, \delta)$. The second
 1070 change is using Lemma A.6 instead of Lemma A.5 to bound the sum of norms of trajectory features.
 1071 Finally, the good events defined in the proof of Theorem 3.4 are slightly modified to account for the
 1072 lazy design matrix and the new choice of comparator policy π'_t .

1073 *Proof.* Let us first define the function $t_{\text{stop}} : \mathbb{N} \rightarrow \mathbb{N}$, which to a time t , assigns the last time
 1074 $t_{\text{stop}}(t) \leq t$ that the update condition (line 4 in Algorithm 2) was met.

1075 We work with the same two filtrations as before $\mathcal{F}_t = \sigma(x_1, y_1, \dots, x_{t-1}, y_{t-1})$ and $\mathcal{F}_t^\theta = \sigma(\mathcal{F}_t, \tilde{\theta}_t)$.
 1076 And we recall that, given \mathcal{F}_t , $\tilde{\theta}_t$ in Algorithm 2 is sampled from $\mathcal{N}(\hat{\theta}_{t_{\text{stop}}(t)}, \beta_{t_{\text{stop}}(t)}^2 V_{t_{\text{stop}}(t)}^{-1})$.

1077 *Step 1 (good events):* Consider $\delta' = \delta/4$, we redefine the high-probability events:

1078 1. Let $\delta'' = \delta'/T$ and $c(\delta'') := \sqrt{2d \log(2d/\delta'')}$. Consider the inflated ellipsoid

$$\mathcal{E}_{t_{\text{stop}}(t)}^{\text{TS}} := \left\{ \theta \in \mathbb{R}^d : \left\| \theta - \hat{\theta}_{t_{\text{stop}}(t)} \right\|_{V_{t_{\text{stop}}(t)}} \leq \beta_{t_{\text{stop}}(t)}(\delta') c(\delta'') \right\}.$$

1079 We define the events $E_t := \hat{E}_t \cap \tilde{E}_t$, $\hat{E}_t := \{\theta^* \in \mathcal{E}_{t_{\text{stop}}(t)}\}$, $\tilde{E}_t := \{\tilde{\theta}_t \in \mathcal{E}_{t_{\text{stop}}(t)}^{\text{TS}}\}$,
 1080 and let $G_1 := \bigcap_{t=1}^T E_t$.

1081 2. Let $C_T := 2\sqrt{T\left(\frac{8(1+C)d}{(1-1/e)^2} \log\left(1 + \frac{4TL^2H_\gamma^2}{d\lambda}\right) + \frac{12dL^2H_\gamma^2}{\log(2)\lambda} \log\left(1 + \frac{4L^2H_\gamma^2}{\log(2)\lambda}\right)\right) + \frac{16LH_\gamma}{\sqrt{\lambda}} \log\left(\frac{2}{\delta'}\right)}$, and
 1082 define G_2 as the event under which it holds that

$$\sum_{t=1}^T \mathbb{E} \left[\left\| \phi(\tau_t) - \phi(\tau'_t) \right\|_{V_{t_{\text{stop}}(t)}^{-1}} \mid \mathcal{F}_t \right] \leq C_T,$$

1083 and

$$\sum_{t=1}^T \mathbb{E} \left[\left\| \phi(\tau_t) - \phi(\tau'_t) \right\|_{V_{t_{\text{stop}}(t)}^{-1}} \mid \mathcal{F}_t^\theta \right] \leq C_T.$$

1084 3. Let $G_3 = \left\{ \forall t \in [1, T] : V_{\theta_t^*}^* - V_{\hat{\theta}_t}^{\pi_t} \leq \varepsilon \right\}$, where $\pi_t = \mathbf{A}_{\text{RL}}^{\text{PAC}}(\tilde{\theta}_t, \varepsilon, \delta'/T)$.

1085 We also define the intersection, $G := \bigcap_{i=1}^3 G_i$, of all good events.

1086 We now compare each of these events to their counterparts in the proof of Theorem 3.4. The event
 1087 G_1 is modified because we replace V_t by $V_{t_{\text{stop}}(t)}$ and $\hat{\theta}_t$ by $\hat{\theta}_{t_{\text{stop}}(t)}$. The event G_1 still holds with
 1088 probability at least $1 - 2\delta'$ using the same concentration arguments as before. The event G_2 is also
 1089 modified to account for the lazy design matrix, and it holds with probability $1 - \delta'$ thanks to Lemma
 1090 A.6. Finally, the event G_3 remains unchanged.

1091 We conclude that G happens with probability at least $1 - \delta$.

1092 *Step 2 (regret decomposition):* Since Algorithm 2 uses a different comparator policy π'_t than Algorithm
 1093 1, we derive a new regret decomposition.

$$\begin{aligned} R(T) &= \frac{1}{2} \sum_{t=1}^T \left(2V_{\theta^*}^* - V_{\theta^*}^{\pi_t} - V_{\theta^*}^{\pi'_t} \right) \\ &= \frac{1}{2} \sum_{t=1}^T \left((V_{\theta^*}^* - V_{\theta^*}^{\pi_t}) + \langle \theta^*, \phi(\pi^*) - \phi(\pi'_t) \rangle \right) \\ &= \frac{1}{2} \sum_{t=1}^T \left((V_{\theta^*}^* - V_{\theta^*}^{\pi_t}) + \langle \theta^*, \phi(\pi^*) - \phi(\pi_t) \rangle + \langle \theta^*, \phi(\pi_t) - \phi(\pi'_t) \rangle \right) \\ &= \frac{1}{2} \sum_{t=1}^T \left(2(V_{\theta^*}^* - V_{\theta^*}^{\pi_t}) + \langle \theta^* - \hat{\theta}_{t_{\text{stop}}(t)}, \phi(\pi_t) - \phi(\pi'_t) \rangle + \langle \hat{\theta}_{t_{\text{stop}}(t)}, \phi(\pi_t) - \phi(\pi'_t) \rangle \right) \\ &\leq \frac{1}{2} \sum_{t=1}^T \left(2 \underbrace{(V_{\theta^*}^* - V_{\theta^*}^{\pi_t})}_{r_t} + \left\| \theta^* - \hat{\theta}_{t_{\text{stop}}(t)} \right\|_{V_{t_{\text{stop}}(t)}} \left\| \phi(\pi_t) - \phi(\pi'_t) \right\|_{V_{t_{\text{stop}}(t)}^{-1}} + \varepsilon \right), \end{aligned}$$

1094 where the last line follows from the Cauchy-Schwarz inequality and because $\pi'_t = \mathbf{A}_{\text{RL}}^{\text{PAC}}(\hat{\theta}_{t_{\text{stop}}(t)}, \varepsilon, \delta)$.

1095 The second term in the decomposition above can be bounded on the good event G as:

$$\begin{aligned} \sum_{t=1}^T \left\| \theta^* - \hat{\theta}_{t_{\text{stop}}(t)} \right\|_{V_{t_{\text{stop}}(t)}} \left\| \phi(\pi_t) - \phi(\pi'_t) \right\|_{V_{t_{\text{stop}}(t)}^{-1}} &\leq \beta_T \sum_{t=1}^T \left\| \phi(\pi_t) - \phi(\pi'_t) \right\|_{V_{t_{\text{stop}}(t)}^{-1}} \\ &\leq \beta_T C_T \end{aligned}$$

1096 where $C_T = 2\sqrt{T \left(\frac{8(1+C)d}{(1-1/e)^2} \log \left(1 + \frac{4TL^2H_1^2}{d\lambda} \right) + \frac{12dL^2H_2^2}{\log(2)\lambda} \log \left(1 + \frac{4L^2H_2^2}{\log(2)\lambda} \right) \right)} + \frac{16LH\gamma}{\sqrt{\lambda}} \log \left(\frac{2}{\delta'} \right)$.

1097 For the first term in the regret decomposition, similarly to Appendix B, we have that:

$$r_t = V_{\theta^*}^* - V_{\theta^*}^{\pi_t} = \underbrace{\Delta_t(\theta^*) - \Delta_t(\tilde{\theta}_t)}_{r_t^{\text{TS}}} + \underbrace{\langle \tilde{\theta}_t - \theta^*, \phi(\pi_t) - \phi(\pi'_t) \rangle}_{r_t^{\text{MLE}}}$$

1098 where we recall the gap function $\Delta_t(\theta) := \max_{\pi} \langle \theta, \phi(\pi) - \phi(\pi'_t) \rangle$.

1099 *Step 3 (bounding r_t^{TS}):* The proof for $\sum_t r_t^{\text{TS}}$ proceeds exactly like Appendix B up to Equation (5),
 1100 this is because the probability of the optimism event O_t and the events E_t and \hat{E}_t is unaffected by
 1101 the change to the algorithm. Then, we have that:

$$R^{\text{TS}}(T) = \sum_{t=1}^T r_t^{\text{TS}} \leq \frac{2\beta_T(\delta')c(\delta'')}{p} \sum_{t=1}^T \mathbb{E} \left[\left\| \phi(\tau_t) - \phi(\pi'_t) \right\|_{V_t^{-1}} \mid \mathcal{F}_t \right] + T\varepsilon.$$

1102 We can then conclude, on the good event G , that

$$R^{\text{TS}}(T) \leq \frac{2\beta_T(\delta')c(\delta'')}{p} C_T + T\varepsilon.$$

1103 *Step 4 (bounding r_t^{MLE}):* This proof is similar to that of the first term in the regret decomposition. The
 1104 only difference is that we use the confidence set for $\tilde{\theta}_t$ instead of $\hat{\theta}_{t_{\text{stop}}(t)}$. Then, it holds that:

$$\begin{aligned} \sum_{t=1}^T r_t^{\text{MLE}} &\leq \beta_T(\delta')(1 + c(\delta'')) \sum_{t=1}^T \|\phi(\pi_t) - \phi(\pi'_t)\|_{V_t^{-1}} \\ &\leq \beta_T(\delta')(1 + c(\delta'')) \sum_{t=1}^T \mathbb{E}[\|\phi(\pi_t) - \phi(\pi'_t)\|_{V_t^{-1}} | \mathcal{F}_t^\theta], \\ &\leq \beta_T(\delta')(1 + c(\delta'')) C_T \end{aligned}$$

1105 where the second inequality follows from the convexity of the norm.

1106 In summary, we conclude that with probability at least $1 - \delta$, the regret can be bounded as follows:

$$\begin{aligned} R(T) &\leq \frac{1}{2} \sum_{t=1}^T \left(2 \underbrace{(V_{\theta^*}^* - V_{\theta^*}^{\pi_t})}_{=r_t^{\text{TS}} + r_t^{\text{MLE}}} + \underbrace{\|\theta^* - \hat{\theta}_{t_{\text{stop}}(t)}\|_{V_{t_{\text{stop}}(t)}}}_{\leq \beta_T} \|\phi(\pi_t) - \phi(\pi'_t)\|_{V_{t_{\text{stop}}(t)}^{-1}} + \varepsilon \right) \\ &\leq \frac{2\beta_T(\delta')c(\delta'')}{p} C_T + T\varepsilon + \beta_T(\delta')(1 + c(\delta'')) C_T + \frac{\beta_T}{2} C_T + T\varepsilon. \end{aligned}$$

1107

□

1108 D Proof of Theorem 3.5

1109 *Proof.* Recall that in Algorithm 1, we set $\varepsilon = 1/\sqrt{T}$, $\delta' = \delta/5$, and $\hat{\pi} = \mathbf{A}_{\text{RL}}^{\text{PAC}}(\bar{\theta}_T, \varepsilon, \delta')$ for
 1110 $\bar{\theta}_T = \frac{1}{T} \sum_{t=1}^T \hat{\theta}_t$. Moreover, as $\alpha = 0$, we further have $\tilde{\theta}_t | \mathcal{F}_t \sim \mathcal{N}(0, \beta_t^2 V_t^{-1})$, and choose the
 1111 policies $\pi_t = \mathbf{A}_{\text{RL}}^{\text{PAC}}(\tilde{\theta}_t)$ and $\pi'_t = \pi_{t-1}$.

1112 In the following, we will denote $\pi^* \in \arg \max_{\pi} \langle \theta^*, \phi(\pi) \rangle$ for an arbitrary optimal policy corre-
 1113 sponding to the ground truth reward parameter θ^* . Under the good event G of Proposition B.1, we
 1114 can bound the suboptimality of $\hat{\pi}$ as follows

$$\begin{aligned} \text{SubOpt}(\hat{\pi}) &= \langle \theta^*, \phi(\pi^*) - \phi(\hat{\pi}) \rangle \\ &\leq \langle \theta^* - \bar{\theta}_T, \phi(\pi^*) - \phi(\hat{\pi}) \rangle + \varepsilon \\ &= \varepsilon + \frac{1}{T} \sum_{t=1}^T \langle \theta^* - \hat{\theta}_t, \phi(\pi^*) - \phi(\hat{\pi}) \rangle, \end{aligned}$$

1115 where we used ε -optimality of $\hat{\pi}$ and the definition of $\bar{\theta}_T$. Now, recall that conditioned on E_t , we
 1116 have $\theta^* \in \mathcal{E}_t$ and $\hat{\theta}_t + \tilde{\theta}_t \in \mathcal{E}_t^{\text{TS}}$. Define $\tilde{\mathcal{E}}_t = \{\theta : \theta + \hat{\theta} \in \mathcal{E}_t\}$. Then, on the good event G it holds
 1117 that:

$$\begin{aligned} \langle \theta^* - \hat{\theta}_t, \phi(\pi^*) - \phi(\hat{\pi}) \rangle \mathbb{1}(E_t) &\leq \max_{\pi, \pi'} \max_{\theta \in \mathcal{E}_t} \langle \theta - \hat{\theta}_t, \phi(\pi) - \phi(\pi') \rangle \mathbb{1}(\tilde{E}_t) \\ &= \max_{\pi, \pi'} \max_{\theta \in \tilde{\mathcal{E}}_t} \langle \theta, \phi(\pi) - \phi(\pi') \rangle \mathbb{1}(\tilde{E}_t) \\ &\stackrel{(i)}{=} \beta_t \max_{\pi, \pi'} \|\phi(\pi) - \phi(\pi')\|_{V_t^{-1}} \mathbb{1}(\tilde{E}_t) \\ &\stackrel{(ii)}{\leq} \beta_t \max_{\pi, \pi'} \left(\|\phi(\pi) - \phi(\pi'_t)\|_{V_t^{-1}} + \|\phi(\pi'_t) - \phi(\pi')\|_{V_t^{-1}} \right) \mathbb{1}(\tilde{E}_t) \\ &= 2\beta_t \max_{\pi} \|\phi(\pi) - \phi(\pi'_t)\|_{V_t^{-1}} \mathbb{1}(\tilde{E}_t) \\ &\stackrel{(iii)}{=} 2 \max_{\pi} \max_{\theta \in \tilde{\mathcal{E}}_t} \langle \theta, \phi(\pi) - \phi(\pi'_t) \rangle \mathbb{1}(\tilde{E}_t) \\ &= 2 \max_{\theta \in \tilde{\mathcal{E}}_t} \max_{\pi} \langle \theta, \phi(\pi) - \phi(\pi'_t) \rangle \mathbb{1}(\tilde{E}_t). \end{aligned}$$

Here, (i) and (iii) follow because $\tilde{\mathcal{E}}_t$ is a centred ellipsoid and (ii) from the triangle inequality. Next, we define the non-negative function $f_t(\theta) := \max_{\pi} \langle \theta, \phi(\pi) - \phi(\pi'_t) \rangle$ and the event $O_t = \{f_t(\tilde{\theta}_t) \geq \max_{\theta \in \tilde{\mathcal{E}}_t} f_t(\theta)\}$. By Proposition A.7 f_t is convex and continuous. Applying Lemma A.3 with $f = f_t$, $\tilde{x} = \tilde{\theta}_t \mid \mathcal{F}_t$, and $\mathcal{E} = \tilde{\mathcal{E}}_t$, it holds that $\Pr [O_t \mid \mathcal{F}_t] \geq p := 1/(4\sqrt{e\pi})$. Hence,

$$\begin{aligned} \langle \theta^* - \hat{\theta}_t, \phi(\pi^*) - \phi(\hat{\pi}) \rangle \mathbb{1}(E_t) &\leq 2 \max_{\theta \in \tilde{\mathcal{E}}_t} \max_{\pi} \langle \theta, \phi(\pi) - \phi(\pi'_t) \rangle \mathbb{1}(\tilde{E}_t) \\ &\leq 2\mathbb{E} \left[\max_{\pi} \langle \tilde{\theta}_t, \phi(\pi) - \phi(\pi'_t) \rangle \mathbb{1}(\tilde{E}_t) \mid O_t, \mathcal{F}_t \right] \\ &\leq 2\mathbb{E} \left[\langle \tilde{\theta}_t, \phi(\pi_t) - \phi(\pi'_t) \rangle \mathbb{1}(\tilde{E}_t) \mid O_t, \mathcal{F}_t \right] + 2\varepsilon \\ &\leq 2\beta_t(\delta')c(\delta'')\mathbb{E} \left[\|\phi(\pi_t) - \phi(\pi'_t)\|_{V_t^{-1}} \mid O_t, \mathcal{F}_t \right] + 2\varepsilon \\ &\leq \frac{2\beta_t(\delta')c(\delta'')}{p} \mathbb{E} \left[\|\phi(\pi_t) - \phi(\pi'_t)\|_{V_t^{-1}} \mid \mathcal{F}_t \right] + 2\varepsilon. \end{aligned}$$

Using Lemma A.5, we conclude that with probability $1 - \delta$, we have that

$$\begin{aligned} \text{SubOpt}(\hat{\pi}) &\leq 3\varepsilon + \frac{2\beta_t(\delta')c(\delta'')}{pT} \sum_{t=1}^T \mathbb{E} \left[\|\phi(\pi_t) - \phi(\pi'_t)\|_{V_t^{-1}} \mid \mathcal{F}_t \right] \\ &\leq \frac{3}{\sqrt{T}} + \frac{2 \left[\sqrt{\kappa \left[\log\left(\frac{5}{\delta}\right) + d \log \left(\max \left\{ e, \frac{4eBLH_\gamma(T-1)}{d} \right\} \right) \right]} + 2\sqrt{\lambda}B \right] \sqrt{2d \log(10dT/\delta)}}{pT} \\ &\quad \cdot \left[2\sqrt{T \left(2d \log \left(1 + \frac{4TL^2H_\gamma^2}{d\lambda} \right) + \frac{12dL^2H_\gamma^2}{\log(2)\lambda} \log \left(1 + \frac{4L^2H_\gamma^2}{\log(2)\lambda} \right) \right)} + \frac{16LH_\gamma}{\sqrt{\lambda}} \log \left(\frac{10}{\delta} \right) \right] \\ &= \mathcal{O} \left(\sqrt{\frac{\kappa d^3}{T} \log \left(\frac{dT}{\delta} \right)^3} \right). \end{aligned}$$

□

E Discussion and background

E.1 Background on optimal experimental design

Given a possibly infinite set of features $\mathcal{X} \subset \mathbb{R}^d$, a D-optimal design is defined as a distribution π such that:

$$\pi \in \arg \max_{\pi \in \Delta_{\mathcal{X}}} \log \det \left(\sum_{x \in \mathcal{X}} \pi(x) x x^\top \right).$$

The Kiefer-Wolfowitz theorem, see [Kiefer and Wolfowitz, 1960], shows that a D-optimal design also ensures that $\max_{x \in \mathcal{X}} \|x\|_{\left(\sum_{x \in \mathcal{X}} \pi(x) x x^\top\right)^{-1}}^2 = d$. A direct consequence is that \mathcal{X} is a subset of the ellipsoid $\{x \in \mathbb{R}^d : \|x\|_{\left(\sum_{x \in \mathcal{X}} \pi(x) x x^\top\right)^{-1}}^2 \leq d\}$. The Kiefer-Wolfowitz theorem can be interpreted by saying that the set $\{x \in \mathbb{R}^d : \|x\|_{\left(\sum_{x \in \mathcal{X}} \pi(x) x x^\top\right)^{-1}}^2 \leq d\}$ is the minimum volume ellipsoid containing \mathcal{X} , see [Lattimore and Szepesvári, 2020, Theorem 21.1].

In the case where we have a budget of n samples for the design, we can define D-optimal design as the best allocation $\{n_x\}_{x \in \mathcal{X}} \in \mathbb{N}$ of the n samples such that $\log \det \left(\sum_{x \in \mathcal{X}} n_x x x^\top \right)$ is maximized and $\sum_{x \in \mathcal{X}} n_x = n$. Finding a D-optimal design with a budget of n is a challenging combinatorial optimization problem. However, there is a key property of the $\log \det$ function that enables an efficient approximation scheme. Namely, for a set $\mathcal{X} \subset \mathbb{R}^d$, the set function $S \subset \mathcal{X} \rightarrow \log \det \left(\lambda I + \sum_{x \in S} x x^\top \right)$ is submodular if λ is greater than one. Submodularity is a property describing decreasing additional benefit, known as diminishing returns. Fortunately, a submodular set function can be approximately maximized using a greedy algorithm, [Nemhauser et al., 1978].

Therefore, using Algorithm 3, we know that if t_{stop} and t'_{stop} are two consecutive update times, then the design matrix satisfies: $\log \det \left(V_{t'_{\text{stop}}} \right) \geq (1 - 1/e) \max_{S \subset \mathcal{X}} \log \det \left(V_{t_{\text{stop}}} + \sum_{x \in S} x x^\top \right)$.

E.2 On the intractability of optimistic approaches

Optimism in the face of uncertainty is a widely used principle for regret minimization and exploration in reinforcement learning. In the bandit setting, optimistic algorithms can be applied directly and yield minimax-optimal regret bounds [Auer, 2002]. However, in more general settings—such as linear bandits or MDPs—these approaches often lead to intractable optimization problems.

Optimism for regret minimization. For regret minimization in RLHF, an optimistic algorithm would choose the policy π_t to maximize the upper confidence bound on the reward difference relative to a comparator policy π'_t , typically the previous policy:

$$\pi_t = \arg \max_{\pi} \max_{\theta \in \mathcal{E}_t} V_{\theta}^{\pi} - V_{\theta}^{\pi'_t} = \arg \max_{\pi} V_{\theta_t}^{\pi} + \beta_t \|\phi(\pi) - \phi(\pi'_t)\|_{V_t^{-1}},$$

where \mathcal{E}_t is a confidence set for θ^* . This leads to the following bound on the instantaneous regret:

$$\begin{aligned} r_t &= \left(V_{\theta^*}^* - V_{\theta^*}^{\pi'_t} \right) - \left(V_{\theta^*}^{\pi_t} - V_{\theta^*}^{\pi'_t} \right) \\ &\leq \left(\max_{\pi} \max_{\theta \in \mathcal{E}_t} V_{\theta}^{\pi} - V_{\theta}^{\pi'_t} \right) - \left(V_{\theta^*}^{\pi_t} - V_{\theta^*}^{\pi'_t} \right) \\ &= \left(V_{\theta_t}^{\pi_t} - V_{\theta_t}^{\pi'_t} \right) - \left(V_{\theta^*}^{\pi_t} - V_{\theta^*}^{\pi'_t} \right) \\ &\leq \left\| \tilde{\theta}_t - \theta^* \right\|_{V_t} \|\phi(\pi_t) - \phi(\pi'_t)\|_{V_t^{-1}} \leq 2\beta_t(\delta') \|\phi(\pi_t) - \phi(\pi'_t)\|_{V_t^{-1}}, \end{aligned}$$

where $\tilde{\theta}_t \in \mathcal{E}_t$. The cumulative regret can then be bounded using standard elliptical potential arguments (see Lemma A.5).

Optimism for pure exploration. Optimistic algorithms are also successful for minimizing the suboptimality gap of the final policy. In this case, algorithms aim to maximize the information gain by selecting either a policy pair (π_t, π'_t) or a single policy π_t relative to a reference policy π'_t to maximize the exploration bonus:

$$\|\phi(\pi) - \phi(\pi')\|_{V_t^{-1}}.$$

These formulations are appealing due to their theoretical guarantees. However, not much can be said about the computational efficiency of the subsequent optimization problems.

Challenge of optimizing the exploration bonus. As shown by Efroni et al. [2021], optimizing these exploration bonuses is intractable even in tabular MDPs with a known transition law. Specifically, maximizing the bonus reduces to the following problem over the set of discounted occupancy measures:

$$\max_{\mu \in \mathcal{M}} \langle r_{\theta}, \mu \rangle + \beta_t \left\| \mathbb{E}_{(s,a) \sim \mu} \phi(s, a) - \phi(\pi'_t) \right\|_{V_t^{-1}},$$

where $\mathcal{M} := \{\mu^{\pi} : \mu^{\pi}(s, a) = \sum_{t=0}^{\infty} \gamma^t \mathbb{P}_{\pi}(s_t = s, a_t = a), \pi : \mathcal{S} \rightarrow \Delta_{\mathcal{A}}\}$ is the set of valid occupancy measures induced by stationary policies [Puterman, 2014]. The norm term is strictly convex in μ due to the strict convexity of the norm and the linearity of expectations. As a result, this is a convex maximization problem, which is NP-hard in general [Atamtürk and Gómez, 2017].

These intractability results highlight a major limitation of optimistic algorithms. Although they yield strong statistical guarantees, their practical implementation in general MDPs remains computationally challenging. This motivates exploring alternative approaches, such as randomized exploration, which can provide almost optimal theoretical guarantees while being computationally tractable.

F Experiments

Here, we provide additional details for experiments on Isaac Lab’s Isaac-Cartpole-v0 environment, as well as experimental results for the pure exploration version of our algorithm.

1175 **Implementation details** All experiments run on Isaac Lab’s unmodified Isaac-Cartpole-v0
 1176 environment using the default PPO configuration. We train over 30 RLHF iterations, using 30 steps
 1177 of PPO at each iteration, and training is repeated for 10 independent seeds. For the randomized
 1178 exploration, we set $\beta_t = 0.01 + \max(1, \log t)$ and $\lambda = 1$, and for lazy updates we set $C = 1$. At
 1179 each RLHF iteration we compare 5 independently sampled trajectories of horizon $H = 150$. For the
 1180 maximum likelihood estimation we perform 500 Adam steps (batch size 64, ℓ_2 penalty $\lambda = 10^{-1}$).
 1181 Experiments were executed on a single machine equipped with an Intel i9-14900KS CPU and an
 1182 NVIDIA RTX 4090 GPU; completing 30 RLHF iterations required approximately 2 min 50 s.

1183 **Pure exploration** Our results for the pure exploration setting with $\alpha = 0$ are shown in Figure 2 and
 1184 3 below. In Figure 2, we see that all three versions of RP0 achieve performance competitive to RL
 1185 with the ground truth reward³, but LRPO-OD needs the least preference queries. Moreover, in Figure 3
 1186 we see that the performance during training is poor, *i.e.* the regret is large, which is to be expected
 1187 due to the pure exploration scheme.

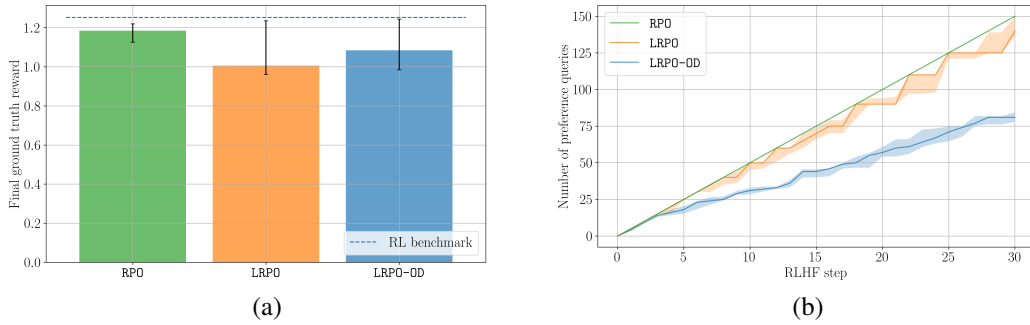


Figure 2: Comparison of RLHF algorithms in terms of (a) the last iterate ground truth reward $V_{\hat{\pi}^*}^{\pi^*}$ (estimated from samples) and (b) number of preference queries performed. In particular, for the choice $\alpha = 0$, we compare RP0 (green, Algorithm 1) with its lazy versions LRPO and LRPO-OD (orange & blue). Here, LRPO and LRPO-OD refer to Algorithm 2 without and with optimal design subroutine. The error bars indicate the 0.2 and 0.8 quantiles, across 10 independent runs. The dashed blue line indicates the mean reward achieved by PPO with the ground truth parameter θ^* .

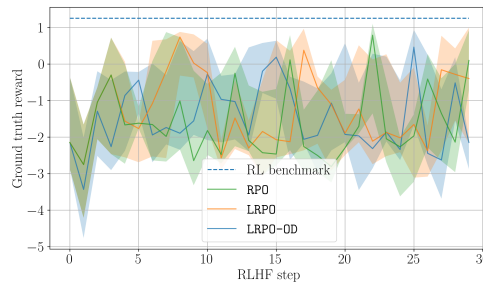


Figure 3: Comparing the ground truth rewards $V_{\hat{\pi}^*}^{\pi^*}$ (estimated from samples) of RLHF algorithms for pure exploration (*i.e.* $\alpha = 0$) during training, using the same color codes as in Figure 2.

³Note that the RL baseline makes $24 \times 4096 \times 50 = 4'915'200$ queries to the ground truth reward, whereas RP0 uses at most 150 binary preference queries.