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

Andreas Schlaginhaufen[*]
SYCAMORE, EPFL

Reda Ouhamma
SYCAMORE, EPFL

Maryam Kamgarpour
SYCAMORE, EPFL

Abstract

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

1 Introduction

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

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

Despite its empirical success, the theoretical foundations of RLHF are still in development. Existing works first studied the simpler setting of dueling bandits. In this context, the learner selects pairs of actions and observes noisy preference feedback [Yue et al., 2012, Komiyama et al., 2015]. Classical algorithms for regret minimization in this setting include approaches based on zeroth-order

[*]Correspondence to andreas.schlaginhaufen@epfl.ch.

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

 https://doi.org/10.52202/085713-0407

optimization [Yue and Joachims, 2009] or the principle of optimism [Ailon et al., 2014]. A key challenge in this setting is reducing the number of preference queries. For this purpose, several recent works propose 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 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, a central challenge remains open: designing tractable algorithms for active preference query selection that reduce the workload of human 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 computational tractability. For instance, Wu and Sun [2023] propose a randomized exploration algorithm with regret guarantees limited to linear dynamics, and Dwaracherla et al. [2024] show empirically that randomized exploration is efficient for fine-tuning of language models. In parallel, Wang et al. [2023] introduce a general reduction from RLHF to standard RL and establish 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. Leveraging randomized exploration for tractable and efficient preference query selection, we provide both online algorithms with regret guarantees and a preference-free algorithm with PAC-style guarantees under RL-oracle assumptions. Our contributions are as follows:

- We propose two meta-algorithms for RLHF using RL oracles: Algorithm 1, optimized for regret minimization, and Algorithm 2, which performs preference-free exploration and defers preference collection to a single batch at the end. For these methods, we establish regret and PAC-style guarantees, respectively, holding in general MDPs.
- We present a second meta-algorithm for regret minimization (Algorithm 3) 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 informative preference queries and improves sample efficiency.
- We provide empirical results showing that: Our algorithms are implementable, competitive with reward-based RL, and substantially outperform a baseline that relies solely on entropy-based exploration; Algorithm 3 achieves comparable performance to Algorithm 1 while significantly reducing the query complexity.[2]

[2]The code is openly accessible at `https://github.com/andrschl/isaac_rlhf`.

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 $\tilde{\mathcal{O}}(n) = \mathcal{O}(n \operatorname{polylog}(\mathrm{n}))$ for suppressing polylogarithmic terms.

Setting We consider an infinite-horizon[3] Markov decision process (MDP) $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, \nu_0, P, r, \gamma\}$ with state and action spaces $\mathcal{S} \subseteq \mathbb{R}^n$ 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 $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ is a continuous 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 $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$ as $\mathbb{P}_\pi$. For a trajectory $\tau = (s_0, a_0, s_1, \ldots) \in \mathcal{T}$, 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 $V_\theta^\pi := \mathbb{E}_{\tau \sim \mathbb{P}_\pi}\left[\sum_{h=0}^\infty \gamma^h r_\theta(s_h, a_h)\right] = \langle \theta, \phi(\pi) \rangle$ and the optimal value by $V_\theta^* := \max_\pi V_\theta^\pi$.

Interaction protocol For each round $t = 1, \ldots, T$ 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 $\mathcal{P} : \mathcal{T} \times \mathcal{T} \to [0, 1]$, 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 $\tau, \tau' \in \mathcal{T}$

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

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]. It captures that the probability of preferring τ over τ' is increasing in the difference between their values $\langle \theta^*, \phi(\tau) - \phi(\tau') \rangle$.

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

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

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.

[3] Our results extend directly to the finite-horizon setting as well.

https://doi.org/10.52202/085713-0407

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

$$\mathsf{SubOpt}(\hat{\pi}) := V^*_{\theta^*} - V^{\hat{\pi}}_{\theta^*}.$$

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.

3 Randomized Preference Optimization

3.1 Algorithm

Both algorithms share the same key ingredients: (i) estimating the reward parameter from preference feedback using maximum likelihood estimation, (ii) sampling reward parameters from a Gaussian distribution, reminiscent of linear Thompson sampling [Abeille and Lazaric, 2017], and (iii) leveraging an RL oracle to find an approximately optimal policy for the sampled reward parameter. In Algorithm 1, we apply these steps iteratively and, at each round, query the preference oracle by comparing a trajectory from the new policy to one from the previous policy to minimize regret. In contrast, Algorithm 2 performs preference-free exploration and defers all preference queries to a single batch at the end.

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 $\hat{\theta}_t = \arg\min_{\|\theta\| \le B} \mathcal{L}_{\mathcal{D}_t}(\theta)$, where

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

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 $\lambda \ge 0$ and define the design matrix at time t given by $V_t = \lambda I + \sum_{k=1}^{t-1} x_k x_k^\top$. Then, with probability $1-\delta$, for all $t \in \mathbb{N}$, the true reward parameter θ^* is contained in the ellipsoid*

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

Here, $\kappa := \max_{\theta\in\mathcal{B}^d(B), x\in\mathcal{B}^d(2LH_\gamma)} 1/\dot{\sigma}(\langle \theta, x \rangle)$ 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 result for likelihood-ratio confidence sets by Lee et al. [2024]. The proof and the precise constants are deferred to Appendix D.

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]. We adopt confidence sets based on V_t, as this facilitates preference-free exploration, and 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 rely on maximizing an exploration bonus of the form $\|\phi(\pi_t) - \phi(\pi'_t)\|_{V_t^{-1}}$ or $\mathbb{E}_{\tau\sim\mathbb{P}_{\pi_t}, \tau'\sim\mathbb{P}_{\pi'_t}} \|\phi(\tau) - \phi(\tau')\|_{V_t^{-1}}^2$. 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). 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, $\mathtt{A}_{\mathtt{RL}}^{\mathtt{PAC}}$, for the RL problem. That is, a polynomial-time algorithm that produces for every $\varepsilon > 0$, $\delta \in (0,1)$, and $\theta \in \mathbb{R}^d$ a policy $\pi = \mathtt{A}_{\mathtt{RL}}^{\mathtt{PAC}}(\theta, \varepsilon, \delta)$ 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 $\mathtt{A}_{\mathtt{RL}}^{\mathtt{PAC}}$ 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 randomized preference optimization (`RPO`) presented below comes in two variants: (i) `RPO-Regret` (Algorithm 1) which balances exploration and exploitation for regret minimization, and (ii) `RPO-Explore` (Algorithm 2) which performs preference-free exploration and collects a single batch of preferences at the end. In `RPO-Regret`, we sample in each round a reward parameter from a confidence set (see Lemma 3.1) inflated by $\sqrt{d}$, compute the policy π_t via the RL oracle, and update the reward estimate by maximum likelihood using the newly collected preference. In contrast, `RPO-Explore` samples reward parameters centered at zero, stores trajectory pairs during exploration, and collects a single batch of preferences at the end.

Algorithm 1: `RPO-Regret` (online preference learning for regret minimization)

Input: $T \in \mathbb{N},\ \delta \in (0,1),\ \lambda > 0$

Initialize: $\varepsilon = 1/\sqrt{T}$; $\delta' = \delta/5$; $V_1 = \lambda I$; $\mathcal{D}_1 = \emptyset$; $\hat{\theta}_1 = 0$; and choose π_0.

for $t = 1, 2, \ldots, T$ **do**

- $\tilde{\theta}_t \sim \mathcal{N}\big(\hat{\theta}_t, \beta_t(\delta')^2 V_t^{-1}\big)$ `// Reward sampling`
- $\pi_t = \mathtt{A}_{\mathtt{RL}}^{\mathtt{PAC}}\big(\tilde{\theta}_t, \varepsilon, \delta'/T\big), \pi_t' = \pi_{t-1}$ `// RL update`
- $\tau_t \sim \mathbb{P}_{\pi_t}, \tau_t' \sim \mathbb{P}_{\pi_t'}$
- $x_t = \phi(\tau_t) - \phi(\tau_t'), V_{t+1} = V_t + x_t x_t^\top$
- $y_t = \mathbb{1}(\tau_t \succ \tau_t'), \mathcal{D}_{t+1} = \mathcal{D}_t \cup \{(x_t, y_t)\}$ `// Preference feedback`
- $\hat{\theta}_{t+1} \in \arg\min_{\|\theta\| \leq B} \mathcal{L}_{\mathcal{D}_{t+1}}(\theta)$ `// Reward estimation`

end

Algorithm 2: `RPO-Explore` (preference-free exploration and batched reward estimation)

Input: $T \in \mathbb{N},\ \delta \in (0,1),\ \lambda > 0$

Initialize: $\varepsilon = 1/\sqrt{T}$; $\delta' = \delta/5$; $V_1 = \lambda I$; $\mathcal{B} = \mathcal{D} = \emptyset$; and choose π_0.

for $t = 1, 2, \ldots, T$ **do**

- $\tilde{\theta}_t \sim \mathcal{N}\big(0, V_t^{-1}\big)$ `// Reward sampling`
- $\pi_t = \mathtt{A}_{\mathtt{RL}}^{\mathtt{PAC}}\big(\tilde{\theta}_t, \varepsilon, \delta'/T\big),\ \pi_t' = \pi_{t-1}$ `// RL update`
- $\tau_t \sim \mathbb{P}_{\pi_t},\ \tau_t' \sim \mathbb{P}_{\pi_t'}$
- $x_t = \phi(\tau_t) - \phi(\tau_t'), V_{t+1} = V_t + x_t x_t^\top$
- $\mathcal{B} = \mathcal{B} \cup \{(\tau_t, \tau_t', x_t)\}$ `// Defer preference feedback`

end

foreach $(\tau, \tau', x) \in \mathcal{B}$ **do**

- $y = \mathbb{1}(\tau \succ \tau'); \mathcal{D} = \mathcal{D} \cup \{(x, y)\}$ `// Preference feedback`

end

Output: Policy $\hat{\pi} = \mathtt{A}_{\mathtt{RL}}^{\mathtt{PAC}}\left(\hat{\theta}, \varepsilon, \delta'\right)$ with $\hat{\theta} \in \arg\min_{\|\theta\| \leq B} \mathcal{L}_{\mathcal{D}}(\theta)$ `// One MLE at end`

https://doi.org/10.52202/085713-0407

3.2 Theoretical results

We analyze the regret of `RPO-Regret` and the suboptimality of the output policy $\hat{\pi}$ of `RPO-Explore`.

3.2.1 Regret analysis

We show that `RPO-Regret` incurs sublinear regret with high probability.

Theorem 3.4. *For any $\delta \in (0,1)$ and $T \in \mathbb{N}$, Algorithm 1 satisfies, with probability at least $1-\delta$,*

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

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 $\widetilde{\mathcal{O}}(\kappa d\sqrt{T})$. Similarly, Chen et al. [2022] considers episodic linear MDPs and derives a regret bound of $\widetilde{\mathcal{O}}(d\sqrt{HT})$, avoiding dependence on κ by assuming a linear preference model. However, these approaches rely on a type of optimism which is computationally intractable in this setting (see Appendix E.2). Similar to us, Wu and Sun [2023] avoid this challenge by resorting to randomized exploration and proves a $\widetilde{\mathcal{O}}(d^3\sqrt{\kappa T})$ 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] considers linear MDPs. Second, their approach is model-based, while Algorithm 1 is oracle-based and can accommodate both model-based and model-free implementations.

Proof idea The full proof of Theorem 3.4 is provided in Appendix A. Analogous to the analysis of linear Thompson sampling [Abeille and Lazaric, 2017], the main idea is to control the regret by showing that randomized exploration ensures a constant probability of optimism. However, compared to the linear bandit analysis, our setting comes with additional challenges: First, due to preference-based learning we require a different regret decomposition accounting for the reference policy. Second, as we observe preference feedback on trajectories rather than policies, we need to apply Freedman's inequality (see Lemma D.9) to control the deviation between expected and observed features. Lastly, as we assume a PAC RL oracle – in place of an exact maximization oracle – we need to carefully track the resulting approximation error.

3.2.2 Suboptimality gap

As `RPO-Explore` collects no preferences during exploration, it may incur linear regret as the policies π_t can be highly suboptimal. However, Theorem 3.5 below shows that the final output policy is $\tilde{\mathcal{O}}(1/\sqrt{T})$-optimal.

Theorem 3.5. *For any $\delta \in (0,1)$ and $T \in \mathbb{N}$, Algorithm 2 satisfies, with probability at least $1-\delta$,*

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

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

Except for the extra $\sqrt{d}$ dependency, which is inherent to approaches based on randomized exploration,[4] we match the last iterate guarantee proposed by Das et al. [2024] for a contextual linear bandit setting, but with an algorithm that (i) collects a single batch of preferences and (ii) remains tractable in a full RL setting with trajectory-level feedback, given the PAC RL oracle is tractable.

[4]Recently, Abeille et al. [2025] showed that, in certain cases, the factor $\sqrt{d}$ can be removed by an improved analysis. Their assumptions do not hold in our setting, so whether $\sqrt{d}$ can be avoided remains open.

Comparison with prior work Algorithm 2 is reminiscent of reward-free RL algorithms [Wang et al., 2020], but in a preference-based setting. To our knowledge, Algorithm 2 is the first tractable algorithm to perform efficient preference-free exploration with trajectory-level preferences, and the first – across preference- or reward-based RL – to do so via randomized exploration. Few other works provide suboptimality guarantees in preference-based RL, but require preference-feedback at every round. Wang et al. [2023] propose a meta-algorithm interfacing with a PAC-RL oracle that outputs an ε-optimal policy after $\widetilde{\mathcal{O}}(\kappa^2 d^3/\varepsilon^2)$ queries. A different approach by Zhan et al. [2024] leverages optimal design to prove a bound of $\widetilde{\mathcal{O}}((|S|^2|A|d + \kappa^2 d^2)/\varepsilon^2)$, but their method relies on an intractable maximization oracle. In comparison, Algorithm 2 achieves a bound of $\widetilde{\mathcal{O}}(\kappa d^3/\varepsilon^2)$, improving the dependence on κ over both prior results. The extra $\sqrt{d}$ compared to Zhan et al. [2024] is expected for randomized (rather than optimistic) exploration [Abeille and Lazaric, 2017]. Finally, note that an online-to-batch conversion yields similar suboptimality bounds for Algorithm 1, but would require preferences at every round [Ménard et al., 2021].

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

3.3 Practical limitations

While Algorithm 1 is tractable and efficient, it presents certain limitations. Issuing preference queries at each round (line 7) is impractical, due to the need for continuous feedback, and requesting a label for all trajectory pairs can be expensive and inefficient, as many comparisons may be uninformative. Moreover, the large regret of Algorithm 2 may be undesirable for certain applications. The next section introduces a refined regret-minimization algorithm that addresses these issues by decoupling trajectory collection from query selection and querying only the most informative comparisons.

4 A practical algorithm with efficient query selection

We present Algorithm 3, an improved method for preference collection and active query selection.

4.1 Algorithm

As discussed earlier, we design Algorithm 3 by using lazy updates to collect a batch of trajectory pairs, then applying optimal design to select the informative queries from the batch.

Algorithm 3: `LRPO-OD-Regret` (lazy randomized preference optimization with optimal design)

Input: $T \in \mathbb{N}$, $\delta \in (0,1)$, $\lambda > 1$, $C > 0$

Initialize: $\varepsilon = 1/\sqrt{T}$; $\delta' = \delta/5$; $V_1 = W_1 = \lambda I$; $\mathcal{D}_1 = \mathcal{D} = \emptyset$; $t_{\text{stop}} = 1$; $\hat{\theta}_1 = 0$; π'.

for $t = 1, 2, \ldots, T$ **do**

  if $\det(W_t) > (1 + C)\det(V_{t_{\text{stop}}})$ **then**

    $\mathcal{D}_{\text{opt}}, V_t = \text{D-OptDes}(\mathcal{D}, V_{t_{\text{stop}}}, W_t)$ `// Optimal design`

    $\mathcal{D}_t = \mathcal{D}_{t_{\text{stop}}}$

    for $(\tau, \tau') \in \mathcal{D}_{opt}$ **do**

      $\mathcal{D}_t = \mathcal{D}_t \cup \{(x, y)\}, x = \phi(\tau) - \phi(\tau'), y = \mathbb{1}(\tau \succ \tau')$ `// Preferences`

    end

    $\hat{\theta}_t \in \arg\min_{\|\theta\| \le B} \mathcal{L}_{\mathcal{D}_t}(\theta)$ `// Reward estimation`

    $t_{\text{stop}} = t$, $\mathcal{D} = \emptyset$, $\pi' = \mathtt{A}_{\mathtt{RL}}^{\mathtt{PAC}}\left(\hat{\theta}_t, \varepsilon, \delta'/T\right)$

  end

  $\tilde{\theta}_t \sim \mathcal{N}(\hat{\theta}_{t_{\text{stop}}}, \beta_{t_{\text{stop}}}(\delta')^2 V_{t_{\text{stop}}}^{-1})$ `// Reward sampling`

  $\pi_t = \mathtt{A}_{\mathtt{RL}}^{\mathtt{PAC}}\left(\tilde{\theta}_t, \varepsilon, \delta'/T\right)$ `// Update policy with RL`

  $\mathcal{D} = \mathcal{D} \cup \{(\tau_t, \tau_t')\}$, $x_t = \phi(\tau_t) - \phi(\tau_t')$ with $\tau_t \sim \mathbb{P}_{\pi_t}, \tau_t' \sim \mathbb{P}_{\pi'}$

  If $t = t_{\text{stop}}$, then $W_{t+1} = V_t + x_t x_t^\top$, else: $W_{t+1} = W_t + x_t x_t^\top$

end

https://doi.org/10.52202/085713-0407

Lazy updates We use an idea from Abbasi-Yadkori et al. [2011] to collect many trajectory pairs without querying the preference oracle. The modification compared to Algorithm 1 is collecting trajectories without updating the MLE $\hat{\theta}_{t_{\text{stop}}}$ until the information gain, represented by $\det(V_t)$, increases by a multiplicative constant; see line 4 of Algorithm 3. We show that this procedure limits the number of batches to $\mathcal{O}(\log(T))$. In other words, the average (over batches) size of a given batch is of order $\mathcal{O}(T/\log(T))$. A key advantage of this lazy update structure is that the preference queries (line 7) can be collected in parallel across all trajectory pairs within a batch. This significantly reduces the workload of the preference oracle, *e.g.* a human annotator, by eliminating the need for round-by-round feedback, and the algorithm no longer pauses at each timestep to wait for preference labels.

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

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

Due to the submodularity of the $\log\det$ function for λ greater than one, the greedy procedure of Algorithm 4 achieves an $(1-1/e)$-approximation to the optimal solution; see [Nemhauser et al., 1978, Krause et al., 2008] and Appendix E.1.

Another key feature of Algorithm 4 is its early stopping rule: the while loop terminates early if $\det(V)$ exceeds the determinant of the naive design W; if this never happens we simply return W. When this early termination is satisfied, Algorithm 3 requires fewer preferences than Algorithm 1. Since the optimal design maximizes the information gain (as measured by $\det(V)$), this termination condition is expected to be satisfied frequently in practice.

Algorithm 4: Greedy D-Optimal Design

Input: Dataset $\mathcal{D}$, previous design matrix V, current design matrix W including $\mathcal{D}$.

Initialize dataset $\mathcal{D}_{\text{opt}} = \emptyset$
while $\det(V) < \det(W)$ *and* $|\mathcal{D}_{opt}| < |\mathcal{D}|$ **do**
  $(\tau, \tau') = \arg\max_{(\tau,\tau')\in\mathcal{D}} \det(V + (\phi(\tau)-\phi(\tau'))(\phi(\tau)-\phi(\tau'))^\top)$
  $V = V + xx^\top$, where $x = \phi(\tau) - \phi(\tau')$; $\mathcal{D}_{\text{opt}} = \mathcal{D}_{\text{opt}} \cup \{(\tau, \tau')\}$
end

Output: $(\mathcal{D}_{\text{opt}}, V)$ if $\det(V) \geq \det(W)$, else $(\mathcal{D}, W)$

4.2 Theoretical result

We now provide our high probability regret bound for Algorithm 3.

Theorem 4.1. *Instantiating Algorithm 3 with $C > 0$ and $\lambda \geq 4H_\gamma^2 L^2$, it holds for any $\delta \in (0,1)$ and $T \in \mathbb{N}$, with probability at least $1-\delta$ that*

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

In addition, the number of times the condition of line 4 holds is at most $\frac{d}{\log(1+C)} \log\left(1 + \frac{T(2LH_\gamma)^2}{d\lambda}\right)$. Therefore, the size of the batches is on average of order $\tilde{\mathcal{O}}(T/(d\log(T)))$.

Compared to Theorem 3.4, the above regret bound increases only by constant factors. Regarding the number of preference queries, Sekhari et al. [2023] shows that at least $\Omega(T)$ queries are required to achieve $\mathcal{O}(\sqrt{T})$ regret in the worst case. However, optimal design may lead to significantly fewer queries in favorable instances, as demonstrated in our experiments, while preserving the worst-case guarantees.

Proof idea We briefly outline the main idea of the proof of Theorem 4.1; the full argument is deferred to Appendix C. The central observation is that similarly as for the standard lazy update analysis [Abbasi-Yadkori et al., 2011], the regret only increases by a constant factor $(1+C)$ given that the optimal design subroutine ensures that $\det V_{t_{\text{stop}}} \geq \det W_{t_{\text{stop}}}$ and $|\mathcal{D}_{\text{opt}}| \leq |\mathcal{D}|$.

5 Experiments

We first validate our theoretical results on regret minimization in a tabular gridworld environment, where our RL oracle assumption provably holds, and then compare Algorithm 1 and 3 on more challenging continuous control tasks. The validation of Algorithm 2 is deferred to Appendix F.

5.1 Tabular environment

We consider a 6×6 gridworld environment with deterministic transitions and 4 actions (up, down, left, right). The agent starts in the center and receives reward 0.5 if it reaches one of two boundary states. The reward features are one-hot features of six boundary states – including the two goal states (see Figure 3 for an illustration of the environment). We compare `RPO-Regret` (Algorithm 1) against an RLHF baseline that explores purely through entropy regularization, using synthetic preferences generated from the ground truth reward parameter θ^*. We compute optimal policies using soft value iteration [Haarnoja et al., 2017], and to reduce variance in the reward estimate, we sample 100 trajectories from π_t and π'_t in each RLHF round.

As shown in Figure 1, `RPO-Regret` attains considerably lower regret with less variance across runs than the baseline. These results underscore that entropy exploration does not necessarily guarantee low regret in MDPs.

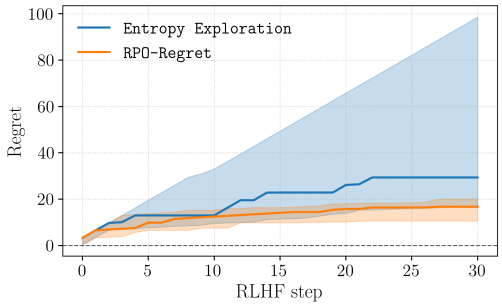
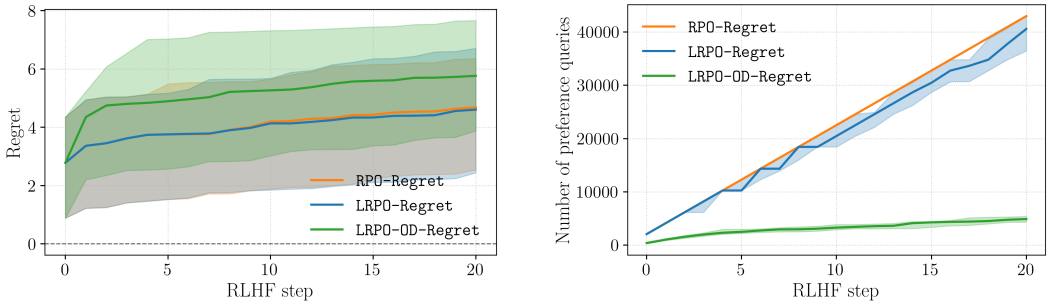


Figure 1: We compare the regret of `RPO-Regret` (orange, Algorithm 1) against a baseline with entropy exploration (blue). The solid lines indicate the median and the shaded areas the 0.2 and 0.8 quantiles, across 20 independent runs. The regret is computed with respect to the ground truth reward parameter θ^*.

5.2 Continuous control

We validate our theoretical results on regret minimization (Theorems 3.4 and 4.1) on the `Isaac-Cartpole-v0` environment from Nvidia Isaac Lab [Mittal et al., 2023]. In this task the goal is to balance a pole on a cart by applying left or right forces, preventing the pole from falling. We compare `RPO-Regret` with its lazy variants (`LRPO-Regret` without optimal design and `LRPO-OD-Regret` with optimal design). To simulate human preferences, we generate synthetic preferences using the built-in task-specific reward function, which is a linear combination of 5 reward



(a) (b)

Figure 2: Comparison of regret minimization algorithms in terms of (a) the cumulative regret (estimated from samples against an RL policy trained with θ^*) and (b) number of preference queries performed. In particular, we compare `RPO-Regret` (orange, Algorithm 1) with its lazy versions `LRPO-Regret` and `LRPO-OD-Regret` (blue & green). Here, `LRPO-Regret` and `LRPO-OD-Regret` refer to Algorithm 3 without and with optimal design subroutine. The solid lines indicate the median and the shaded areas the 0.2 and 0.8 quantiles, across 20 independent runs.

https://doi.org/10.52202/085713-0407

terms (see Appendix F). Furthermore, we adopt PPO [Schulman et al., 2017] as our RL oracle using 30 PPO steps per iteration of Algorithm 1 and 3. Again, we sample 100 trajectories from π_t and π'_t in each RLHF round.

Figure 2 shows that for all three algorithms the regret slope flattens after a few RLHF rounds, demonstrating sublinear regret and performance competitive with RL using ground truth rewards. Although `LRPO-OD-Regret` suffers slightly higher regret initially, it quickly reaches the same performance as `RPO-Regret` and `LRPO-Regret`, while reducing the number of preference queries considerably. 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.

6 Conclusion

We presented two simple meta-algorithms for reinforcement learning from human feedback (RLHF) that combine an RL oracle and randomized exploration, achieving complementary guarantees: regret bounds for Algorithm 1 and PAC-style guarantees for Algorithm 2. Algorithm 2, to our knowledge, is the first tractable method to perform preference-free exploration with trajectory-level preferences. Building on this framework, we introduced Algorithm 3, a practical regret-minimization algorithm that combines lazy updates to enable parallelization with optimal design to reduce query complexity, while maintaining regret guarantees. 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 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. Second, 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 tabular settings and 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.

Acknowledgments Andreas Schlaginhaufen is funded by a PhD fellowship from the Swiss Data Science Center.

References

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

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

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

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

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

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

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

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

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

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

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

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

Vikranth Dwaracherla, Seyed Mohammad Asghari, Botao Hao, and Benjamin Van Roy. Efficient exploration for llms. In *International Conference on Machine Learning*, pages 12215–12227. PMLR, 2024.

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

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

Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International conference on machine learning*, pages 1352–1361. PMLR, 2017.

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

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

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

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

Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, and Michal Valko. Adaptive reward-free exploration. In *Algorithmic Learning Theory*, pages 865–891. PMLR, 2021.

Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

Junpei Komiyama, Junya Honda, Hisashi Kashima, and Hiroshi Nakagawa. Regret lower bound and optimal algorithm in dueling bandit problem. In *Conference on learning theory*, pages 1141–1154. PMLR, 2015.

 https://doi.org/10.52202/085713-0407

Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(2), 2008.

Tor Lattimore and Csaba Szepesvári. *Bandit algorithms*. Cambridge University Press, 2020.

Junghyun Lee, Se-Young Yun, and Kwang-Sung Jun. A unified confidence sequence for generalized linear models, with applications to bandits. *Advances in Neural Information Processing Systems*, 37:124640–124685, 2024.

Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv preprint arXiv:2302.12192*, 2023.

Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

Pangpang Liu, Chengchun Shi, and Will Wei Sun. Dual active learning for reinforcement learning from human feedback. *arXiv preprint arXiv:2410.02504*, 2024.

Yiren Lu, Justin Fu, George Tucker, Xinlei Pan, Eli Bronstein, Rebecca Roelofs, Benjamin Sapp, Brandyn White, Aleksandra Faust, Shimon Whiteson, et al. Imitation is not enough: Robustifying imitation with reinforcement learning for challenging driving scenarios. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7553–7560. IEEE, 2023.

R Duncan Luce et al. *Individual choice behavior*, volume 4. Wiley New York, 1959.

Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *International Conference on Machine Learning*, pages 7599–7608. PMLR, 2021.

Mayank Mittal, Calvin Yu, Qinxi Yu, Jingzhou Liu, Nikita Rudin, David Hoeller, Jia Lin Yuan, Ritvik Singh, Yunrong Guo, Hammad Mazhar, Ajay Mandlekar, Buck Babich, Gavriel State, Marco Hutter, and Animesh Garg. Orbit: A unified simulation framework for interactive robot learning environments. *IEEE Robotics and Automation Letters*, 8(6):3740–3747, 2023. doi: 10.1109/LRA.2023.3270034.

Subhojyoti Mukherjee, Anusha Lalitha, Kousha Kalantari, Aniket Deshmukh, Ge Liu, Yifei Ma, and Branislav Kveton. Optimal design for human feedback. *arXiv preprint arXiv:2404.13895*, 2024.

Mojmir Mutny, Tadeusz Janik, and Andreas Krause. Active exploration via experiment design in markov chains. In *International Conference on Artificial Intelligence and Statistics*, pages 7349–7374. PMLR, 2023.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14:265–294, 1978.

Ellen Novoseller, Yibing Wei, Yanan Sui, Yisong Yue, and Joel Burdick. Dueling posterior sampling for preference-based reinforcement learning. In *Conference on Uncertainty in Artificial Intelligence*, pages 1029–1038. PMLR, 2020.

Reda Ouhamma, Debabrota Basu, and Odalric Maillard. Bilinear exponential family of MDPs: frequentist regret bound with tractable exploration & planning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 9336–9344, 2023.

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

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

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

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

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

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

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

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

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

Andreas Schlaginhaufen and Maryam Kamgarpour. Identifiability and generalizability in constrained inverse reinforcement learning. In *International Conference on Machine Learning*, pages 30224–30251. PMLR, 2023.

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

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

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

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

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

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

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

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

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

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

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

 https://doi.org/10.52202/085713-0407

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

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

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

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

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

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

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

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

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

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

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

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

Appendix

Table of Contents

 https://doi.org/10.52202/085713-0407

A Proof of Theorem 3.4

Theorem 3.4. *For any $\delta \in (0,1)$ and $T \in \mathbb{N}$, Algorithm 1 satisfies, with probability at least $1-\delta$,*

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

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

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

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

1. Let $\delta'' = \delta'/T$ and $c(\delta'') := \sqrt{2d\log(2d/\delta'')} = \sqrt{2d\log(10dT/\delta)}$. Consider the inflated ellipsoid
$$\mathcal{E}_t^{\mathrm{TS}} := \left\{\theta \in \mathbb{R}^d : \left\|\theta - \hat{\theta}_t\right\|_{V_t} \le \beta_t(\delta') c(\delta'')\right\} = c(\delta'')\mathcal{E}_t(\delta').$$
We define the events
$$\widehat{E}_t := \{\theta^* \in \mathcal{E}_t(\delta')\}, \quad \widetilde{E}_t := \left\{\tilde{\theta}_t \in \mathcal{E}_t^{\mathrm{TS}}\right\}, \quad \text{and } E_t := \widehat{E}_t \cap \widetilde{E}_t$$
and let $G_1 := \bigcap_{t=1}^T E_t$. This event implies that θ^* lies in the confidence set uniformly over times $t \in [1,T]$, and the sampled parameter is close to $\hat{\theta}_t$.
2. Let G_2 denote the event that for
$$C_T := 2\sqrt{T\left(2d\log\left(1+\frac{4TL^2H_\gamma^2}{d\lambda}\right)+\frac{12dL^2H_\gamma^2}{\log(2)\lambda}\log\left(1+\frac{4L^2H_\gamma^2}{\log(2)\lambda}\right)\right)} + \frac{16LH_\gamma}{\sqrt{\lambda}}\log\left(\frac{2}{\delta'}\right),$$
it holds that
$$\sum_{t=1}^T \mathbb{E}\left[\|\phi(\tau_t) - \phi(\tau_t')\|_{V_t^{-1}} \mid \mathcal{F}_{t-1}\right] \le C_T,$$
and that
$$\sum_{t=1}^T \mathbb{E}\left[\|\phi(\tau_t) - \phi(\tau_t')\|_{V_t^{-1}} \mid \mathcal{F}^{\theta}_{t-1}\right] \le C_T.$$
3. Let $A_t = \left\{V^*_{\tilde{\theta}_t} - V^{\pi_t}_{\tilde{\theta}_t} \le \varepsilon\right\}$, where $\pi_t = \mathtt{A}^{\mathtt{PAC}}_{\mathtt{RL}}\left(\tilde{\theta}_t, \varepsilon, \delta'/T\right)$, and let $G_3 = \bigcap_{t=1}^T A_t$.

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

Proposition A.1. *Let $G := \bigcap_{i=1}^3 G_i$. It holds that $\Pr[G] \ge 1-\delta$.*

Proof. For the event G_1, Lemma 3.1 and D.1 imply that $\Pr\left[\bigcup_{t=1}^T \widehat{E}_t^{\complement}\right] \le \delta'$ and $\Pr[\widetilde{E}_t^{\complement}] \le \delta'/T$ for any $t \in [1,T]$. Hence, a union bound yields $\Pr\left[G_1^{\complement}\right] \le 2\delta'$. Furthermore, by Lemma D.4 2), we have $\Pr\left[G_2^{\complement}\right] \le \delta'$. Moreover, by union bound and the definition of $\mathtt{A}^{\mathtt{PAC}}_{\mathtt{RL}}$, we have $\Pr\left[G_3^{\complement}\right] \le \delta'$. Hence, we have $\Pr[G] \ge 1 - \sum_{i=1}^3 \Pr\left[G_i^{\complement}\right] \ge 1 - 4\delta' \ge 1-\delta$. □

Step 2 (regret decomposition): Recall that in algorithm 1, we choose $\pi_t' = \pi_{t-1}$. Hence, we can upper 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

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

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 D.2.

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

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

for some $\underline{\theta}_t \in \mathcal{E}_t^{\mathrm{TS}}$. 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

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

Putting this together, while keeping track of the events E_t and A_t, we have

$$\begin{aligned} r_t^{\mathrm{TS}}\mathbb{1}(E_t\cap A_t) &\leq \mathbb{E}\left[\left(\Delta_t(\tilde{\theta}_t) - \underline{\Delta}_t\right)\mathbb{1}(E_t\cap A_t) \mid \mathcal{F}_{t-1}, O_t\right] \\ &\overset{(i)}{\leq} \mathbb{E}\left[\left(\langle\tilde{\theta}_t, \phi(\pi_t) - \phi(\pi_t')\rangle + \varepsilon - \max_{\pi}\langle\underline{\theta}_t, \phi(\pi) - \phi(\pi_t')\rangle\right)\mathbb{1}(E_t\cap A_t) \mid \mathcal{F}_{t-1}, O_t\right] \\ &\overset{(ii)}{\leq} \mathbb{E}\left[\left(\langle\tilde{\theta}_t, \phi(\pi_t) - \phi(\pi_t')\rangle - \langle\underline{\theta}_t, \phi(\pi_t) - \phi(\pi_t')\rangle\right)\mathbb{1}(E_t\cap A_t) \mid \mathcal{F}_{t-1}, O_t\right] + \varepsilon \\ &\overset{(iii)}{\leq} 2\beta_t(\delta')c(\delta'')\mathbb{E}\left[\|\phi(\pi_t) - \phi(\pi_t')\|_{V_t^{-1}} \mid \mathcal{F}_{t-1}, \widehat{E}_t, O_t\right]\Pr\left[\widehat{E}_t \mid \mathcal{F}_{t-1}\right] + \varepsilon, \end{aligned}$$

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

$$\mathbb{E}\left[\|\phi(\pi_t) - \phi(\pi_t')\|_{V_t^{-1}} \mid \mathcal{F}_{t-1}, \widehat{E}_t, O_t\right] \leq \mathbb{E}\left[\|\phi(\pi_t) - \phi(\pi_t')\|_{V_t^{-1}} \mid \mathcal{F}_{t-1}, \widehat{E}_t\right] / \Pr\left[O_t|\mathcal{F}_{t-1}, \widehat{E}_t\right].$$

https://doi.org/10.52202/085713-0407

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

$$\Pr\left[O_t \mid \mathcal{F}_{t-1}, \widehat{E}_t\right] \geq \Pr\left[\Delta_t(\tilde{\theta}_t) \geq \max_{\theta \in \mathcal{E}_t} \Delta_t(\theta) \mid \mathcal{F}_{t-1}\right].$$

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 D.7. Then, applying Lemma D.2 with $f(\theta) = \Delta_t(\theta)$, $\tilde{x} = \tilde{\theta}_t \mid \mathcal{F}_{t-1}$, and $\mathcal{E} = \mathcal{E}_t$, we have

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

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

$$\begin{aligned} r_t^{\text{TS}} \mathbb{1}(E_t \cap A_t) &\leq \frac{2\beta_t(\delta')c(\delta'')\mathbb{E}\left[\|\phi(\pi_t) - \phi(\pi_t')\|_{V_t^{-1}} \mid \mathcal{F}_{t-1}\right]}{p} + \varepsilon \\ &\leq \frac{2\beta_t(\delta')c(\delta'')\mathbb{E}\left[\|\phi(\tau_t) - \phi(\tau_t')\|_{V_t^{-1}} \mid \mathcal{F}_{t-1}\right]}{p} + \varepsilon, \end{aligned}$$

where the second inequality follows from the convexity of norms. Applying Lemma D.4 2), we have on the good event G that

$$\begin{aligned} \sum_{t=1}^T r_t^{\text{TS}} &\leq \frac{2\beta_T(\delta')c(\delta'')}{p} \sum_{t=1}^T \mathbb{E}\left[\|\phi(\tau_t) - \phi(\tau_t')\|_{V_t^{-1}} \mid \mathcal{F}_{t-1}\right] + T\varepsilon \\ &\leq \frac{2\beta_T(\delta')c(\delta'')}{p} C_T + T\varepsilon, \end{aligned} \tag{3}$$

for the constant

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

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

$$\begin{aligned} r_t^{\text{MLE}} &= \langle \tilde{\theta}_t - \theta^*, \phi(\pi_t) - \phi(\pi_t')\rangle \\ &= \langle \tilde{\theta}_t - \hat{\theta}_t, \phi(\pi_t) - \phi(\pi_t')\rangle + \langle \hat{\theta}_t - \theta^*, \phi(\pi_t) - \phi(\pi_t')\rangle \\ &\leq \beta_t(\delta')(1 + c(\delta''))\, \|\phi(\pi_t) - \phi(\pi_t')\|_{V_t^{-1}} \\ &\leq \beta_t(\delta')(1 + c(\delta''))\mathbb{E}\left[\|\phi(\tau_t) - \phi(\tau_t')\|_{V_t^{-1}} |\mathcal{F}_{t-1}^\theta\right], \end{aligned}$$

where the last inequality follows from the convexity of norms. From here, we can again apply the second result of Lemma D.4. We deduce that on the good event G, we have

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

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

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

□

B Proof of Theorem 3.5

Theorem 3.5. *For any $\delta \in (0,1)$ and $T \in \mathbb{N}$, Algorithm 2 satisfies, with probability at least $1-\delta$,*

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

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

Proof. Similar to the proof of Theorem 4.1, we start by defining a set of good events.

Step 1 (good events): Recall that in Algorithm 2 we set $\varepsilon = 1/\sqrt{T}$ and $\delta' = \delta/5$; at round t, conditional on $\mathcal{F}_{t-1} = \sigma(x_1, \ldots, x_{t-1})$, we sample $\tilde{\theta}_t \sim \mathcal{N}(0, V_t^{-1})$ with $V_t = \lambda I + \sum_{k=1}^{t-1} x_k x_k^\top$ (where $x_k = \phi(\tau_k) - \phi(\tau_k')$), and $\hat{\theta}$ is the MLE estimate at time $T+1$ using all the data from round 1 to T. We now define the following high probability events.

1. Let $\delta'' = \delta'/T$ and $c(\delta'') := \sqrt{2d\log(2d/\delta'')} = \sqrt{2d\log(10dT/\delta)}$. Consider the centered ellipsoid
$$
\bar{\mathcal{E}}_t^{\mathrm{TS}} := \left\{\theta \in \mathbb{R}^d : \|\theta\|_{V_t} \leq c(\delta'')\right\}.
$$
We define the events
$$
\widehat{E} := \left\{\|\theta^* - \hat{\theta}\|_{V_{T+1}} \leq \beta_{T+1}(\delta')\right\}, \quad \widetilde{E}_t := \left\{\tilde{\theta}_t \in \bar{\mathcal{E}}_t^{\mathrm{TS}}\right\},
$$
and let $G_1 := \bigcap_{t=1}^{T} \widetilde{E}_t \cap \widehat{E}$. This event implies that θ^* lies in the confidence ellipsoid $\mathcal{E}_{T+1}(\delta')$, and the sampled parameter lies in $\bar{\mathcal{E}}_t^{\mathrm{TS}}$ for all $t \in [1,T]$.
2. Let G_2 denote the event that for
$$
C_T := 2\sqrt{T\left(2d\log\left(1+\frac{4TL^2H_\gamma^2}{d\lambda}\right)+\frac{12dL^2H_\gamma^2}{\log(2)\lambda}\log\left(1+\frac{4L^2H_\gamma^2}{\log(2)\lambda}\right)\right)} + \frac{16LH_\gamma}{\sqrt{\lambda}}\log\left(\frac{2}{\delta'}\right),
$$
it holds that
$$
\sum_{t=1}^{T} \mathbb{E}\left[\|\phi(\tau_t) - \phi(\tau_t')\|_{V_t^{-1}} \mid \mathcal{F}_{t-1}\right] \leq C_T.
$$

https://doi.org/10.52202/085713-0407

3. Let $A_t = \left\{ V^*_{\tilde{\theta}_t} - V^{\pi_t}_{\tilde{\theta}_t} \leq \varepsilon \right\}$, where $\pi_t = \mathtt{A}^{\mathtt{PAC}}_{\mathtt{RL}}\left(\tilde{\theta}_t, \varepsilon, \delta'/T\right)$, and let $G_3 = \bigcap_{t=1}^{T} A_t$.

4. Let G_4 denote the event that $V^*_{\hat{\theta}} - V^{\hat{\pi}}_{\hat{\theta}} \leq \varepsilon$ where $\hat{\pi} = \mathtt{A}^{\mathtt{PAC}}_{\mathtt{RL}}\left(\hat{\theta}, \varepsilon, \delta'\right)$.

Analogously to Proposition A.1, it follows that $\Pr[G_1^{\complement}] \leq 2\delta'$ and $\Pr[G_k^{\complement}] \leq \delta'$ for $k = 2, 3, 4$. Thus, by a union bound, the good event $G := \bigcap_{i=1}^{4} G_i$ happens with probability at least $1 - \delta$.

Step 2 (suboptimality bound): In the following, we will denote $\pi^* \in \arg\max_\pi \langle \theta^*, \phi(\pi) \rangle$ for an arbitrary optimal policy corresponding to the ground truth reward parameter θ^*. Under the above good event G, we can bound the suboptimality of $\hat{\pi}$ as follows

$$\begin{aligned}\mathsf{SubOpt}(\hat{\pi}) &= \langle \theta^*, \phi(\pi^*) - \phi(\hat{\pi}) \rangle \leq \langle \theta^* - \hat{\theta}, \phi(\pi^*) - \phi(\hat{\pi}) \rangle + \varepsilon \\ &\leq \max_{\pi,\pi'} \langle \theta^* - \hat{\theta}, \phi(\pi) - \phi(\pi') \rangle + \varepsilon \leq \beta_{T+1}(\delta') \max_{\pi,\pi'} \|\phi(\pi) - \phi(\pi')\|_{V_{T+1}^{-1}} + \varepsilon. \quad (4)\end{aligned}$$

Next, we continue with bounding the term $\max_{\pi,\pi'} \|\phi(\pi) - \phi(\pi')\|_{V_{T+1}^{-1}}$.

Step 3 (approximate G-optimal design): Due to the matrix inequalities $V_{T+1} \succeq V_T \succeq \ldots \succeq V_1$, we have

$$\|\cdot\|_{V_{T+1}^{-1}} \leq \|\cdot\|_{V_T^{-1}} \leq \ldots \leq \|\cdot\|_{V_1^{-1}}.$$

Therefore, we can bound the right-hand-side of (4) as follows

$$\begin{aligned}\max_{\pi,\pi'} \|\phi(\pi) - \phi(\pi')\|_{V_{T+1}^{-1}} &\leq \frac{1}{T} \sum_{t=1}^{T} \max_{\pi,\pi'} \|\phi(\pi) - \phi(\pi')\|_{V_{T+1}^{-1}} \\ &\leq \frac{1}{T} \sum_{t=1}^{T} \max_{\pi,\pi'} \|\phi(\pi) - \phi(\pi')\|_{V_t^{-1}} \\ &\leq \frac{2}{T} \sum_{t=1}^{T} \max_{\pi} \|\phi(\pi) - \phi(\pi'_t)\|_{V_t^{-1}},\end{aligned}$$

where we used the triangle inequality in the last step. If it were tractable to compute $\pi_t = \arg\max_\pi \|\phi(\pi) - \phi(\pi'_t)\|_{V_t^{-1}}$, we could invoke Lemma D.4 directly from here. Instead, we leverage randomized exploration to keep our algorithm tractable. In particular, we consider the function $f_t(\theta) = \max_\pi \langle \theta, \phi(\pi) - \phi(\pi'_t) \rangle$, the ellipsoid $\bar{\mathcal{E}}_t = \left\{\theta : \|\theta\|_{V_t} \leq 1\right\}$, and the event $O_t = \{f_t(\tilde{\theta}_t) \geq \max_{\theta \in \bar{\mathcal{E}}_t} f_t(\theta)\}$. By Proposition D.7 f_t is convex and continuous. Applying Lemma D.2 with $f = f_t$, $\tilde{\theta} = \tilde{\theta}_t \mid \mathcal{F}_{t-1}$, and $\mathcal{E} = \bar{\mathcal{E}}_t$, it holds that $\Pr\left[O_t \mid \mathcal{F}_{t-1}\right] \geq p := 1/(4\sqrt{e\pi})$. Hence, we can proceed analogously to the proof of Theorem 3.4:

$$\begin{aligned}\max_{\pi} \|\phi(\pi) - \phi(\pi'_t)\|_{V_t^{-1}} \mathbb{1}\left(\widetilde{E}_t \cap A_t\right) &\overset{(i)}{=} \max_{\theta \in \bar{\mathcal{E}}_t} f_t(\theta) \mathbb{1}\left(\widetilde{E}_t \cap A_t\right) \\ &\leq \mathbb{E}\left[f_t(\tilde{\theta}_t) \mathbb{1}\left(\widetilde{E}_t \cap A_t\right) \mid O_t, \mathcal{F}_{t-1}\right] \\ &\leq \mathbb{E}\left[\langle \tilde{\theta}_t, \phi(\pi_t) - \phi(\pi'_t) \rangle \mathbb{1}\left(\widetilde{E}_t \cap A_t\right) \mid O_t, \mathcal{F}_{t-1}\right] + \varepsilon \\ &\leq c(\delta'') \mathbb{E}\left[\|\phi(\pi_t) - \phi(\pi'_t)\|_{V_t^{-1}} \mid O_t, \mathcal{F}_{t-1}\right] + \varepsilon \\ &\overset{(ii)}{\leq} \frac{c(\delta'')}{p} \mathbb{E}\left[\|\phi(\pi_t) - \phi(\pi'_t)\|_{V_t^{-1}} \mid \mathcal{F}_{t-1}\right] + \varepsilon \\ &\leq \frac{c(\delta'')}{p} \mathbb{E}\left[\|\phi(\tau_t) - \phi(\tau'_t)\|_{V_t^{-1}} \mid \mathcal{F}_{t-1}\right] + \varepsilon.\end{aligned}$$

Here, (i) follows because $\bar{\mathcal{E}}_t$ is a centered ellipsoid, and (ii) from the total law of probability.

Using Lemma D.4, we conclude that with probability $1-\delta$, we have that

$$\mathsf{SubOpt}(\hat{\pi}) \le \varepsilon(1+2\beta_{T+1}(\delta')) + \frac{2\beta_{T+1}(\delta')c(\delta'')}{pT} \underbrace{\sum_{t=1}^{T} \mathbb{E}\left[\|\phi(\tau_t)-\phi(\tau_t')\|_{V_t^{-1}} \mid \mathcal{F}_{t-1}\right]}_{\le C_T}$$

$$\le \left[\sqrt{\kappa\left[\log\left(\frac{5}{\delta}\right) + d\log\left(\max\left\{e, \frac{4eBLH_\gamma T}{d}\right\}\right)\right]} + 2\sqrt{\lambda}B\right]\left\{\frac{3}{\sqrt{T}} + \frac{2\sqrt{2d\log(10dT/\delta)}}{pT}\right.$$

$$\left.\cdot\left[2\sqrt{T\left(2d\log\left(1+\frac{4TL^2H_\gamma^2}{d\lambda}\right)+\frac{12dL^2H_\gamma^2}{\log(2)\lambda}\log\left(1+\frac{4L^2H_\gamma^2}{\log(2)\lambda}\right)\right)} + \frac{16LH_\gamma}{\sqrt{\lambda}}\log\left(\frac{10}{\delta}\right)\right]\right\}$$

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

□

Remark B.1 (Connection to G-optimal design). Given a subset $\mathcal{X} \subset \mathbb{R}^d$, a G-optimal design selects $x_1, \dots, x_T \in \mathcal{X}$ to minimize $\max_{x\in\mathcal{X}} \|x\|_{V^{-1}}$ with $V = \sum_{t=1}^{T} x_t x_t^\top$. Under compactness and if $\operatorname{span}(\mathcal{X}) = \mathbb{R}^d$, the Kiefer-Wolfowitz theorem [Kiefer and Wolfowitz, 1960] yields the lower bound $\max_{x\in\mathcal{X}} \|x\|_{V^{-1}} \ge \sqrt{d/T}$, which is tight up to rounding. In the proof above, we show that Algorithm 2 guarantees $\max_{\pi,\pi'} \|\phi(\pi)-\phi(\pi')\|_{V_{T+1}^{-1}} = \mathcal{O}(d\log(dT/\delta)/\sqrt{T})$. Thus, ignoring the noise in $\phi(\tau)$ (handled via Freedman) and, for simplicity, the regularization λ, Algorithm 2 achieves a $\mathcal{O}(\sqrt{d}\log(dT/\delta))$-approximate G-optimal design with probability at least $1-\delta$.

C Proof of Theorem 4.1

Theorem 4.1. *Instantiating Algorithm 3 with $C > 0$ and $\lambda \ge 4H_\gamma^2 L^2$, it holds for any $\delta \in (0,1)$ and $T \in \mathbb{N}$, with probability at least $1-\delta$ that*

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

In addition, the number of times the condition of line 4 holds is at most $\frac{d}{\log(1+C)}\log\left(1+\frac{T(2LH_\gamma)^2}{d\lambda}\right)$. Therefore, the size of the batches is on average of order $\tilde{\mathcal{O}}(T/(d\log(T)))$.

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

Lemma C.1 (Number of design matrix updates)**.** *Using Algorithm 3 with a parameter $C > 0$, it holds that:*

$$\sum_{t=1}^{T} \mathbb{1}\left[V_{t+1} \ne V_t\right] \le \frac{d}{\log(1+C)}\log\left(1+\frac{T(2LH_\gamma)^2}{d\lambda}\right).$$

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

Proof. Denote $N_T = \sum_{t=1}^{T} \mathbb{1}\left[V_{t+1} \ne V_t\right]$, and let $t_{\text{stop}} \le T$ be the last time the matrix V_t was updated. Then,

$$\det(V_T) = \det(V_{t_{\text{stop}}}) \ge (1+C)^{N_T}\det(\lambda I).$$

Here, we used that for two consecutive update rounds $t_{\text{stop}} \ge t'_{\text{stop}}$ we have that $\det V_{t_{\text{stop}}} \ge (1+C)\det V_{t'_{\text{stop}}}$. Then, using the trace-determinant inequality $\det A \le (\operatorname{tr}(A)/d)^d$, we have

$$(1+C)^{N_T}\lambda^d \le \left(\frac{d\lambda + T(2LH_\gamma)^2}{d}\right)^d,$$

and

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

□

https://doi.org/10.52202/085713-0407

We now present the proof for the regret bound, which proceeds similarly to that of Theorem 3.4 up to some modifications. The first change is in the regret decomposition, which needs to be adapted because we no longer compare to the past policy but rather to $\pi'_t = \mathtt{A}_{\mathtt{RL}}^{\mathtt{PAC}}(\hat{\theta}_{t_{\mathrm{stop}}(t)}, \varepsilon, \delta)$. The second change is using Lemma D.5 instead of Lemma D.4 to bound the sum of norms of trajectory features. Finally, the good events defined in the proof of Theorem 3.4 are slightly modified to account for the lazy design matrix and the new choice of comparator policy π'_t.

Proof. Let us first define the function $t_{\mathrm{stop}} : \mathbb{N} \to \mathbb{N}$, which to a time t, assigns the last time $t_{\mathrm{stop}}(t) \leq t$ that the update condition (line 4 in Algorithm 3) was met.

We work with the same two filtrations as before $\mathcal{F}_{t-1} = \sigma(x_1, y_1, \ldots, x_{t-1}, y_{t-1})$ and $\mathcal{F}^{\theta}_{t-1} = \sigma(\mathcal{F}_{t-1}, \tilde{\theta}_t)$. And we recall that, given $\mathcal{F}_{t-1}$, $\tilde{\theta}_t$ in Algorithm 3 is sampled from $\mathcal{N}(\hat{\theta}_{t_{\mathrm{stop}}(t)}, \beta^2_{t_{\mathrm{stop}}(t)} V^{-1}_{t_{\mathrm{stop}}(t)})$.

Step 1 (good events): Consider $\delta' = \delta/5$, we redefine the high-probability events:

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

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

   We define the events $E_t := \widehat{E}_t \cap \widetilde{E}_t, \quad \widehat{E}_t := \left\{ \theta^* \in \mathcal{E}_{t_{\mathrm{stop}}(t)} \right\}, \quad \widetilde{E}_t := \left\{ \tilde{\theta}_t \in \mathcal{E}^{\mathrm{TS}}_{t_{\mathrm{stop}}(t)} \right\}$, and let $G_1 := \bigcap_{t=1}^{T} E_t$.

2. Let $C_T := 2\sqrt{\frac{8(1+C)\, dT\, L^2 H^2_\gamma}{\lambda} \log\!\left(1 + \frac{4TL^2H^2_\gamma}{\lambda d}\right)} + \frac{16LH_\gamma}{\sqrt{\lambda}} \log(2/\delta)$, and define G_2 as the event under which it holds that

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

   and

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

3. Let $A_t = \left\{ V^*_{\tilde{\theta}_t} - V^{\pi_t}_{\tilde{\theta}_t} \leq \varepsilon \right\}$, where $\pi_t = \mathtt{A}_{\mathtt{RL}}^{\mathtt{PAC}}\left(\tilde{\theta}_t, \varepsilon, \delta'/T\right)$, and let $G_3 = \bigcap_{t=1}^{T} A_t$.

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

We now compare each of these events to their counterparts in the proof of Theorem 3.4. The event G_1 is modified because we replace V_t by $V_{t_{\mathrm{stop}}(t)}$ and $\hat{\theta}_t$ by $\hat{\theta}_{t_{\mathrm{stop}}(t)}$. The event G_1 still holds with probability at least $1 - 2\delta'$ using the same concentration arguments as before. The event G_2 is also modified to account for the lazy design matrix, and it holds with probability $1 - \delta'$ thanks to Lemma D.5. Finally, the event G_3 remains unchanged.

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

Step 2 (regret decomposition): Since Algorithm 3 uses a different comparator policy π'_t than Algorithm 1, we derive a new regret decomposition. On the good event G, we have

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

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

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

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

where $C_T = 2\sqrt{\frac{8(1+C)\,dT\,L^2H^2_\gamma}{\lambda}\log\!\left(1 + \frac{4TL^2H^2_\gamma}{\lambda d}\right)} + \frac{16LH_\gamma}{\sqrt{\lambda}}\log(2/\delta)$.

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

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

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

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

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

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

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

Step 4 (bounding r^{MLE}_t): This step is analogous to Appendix A. We have:

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

https://doi.org/10.52202/085713-0407

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

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

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

□

D Technical results

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

D.1 Confidence sequence

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

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

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

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

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

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

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

$$\begin{aligned} G_t(\hat{\theta}_t, \theta^*) &= \int_0^1 (1-\tau)\left(\sum_{k=1}^{t-1}\dot{\sigma}(\langle\hat{\theta}_t + \tau(\theta^* - \hat{\theta}_t), x_k\rangle)x_k x_k^\top\right)\mathrm{d}\tau \\ &= \sum_{k=1}^{t-1}\left[\int_0^1 (1-\tau)\dot{\sigma}(\langle\hat{\theta}_t + \tau(\theta^* - \hat{\theta}_t), x_k\rangle)\,\mathrm{d}\tau\right]x_k x_k^\top \\ &\succeq \kappa^{-1}\sum_{k=1}^{t-1} x_k x_k^\top. \end{aligned} \tag{5}$$

Rearranging terms gives

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

where (i) follows from the first order optimality condition for $\hat{\theta}_t$ and (ii) from the lower bound in (5). Rearranging again and applying Theorem 3.1 by Lee et al. [2024] for likelihood ratio confidence sequences with the Lipschitz constant of $\mathcal{L}_{\mathcal{D}_t}$ equal to $L_t = 2LH_\gamma(t-1)$, we get

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

□

D.2 Optimism with constant probability

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

Lemma D.1 (Appendix A of [Abeille and Lazaric, 2017])**.** *Let $z \sim \mathcal{N}(0, I)$ be a d-dimensional Gaussian random vector. Then, we have:*

1. *Anti-concentration: For any $u \in \mathcal{B}^d(1)$, we have $\Pr[\langle u, z\rangle \geq 1] \geq \frac{1}{4\sqrt{e\pi}}$.*
2. *Concentration: $\Pr\left[\|z\| \leq \sqrt{2d\log(2d/\delta)}\right] \geq 1-\delta$.*

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

Lemma D.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a continuous and convex function, and consider the ellipsoid $\mathcal{E} := \left\{\theta \in \mathbb{R}^d : \|\theta - \theta_0\|_V \leq b\right\}$ for a positive definite matrix V and $b > 0$. If $\tilde{\theta} \sim \mathcal{N}(\theta_0, b^2V^{-1})$, then $\Pr\left[f(\tilde{\theta}) \geq \max_{\theta\in\mathcal{E}} f(\theta)\right] \geq 1/\left(4\sqrt{e\pi}\right)$.*

Proof. Note that by definition of $\tilde{\theta}$ we have $\tilde{\theta} \overset{d}{=} \theta_0 + bV^{-1/2}\tilde{z}$, where $\tilde{z} \sim \mathcal{N}(0, I)$. Hence, considering $g(z) := f(\theta_0 + bV^{-1/2}z)$, we have

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

where we used that $\theta_0 + bV^{-1/2}z \in \mathcal{E}$ if and only if $z \in \mathcal{B}^d(1)$. Since g is a continuous convex 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

$$\partial g(\bar{z}) \subseteq N_{\mathcal{B}^d(1)}(\bar{z}), \tag{6}$$

where $N_{\mathcal{B}^d(1)}(\bar{z}) = \left\{h \in \mathbb{R}^d : 0 \geq \langle h, z - \bar{z}\rangle, \forall z \in \mathcal{B}^d(1)\right\} = \left\{h \in \mathbb{R}^d : h = \lambda\bar{z}, \lambda \geq 0\right\}$ denotes the normal cone to $\mathcal{B}^d(1)$ at $\bar{z}$. While this optimality condition for convex function maximization is somewhat standard (see *e.g.* [Rockafellar, 1997, Theorem 32.4]), we can verify directly that by optimality of $\bar{z}$ we have

$$\begin{aligned}
\partial g(\bar{z}) &:= \left\{h \in \mathbb{R}^d : g(z) - g(\bar{z}) \geq \langle h, z - \bar{z}\rangle, \forall z\right\} \\
&\subseteq \left\{h \in \mathbb{R}^d : g(z) - g(\bar{z}) \geq \langle h, z - \bar{z}\rangle, \forall z \in \mathcal{B}^d(1)\right\} \\
&\subseteq \left\{h \in \mathbb{R}^d : 0 \geq \langle h, z - \bar{z}\rangle, \forall z \in \mathcal{B}^d(1)\right\} =: N_{\mathcal{B}^d(1)}(\bar{z}).
\end{aligned}$$

Since g is convex and finite on all of $\mathbb{R}^d$, we have $\partial g(\bar{z}) \neq \emptyset$. Therefore, the inclusion (6) implies that

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

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

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

In light of Lemma D.1, this establishes $p \geq 1/(4\sqrt{e\pi})$. □

The above Lemma is helpful in the following way: Let $\theta^* \in \mathcal{E}$, $\tilde{\theta} \sim \mathcal{N}(\theta_0, b^2V^{-1})$, and $x_{\tilde{\theta}} = \arg\max_{x\in\mathcal{X}}\langle\tilde{\theta}, x\rangle$ for some bounded subset $\mathcal{X} \subset \mathbb{R}^d$ (assuming the maximum exists). If f is the support function of $\mathcal{X}$, *i.e.* $f(\theta) = \max_{x\in\mathcal{X}}\langle\theta, x\rangle$, then we have with probability at least $1/\left(4\sqrt{e\pi}\right)$ that

$$\max_{x\in\mathcal{X}}\langle\theta^*, x\rangle = f(\theta^*) \leq \max_{\theta\in\mathcal{E}} f(\theta) \leq f(\tilde{\theta}) = \langle\tilde{\theta}, x_{\tilde{\theta}}\rangle.$$

Moreover, by part 2. of Lemma D.1, we still have that $\left\|\tilde{\theta} - \theta_0\right\|_V \leq \tilde{\mathcal{O}}\left(\sqrt{d}b\right)$ with high probability. This idea will be key to the proofs of Theorems 3.4, 3.5, and 4.1.

https://doi.org/10.52202/085713-0407

D.3 Elliptical potential bounds

Another key component of the convergence proofs are the following so-called elliptical potential lemmas that provide an upper bound on the sum of norms of sequentially observed vectors in the norm induced by their design matrix.

Lemma D.3 (Lemma 19.4 of Lattimore and Szepesvári [2020]). *Let $(x_t)_{t\geq 1} \subset \mathbb{R}^d$ and for all $t \geq 1$, $\|x_t\| \leq L$, let $V_t = \lambda I + \sum_{k=1}^{t-1} x_k x_k^\top$ for some $\lambda > 0$. Then,*

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

In the analysis of our algorithms, a central challenge is to control the norms of policy feature differences $\phi(\pi_t) - \phi(\pi'_t)$. However, the learner only observes the trajectory-level differences $x_t = \phi(\tau_t) - \phi(\tau'_t)$, which are random variables with mean $\phi(\pi_t) - \phi(\pi'_t)$. To overcome this, we build on Lemma D.3 and introduce new tools to bound the sum of norms of policy feature differences.

Lemma D.4 (Elliptical lemma). *Let $\{x_t\}_{t\geq 1} \subset \mathbb{R}^d$ be a sequence of random vectors adapted to a filtration $\{\mathcal{F}_t\}_{t\geq 1}$, and let $\|x_t\| \leq L$ almost surely for all $t \geq 1$. Let $V_t = \lambda I + \sum_{k=1}^{t-1} x_k x_k^\top$ for some $\lambda > 0$. Then, the following holds:*

1) Deterministic bound. *For all $T \in \mathbb{N}$, almost surely,*

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

2) High-probability bound. *For all $\delta \in (0,1)$, with probability at least $1-\delta$,*

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

The first statement is a small improvement over Lemma D.3 because it involves $\|x_t\|^2_{V_t^{-1}}$ instead of $\min\{1, \|x_t\|^2_{V_t^{-1}}\}$ and maintains a similar upper bound. In the second statement, $\{x_t\}_{t\geq 1}$ represent trajectory features and their expected values are policy features. Hence, the second statement of the lemma above allows us to control the elliptical potentials of the policy features while only observing trajectory features.

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

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

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

where the first term of the last inequality follows by definition of $\mathcal{T}_T$. The second term follows 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 discussed, and because $\|x_t\|^2_{V_t^{-1}} \leq L^2/\lambda$. Then, the proof is concluded by bounding the first sum on the right-hand side using Lemma D.3.

Second statement: The proof proceeds by using Lemma D.9. We have for any $\delta \in (0,1)$ that with probability $1-\delta$:

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

where the first inequality uses Lemma D.9, the second uses the Cauchy-Schwarz inequality, and the last follows from the first result of the lemma. □

We now present a variant of the elliptical potential lemma above, adapted for the case where the design matrix is updated with lazy updates and optimal design; see Algorithm 3 for more details.

Lemma D.5 (Lazy elliptical lemma)**.** *Let $(x_t)_{t\ge1} \subset \mathbb{R}^d$ be a sequence of random vectors adapted to a filtration $(\mathcal{F}_t)_{t\ge1}$, with $\|x_t\| \le L$ almost surely. Fix $\lambda \ge L^2$ and $C > 0$.*

Define $(V_t, W_t)_{t\ge1}$ as in Algorithm 3:

- $V_1 = W_1 = \lambda I$, $\mathcal{D}_1 = \emptyset$, $t_{\text{stop}} = 1$.
- *For $t = 1, \dots, T$:*
  - *If $\det(W_t) > (1+C)\det(V_{t_{\text{stop}}})$:*
    - *Let $V_t = V_{t-1} + \sum_{x\in\mathcal{D}_{opt}} xx^\top$ be such that $\det V_t \ge \det W_t$ and $|\mathcal{D}_{opt}| \le |\mathcal{D}|$.*
    - *Set $t_{\text{stop}} = t$, $\mathcal{D} = \emptyset$*
  - *Observe x_t and append $\mathcal{D} = \mathcal{D} \cup \{x_t\}$.*
  - *Update*

$$W_{t+1} = \begin{cases} V_t + x_t x_t^\top, & \text{if } t = t_{\text{stop}}, \\ W_t + x_t x_t^\top, & \text{otherwise,} \end{cases} \qquad V_{t+1} = V_t.$$

Then, the following holds:

1) Deterministic bound. *For all $T \in \mathbb{N}$,*

$$\sum_{t=1}^{T} \|x_t\|^2_{V_t^{-1}} \le \frac{2(1+C)\,d\,L^2}{\lambda} \log\Big(1 + \frac{TL^2}{\lambda d}\Big).$$

2) High-probability bound. *For a fixed $T \in \mathbb{N}$ and any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sum_{t=1}^{T} \mathbb{E}\big[\|x_t\|_{V_t^{-1}} \mid \mathcal{F}_{t-1}\big] \le 2\sqrt{\frac{2(1+C)\,dT\,L^2}{\lambda}\log\Big(1+\frac{TL^2}{\lambda d}\Big)} + \frac{8L}{\sqrt{\lambda}}\log(1/\delta).$$

Proof. **First statement:** First note that $V_t \succeq \lambda I$ for all t, hence $\|x_t\|^2_{V_t^{-1}} \le L^2/\lambda$. For non-update steps, $t \ne t_{\text{stop}}$, we have $V_t \preceq W_t$ and $\det(W_t) \le (1+C)\det(V_t)$. Hence, by Lemma D.8

$$\forall x \in \mathbb{R}^d, \quad \|x\|^2_{V_t^{-1}} \le \frac{\det(V_t^{-1})}{\det(W_t^{-1})}\|x\|^2_{W_t^{-1}} \le (1+C)\|x\|^2_{W_t^{-1}}. \tag{7}$$

Using the inequality $\min\{1,z\} \le 2\log(1+z)$ for $z \ge 0$, we get

$$\begin{aligned}\min\{1, \|x_t\|^2_{V_t^{-1}}\} &\le (1+C)\,\min\{1, \|x_t\|^2_{W_t^{-1}}\} \le 2(1+C)\log\left(1+\|x_t\|^2_{W_t^{-1}}\right) \\ &= 2(1+C)\log\frac{\det(W_{t+1})}{\det(W_t)},\end{aligned}$$

 https://doi.org/10.52202/085713-0407

where the last equality follows from $W_{t+1} = W_t + x_t x_t^\top$ and the matrix determinant lemma (see [Abbasi-Yadkori et al., 2011, Lemma 11]).

For update steps, $t = t_{\text{stop}}$, we have by construction $\det(V_t) \geq \det(W_t)$ and $W_{t+1} = V_t + x_t x_t^\top$. Therefore, with the same reasoning as above,

$$\min\{1, \|x_t\|^2_{V_t^{-1}}\} \leq 2 \log\left(1 + \|x_t\|^2_{V_t^{-1}}\right) = 2 \log \frac{\det(W_{t+1})}{\det(V_t)} \leq 2 \log \frac{\det(W_{t+1})}{\det(W_t)}.$$

By summing over t and telescoping, we therefore get

$$\sum_{t=1}^{T} \min\{1, \|x_t\|^2_{V_t^{-1}}\} \leq 2(1+C) \log \frac{\det(W_{T+1})}{\det(W_1)} \leq 2(1+C) d \log\left(1 + \frac{TL^2}{\lambda d}\right).$$

Using that $\|x_t\|^2_{V_t^{-1}} \leq (L^2/\lambda) \min\{1, \|x_t\|^2_{V_t^{-1}}\}$ concludes the proof of the first result.

Second statement: Analogously to Lemma D.4, the second statement follows from the first statement and Lemma D.9. □

Remark D.6. Note that we have not used the trick from [Lattimore and Szepesvári, 2020, Exercise 19.3] in the proof above, as it doesn't trivially extend to our lazy setting.

D.4 Miscellaneous

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

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

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

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

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

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

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

Lemma D.9 (Lemma 2 of Zhu and Nowak [2022])**.** *Let $(z_t)_{t \leq T}$ be a real-valued sequence of random variables adapted to a filtration $\mathcal{F}_t$. If $0 \leq z_t \leq B$ almost surely, then with probability at least $1 - \delta$,*

$$\sum_{t=1}^{T} z_t \leq \frac{3}{2} \sum_{t=1}^{T} \mathbb{E}\left[z_t \mid \mathcal{F}_{t-1}\right] + 4B \log\left(2\delta^{-1}\right)$$

and

$$\sum_{t=1}^{T} \mathbb{E}\left[z_t \mid \mathcal{F}_{t-1}\right] \leq 2 \sum_{t=1}^{T} z_t + 8B \log\left(2\delta^{-1}\right).$$

E Discussion and background

E.1 Background on optimal experimental design

Continuous designs Given a (possibly infinite) set of features $\mathcal{X} \subset \mathbb{R}^d$, a *D-optimal design* is a probability measure $\xi^* \in \Delta_{\mathcal{X}}$ maximizing the log-determinant of the (normalized) information matrix

$$\xi^* \in \operatorname*{arg\,max}_{\xi \in \Delta_{\mathcal{X}}} \log\det V(\xi), \qquad V(\xi) := \int_{\mathcal{X}} x x^\top \, \mathrm{d}\xi(x).$$

The Kiefer–Wolfowitz equivalence theorem [Kiefer and Wolfowitz, 1960] further implies that any D-optimal design also minimizes the *G-criterion*

$$\xi^* \in \arg\min_{\xi \in \Delta_{\mathcal{X}}} \max_{x \in \mathcal{X}} \|x\|^2_{V(\xi)^{-1}} \quad \text{with} \quad \max_{x \in \mathcal{X}} \|x\|^2_{V(\xi^*)^{-1}} \;=\; d. \tag{8}$$

Consequently, $\mathcal{X}$ is contained in the ellipsoid induced by $V(\xi^*)$:

$$\mathcal{X} \subseteq \mathcal{E}(\xi^*) \quad \text{where} \quad \mathcal{E}(\xi^*) := \Big\{ x \in \mathbb{R}^d : \|x\|^2_{V(\xi^*)^{-1}} \leq d \Big\}.$$

This characterization can be interpreted geometrically: $\mathcal{E}(\xi^*)$ is the minimum-volume centered ellipsoid containing $\mathcal{X}$; see [Lattimore and Szepesvári, 2020, Theorem 21.1]. Moreover, although $\mathcal{X}$ may be infinite, a D-optimal design can be chosen with finite support, in which case $V(\xi^*) = \sum_{i=1}^m w_i x_i x_i^\top$ for some $x_i \in \mathcal{X}$ and weights $w_i \geq 0$, $\sum_i w_i = 1$.

Discrete designs. When a budget of n samples is available, a (discrete) D-optimal design may be posed as the allocation problem

$$\max_{\{n_x\}_{x \in \mathcal{X}}} \log\det\Big(\sum_{x \in \mathcal{X}} n_x \, x x^\top \Big) \qquad \text{s.t.} \qquad n_x \in \mathbb{N}, \;\; \sum_{x \in \mathcal{X}} n_x = n,$$

equivalently as selecting a multiset S of size n to maximize $\log\det\big(\sum_{x \in S} x x^\top\big)$. This is a challenging combinatorial optimization problem. However, if $\lambda \geq 1$, the set function

$$f_\lambda(S) := \log\det\Big(\lambda I + \sum_{x \in S} x x^\top \Big), \qquad S \subseteq \mathcal{X},$$

is a non-negative, monotone, and submodular function. Therefore, under a cardinality constraint, $|S| \leq n$, the standard greedy algorithm achieves a $(1 - 1/e)$-approximation guarantee [Nemhauser et al., 1978]. Hence, Algorithm 4 provides a principled way to select informative queries between two consecutive update steps t_{stop} and t'_{stop}.

E.2 On the intractability of optimistic approaches

Optimism for regret minimization Optimism in the face of uncertainty is a widely used principle for regret minimization in reinforcement learning. In the bandit setting, optimistic algorithms can be applied directly and yield minimax-optimal regret bounds [Auer, 2002]. 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:

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

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

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

where $\tilde{\theta}_t \in \mathcal{E}_t$. The cumulative regret can then be bounded using standard elliptical potential arguments (see Lemma D.4).

Optimism for preference-free exploration As is evident from the proof of Theorem 3.5, algorithms that solely maximize an exploration bonus of the form

$$\|\phi(\pi) - \phi(\pi'_t)\|_{V_t^{-1}} \quad \text{or} \quad \mathbb{E}_{\tau \sim \mathbb{P}_\pi, \tau' \sim \mathbb{P}_{\pi'_t}} \|\phi(\tau) - \phi(\tau')\|^p_{V_t^{-1}}, p = 1, 2,$$

are also effective for preference-free exploration.

https://doi.org/10.52202/085713-0407

Challenge of optimizing the exploration bonus The key problem of the above approaches is that for trajectory-level feedback – preferences or reward – they lead to optimization problems over policies that cannot be framed in terms of state-action rewards. Efroni et al. [2021] therefore conjecture that exactly solving these problems is intractable even in tabular settings. This stands in contrast to standard reinforcement learning with state–action feedback, where optimistic algorithms, and approaches based on optimal design [Wagenmaker et al., 2022, Mutny et al., 2023], involve (or can be reduced to) maximization of bonuses of the form

$$\mathbb{E}_{(s,a)\sim\mu_\pi} \|\phi(s,a)\|^p_{V^{-1}}, p = 1, 2,$$

which are linear in the occupancy measure μ_π and can be maximized via dynamic programming.

These computational challenges associated with exploration under trajectory-level feedback motivate alternative approaches, such as randomized exploration, which can provide near-optimal theoretical guarantees while remaining computationally tractable.

F Experiments

F.1 Gridworld environment

As illustrated in Figure 3 below, the gridworld consists of 36 grid cells and the initial state lies in the center. The agent can choose the actions up, down, left, right, and will deterministically move in that direction or stay if it hits a boundary. The reward features are one-hot features for the six boundary states in blue, and the ground truth reward is 0.5 for two of these states as indicated in Figure 3. Moreover, rewards are discounted with $\gamma = 0.9$. Our gridworld implementation builds on the code by Schlaginhaufen and Kamgarpour [2023].

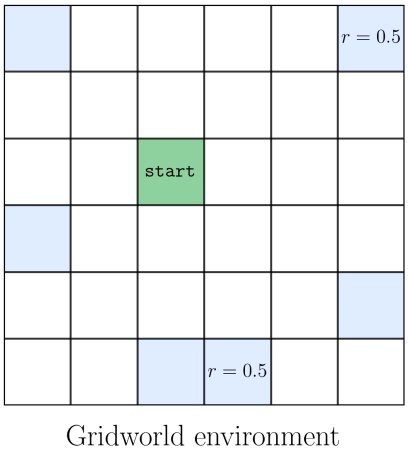


Figure 3: Illustration of the gridworld environment.

F.2 Cartpole environment

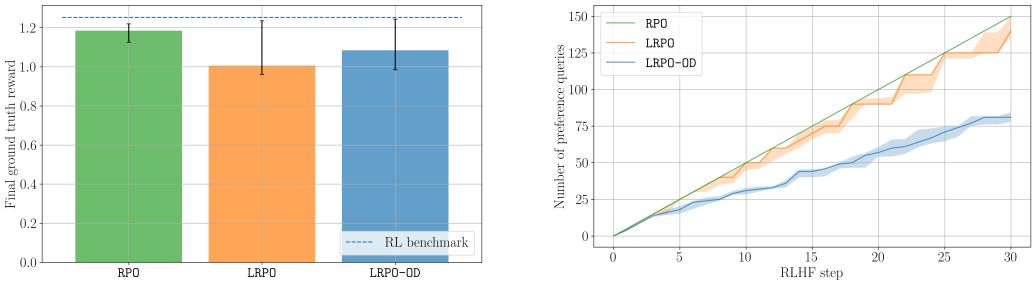
Here, we provide additional details for experiments on Isaac Lab's `Isaac-Cartpole-v0` environment, as well as experimental results for the pure exploration version of our algorithm.

Implementation details All experiments run on Isaac Lab's unmodified `Isaac-Cartpole-v0` environment using the default PPO configuration. As reward features, we use the pre-defined reward terms. For cartpole these are: 1) Alive term: equal to 1 for all non-terminal states; 2) Termination term: equal to 1 for terminal states. A state is terminal if the cart goes out of bounds; 3) Goal tracking term: absolute value of pole angle measured from the upright position; 4) Cart velocity term: absolute value of cart velocity; 5) Joint velocity term: absolute value of pole angular velocity. However, our implementation supports any Isaac Lab manager-based tasks.

We train over 30 RLHF iterations, using 30 steps of PPO at each iteration, and training is repeated for 20 independent seeds. For the randomized exploration, we set $\beta_t = 0.001 + 0.1 \max(1, \log t)$ and

$\lambda = 1$, and for lazy updates we set $C = 0.5$. At each RLHF iteration we compare 100 independently sampled trajectories. For the maximum likelihood estimation we perform 50 Adam steps (batch size 64, ℓ_2 penalty $\lambda = 10^{-1}$). Experiments were executed on a single machine equipped with an Intel i9-14900KS CPU and an NVIDIA RTX 4090 GPU; completing 30 RLHF iterations required approximately 2 min 50 s.

Preference-free exploration Our results for the preference-free exploration algorithm `RPO-Explore` are shown in Figure 4 and 5 below. In Figure 4, we see that all three versions of `RPO-Explore` achieve performance competitive to RL with the ground truth reward[5], but `RPO-OD-Explore` needs the least preference queries. Moreover, in Figure 5 we see that the performance during training is poor, *i.e.* the regret is large, which is to be expected due to the pure exploration scheme.



(a) (b)

Figure 4: Comparison of RLHF algorithms in terms of (a) the last iterate ground truth reward $V_{\theta^*}^{\hat{\pi}}$ (estimated from samples) and (b) number of preference queries performed. In particular, we compare `RPO-Explore` (green, Algorithm 1) with its lazy versions `LRPO-Explore` and `LRPO-OD-Explore` (orange & blue). Here, `LRPO-Explore` and `LRPO-OD-Explore` refer to Algorithm 3 without and with optimal design subroutine. The error bars indicate the 0.2 and 0.8 quantiles, across 20 independent runs. The dashed blue line indicates the mean reward achieved by PPO with the ground truth parameter θ^*.

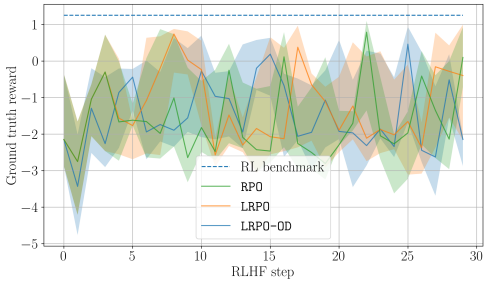


Figure 5: Comparing the ground truth rewards $V_{\theta^*}^{\pi_t}$ (estimated from samples) of RLHF algorithms for reward-free exploration during training, using the same color codes as in Figure 4.

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

https://doi.org/10.52202/085713-0407

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?

https://doi.org/10.52202/085713-0407

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, hyperparameters, 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 20 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.

https://doi.org/10.52202/085713-0407

- 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**

https://doi.org/10.52202/085713-0407

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

Answer: [No]

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

Guidelines:

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