Value-Incentivized Preference Optimization: A Unified Approach to Online and Offline RLHF

Shicong Cen* CMU | Jincheng Mei† Google | Katayoon Goshvadi Google | Hanjun Dai Google | Tong Yang* CMU

Sherry Yang Google | Dale Schuurmans Google | Yuejie Chi* CMU | Bo Dai Google

May 2024

Abstract

Reinforcement learning from human feedback (RLHF) has demonstrated great promise in aligning large language models (LLMs) with human preference. Depending on the availability of preference data, both online and offline RLHF are active areas of investigation. A key bottleneck is understanding how to incorporate uncertainty estimation in the reward function learned from the preference data for RLHF, regardless of how the preference data is collected. While the principles of optimism or pessimism under uncertainty are well-established in standard reinforcement learning (RL), a practically-implementable and theoretically-grounded form amenable to large language models is not yet available, as standard techniques for constructing confidence intervals become intractable under arbitrary policy parameterizations.

In this paper, we introduce a unified approach to online and offline RLHF — value-incentivized preference optimization (VPO) — which regularizes the maximum-likelihood estimate of the reward function with the corresponding value function, modulated by a *sign* to indicate whether the optimism or pessimism is chosen. VPO also directly optimizes the policy with implicit reward modeling, and therefore shares a simpler RLHF pipeline similar to direct preference optimization. Theoretical guarantees of VPO are provided for both online and offline settings, matching the rates of their standard RL counterparts. Moreover, experiments on text summarization and dialog verify the practicality and effectiveness of VPO.

1 Introduction

Fine-tuning large language models (LLMs) by *reinforcement learning from human feedback* (RLHF) (Ziegler et al., 2019) has been shown to significantly improve the helpfulness, truthfulness and controllability of LLMs, as illustrated by InstructGPT (Ouyang et al., 2022) and many follow-ups. Roughly speaking, there are two critical components of RLHF: (1) *reward modeling*, which maps human preference rankings into a quantitative reward function that can guide policy improvement; and (2) *RL fine-tuning*, which seeks to adjust LLM output to align with human preferences by leveraging the learned reward function, i.e., increasing the probability of preferred answers and decreasing the probability of unfavored answers.

Evidently, the curation of preference data is instrumental in the performance of RLHF, which is commonly modeled as pairwise comparisons from a Bradley-Terry ranking model (Bradley and Terry, 1952). In particular, given a query x, human annotators choose a preferred answer from two candidate answers y_1 and y_2 generated by an LLM. Despite the simple form, collecting large-scale and high-quality preference data can be expensive and time-consuming. Depending on the availability of preference data, two paradigms of RLHF are considered: (1) *offline* RLHF, where only a pre-collected preference dataset is available, possibly generated from a

*Carnegie Mellon University; emails: `{shicongc,tongyang,yuejiec}@andrew.cmu.edu`.

†Google Research; emails: `{jcmei,kgoshvadi,hadai,sherryy,schuurmans,bodai}@google.com`.

pre-trained LLM after supervised fine-tuning (SFT); and (2) *online* RLHF, where additional preference data can be collected adaptively to improve alignment. While initial work on RLHF focused on the offline setting, the online setting has also begun to receive considerable attention, as even a small amount of additional preference data has been shown to greatly boost performance.

There has been significant work on the theoretical underpinnings of RLHF that seeks to uncover algorithmic improvements. Notably, while the original RLHF pipeline decouples reward modeling from RL fine-tuning, direct preference optimization (DPO) (Rafailov et al., 2023) integrates these as a single step in the *offline* setting, leveraging a closed-form solution for the optimal policy in the RL fine-tuning phase. This has led to a welcome simplification of the RLHF pipeline, allowing direct optimization of the policy (i.e., the LLM) from preference data.

Nevertheless, significant challenges remain in RLHF, particularly concerning how to incorporate estimates of reward *uncertainty* in direct preference optimization when parameterizing policies with large-scale neural networks — such as LLMs — in a theoretically and practically effective manner. In standard reinforcement learning (RL), managing uncertainty when an agent interacts with an environment is a critical aspect in achieving near-optimal performance (Sutton and Barto, 2018), when using methods that range from policy-based (Schulman et al., 2017; Xiao et al., 2021), value-based (Mnih et al., 2015; Kumar et al., 2020), and actor-critic methods (Mnih et al., 2016). One dominant approach in the bandit setting, for example, is to construct confidence intervals of the reward estimates, then acting according to the upper and lower confidence bounds — following the principles of optimism and pessimism in the online and offline settings respectively (Lattimore and Szepesvári, 2020; Lai et al., 1985; Rashidinejad et al., 2022).

Despite the fact that uncertainty estimation is even more critical in RLHF, due to the coarse nature of preference data, effective implementations of theoretically justified optimistic and pessimistic principles have yet to be developed in the RLHF literature. For example, existing online preference alignment methods, such as Nash-MD (Munos et al., 2023) and OAIF (Guo et al., 2024), do not incorporate exploration; similarly, pessimism is also not implemented in offline preference alignment methods, such as DPO (Rafailov et al., 2023) and IPO (Azar et al., 2024). A key reason for these omissions is that it is extremely difficult to construct confidence intervals for arbitrary neural networks (Gawlikowski et al., 2021), let alone LLMs. Since optimism for online exploration and pessimism for offline RL both require uncertainty estimation, and given the difficulty of conducting uncertainty estimation for large-scale neural networks, a natural and important question arises:

Can we implement the optimistic/pessimistic principles under uncertainty in a practically efficient manner for online/offline preference alignment in LLMs while retaining theoretical guarantees?

1.1 Our contributions

In this paper, we provide affirmative answer to the question. Our major contributions are as follows.

- **(i)** We propose value-incentivized preference optimization (VPO) for both online and offline RLHF, a unified algorithmic framework that *directly optimizes the LLM policy* with the optimistic/pessimistic principles under uncertainty. Avoiding explicit uncertainty estimation, VPO regularizes maximum likelihood estimation of the reward function toward (resp. against) responses that lead to the highest value in the online (resp. offline) setting, hence implementing optimism (resp. pessimism). Theoretical regret guarantees of VPO are developed for both online and offline RLHF, matching their corresponding rates in the standard RL literature with explicit uncertainty estimation.
- **(ii)** In addition, VPO reveals the critical role of reward calibration, where the shift ambiguity of the reward model inherent in the Bradley-Terry model (Bradley and Terry, 1952) can be exploited to implement additional behavior regularization (Pal et al., 2024; Ethayarajh et al., 2024). This allows VPO to provide a theoretical foundation for popular conservative offline RL methods (e.g., (Kumar et al., 2020)), as well as regularized RLHF methods (e.g., DPOP (Pal et al., 2024)).
- **(iii)** VPO admits a practically-implementable form suitable for RLHF on LLMs, and more generally, deep-learning architectures. We conduct extensive experimental studies using `TL;DR` and `ARC-Challenge`

tasks in online and offline settings with optimistic and pessimistic bias, respectively. The results demonstrate improved empirical performance.

1.2 Related work

RLHF. Since the introduction of the original RLHF framework, there have been many proposed simplifications of the preference alignment procedure and attempts to improve performance, including SLiC (Zhao et al., 2023), GSHF (Xiong et al., 2023), DPO (Rafailov et al., 2023), and its variants, such as Nash-MD (Munos et al., 2023), IPO (Azar et al., 2024), OAIF (Guo et al., 2024), SPO (Swamy et al., 2024), GPO (Tang et al., 2024), and DPOP (Pal et al., 2024). These methods can roughly be grouped into online and offline variants, depending on whether preference data is collected before training (offline) or by using the current policy during training (online).

In offline preference alignment, identity preference optimization (IPO, (Azar et al., 2024)) argues that it is problematic to use the Bradley-Terry model in DPO to convert pairwise preferences into pointwise reward values, and proposes an alternative objective function to bypass the use of the Bradley-Terry model. DPO-Positive (DPOP, (Pal et al., 2024)) observes a failure mode of DPO that the standard DPO loss can reduce the model's likelihood on preferred answers, and proposes to add a regularization term to the DPO objective to avoid such a failure mode. On the other hand, online AI feedback (OAIF, (Guo et al., 2024)) proposes an online version of DPO, where online preference data from LLM annotators is used to evaluate and update the current LLM policy in an iterative manner. Iterative reasoning preference optimization (Iterative RPO, (Yuanzhe Pang et al., 2024)) proposes to add an additional negative log-likelihood term in the DPO loss to improve performances on reasoning tasks. Finally, (Chang et al., 2024) proposes to reuse the offline preference data via reset.

Uncertainty estimation in RL. The principles of optimism and pessimism are typically implemented via constructing confidence intervals or posterior sampling, which have been demonstrated to be provably efficient in tabular settings (Jin et al., 2018; Shi et al., 2022). Yet, these approaches have had limited success in conjunction with deep learning architectures (Gawlikowski et al., 2021), and many empirical heuristics in turn lack theoretical validation (Kumar et al., 2020). VPO draws inspiration from reward-biased exploration (Kumar and Becker, 1982; Liu et al., 2020, 2024; Hung et al., 2021; Mete et al., 2021) in the standard online RL literature, but significantly broadens its scope to the offline setting and RLHF for the first time.

2 Preliminaries

In RLHF, a language model is described by a policy π, which generates an answer $y \in \mathcal{Y}$ given prompt $x \in \mathcal{X}$ according to the conditional probability distribution $\pi(\cdot|x)$. The standard RLHF process consists of four stages: supervised fine-tuning (SFT), preference data generation, reward modeling, and RL fine-tuning. In the SFT stage, a language model $\pi_{\rm sft}$ is obtained by fine-tuning a pre-trained LLM with supervised learning. The remaining stages continue training by leveraging the preference data, which we elaborate below.

Reward modeling from preference data. An oracle (e.g., a human labeler or a scoring model) evaluates the quality of two answers y_1 and y_2 given prompt x and reveals its preference. A widely used approach for modelling the probability of pairwise preferences is the Bradley–Terry model (Bradley and Terry, 1952):

$$\mathbb{P}(y_1 \succ y_2|x) = \frac{\exp(r^\star(x, y_1))}{\exp(r^\star(x, y_1)) + \exp(r^\star(x, y_2))} = \sigma(r^\star(x, y_1) - r^\star(x, y_2)), \tag{1}$$

where $y_1 \succ y_2$ indicates that y_1 is preferred over y_2, $r^\star : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ is the ground truth reward function, and $\sigma : \mathbb{R} \to (0, 1)$ is the logistic function. A preference data sample is denoted by a tuple (x, y_+, y_-), where y_+ (resp. y_-) is the preferred (resp. unpreferred) answer in the comparison.

Given a preference dataset $\mathcal{D} = \{(x^i, y_+^i, y_-^i)\}$ composed of independent samples, the reward function r can be estimated by maximum likelihood estimation (MLE):

$$r_{\mathsf{MLE}} = \arg\min_r \ \ell(r, \mathcal{D}), \tag{2}$$

where $\ell(r, \mathcal{D})$ is the negative log-likelihood of $\mathcal{D}$, given as

$$\ell(r, \mathcal{D}) \coloneqq - \sum_{(x^i, y_+^i, y_-^i) \in \mathcal{D}} \log \sigma(r(x^i, y_+^i) - r(x^i, y_-^i)). \tag{3}$$

RL fine-tuning. Given a reward model r, we seek to fine-tune the policy π to achieve an ideal balance between the expected reward and its distance from an initial policy π_{ref}, which is typically the same as π_{sft}. This is achieved by maximizing the KL-regularized value function $J(r, \pi)$, defined as

$$J(r, \pi) = \mathop{\mathbb{E}}_{x \sim \rho, y \sim \pi(\cdot|x)} [r(x, y)] - \beta \mathop{\mathbb{E}}_{x \sim \rho} \left[\mathsf{KL}\big(\pi(\cdot|x) \,\|\, \pi_{\mathrm{ref}}(\cdot|x)\big)\right], \tag{4}$$

where $\mathsf{KL}\big(\pi_1 \,\|\, \pi_2\big)$ is the KL divergence from π_1 to π_2, and $\beta > 0$ is a regularization parameter. Consequently, the RL fine-tuned policy π_r with respect to the reward r satisfies

$$\pi_r \coloneqq \arg\max_\pi J(r, \pi), \tag{5}$$

which admits a closed-form solution (Rafailov et al., 2023), i.e.,

$$\forall (x \times y) \in \mathcal{X} \times \mathcal{Y}: \qquad \pi_r(y|x) = \frac{\pi_{\mathrm{ref}}(y|x) \exp(r(x, y)/\beta)}{Z(r, x)}. \tag{6}$$

Here, $Z(r, x)$ is a normalization factor given by

$$Z(r, x) = \sum_{y' \in \mathcal{Y}} \pi_{\mathrm{ref}}(y'|x) \exp(r(x, y')/\beta). \tag{7}$$

Direct preference optimization. The closed-form solution (6) allows us to write the reward function r in turn as

$$r(x, y) = \beta(\log \pi_r(y|x) - \log \pi_{\mathrm{ref}}(y|x) + \log Z(r, x)). \tag{8}$$

Plugging the above equation into the reward MLE (2), we obtain the seminal formulation of direct preference optimization (DPO) over the policy space (Rafailov et al., 2023),

$$\pi_{\mathsf{DPO}} = \arg\min_\pi \ - \sum_{(x^i, y_+^i, y_-^i) \in \mathcal{D}} \log \sigma \left(\beta \left(\log \frac{\pi(y_+^i|x)}{\pi_{\mathrm{ref}}(y_+^i|x)} - \log \frac{\pi(y_-^i|x)}{\pi_{\mathrm{ref}}(y_-^i|x)} \right) \right), \tag{9}$$

which avoids explicitly learning the reward model.

3 Value-Incentivized Preference Optimization

A major caveat of the standard RLHF framework concerns the lack of accounting for reward uncertainty, which is known to be indispensable in the success of standard RL paradigms in both online and offline settings (Cesa-Bianchi et al., 2017; Rashidinejad et al., 2022). This motivates us to investigate a principled mechanism that be easily integrated into the RLHF pipeline, while bypassing the difficulties of explicit uncertainty estimation in LLMs.

3.1 General framework

In view of the sub-optimality of naive MLE for reward estimation (Cesa-Bianchi et al., 2017; Rashidinejad et al., 2022), and motivated by the effectiveness of reward-biased MLE in online RL (Kumar and Becker, 1982; Liu et al., 2020, 2024), we propose to regularize the reward estimate via

$$J^\star(r) = \max_\pi J(r, \pi), \tag{10}$$

which measures the resulting value function for the given reward if one acts according to its optimal policy. However, in RLHF, by the definition (1), the reward function $r^\star$ is only identifiable up to a prompt-dependent global shift. Specifically, letting $r_1(x, y) = r_2(x, y) + c(x)$ be two reward functions that only differ by a prompt-dependent shift $c(x)$, we have $r_1(x, y_1) - r_1(x, y_2) = r_2(x, y_1) - r_2(x, y_2)$, which leads to $J^\star(r_1) = J^\star(r_2) + \mathbb{E}_{x\sim\rho}[c(x)]$. To resolve this challenge, we introduce the following equivalent class of reward functions for the Bradley-Terry model to eliminate the shift ambiguity, which also has the calibration effect of centering the reward function while offering a regularization mechanism to incorporate additional policy preferences.

Assumption 1 *We assume that $r^\star \in \mathcal{R}$, where*

$$\mathcal{R} = \left\{ r : \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{cal}(\cdot|x)} [r(x, y)] = 0. \right\}. \tag{11}$$

Here, ρ is the prompt distribution and π_{cal} is a fixed calibration distribution independent of the algorithm.

The proposed regularized MLE of the Bradley-Terry model (2) appends a bias term to the negative likelihood

$$r_{\mathsf{VPO}} = \arg\min_{r\in\mathcal{R}} \{\ell(r, \mathcal{D}) - \mathsf{sign} \cdot \alpha \cdot J^\star(r)\}, \tag{12}$$

incentivizing the algorithm to favor (resp. avoid) reward models with higher value $J^\star(r)$ in the online (resp. offline) setting. Here, $\alpha > 0$ is a constant controlling the strength of regularization, and sign is set to 1 in the online setting and -1 in the offline setting.

At first glance, the objective function for VPO (12) does not immediately imply a computationally-efficient algorithm due to the presence of $J^\star(r)$. However, by exploiting the same closed-form solution for the optimal policy given the reward in (6), and the reward representation inferred from the policy via (8), we can explicitly express $J^\star(r)$ as

$$\begin{aligned}
J^\star(r) &= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} [r(x, y) - \beta(\log \pi_r(y|x) - \log \pi_{\text{ref}}(y|x))] \\
&= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_r(\cdot|x)} [\log Z(r, x)] \\
&= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} [\log Z(r, x)] \\
&= \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} [r(x, y) - \beta(\log \pi_r(y|x) - \log \pi_{\text{ref}}(y|x))] \\
&= -\beta \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} [\log \pi_r(y|x) - \log \pi_{\text{ref}}(y|x)],
\end{aligned} \tag{13}$$

where the second step follows because the bracketed term is independent of y (c.f. (6)) and the last step follows from (11) whenever $r \in \mathcal{R}$. Given this key ingredient, we can then rewrite (12) to directly optimize the LLM policy, in a flavor similar to DPO, as

$$\begin{aligned}
\pi_{\mathsf{VPO}} &= \mathop{\mathrm{argmin}}_{\pi_r :\, r\in\mathcal{R}} \{\ell(r, \mathcal{D}) - \mathsf{sign} \cdot \alpha \cdot J^\star(r)\} \\
&= \mathop{\mathrm{argmin}}_{\pi_r :\, r\in\mathcal{R}} \Big\{ - \sum_{(x^i, y^i_+, y^i_-)\in\mathcal{D}} \log \sigma\Big(\beta \log \frac{\pi_r(y^i_+|x^i)}{\pi_{\text{ref}}(y^i_+|x^i)} - \beta \log \frac{\pi_r(y^i_-|x^i)}{\pi_{\text{ref}}(y^i_-|x^i)}\Big)
\end{aligned}$$

$$
\begin{aligned}
&\quad + \mathsf{sign} \cdot \alpha\beta \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} [\log \pi_r(y|x) - \log \pi_{\text{ref}}(y|x)] \Big\} \\
&= \arg\min_{\pi} \Big\{ - \sum_{(x^i, y_+^i, y_-^i)\in\mathcal{D}} \log \sigma\Big(\beta \log \frac{\pi(y_+^i|x^i)}{\pi_{\text{ref}}(y_+^i|x^i)} - \beta \log \frac{\pi(y_-^i|x^i)}{\pi_{\text{ref}}(y_-^i|x^i)}\Big) \\
&\quad + \mathsf{sign} \cdot \alpha\beta \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} [\log \pi(y|x) - \log \pi_{\text{ref}}(y|x)] \Big\},
\end{aligned} \tag{14}
$$

where we drop the constraint on $r \in \mathcal{R}$, since for any policy π there exists $r \in \mathcal{R}$ such that $\pi = \pi_r$.

Observing that the reference policy $\pi_{\text{ref}}(y|x)$ in the last term of (14) $\mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} [\log \pi(y|x) - \log \pi_{\text{ref}}(y|x)]$ does not impact the optimization solution, we can change it to $\mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} [\log \pi(y|x) - \log \pi_{\text{cal}}(y|x)] \Big\} = - \mathop{\mathbb{E}}_{x\sim\rho} \big[\mathsf{KL}\big(\pi_{\text{cal}}(\cdot|x) \,\|\, \pi(\cdot|x)\big)\big]$, which amounts to adding a KL regularization to the original DPO, and offers an interesting interpretation as pushing π against/towards π_{cal} in the online/offline settings respectively, unveiling the role of reward calibration in RLHF.

In what follows, we elaborate the development of VPO in both the online and offline settings with corresponding theoretical guarantees under linear function approximation.

3.2 Online RLHF: algorithm and theory

The online RLHF procedure extends training by performing reward learning and policy learning iteratively, with a growing preference dataset collected by using the current policy. We use $\pi^{(t)}$ to denote the policy used in the t-th iteration, where the superscript ${}^{(t)}$ indicates iteration t in the online setting. The t-th iteration of VPO for online RLHF consists of the following steps:

1. **New preference data generation.** We sample a new prompt $x^{(t)} \sim \rho$ and two answers $y_1^{(t)}, y_2^{(t)} \sim \pi^{(t)}(\cdot|x^{(t)})$, query the preference oracle and append $(x^{(t)}, y_+^{(t)}, y_-^{(t)})$ to the preference dataset.

2. **Reward learning.** We train a reward model with preference data $\mathcal{D}^{(t)} := \{(x^{(s)}, y_+^{(s)}, y_-^{(s)})\}_{s=1}^t$ by minimizing the regularized negative log-likelihood, i.e.,

$$
r^{(t+1)} = \arg\min_{r\in\mathcal{R}} \{\ell(r, \mathcal{D}^{(t)}) - \alpha \cdot J^\star(r)\}. \tag{15}
$$

3. **Policy learning.** This step trains the policy by solving the RL fine-tuning problem:

$$
\pi^{(t+1)} = \arg\max_{\pi} J(r^{(t+1)}, \pi). \tag{16}
$$

We summarize the detailed procedure in Algorithm 1.

Theoretical analysis. Encouragingly, VPO admits appealing theoretical guarantees under function approximation. For simplicity, we restrict attention to linear approximation of the reward model.

Assumption 2 (Linear Reward) *We parameterize the reward model by*

$$
r_\theta(x, y) = \big\langle \phi(x, y), \theta \big\rangle, \quad \forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \tag{18}
$$

where $\phi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^d$ is a fixed feature mapping and $\theta \in \mathbb{R}^d$ is the parameters. We assume that $\|\phi(x, y)\|_2 \le 1$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, and that $r^\star(x, y) = \big\langle \phi(x, y), \theta^\star \big\rangle$ for some $\theta^\star$.

Under Assumption 1 and 2, it is sufficient to focus on $\theta \in \Theta$ where

$$
\Theta = \Big\{\theta \in \mathbb{R}^d : \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} \big[\big\langle \phi(x, y), \theta \big\rangle\big] = 0\Big\}. \tag{19}
$$

Algorithm 1 VPO for online RLHF

initialization: $\pi^{(0)}$.

for $t = 0, 1, 2, \cdots$ **do**

Sample $x^{(t)} \sim \rho$, $y_1^{(t)}, y_2^{(t)} \sim \pi^{(t)}(\cdot|x^{(t)})$.

Obtain the preference between $(x^{(t)}, y_1^{(t)})$ and $(x^{(t)}, y_2^{(t)})$ from some oracle. Denote the comparison outcome by $(x^{(t)}, y_+^{(t)}, y_-^{(t)})$.

Update policy π as

$$\pi^{(t+1)} = \underset{\pi}{\operatorname{argmin}} \Big\{ -\sum_{s=1}^{t} \log\sigma\Big(\beta \log \frac{\pi(y_+^{(s)}|x^{(s)})}{\pi_{\text{ref}}(y_+^{(s)}|x^{(s)})} - \beta \log \frac{\pi(y_-^{(s)}|x^{(s)})}{\pi_{\text{ref}}(y_-^{(s)}|x^{(s)})}\Big) + \alpha\beta \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} [\log\pi(y|x) - \log\pi_{\text{ref}}(y|x)] \Big\}. \tag{17}$$

end

The next theorem demonstrates that Algorithm 1 achieves $\widetilde{\mathcal{O}}(\sqrt{T})$ cumulative regret under mild assumptions. The proof is provided in Appendix A. The proof logic follows from that of (Liu et al., 2024).

Theorem 1 *Under Assumptions 1 and 2, let $r_{\theta^{(t)}} \in \Theta$ denote the corresponding reward model for $\pi^{(t)}$. Assume that $\|\theta^\star\|_2 \le C$ and $\|\theta^{(t)}\|_2 \le C, \forall t \ge 0$ for some $C > 0$. Then with probability $1-\delta$ we have*

$$\mathsf{Regret} \coloneqq \sum_{t=1}^{T} \big[J^\star(r^\star) - J(r^\star, \pi^{(t)})\big] \le \widetilde{\mathcal{O}}(\exp(2C + C/\beta)\sqrt{\kappa d T}),$$

with $\alpha = \dfrac{1}{\exp(2C + C/\beta)} \sqrt{\dfrac{T}{\kappa \min\{d\log T, T\}}}$ *and* $\kappa = \sup_{x,y} \dfrac{\pi_{cal}(y|x)}{\pi_{ref}(y|x)}$.

Theorem 1 shows that VPO achieves the same $\widetilde{O}(\sqrt{T})$ regret for online RLHF as its counterparts in standard contextual bandits with scalar rewards and using UCB for exploration (Lattimore and Szepesvári, 2020).

Remark 1 *The analysis naturally extends to allowing mini-batch samples of size M in every iteration, yielding an improved regret bound scaled by $1/\sqrt{M}$ and α scaled by $\sqrt{M}$.*

3.3 Offline RLHF: algorithm and theory

In offline RLHF, a fixed offline preference dataset is collected $\mathcal{D} \coloneqq \{x^i, y_+^i, y_-^i\}_{i=1}^N$, where $x^i \sim \rho$, $y^i \sim \pi_{\mathsf{b}}(\cdot|x)$ are sampled from a behavior policy π_{b}, such as π_{sft} from SFT. The proposed VPO for offline RLHF consists of one pass through the reward and policy learning phases, i.e.,

$$\widehat{r} = \arg\min_{r\in\mathcal{R}} \{\ell(r, \mathcal{D}) + \alpha \cdot J^\star(r)\} \qquad \text{and} \qquad \widehat{\pi} = \arg\max_{\pi} J(\widehat{r}, \pi), \tag{20}$$

which discourages over-optimization of the reward function given the limited offline preference data. In the same vein as deriving (17), and by leveraging (13), we obtain the direct policy update rule:

$$\widehat{\pi} = \arg\min_{\pi} \Big\{ -\sum_{i=1}^{N} \log\sigma\Big(\beta \log \frac{\pi(y_+^i|x^i)}{\pi_{\text{ref}}(y_+^i|x^i)} - \beta \log \frac{\pi(y_-^i|x^i)}{\pi_{\text{ref}}(y_-^i|x^i)}\Big) - \alpha\beta \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\text{cal}}(\cdot|x)} [\log\pi(y|x) - \log\pi_{\text{ref}}(y|x)] \Big\}. \tag{21}$$

We summarize the detailed procedure in Algorithm 2. When π_{cal} is set to π_{ref}, the regularization term becomes the KL divergence between π and π_{ref}, which is reminiscent of a popular choice in offline RL practice (Kumar et al., 2020). Another heuristic choice is to set π_{cal} to the marginalized positive answer distribution from the dataset, *i.e.*, $(x, y_+) \sim \mathcal{D}$, which leads to a similar objective in (Pal et al., 2024).

Algorithm 2 VPO for offline RLHF

input: offline preference data $\mathcal{D}$ of size N.

Get policy $\widehat{\pi}$ by optimizing

$$\widehat{\pi} = \arg\min_{\pi} \Big\{ -\sum_{i=1}^{N} \log \sigma\Big(\beta \log \frac{\pi(y_+^i|x^i)}{\pi_{\mathrm{ref}}(y_+^i|x^i)} - \beta \log \frac{\pi(y_-^i|x^i)}{\pi_{\mathrm{ref}}(y_-^i|x^i)}\Big) - \alpha\beta \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{\mathrm{cal}}(\cdot|x)} [\log \pi(y|x) - \log \pi_{\mathrm{ref}}(y|x)] \Big\}.$$

Saddle-point characterization and pessimism. We first illustrate that VPO indeed executes the principle of pessimism in a complementary manner to the standard approach of pessimism, which finds a policy that maximizes the worst-case value function over a confidence set. In particular, this strategy (Uehara and Sun, 2021) obtains a policy by solving

$$\widehat{\pi}_{\mathsf{LCB}} = \arg\max_{\pi} \min_{r\in\mathcal{R}_\delta} J(r, \pi) \tag{22}$$

where the confidence set $\mathcal{R}_\delta$ is typically set to $\{r : \ell(r, \mathcal{D}) \le \ell(r_{\mathsf{MLE}}, \mathcal{D}) + \delta\}$ or $\{r : \mathsf{dist}(r, r_{\mathsf{MLE}}) \le \delta\}$ for some $\delta > 0$ and s distance measure dist. Turning to VPO, note that by (20) we have

$$\widehat{r} = \arg\min_{r} \big\{\ell(r, \mathcal{D}) + \alpha J^\star(r)\big\} = \arg\min_{r} \max_{\pi} \big\{\ell(r, \mathcal{D}) + \alpha J(r, \pi)\big\}. \tag{23}$$

Since $\ell(r, \mathcal{D}) + \alpha J(r, \pi)$ is strongly concave over π, and convex over r, it allows us to formulate $(\widehat{r}, \widehat{\pi})$ as a saddle point in the following lemma. The proof is given in Appendix B.1.

Lemma 1 *$(\widehat{r}, \widehat{\pi})$ is a saddle point of the objective $\ell(r, \mathcal{D}) + \alpha J(r, \pi)$,* i.e.*, for any (r', π'), we have*

$$\begin{cases} \ell(\widehat{r}, \mathcal{D}) + \alpha J(\widehat{r}, \widehat{\pi}) \le \ell(r', \mathcal{D}) + \alpha J(r', \widehat{\pi}) \\ \ell(\widehat{r}, \mathcal{D}) + \alpha J(\widehat{r}, \widehat{\pi}) \ge \ell(\widehat{r}, \mathcal{D}) + \alpha J(\widehat{r}, \pi') \end{cases}.$$

As such, the policy obtained by VPO can be equivalently written as

$$\widehat{\pi} \in \arg\max_{\pi} \min_{r} \Big\{J(r, \pi) + \frac{1}{\alpha}\ell(r, \mathcal{D})\Big\} = \arg\max_{\pi} \min_{r\in\mathcal{R}_{\delta(\pi,\alpha)}} J(r, \pi), \tag{24}$$

where $\mathcal{R}_{\delta(\pi,\alpha)}$ is the constraint set $\{r : \ell(r, \mathcal{D}) \le \ell(r_{\mathsf{MLE}}, \mathcal{D}) + \delta(\pi, \alpha)\}$ such that the constrained optimization problem $\min_{r\in\mathcal{R}_{\delta(\pi,\alpha)}} J(r, \pi)$ is equivalent to the regularized problem $\min_r \big\{J(r, \pi) + \frac{1}{\alpha}\ell(r, \mathcal{D})\big\}$. In view of the similarity between the formulations (22) and (24), we conclude that VPO implements the pessimism principle (22) in an oblivious manner without explicitly estimating the uncertainty level, justifying popular practice as a valid approach to pessimism (Kumar et al., 2020).

Theoretical analysis. The next theorem establishes the sub-optimality gap of VPO with linear function approximation under mild assumptions. The proof is given in Appendix B.

Theorem 2 *Under Assumptions 1 and 2, let $\widehat{\theta} \in \Theta$ denote the corresponding reward model for $\widehat{\pi}$. Assume that $\|\theta^\star\|_2 \le C$ and $\|\widehat{\theta}\|_2 \le C$ for some $C > 0$. Let $\alpha = \sqrt{N}$ and $\delta \in (0, 1)$. With probability $1 - \delta$, we have*

$$J^\star(r^\star) - J(r^\star, \widehat{\pi}) \le \mathcal{O}\Bigg(\frac{C_1}{\sqrt{N}} \cdot \Big\| \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi^\star(\cdot|x)}} [\phi(x, y)] \Big\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} + \frac{C_2}{\sqrt{N}}\Bigg),$$

where $\Sigma_{\mathcal{D}} = \frac{1}{N}\sum_{i=1}^{N}(\phi(x^i, y_+^i) - \phi(x^i, y_-^i))(\phi(x^i, y_+^i) - \phi(x^i, y_-^i))^\top$ *is the feature sample covariance matrix,* $\lambda = 1/N$, $C_1 = \exp(C)\left(\sqrt{d + \log(1/\delta)} + \kappa_{\mathcal{D}}\right) + C$ *and* $C_2 = \exp(C)\kappa_{\mathcal{D}}^2 + C\kappa_{\mathcal{D}} + 1$. *Here,*

$$\kappa_{\mathcal{D}} = \left\| \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\widehat{\pi}(\cdot|x)}}[\phi(x,y)] - \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{cal}(\cdot|x)}}[\phi(x,y)] \right\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} \leq 4(\lambda_{\min}(\Sigma_{\mathcal{D}}) + \lambda)^{-1}.$$

Theorem 2 establishes that VPO achieves the same rate of $\widetilde{\mathcal{O}}(1/\sqrt{N})$ as standard offline RL, as long as the offline dataset $\mathcal{D}$ has sufficient coverage. We remark that $\left\| \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi^\star(\cdot|x)}}[\phi(x,y)] \right\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}$ is reminiscent of the standard single-policy concentratability coefficient in offline RL, which measures the distribution shift between the offline dataset and the optimal policy (Zhu et al., 2023).

4 Token-level VPO

Recently, Rafailov et al. (2024) offered an interpretation of DPO using the token-level Markov Decision Process (MDP), aiming at reconciling the gap between the practical fine-tuning of LLMs at the token level and the theoretical formulation of DPO at the sentence level. Fortunately, this interpretation does not require algorithmic modifications. Below we provide a short discussion and highlight that VPO can be similarly interpreted using the token-level MDP.

Token-level MDP and preference modeling. Recall that in LLMs, the prompt x can be broken into a sequence of tokens, e.g., $x = (x_0, \ldots, x_m)$, from a fixed discrete vocabulary $\mathcal{A}$. We define the token-level MDP as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r^\star, H)$, where H is the horizon length, i.e., the longest possible number of tokens in a sentence. The state space $\mathcal{S}$ consists of all the possible token combinations of length H, and the transition kernel is deterministically defined as follows.

1. The initial state is defined by the prompt x as $s_0 = \{x_0, \cdots, x_m\}$;
2. Given the response $y = \{y_0, \cdots, y_{i-1}\}$ up to the i-th token, the state at step i is defined as $s_i = \{x_0, \cdots, x_m, y_0, \cdots, y_{i-1}\}$;
3. Upon an action of the LLM for the next token $a_i = y_i$, the next state at the token-level MDP deterministically becomes $s_{i+1} = (s_i, a_i) = (x_0, \cdots, x_m, y_0, \cdots, y_i)$.

We assume that the last token of a sentence, the EOS token, is absorbing, such that the token-level MDP stays in the corresponding state as soon as the last action is the EOS token. With slight abuse of notation from earlier sections, the reward function $r^\star(s, a)$ defines the ground truth reward at state s upon action a.

Given a pair of trajectories $\tau_1 = \{s_0, a_0^1, \ldots, s_{H-1}^1, a_{H-1}^1, s_H^1\}$ and $\tau_2 = \{s_0, a_0^2, \ldots, s_{H-1}^2, a_{H-1}^2, s_H^2\}$, the corresponding Bradley-Terry preference model (Bradley and Terry, 1952) is

$$\begin{aligned}\mathbb{P}(\tau_1 \succ \tau_2) &= \frac{\exp\left(\sum_{i=0}^{H-1} r^\star(s_i^1, a_i^1)\right)}{\exp\left(\sum_{i=0}^{H-1} r^\star(s_i^1, a_i^1)\right) + \exp\left(\sum_{i=0}^{H-1} r^\star(s_i^2, a_i^2)\right)}\\ &= \sigma\left(\sum_{i=0}^{H-1} r^\star(s_i^1, a_i^1) - \sum_{i=0}^{H-1} r^\star(s_i^2, a_i^2)\right),\end{aligned}$$

A preference data sample is denoted by a tuple (x, τ_+, τ_-), where τ_+ (resp. τ_-) is the preferred (resp. unpreferred) answer in the comparison. Given a preference dataset $\mathcal{D}$ composed of independent samples, The negative log-likelihood can be defined as

$$\ell(r, \mathcal{D}) := -\sum_{(\tau_+,\tau_-)\in\mathcal{D}} \log\sigma\left(\sum_{i=0}^{H-1} r(s_i^+, a_i^+) - \sum_{i=0}^{H-1} r(s_i^-, a_i^-)\right), \tag{25}$$

where (s_i^+, a_i^+) (resp. (s_i^-, a_i^-)) are the state-action pairs in the trajectory τ_+ (resp. τ_-).

Token-level RL fine-tuning. Let the entropy of policy π under initial state distribution $s_0 \sim \rho$ be defined as

$$\mathcal{H}(\rho, \pi) \coloneqq - \mathop{\mathbb{E}}_{\substack{s_0 \sim \rho, \\ a_i \sim \pi(\cdot|s_i)}} \left[\sum_{i=0}^{H-1} \log \pi(a_i|s_i) \right],$$

and $\mathcal{H}(s, \pi)$ be the entropy when the initial state $s_0 = s$. Given a reward function r, we define the KL-constrained RL objective against the reference policy π_{ref} as (Rafailov et al., 2024):

$$\pi_r \coloneqq \operatorname*{argmax}_{\pi} J(r, \pi) \coloneqq \mathop{\mathbb{E}}_{\substack{s_0 \sim \rho, \\ a_i \sim \pi(\cdot|s_i)}} \Big[\sum_{i=0}^{H-1} \underbrace{\big(r(s_i, a_i) + \beta \log \pi_{\text{ref}}(a_i|s_i)\big)}_{r_\beta(s_i, a_i)} \Big] + \beta \mathcal{H}(\rho, \pi), \tag{26}$$

where we denote $r_\beta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ as follows:

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} : \quad r_\beta(s, a) \coloneqq r(s, a) + \beta \log \pi_{\text{ref}}(a|s), \tag{27}$$

which can be seen as the actual token-wise reward function optimized by the LLM.

Token-level DPO. The KL-constrained RL objective (26) has a closed-form solution (Nachum et al., 2017; Cen et al., 2022) given by

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} : \quad \pi_r(a|s) = \exp((Q_\beta^\star(s, a) - V_\beta^\star(s))/\beta), \tag{28}$$

where $V_\beta^\star : \mathcal{S} \to \mathbb{R}$ and $Q_\beta^\star : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ are the optimal soft value and Q functions, respectively,

$$\forall s \in \mathcal{S} : \quad V_\beta^\star(s) \coloneqq \mathop{\mathbb{E}}_{a_i \sim \pi_r(\cdot|s_i)} \left[\sum_{i=0}^{H-1} r_\beta(s_i, a_i) | s_0 = s \right] + \beta \mathcal{H}(s, \pi_r) \tag{29}$$

denote the optimal soft value function w.r.t. the reward function r_β, and

$$\forall (s, a) \in \mathcal{S} \times \mathcal{A} : \quad Q_\beta^\star(s, a) \coloneqq r_\beta(s, a) + V_\beta^\star(s'), \tag{30}$$

where $s' = (s, a)$ is the deterministic next state. Plugging (30) and (27) into (28) implies that, for any trajectory $\tau = \{s_0, a_0, \cdots, a_{H-1}, s_H\}$, Rafailov et al. (2024) shows

$$\begin{aligned}
\sum_{i=0}^{H-1} r(s_i, a_i) &= \sum_{i=0}^{H-1} \big(Q_\beta^\star(s_i, a_i) - \beta \log \pi_{\text{ref}}(a_i|s_i) - V_\beta^\star(s_{i+1})\big) \\
&= Q_\beta^\star(s_0, a_0) - \beta \log \pi_{\text{ref}}(a_0|s_0) + \sum_{i=1}^{H-1} \big(Q_\beta^\star(s_i, a_i) - V_\beta^\star(s_i) - \beta \log \pi_{\text{ref}}(a_i|s_i)\big) \\
&= V_\beta^\star(s_0) + \beta \sum_{i=0}^{H-1} \log \frac{\pi_r(a_i|s_i)}{\pi_{\text{ref}}(a_i|s_i)},
\end{aligned} \tag{31}$$

where the second line uses $V_\beta^\star(s_H) = 0$ at the terminal state. Thus the DPO loss (which is the negative log-likelihood loss) could be written as

$$\mathcal{L}(\pi, \mathcal{D}) = - \sum_{(\tau_+, \tau_-) \in \mathcal{D}} \log \sigma \left(\beta \sum_{i=0}^{H-1} \log \frac{\pi(a_i^+|s_i^+)}{\pi_{\text{ref}}(a_i^+|s_i^+)} - \beta \sum_{i=0}^{H-1} \log \frac{\pi(a_i^-|s_i^-)}{\pi_{\text{ref}}(a_i^-|s_i^-)} \right). \tag{32}$$

Token-level VPO. With slight abuse of notation, define

$$J^\star(r) \coloneqq \max_\pi J(r,\pi), \tag{33}$$

which is used as the bias term in regularizing the reward estimation in VPO. Again, we impose the following assumption to deal with the shift ambiguity issue caused by the Bradley-Terry model:

Assumption 3 *We assume that $r^\star \in \mathcal{R}$, where*

$$\mathcal{R} = \left\{ r : \mathop{\mathbb{E}}_{\substack{s_0\sim\rho,\\ a_i\sim\pi_{cal}(\cdot|s_i)}} \sum_{i=0}^{H-1} r(s_i,a_i) = 0 \right\}. \tag{34}$$

Here, ρ is the prompt distribution and π_{cal} is a fixed calibration distribution independent of the algorithm.

Combining (26) with (29), similar to previous derivations, we have

$$\begin{aligned}
J^\star(r) &= \mathop{\mathbb{E}}_{s_0\sim\rho}\left[V^\star_\beta(s_0)\right] \\
&= \mathop{\mathbb{E}}_{\substack{s_0\sim\rho,\\ a_i\sim\pi_{\mathrm{cal}}(\cdot|s_i)}}\left[V^\star_\beta(s_0)\right] \\
&= \mathop{\mathbb{E}}_{\substack{s_0\sim\rho,\\ a_i\sim\pi_{\mathrm{cal}}(\cdot|s_i)}}\left[\sum_{i=0}^{H-1} r(s_i,a_i) - \beta\sum_{i=0}^{H-1}\log\frac{\pi_r(a_i|s_i)}{\pi_{\mathrm{ref}}(a_i|s_i)}\right] \\
&= -\beta \mathop{\mathbb{E}}_{\substack{s_0\sim\rho,\\ a_i\sim\pi_{\mathrm{cal}}(\cdot|s_i)}}\left[\sum_{i=0}^{H-1}\log\frac{\pi_r(a_i|s_i)}{\pi_{\mathrm{ref}}(a_i|s_i)}\right],
\end{aligned} \tag{35}$$

where the penultimate line uses (31), and the last line uses Assumption 3.

Consequently, the token-level VPO can be rewritten as

$$\begin{aligned}
\pi_{\mathsf{VPO}} = \arg\min_\pi \Bigg\{ &-\sum_{(\tau_+,\tau_-)\in\mathcal{D}} \log\sigma\left(\beta\sum_{i=0}^{H-1}\log\frac{\pi(a_i^+|s_i^+)}{\pi_{\mathrm{ref}}(a_i^-|s_i^-)} - \beta\sum_{i=0}^{H-1}\log\frac{\pi(a_i^-|s_i^-)}{\pi_{\mathrm{ref}}(a_i^-|s_i^-)}\right) \\
&+ \mathsf{sign}\cdot\alpha\beta \mathop{\mathbb{E}}_{\substack{s_0\sim\rho,\\ a_i\sim\pi_{\mathrm{cal}}(\cdot|s_i)}}\left[\sum_{i=0}^{H-1}\log\frac{\pi(a_i|s_i)}{\pi_{\mathrm{ref}}(a_i|s_i)}\right]\Bigg\}.
\end{aligned} \tag{36}$$

5 Experiments

In this section, we evaluate the proposed VPO on both synthetic multi-armed bandit (MAB), and RLHF for LLMs, in online and offline settings.

5.1 Synthetic multi-armed bandits

We evaluate the proposed methods on a synthetic dataset of size $|\mathcal{X}| = 1$ and $|\mathcal{Y}| = 10$. We set $\pi_{\mathrm{ref}} = \pi_{\mathsf{b}} = \pi_{\mathrm{cal}}$, where $\pi_{\mathrm{ref}} = \mathsf{softmax}(\theta_{\mathrm{ref}})$ with $\theta_{\mathrm{ref}}(x,y)$ sampled i.i.d. from $\mathcal{N}(0,1)$. The ground truth reward $r^\star$ is randomly generated i.i.d. according to $r^\star(x,y) \sim U([0,1])$. We approximately solve the optimization problems by performing 20 AdamW optimization steps with learning rate 0.01 and weight decay rate 0.01 in every iteration for the online setting and 1000 steps for the offline setting.

We plot the average results over 10 independent runs in Figure 1. As demonstrated in the left panel of Figure 1, an appropriate choice of α allows our method to outperform the model-based MAB with MLE baseline in the long-term performance of cumulative regret, at the cost of slightly increased cumulative regret in the first 100 iterations. This highlights the effectiveness of the VPO in achieving more principled exploration-exploitation trade-off. For the offline setting, the right panel of Figure 1 demonstrates that the performance of both MLE-MAB and VPO improves as the number of offline data increases. However, VPO achieves a consistently lower sub-optimality gap compared with that of MLE-MAB.

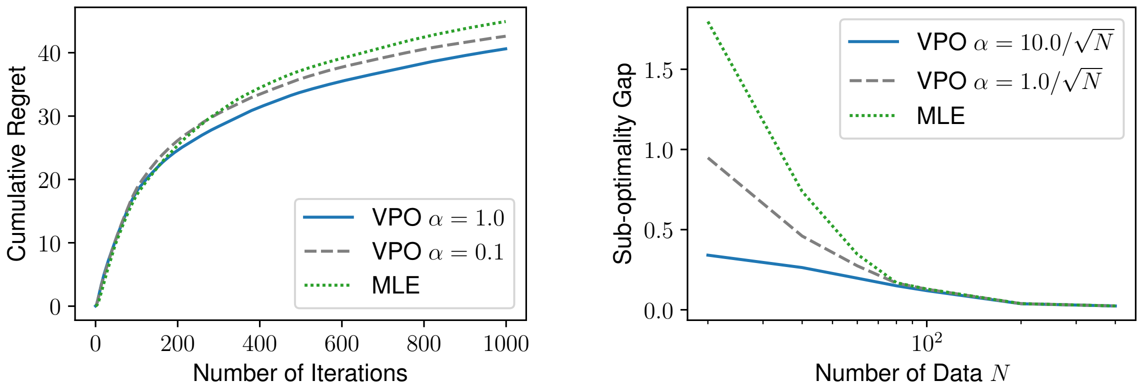


Figure 1: The cumulative regret v.s. number of iterations plot (left panel) and sub-optimality gap v.s. number of data plot (right panel) of VPO and MLE-MAB methods in the online and offline settings, respectively.

5.2 RLHF for LLMs

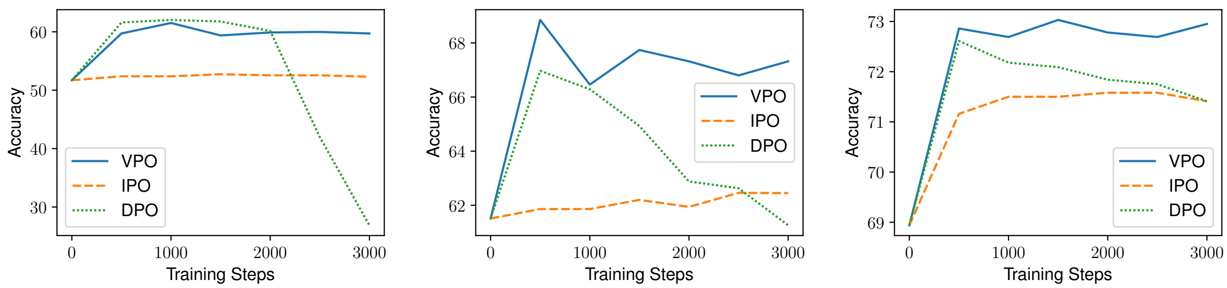
We further evaluate the pessimistic/optimistic VPO for LLMs in offline and online setting, respectively. In both settings, the proposed VPO demonstrates strong performances over the baselines.

Offline setting. In this setting, we test pessimistic VPO on `ARC-Challenge` task (Clark et al., 2018), which contains 7787 multiple-choices questions from multiple science subjects. We evaluate the performances on the `ARC-Challenge` test set, which contains 1172 questions. The data set only provides ground truth answer for each question. To construct the preference pairs and their labels, for each correct response in the training split, we create three pairs of comparison between the correct answer and each incorrect answer.

We emphasize that our goal is to evaluate the RLHF algorithm designs for LLMs, rather than pushing LLM towards state-of-the-art performance. To demonstrate the advantages of the proposed VPO, we conduct comparison with several offline RLHF baselines (DPO (Rafailov et al., 2023) and IPO (Azar et al., 2024)) on several LLMs, including Llama2-7b-chat, Llama2-13b-chat (Touvron et al., 2023) and Flan-T5-xl (Chung et al., 2022). For fair comparison, we keep all the experiment settings and prompts the same for every RLHF algorithm. We did not apply any additional chain-of-thought reasoning to avoid compounding factors affecting the RLHF performances. We tuned the hyperparameters for both the proposed VPO and the baselines on the validation set to achieve their best performances. For detailed hyperparameters setup, please refer to Appendix C.

The performances are illustrated in Figure 2. As we can see, the proposed VPO method demonstrates significantly better performance over the existing baselines on the three models, verifying the benefits across different models. In particular, the performance benefit becomes more evident for larger models. Another important observation is that the proposed VPO method is more robust to over-optimization (Gao et al., 2023). In the experiment, the performances of DPO significantly drops after 1000 iterations, and the longer DPO is trained, the worse it performs. In contrast, VPO consistently maintains the performances, avoiding the overoptimization issue and justifying the implicit robustness of pessimism as we revealed in (23).

Online setting. In this setting, we evaluate the performance of VPO on the `TL;DR` task (Stiennon et al., 2020). We prepare the prompts dataset by extracting the input prompts from the preference data. Recall we are evaluating the algorithm performance in online setting, we only compare to the online RLHF baselines (Guo et al., 2024) for fairness. We adopt PaLM2 (Anil et al., 2023) as the language model and also the LLM annotator. We conduct VPO and online DPO to the same PaLM2-XXS as the policy, which is initialized by supervised finetuning, denoted as SFT model. We exploit another PaLM2-XS model as the LLM annotator



(a) LLAMA2-7B-CHAT (b) LLAMA2-13B-CHAT (c) FLAN-T5-XL

Figure 2: The accuracy of the LLAMA2-7B-CHAT, LLAMA2-13B-CHAT and FLAN-T5-XL policies trained by VPO and other baselines (DPO and IPO) on `ARC-challenge`, respectively. The proposed pessimistic VPO performs consistently strong, and avoids over-optimization.

to provide online feedbacks. Similar to (Guo et al., 2024), we use *Detailed 0-shot* prompt from Lee et al. (2023). The prompts we used and how we get preference scores are detailed in Appendix C. We emphasize our algorithm is agnostic to human or AI feedback.

As a sanity check, we track the win rate of VPO and online DPO against the SFT baseline on `TL;DR` during training in Figure 3a. For ablation purpose, we varies the exploration weight $\alpha = \{0.01, 0.1\}$ in the optimistic VPO. One significant observation is that although all the online RLHF algorithms follow the increase trend, the win-rate against SFT of the optimistic VPO has larger oscillation, comparing to online DPO. And the oscillation reduces, with α diminishing. Our conjecture is that this behavior is encouraged by the optimistic term in VPO, for collecting more unexplored data, which may delay the learning due to the diversity in data. However, as the learning proceeds, the proposed VPO outperforms the competitors, because of the coverage of the collected data.

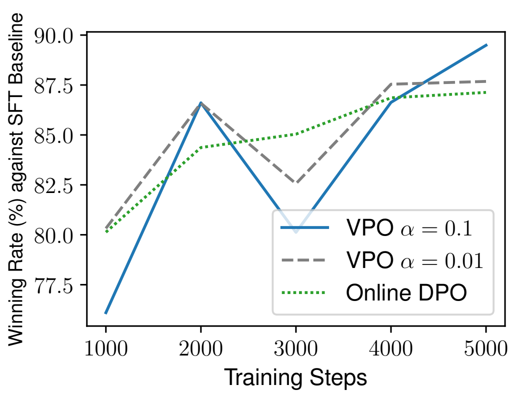


Figure 3(a): Win rate of VPO and online DPO against the SFT baseline on `TL;DR` task.

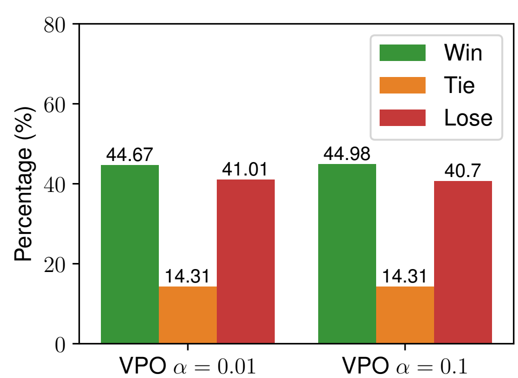


Figure 3(b): Win/tie/loss rate of VPO with different exploration rate $\alpha = \{0.01, 0.1\}$, directly against online DPO.

To demonstrate the advantages of optimistic VPO in online setting more directly, we evaluate the win/tie/loss rate against online DPO head-to-head, as shown in Figure 3b. This clearly shows that the optimistic VPO achieves better performances with larger exploration preference, and thus, consolidates our conclusion that *i)*, the simple value-incentivized term makes the exploration practical without uncertainty estimation; and *ii)*, exploration is potentially beneficial for better model.

6 Conclusion and Discussion

In this work, we develop a unified approach to achieving principled optimism and pessimism in online and offline RLHF, which enables a practical computation scheme by incorporating uncertainty estimation implicitly within reward-biased maximum likelihood estimation. Theoretical analysis indicates that the proposed methods mirror the guarantees of their standard RL counterparts, which is furthermore corroborated by numerical results. Important future directions include investigating adaptive rules for selecting α without prior information and more refined analysis on the choice of π_{cal}. This work also hints at a general methodology of designing practical algorithms with principled optimism/pessimism under more general RL setups.

Acknowledgement

The work of S. Cen and Y. Chi is supported in part by the grants ONR N00014-19-1-2404, NSF DMS-2134080, CCF-2106778 and CNS-2148212. S. Cen is also gratefully supported by Wei Shen and Xuehong Zhang Presidential Fellowship and JP Morgan AI Research Fellowship.

References

Abbasi-Yadkori, Y., Pál, D., and Szepesvári, C. (2011). Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24.

Anil, R., Dai, A. M., Firat, O., Johnson, M., Lepikhin, D., Passos, A., Shakeri, S., Taropa, E., Bailey, P., Chen, Z., et al. (2023). Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Azar, M. G., Guo, Z. D., Piot, B., Munos, R., Rowland, M., Valko, M., and Calandriello, D. (2024). A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR.

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Cen, S., Cheng, C., Chen, Y., Wei, Y., and Chi, Y. (2022). Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research*, 70(4):2563–2578.

Cesa-Bianchi, N., Gentile, C., Lugosi, G., and Neu, G. (2017). Boltzmann exploration done right. *Advances in neural information processing systems*, 30.

Chang, J. D., Shan, W., Oertell, O., Brantley, K., Misra, D., Lee, J. D., and Sun, W. (2024). Dataset reset policy optimization for RLHF. *arXiv preprint arXiv:2404.08495*.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, E., Wang, X., Dehghani, M., Brahma, S., et al. (2022). H. chi, jeff dean, jacob devlin, adam roberts, denny zhou, quoc v. le, and jason wei. 2022. scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. (2018). Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Ethayarajh, K., Xu, W., Muennighoff, N., Jurafsky, D., and Kiela, D. (2024). KTO: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*.

Gao, L., Schulman, J., and Hilton, J. (2023). Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR.

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., Kruspe, A., Triebel, R., Jung, P., Roscher, R., et al. (2021). A survey of uncertainty in deep neural networks. *arXiv preprint arXiv:2107.03342*.

Guo, S., Zhang, B., Liu, T., Liu, T., Khalman, M., Llinares, F., Rame, A., Mesnard, T., Zhao, Y., Piot, B., et al. (2024). Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*.

Hung, Y.-H., Hsieh, P.-C., Liu, X., and Kumar, P. (2021). Reward-biased maximum likelihood estimation for linear stochastic bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7874–7882.

Jin, C., Allen-Zhu, Z., Bubeck, S., and Jordan, M. I. (2018). Is Q-learning provably efficient? *Advances in neural information processing systems*, 31.

Jin, C., Liu, Q., and Yu, T. (2022). The power of exploiter: Provable multi-agent rl in large state spaces. In *International Conference on Machine Learning*, pages 10251–10279. PMLR.

Kumar, A., Zhou, A., Tucker, G., and Levine, S. (2020). Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191.

Kumar, P. and Becker, A. (1982). A new family of optimal adaptive controllers for markov chains. *IEEE Transactions on Automatic Control*, 27(1):137–146.

Lai, T. L., Robbins, H., et al. (1985). Asymptotically efficient adaptive allocation rules. *Advances in applied mathematics*, 6(1):4–22.

Lattimore, T. and Szepesvári, C. (2020). *Bandit algorithms*. Cambridge University Press.

Lee, H., Phatale, S., Mansoor, H., Lu, K., Mesnard, T., Bishop, C., Carbune, V., and Rastogi, A. (2023). Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*.

Liu, X., Hsieh, P.-C., Hung, Y. H., Bhattacharya, A., and Kumar, P. (2020). Exploration through reward biasing: Reward-biased maximum likelihood estimation for stochastic multi-armed bandits. In *International Conference on Machine Learning*, pages 6248–6258. PMLR.

Liu, Z., Lu, M., Xiong, W., Zhong, H., Hu, H., Zhang, S., Zheng, S., Yang, Z., and Wang, Z. (2024). Maximize to explore: One objective function fusing estimation, planning, and exploration. *Advances in Neural Information Processing Systems*, 36.

Mete, A., Singh, R., Liu, X., and Kumar, P. (2021). Reward biased maximum likelihood estimation for reinforcement learning. In *Learning for Dynamics and Control*, pages 815–827. PMLR.

Mnih, V., Badia, A. P., Mirza, M., Graves, A., Lillicrap, T., Harley, T., Silver, D., and Kavukcuoglu, K. (2016). Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

Munos, R., Valko, M., Calandriello, D., Azar, M. G., Rowland, M., Guo, Z. D., Tang, Y., Geist, M., Mesnard, T., Michi, A., et al. (2023). Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*.

Nachum, O., Norouzi, M., Xu, K., and Schuurmans, D. (2017). Bridging the gap between value and policy based reinforcement learning. *Advances in neural information processing systems*, 30.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Pal, A., Karkhanis, D., Dooley, S., Roberts, M., Naidu, S., and White, C. (2024). Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*.

Rafailov, R., Hejna, J., Park, R., and Finn, C. (2024). From r to $q^\star$: Your language model is secretly a Q-function. *arXiv preprint arXiv:2404.12358*.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Rashidinejad, P., Zhu, B., Ma, C., Jiao, J., and Russell, S. (2022). Bridging offline reinforcement learning and imitation learning: A tale of pessimism. *IEEE Transactions on Information Theory*, 68(12):8156–8196.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Shi, L., Li, G., Wei, Y., Chen, Y., and Chi, Y. (2022). Pessimistic Q-learning for offline reinforcement learning: Towards optimal sample complexity. In *International conference on machine learning*, pages 19967–20025. PMLR.

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. F. (2020). Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT Press.

Swamy, G., Dann, C., Kidambi, R., Wu, Z. S., and Agarwal, A. (2024). A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*.

Tang, Y., Guo, Z. D., Zheng, Z., Calandriello, D., Munos, R., Rowland, M., Richemond, P. H., Valko, M., Pires, B. Á., and Piot, B. (2024). Generalized preference optimization: A unified approach to offline alignment. *arXiv preprint arXiv:2402.05749*.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Uehara, M. and Sun, W. (2021). Pessimistic model-based offline reinforcement learning under partial coverage. *arXiv preprint arXiv:2107.06226*.

Xiao, C., Wu, Y., Mei, J., Dai, B., Lattimore, T., Li, L., Szepesvari, C., and Schuurmans, D. (2021). On the optimality of batch policy optimization algorithms. In *International Conference on Machine Learning*, pages 11362–11371. PMLR.

Xiong, W., Dong, H., Ye, C., Zhong, H., Jiang, N., and Zhang, T. (2023). Gibbs sampling from human feedback: A provable KL-constrained framework for RLHF. *arXiv preprint arXiv:2312.11456*.

Yuanzhe Pang, R., Yuan, W., Cho, K., He, H., Sukhbaatar, S., and Weston, J. (2024). Iterative reasoning preference optimization. *arXiv e-prints*, pages arXiv–2404.

Zhan, W., Uehara, M., Kallus, N., Lee, J. D., and Sun, W. (2023). Provable offline preference-based reinforcement learning. In *The Twelfth International Conference on Learning Representations*.

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

Zhao, Y., Joshi, R., Liu, T., Khalman, M., Saleh, M., and Liu, P. J. (2023). Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*.

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

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., and Irving, G. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

A Analysis for the online setting

A.1 Proof of Theorem 1

For ease of presentation, we assume that $\mathcal{R}$ is finite, i.e., $|\mathcal{R}| < \infty$. The general case can be directly obtained using a covering number argument, which we refer to (Liu et al., 2024; Jin et al., 2022) for interested readers.

We start by decomposing the regret into two parts:

$$\begin{aligned}\mathsf{Regret} &:= \sum_{t=1}^{T}\big[J^\star(r^\star) - J(r^\star,\pi^{(t)})\big]\\ &= \underbrace{\sum_{t=1}^{T}\big[J^\star(r^\star) - J^\star(r^{(t)})\big]}_{\textbf{Term (i)}} + \underbrace{\sum_{t=1}^{T}\big[J(r^{(t)},\pi^{(t)}) - J(r^\star,\pi^{(t)})\big]}_{\textbf{Term (ii)}}. \end{aligned} \tag{37}$$

Step 1: bounding term (i). By the choice of $r^{(t)}$, we have

$$\ell(r^{(t)},\mathcal{D}^{(t-1)}) - \alpha J^\star(r^{(t)}) \le \ell(r^\star,\mathcal{D}^{(t-1)}) - \alpha J^\star(r^\star). \tag{38}$$

Rearranging terms,

$$J^\star(r^\star) - J^\star(r^{(t)}) \le \frac{1}{\alpha}\big[\ell(r^\star,\mathcal{D}^{(t-1)}) - \ell(r^{(t)},\mathcal{D}^{(t-1)})\big]. \tag{39}$$

The following lemma is adapted from (Liu et al., 2024, Proposition 5.3), whose proof is deferred to Appendix A.2.

Lemma 2 *Let $\delta \in (0,1)$. With probability $1-\delta$, we have*

$$\begin{aligned}&\ell(r^\star,\mathcal{D}^{(t-1)}) - \ell(r^{(t)},\mathcal{D}^{(t-1)})\\ &\le -2\sum_{s=1}^{t-1}\mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ (y_1,y_2)\sim\pi^{(s)}(\cdot|x)}}\Big[D_H^2(\mathbb{P}_{r^{(t)}}(\cdot|x,y_1,y_2)\,\|\,\mathbb{P}_{r^\star}(\cdot|x,y_1,y_2))\Big] + 2\log(|\mathcal{R}|/\delta).\end{aligned} \tag{40}$$

Here, $D_H(\cdot\|\cdot)$ is the Hellinger distance, $\mathbb{P}_r(\cdot|x,y_1,y_2)$ denotes the Bernoulli distribution of the comparison result of (x,y_1) and (x,y_2) under reward model r.

Putting the above inequalities together, it holds with probability $1-\delta$ that

$$\begin{aligned}\textbf{Term (i)} \le -\frac{2}{\alpha}\sum_{t=1}^{T}\sum_{s=1}^{t-1}&\mathop{\mathbb{E}}_{\substack{x^{(s)}\sim\rho,\\ (y_1^{(s)},y_2^{(s)})\sim\pi^{(s)}(\cdot|x^{(s)})}}\Big[D_{\mathrm{H}}^2(\mathbb{P}_{r^{(t)}}(\cdot|x^{(s)},y_1^{(s)},y_2^{(s)})\,\|\,\mathbb{P}_{r^\star}(\cdot|x^{(s)},y_1^{(s)},y_2^{(s)}))\Big]\\ &+ 2\alpha^{-1}T\log(|\mathcal{R}|/\delta).\end{aligned} \tag{41}$$

Step 2: breaking down term (ii) with the elliptical potential lemma. The linear function approximation form (18) allows us to write

$$\mathop{\mathbb{E}}_{x\sim\rho,y\sim\pi_{r_2}(\cdot|x)}[r_1(x,y) - r^\star(x,y)] = \big\langle W(r_1), X(r_2)\big\rangle, \tag{42}$$

where $X, W : \mathcal{R} \to \mathbb{R}^d$ is given by

$$X(r_\theta) = 2C \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_{r_\theta}(\cdot|x)} [\phi(x, y)], \qquad W(r_\theta) = \frac{\theta - \theta^\star}{2C}. \tag{43}$$

Let

$$\Sigma_t = \epsilon I + \sum_{s=1}^{t-1} X(r^{(t)})X(r^{(t)})^\top \tag{44}$$

for some $\epsilon > 0$. We begin by decomposing term (ii) as

$$\begin{aligned}
\textbf{Term (ii)} &= \sum_{t=1}^{T} \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi^{(t)}(\cdot|x)} \left[r^{(t)}(x, y) - r^\star(x, y) \right] \\
&= \sum_{t=1}^{T} \left\langle W(r^{(t)}), X(r^{(t)}) \right\rangle \\
&= \sum_{t=1}^{T} \left\langle W(r^{(t)}), X(r^{(t)}) \right\rangle \mathbf{1}\{\|X(r^{(t)})\|_{\Sigma_t^{-1}}\} \le 1\} \\
&\qquad + \sum_{t=1}^{T} \left\langle W(r^{(t)}), X(r^{(t)}) \right\rangle \mathbf{1}\{\|X(r^{(t)})\|_{\Sigma_t^{-1}}\} > 1\},
\end{aligned} \tag{45}$$

where $\mathbf{1}\{A\}$ is an indicator function of event A. To proceed, we recall the elliptical potential lemma for controlling the cumulative sum of $\min\{\|X(r^{(t)})\|^2_{\Sigma_t^{-1}}, 1\}$.

Lemma 3 ((Abbasi-Yadkori et al., 2011, Lemma 11)) *Let $\{X_t\}$ be a sequence in $\mathbb{R}^d$ and $\Lambda_0 \in \mathbb{R}^{d\times d}$ a positive definite matrix. Define $\Lambda_t = \Lambda_0 + \sum_{s=1}^{t} X_s X_s^\top$. Assume $\|X_t\| \le L$ for all t. It holds that*

$$\begin{aligned}
\sum_{t=1}^{T} \min\{\|X_t\|^2_{\Lambda_t^{-1}}, 1\} &\le 2\log\left(\frac{det(\Lambda_T)}{det(\Lambda_0)}\right) \\
&\le 2(d\log((trace(\Lambda_0) + TL^2)/d) - \log det(\Lambda_0)).
\end{aligned}$$

Applying the above lemma yields

$$\sum_{t=1}^{T} \min\{\|X(r^{(t)})\|^2_{\Sigma_t^{-1}}, 1\} \le \min\left\{2d\log\left(\frac{4C^4T/d + \epsilon}{\epsilon}\right), T\right\} := d(\epsilon). \tag{46}$$

We now control the two terms in (45).

- The first term of (45) can be bounded by

$$\begin{aligned}
&\sum_{t=1}^{T} \left\langle W(r^{(t)}), X(r^{(t)}) \right\rangle \mathbf{1}\{\|X(r^{(t)})\|_{\Sigma_t^{-1}}\} \le 1\} \\
&\le \sum_{t=1}^{T} \|W(r^{(t)})\|_{\Sigma_t} \|X(r^{(t)})\|_{\Sigma_t^{-1}} \mathbf{1}\{\|X(r^{(t)})\|_{\Sigma_t^{-1}}\} \le 1\} \\
&\le \sum_{t=1}^{T} \|W(r^{(t)})\|_{\Sigma_t} \min\left\{\|X(r^{(t)})\|_{\Sigma_t^{-1}}, 1\right\} \\
&= \sum_{t=1}^{T} \left[\epsilon\|W(r^{(t)})\|_2^2 + \sum_{s=1}^{t-1} \left\langle W(r^{(t)}), X(r^{(s)}) \right\rangle^2\right]^{1/2} \min\left\{\|X(r^{(t)})\|^2_{\Sigma_t^{-1}}, 1\right\}^{1/2}
\end{aligned}$$

$$
\begin{aligned}
&\overset{\text{(i)}}{\leq} \left\{ \sum_{t=1}^{T} \left[\epsilon \| W(r^{(t)}) \|_2^2 + \sum_{s=1}^{t-1} \left\langle W(r^{(t)}), X(r^{(s)}) \right\rangle^2 \right] \right\}^{1/2} \left\{ \sum_{t=1}^{T} \min \left\{ \| X(r^{(t)}) \|_{\Sigma_t^{-1}}^2, 1 \right\} \right\}^{1/2} \\
&\overset{\text{(ii)}}{\leq} \sqrt{d(\epsilon)} \left\{ \sum_{t=1}^{T} \sum_{s=1}^{t-1} \left\langle W(r^{(t)}), X(r^{(s)}) \right\rangle^2 \right\}^{1/2} + \sqrt{d(\epsilon)\epsilon T} \\
&\overset{\text{(iii)}}{\leq} \frac{d(\epsilon)}{2\mu} + \frac{\mu}{2} \sum_{t=1}^{T} \sum_{s=1}^{t-1} \left\langle W(r^{(t)}), X(r^{(s)}) \right\rangle^2 + \sqrt{d(\epsilon)\epsilon T}. \qquad (47)
\end{aligned}
$$

Here, (i) is due to Cauchy–Schwarz inequality, (ii) is due to $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $\forall a, b \geq 0$, and (iii) results from Young's inequality. We leave the constant $\mu > 0$ to be determined later.

- The second term of (45) can be bounded by

$$
\begin{aligned}
\sum_{t=1}^{T} \left\langle W(r^{(t)}), X(r^{(t)}) \right\rangle \mathbf{1}\{ \| X(r^{(t)}) \|_{\Sigma_t^{-1}} \} > 1 \} &\leq C \sum_{t=1}^{T} \mathbf{1}\{ \| X(r^{(t)}) \|_{\Sigma_t^{-1}} \} > 1 \} \\
&\leq C \sum_{t=1}^{T} \min\{ \| X(r^{(t)}) \|_{\Sigma_t^{-1}}^2, 1 \} \leq C d(\epsilon), \qquad (48)
\end{aligned}
$$

where the first inequality follows from $\|X(r^{(t)})\|_2 \leq 2C$ and $\|W(r^{(t)})\|_2 \leq 1/2$ since $\|\phi(x,y)\|_2 \leq 1$.

Putting (45), (47) and (48) together, we arrive at

$$
\textbf{Term (ii)} \leq \frac{d(\epsilon)}{2\mu} + \frac{\mu}{2} \sum_{t=1}^{T} \sum_{s=1}^{t-1} \left\langle W(r^{(t)}), X(r^{(s)}) \right\rangle^2 + \sqrt{d(\epsilon)\epsilon T} + C d(\epsilon). \qquad (49)
$$

Step 3: continuing bounding term (ii). It boils down to control $\left\langle W(r^{(t)}), X(r^{(s)}) \right\rangle^2$. We have

$$
\begin{aligned}
\left\langle W(r^{(t)}), X(r^{(s)}) \right\rangle &= \mathop{\mathbb{E}}_{\substack{x \sim \rho, \\ y \sim \pi^{(s)}(\cdot|x)}} \left[r^{(t)}(x,y) - r^\star(x,y) \right] \\
&= \mathop{\mathbb{E}}_{\substack{x \sim \rho, \\ y_1 \sim \pi^{(s)}(\cdot|x)}} \left[r^{(t)}(x,y_1) - r^\star(x,y_1) \right] - \mathop{\mathbb{E}}_{\substack{x \sim \rho, \\ y_2 \sim \pi_{\mathrm{cal}}(\cdot|x)}} \left[r^{(t)}(x,y_2) - r^\star(x,y_2) \right] \\
&= \mathop{\mathbb{E}}_{\substack{x \sim \rho, \\ y_1 \sim \pi^{(s)}(\cdot|x), \\ y_2 \sim \pi_{\mathrm{cal}}(\cdot|x)}} \left[\delta_x(r^{(t)}, r^\star, y_1, y_2) \right], \qquad (50)
\end{aligned}
$$

where $\delta_x(r_1, r_2, y_1, y_2) \coloneqq r_1(x,y_1) - r_1(x,y_2) - (r_2(x,y_1) - r_2(x,y_2))$. Therefore,

$$
\begin{aligned}
\left\langle W(r^{(t)}), X(r^{(s)}) \right\rangle^2 &= \mathop{\mathbb{E}}_{\substack{x \sim \rho, \\ y_1 \sim \pi^{(s)}(\cdot|x), \\ y_2 \sim \pi_{\mathrm{cal}}(\cdot|x)}} \left[\delta_x(r^{(t)}, r^\star, y_1, y_2) \right]^2 \\
&= \mathop{\mathbb{E}}_{\substack{x \sim \rho, \\ y_1 \sim \pi^{(s)}(\cdot|x), \\ y_2 \sim \pi_{\mathrm{cal}}(\cdot|x)}} \left[\delta_x(r^{(t)}, r^\star, y_1, y_2)^2 \right] - \mathop{\mathrm{Var}}_{\substack{x \sim \rho, \\ y_1 \sim \pi^{(s)}(\cdot|x), \\ y_2 \sim \pi_{\mathrm{cal}}(\cdot|x)}} [\delta_x(r^{(t)}, r^\star, y_1, y_2)^2] \\
&\leq \mathop{\mathbb{E}}_{\substack{x \sim \rho, \\ y_1 \sim \pi^{(s)}(\cdot|x), \\ y_2 \sim \pi_{\mathrm{cal}}(\cdot|x)}} \left[\delta_x(r^{(t)}, r^\star, y_1, y_2)^2 \right]
\end{aligned}
$$

$$\le \sup_{x,y} \frac{\pi_{\text{cal}}(y|x)}{\pi^{(s)}(y|x)} \cdot \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y_1,y_2\sim\pi^{(s)}(\cdot|x)}} \left[\delta_x(r^{(t)}, r^\star, y_1, y_2)^2\right]$$

$$\le \sup_{x,y} \frac{\pi_{\text{ref}}(y|x)}{\pi^{(s)}(y|x)} \cdot \sup_{x,y} \frac{\pi_{\text{cal}}(y|x)}{\pi_{\text{ref}}(y|x)} \cdot \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y_1,y_2\sim\pi^{(s)}(\cdot|x)}} \left[\delta_x(r^{(t)}, r^\star, y_1, y_2)^2\right]. \tag{51}$$

Recall from (6) that $\pi^{(s)}(y|x) \propto \pi_{\text{ref}}(y|x)\exp(r^{(s)}(x,y)/\beta)$. It follows that $|\log\pi^{(s)}(y|x) - \log\pi_{\text{ref}}(y|x)| \le 2\|r^{(s)}(x,\cdot)\|_\infty \le 2C/\beta$ (see *e.g.*, (Cen et al., 2022, Appendix A.2)), and hence $\sup_{x,y} \frac{\pi_{\text{ref}}(y|x)}{\pi^{(s)}(y|x)} \le \exp(2C/\beta)$. To proceed, we demonstrate in the following lemma that δ^2 can be upper bounded by the corresponding Hellinger distance, whose proof is deferred to Appendix A.3.

Lemma 4 *Assume bounded reward $\|r_1\|_\infty \le C$, $\|r_2\|_\infty \le C$. We have*

$$\delta_x(r_1, r_2, y_1, y_2)^2 \le 2(3+\exp(2C))^2 D_H^2(\mathbb{P}_{r_1}(\cdot|x,y_1,y_2) \,\|\, \mathbb{P}_{r_2}(\cdot|x,y_1,y_2)).$$

With the above lemma we arrive at

$$\begin{aligned}&\left\langle W(r^{(t)}), X(r^{(s)})\right\rangle^2 \\ &\le 2(3+\exp(2C))^2\exp(2C/\beta)\kappa \cdot \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y_1,y_2\sim\pi^{(s)}(\cdot|x)}} \left[D_{\mathrm{H}}^2(\mathbb{P}_{r^{(t)}}(\cdot|x,y_1,y_2) \,\|\, \mathbb{P}_{r^\star}(\cdot|x,y_1,y_2))\right].\end{aligned}$$

where we denote $\kappa = \sup_{x,y} \dfrac{\pi_{\text{cal}}(y|x)}{\pi_{\text{ref}}(y|x)}$. Plugging the above bound into (49), we get

$$\begin{aligned}&\textbf{Term (ii)} \\ &\le \frac{d(\epsilon)}{2\mu} + \mu(3+\exp(2C))^2\exp(2C/\beta)\kappa \cdot \sum_{t=1}^{T}\sum_{s=1}^{t-1} \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y_1,y_2\sim\pi^{(s)}(\cdot|x)}} \left[D_{\mathrm{H}}^2(\mathbb{P}_{r^{(t)}}(\cdot|x,y_1,y_2) \,\|\, \mathbb{P}_{r^\star}(\cdot|x,y_1,y_2))\right] \\ &\quad + 2B\sqrt{d(\epsilon)\epsilon T} + Cd(\epsilon).\end{aligned} \tag{52}$$

Step 4: finishing up. Combining (37), (41) and (52), with probability $1-\delta$ we have

$$\mathsf{Regret} \le \frac{2T\log(|\mathcal{R}|/\delta)}{\alpha} + \frac{d(\epsilon)}{2\mu} + \sqrt{d(\epsilon)\epsilon T} + Cd(\epsilon) \tag{53}$$

as long as $\alpha\mu(3+\exp(2C))^2\exp(2C/\beta)\kappa \le 2$. Setting $\alpha \asymp \frac{1}{\exp(2C+C/\beta)}\sqrt{\frac{T}{\kappa d(\epsilon)}}$, $\mu \asymp \frac{1}{\exp(2C+C/\beta)}\sqrt{\frac{d(\epsilon)}{\kappa T}}$, and $\epsilon = 1$, we arrive at

$$\mathsf{Regret} \le \widetilde{\mathcal{O}}((\exp(2C+C/\beta))\sqrt{\kappa d T})$$

as claimed.

A.2 Proof of Lemma 2

To begin, we have

$$\ell(r^\star, \mathcal{D}^{(t-1)}) - \ell(r^{(t)}, \mathcal{D}^{(t-1)}) = -\log\frac{\mathbb{P}(\mathcal{D}^{(t-1)}|r^\star)}{\mathbb{P}(\mathcal{D}^{(t-1)}|r^{(t)})} = -\sum_{s=1}^{t-1} X^s_{r^{(t)}}, \tag{54}$$

where we denote

$$X_r^s = \log\frac{\mathbb{P}_{r^\star}(y_+^{(s)} \succ y_-^{(s)}|x^{(s)})}{\mathbb{P}_r(y_+^{(s)} \succ y_-^{(s)}|x^{(s)})}. \tag{55}$$

To proceed, we recall a useful martingale exponential inequality.

Lemma 5 ((Zhang, 2023, Theorem 13.2),(Liu et al., 2024, Lemma D.1)) *Let $\{X_t\}_{t=1}^{\infty}$ be a sequence of real-valued random variables adapted to filtration $\{\mathcal{F}_t\}_{t=1}^{\infty}$. It holds with probability $1-\delta$ such that for any $t \geq 1$,*

$$-\sum_{s=1}^{t} X_s \leq \sum_{s=1}^{t} \log \mathbb{E}\left[\exp(-X_s)|\mathcal{F}_{s-1}\right] + \log(1/\delta).$$

Applying the above lemma to $\{\frac{1}{2}X_r^t\}_{t=1}^{\infty}$ along with the filtration $\{\mathcal{F}_t\}_{t=1}^{\infty}$ with $\mathcal{F}_t$ given by the σ-algebra of $\{(x^{(s)}, y_+^{(s)}, y_-^{(s)}) : s \leq t\}$, we conclude that it holds with probability $1-\delta$ that

$$\begin{aligned}
-\frac{1}{2}\sum_{s=1}^{t-1} X_r^s &\leq \sum_{s=1}^{t-1} \log \mathbb{E}\left[\exp\Big\{-\frac{1}{2}X_r^s\Big\}\Big|\mathcal{F}_{s-1}\right] + \log(|\mathcal{R}|/\delta) \\
&\leq \sum_{s=1}^{t-1}\left(\mathbb{E}\left[\exp\Big\{-\frac{1}{2}X_r^s\Big\}\Big|\mathcal{F}_{s-1}\right] - 1\right) + \log(|\mathcal{R}|/\delta),
\end{aligned} \tag{56}$$

where the last step results from the inequality $\log(1+x) \leq x$ for all $x \geq -1$. To proceed, note that

$$\begin{aligned}
&\mathbb{E}\left[\exp\Big\{-\frac{1}{2}X_r^s\Big\}\Big|\mathcal{F}_{s-1}\right] \\
&= \mathbb{E}\left[\sqrt{\frac{\mathbb{P}_r(y_+^{(s)} \succ y_-^{(s)}|x^{(s)})}{\mathbb{P}_{r^\star}(y_+^{(s)} \succ y_-^{(s)}|x^{(s)})}}\,\Big|\,\mathcal{F}_{s-1}\right] \\
&= \mathop{\mathbb{E}}_{\substack{x^{(s)}\sim\rho,\\ (y_1^{(s)},y_2^{(s)})\sim\pi^{(s)}(\cdot|x^{(s)}),\\ (+,-)\sim\mathbb{P}_{r^\star}}}\left[\sqrt{\frac{\mathbb{P}_r(y_+^{(s)} \succ y_-^{(s)}|x^{(s)})}{\mathbb{P}_{r^\star}(y_+^{(s)} \succ y_-^{(s)}|x^{(s)})}}\right] \\
&= \mathop{\mathbb{E}}_{\substack{x^{(s)}\sim\rho,\\ (y_1^{(s)},y_2^{(s)})\sim\pi^{(s)}(\cdot|x^{(s)})}}\left[\sum_{(+,-)}\sqrt{\mathbb{P}_r(y_+^{(s)} \succ y_-^{(s)}|x^{(s)})\cdot\mathbb{P}_{r^\star}(y_+^{(s)} \succ y_-^{(s)}|x^{(s)})}\right] \\
&= 1 - \frac{1}{2}\mathop{\mathbb{E}}_{\substack{x^{(s)}\sim\rho,\\ (y_1^{(s)},y_2^{(s)})\sim\pi^{(s)}(\cdot|x^{(s)})}}\left[\sum_{(+,-)}\left(\sqrt{\mathbb{P}_r(y_+^{(s)} \succ y_-^{(s)}|x^{(s)})} - \sqrt{\mathbb{P}_{r^\star}(y_+^{(s)} \succ y_-^{(s)}|x^{(s)})}\right)^2\right] \\
&= 1 - \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ (y_1,y_2)\sim\pi^{(s)}(\cdot|x)}}\left[D_{\mathrm{H}}^2(\mathbb{P}_r(\cdot|x,y_1,y_2\,\|\,\mathbb{P}_{r^\star}(\cdot|x,y_1,y_2)\right],
\end{aligned}$$

where we denote by $\sum_{(+,-)}$ the summation over different comparison results. Plugging the above equality into (56) completes the proof.

A.3 Proof of Lemma 4

By the mean value theorem, we have

$$\begin{aligned}
\left|\mathbb{P}_{r_1}(y_1 \succ y_2|x) - \mathbb{P}_{r_2}(y_1 \succ y_2|x)\right| &= \left|\sigma(r_1(x,y_1) - r_1(x,y_2)) - \sigma(r_2(x,y_1) - r_2(x,y_2))\right| \\
&= \left|\delta_x(r_1,r_2,y_1,y_2)\cdot\sigma'(\xi)\right| \\
&= \left|\delta_x(r_1,r_2,y_1,y_2)\right|\cdot\sigma(\xi)(1-\sigma(\xi))
\end{aligned}$$

for some ξ between $r_1(x,y_1) - r_1(x,y_2)$ and $r_2(x,y_1) - r_2(x,y_2)$. Since $|\xi| \leq 2C$, we have

$$\sigma(\xi)(1-\sigma(\xi)) \geq \sigma(2C)(1-\sigma(2C)) \geq \frac{1}{3+\exp(2C)}. \tag{57}$$

Putting together,

$$\begin{aligned}\left|\delta_x(r_1,r_2,y_1,y_2)\right| &\le (3+\exp(2C))\left|\mathbb{P}_{r_1}(y_1\succ y_2|x)-\mathbb{P}_{r_2}(y_1\succ y_2|x)\right|\\ &= (3+\exp(2C))\mathrm{TV}(\mathbb{P}_{r_1}(\cdot|x,y_1,y_2),\mathbb{P}_{r_2}(\cdot|x,y_1,y_2))\\ &\le (3+\exp(2C))\sqrt{2}D_{\mathrm{H}}(\mathbb{P}_{r_1}(\cdot|x,y_1,y_2)\,\|\,\mathbb{P}_{r_2}(\cdot|x,y_1,y_2)).\end{aligned}$$

B Analysis for the offline setting

B.1 Proof of Lemma 1

By definition, the objective function $\ell(r,\mathcal{D})+\alpha J(r,\pi)$ is strongly concave over π, and convex over r. By Danskin's theorem, we have

$$\nabla_r\big(\max_{\pi}[\ell(\widehat{r},\mathcal{D})+\alpha J(\widehat{r},\pi)]\big)=\nabla_r\big(\ell(\widehat{r},\mathcal{D})+\alpha J(\widehat{r},\widehat{\pi})\big).$$

Therefore, for any r', by convexity of the objective function we have

$$\begin{aligned}\ell(r',\mathcal{D})+\alpha J(r',\widehat{\pi}) &\ge \ell(\widehat{r},\mathcal{D})+\alpha J(\widehat{r},\widehat{\pi})+\big\langle r'-\widehat{r},\nabla_r\big(\ell(\widehat{r},\mathcal{D})+\alpha J(\widehat{r},\widehat{\pi})\big)\big\rangle\\ &= \ell(\widehat{r},\mathcal{D})+\alpha J(\widehat{r},\widehat{\pi})+\big\langle r'-\widehat{r},\nabla_r\big(\max_{\pi}[\ell(\widehat{r},\mathcal{D})+\alpha J(\widehat{r},\pi)]\big)\big\rangle\\ &\ge \ell(\widehat{r},\mathcal{D})+\alpha J(\widehat{r},\widehat{\pi}).\end{aligned}$$

The last line is due to the definition of $\widehat{r}$ (c.f. (23)). The other relation, $\ell(\widehat{r},\mathcal{D})+\alpha J(\widehat{r},\widehat{\pi})\ge \ell(\widehat{r},\mathcal{D})+\alpha J(\widehat{r},\pi')$, follows directly from the definition of $\widehat{\pi}$ (c.f. (20)).

B.2 Proof of Theorem 2

We decompose the sub-optimality gap of $\widehat{\pi}$ by

$$\begin{aligned}&J^\star(r^\star)-J(r^\star,\widehat{\pi})\\ &=\big[J(r^\star,\pi^\star)-J(\widehat{r},\pi^\star)\big]+\big[J(\widehat{r},\pi^\star)-J(\widehat{r},\widehat{\pi})\big]+\big[J(\widehat{r},\widehat{\pi})-J(r^\star,\widehat{\pi})\big]\\ &\le \underbrace{\big[J(r^\star,\pi^\star)-J(\widehat{r},\pi^\star)\big]}_{\textbf{Term (i)}}+\underbrace{\big[J(\widehat{r},\widehat{\pi})-J(r^\star,\widehat{\pi})\big]}_{\textbf{Term (ii)}},\end{aligned}\tag{58}$$

where the last line is due to $J(\widehat{r},\pi^\star)\le J(\widehat{r},\widehat{\pi})$ according to the definition of $\widehat{\pi}$ (c.f. (20)). We proceed to bound the two terms separately. Here we have written $\widehat{r}=r_{\widehat{\theta}}$ for notational simplicity. In addition, we denote the MLE estimate by $r_{\mathsf{MLE}}=r_{\theta_{\mathsf{MLE}}}$.

By the definition of $J(r,\pi)$ (cf. (4)), it follows that term (i) in (58) can be further decomposed as

$$\begin{aligned}\textbf{Term (i)} &= \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi^\star(\cdot|x)}}[r^\star(x,y)-\widehat{r}(x,y)]\\ &= \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi^\star(\cdot|x)}}\Big[\big\langle\phi(x,y),\theta^\star-\widehat{\theta}\big\rangle\Big]\\ &= \underbrace{\mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi^\star(\cdot|x)}}\big[\big\langle\phi(x,y),\theta^\star-\theta_{\mathsf{MLE}}\big\rangle\big]}_{\textbf{Term (ia)}}+\underbrace{\mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi^\star(\cdot|x)}}\Big[\big\langle\phi(x,y),\theta_{\mathsf{MLE}}-\widehat{\theta}\big\rangle\Big]}_{\textbf{Term (ib)}},\end{aligned}\tag{59}$$

where $r_{\mathsf{MLE}}(x,y)=\langle\phi(x,y),\theta_{\mathsf{MLE}}\rangle$.

Step 1: bounding term (ia). To continue, we recall a useful lemma from (Zhu et al., 2023).

Lemma 6 ((Zhu et al., 2023, Lemma 3.1)) *For any $\lambda > 0$ and $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\|\theta_{\mathsf{MLE}} - \theta^\star\|_{\Sigma_{\mathcal{D}}+\lambda I} \le \mathcal{O}\left((3+\exp(C))\sqrt{\frac{d+\log(1/\delta)}{N}} + \sqrt{\lambda C^2}\right).$$

In addition, we have

$$\frac{1}{3+\exp(C)}\Sigma_{\mathcal{D}} \preceq \frac{1}{N}\nabla_\theta^2 \ell(r_\theta, \mathcal{D}) \preceq \frac{1}{4}\Sigma_{\mathcal{D}} \tag{60}$$

for all θ such that $\|r_\theta\|_\infty \le C$.

The first term of (59) can be bounded with Lemma 6 as

$$\begin{aligned}
\textbf{Term (ia)} &\le \|\theta^\star - \theta_{\mathsf{MLE}}\|_{\Sigma_{\mathcal{D}}+\lambda I} \cdot \Big\| \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi^\star(\cdot|x)}}[\phi(x,y)] \Big\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} \\
&\le \mathcal{O}\left(\left((3+\exp(C))\sqrt{\frac{d+\log(1/\delta)}{N}} + \sqrt{\lambda C^2}\right) \cdot \Big\| \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi^\star(\cdot|x)}}[\phi(x,y)] \Big\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}\right).
\end{aligned} \tag{61}$$

Step 2: bounding term (ib). For the second term of (59), recall that

$$\widehat{r} = \arg\min_{r\in\mathcal{R}} \big\{\ell(r,\mathcal{D}) + \alpha J(r,\widehat{\pi})\big\},$$

or equivalently

$$\widehat{\theta} = \arg\min_{\theta\in\Theta} \big\{\ell(r_\theta,\mathcal{D}) + \alpha J(r_\theta,\widehat{\pi})\big\},$$

and that

$$\theta_{\mathsf{MLE}} = \arg\min_{\theta\in\Theta} \ell(r_\theta,\mathcal{D}).$$

With linear constraint (19), by KKT condition we have

$$\nabla_\theta \ell(\widehat{r},\mathcal{D}) + \alpha \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\widehat{\pi}(\cdot|x)}}[\phi(x,y)] + \lambda_1 \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}}[\phi(x,y)] = 0$$

for some $\lambda_1 \in \mathbb{R}$, and

$$\nabla_\theta \ell(r_{\mathsf{MLE}},\mathcal{D}) + \lambda_2 \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}}[\phi(x,y)] = 0$$

for some $\lambda_2 \in \mathbb{R}$. By strong monotonicity of $\nabla_\theta \ell$ (cf. (60)), we have

$$\begin{aligned}
\frac{N}{3+\exp(C)}\big\|\widehat{\theta} - \theta_{\mathsf{MLE}}\big\|_{\Sigma_{\mathcal{D}}}^2 &\le \big\langle \nabla_\theta \ell(\widehat{r},\mathcal{D}) - \nabla_\theta \ell(r_{\mathsf{MLE}},\mathcal{D}), \widehat{\theta} - \theta_{\mathsf{MLE}} \big\rangle \\
&= \Big\langle -\alpha \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\widehat{\pi}(\cdot|x)}}[\phi(x,y)] - (\lambda_1 - \lambda_2) \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}}[\phi(x,y)], \widehat{\theta} - \theta_{\mathsf{MLE}} \Big\rangle \\
&= -\alpha \Big\langle \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\widehat{\pi}(\cdot|x)}}[\phi(x,y)] - \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}}[\phi(x,y)], \widehat{\theta} - \theta_{\mathsf{MLE}} \Big\rangle \\
&\le \alpha \Big\| \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\widehat{\pi}(\cdot|x)}}[\phi(x,y)] - \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}}[\phi(x,y)] \Big\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} \big\|\widehat{\theta} - \theta_{\mathsf{MLE}}\big\|_{\Sigma_{\mathcal{D}}+\lambda I}
\end{aligned}$$

$$\leq \alpha\kappa_{\mathcal{D}}\big\|\widehat{\theta}-\theta_{\mathsf{MLE}}\big\|_{\Sigma_{\mathcal{D}}+\lambda I},$$

where we denote

$$\kappa_{\mathcal{D}} = \Big\| \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\widehat{\pi}(\cdot|x)}}[\phi(x,y)] - \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}}[\phi(x,y)]\Big\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}. \tag{62}$$

The penultimate step results from $\widehat{\theta},\theta_{\mathsf{MLE}}\in\Theta$, which ensures

$$\Big\langle \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}}[\phi(x,y)],\widehat{\theta}\Big\rangle = \Big\langle \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}}[\phi(x,y)],\theta_{\mathsf{MLE}}\Big\rangle = 0$$

It follows that

$$\begin{aligned}\frac{N}{3+\exp(C)}\big\|\widehat{\theta}-\theta_{\mathsf{MLE}}\big\|^2_{\Sigma_{\mathcal{D}}+\lambda I} &\leq \frac{N}{3+\exp(C)}\big\|\widehat{\theta}-\theta_{\mathsf{MLE}}\big\|^2_{\Sigma_{\mathcal{D}}} + \frac{N}{3+\exp(C)}\big\|\widehat{\theta}-\theta_{\mathsf{MLE}}\big\|^2_{\lambda I}\\ &\leq \alpha\kappa_{\mathcal{D}}\big\|\widehat{\theta}-\theta_{\mathsf{MLE}}\big\|_{\Sigma_{\mathcal{D}}+\lambda I} + \frac{N\lambda C^2}{3+\exp(C)}.\end{aligned}$$

The above inequality allows us to bound

$$\big\|\widehat{\theta}-\theta_{\mathsf{MLE}}\big\|_{\Sigma_{\mathcal{D}}+\lambda I} \leq \frac{\alpha(3+\exp(C))}{N}\kappa_{\mathcal{D}} + 2\sqrt{\lambda C^2}. \tag{63}$$

Therefore, the second term of (59) can be bounded as

$$\begin{aligned}\textbf{Term (ib)} &\leq \big\|\widehat{\theta}-\theta_{\mathsf{MLE}}\big\|_{\Sigma_{\mathcal{D}}+\lambda I}\Big\| \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi^\star(\cdot|x)}}[\phi(x,y)]\Big\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}\\ &\leq \bigg(\frac{\alpha(3+\exp(C))}{N}\kappa_{\mathcal{D}} + 2\sqrt{\lambda C^2}\bigg)\Big\| \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi^\star(\cdot|x)}}[\phi(x,y)]\Big\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}. \end{aligned}\tag{64}$$

Putting (61) and (64) together, we have

$$\textbf{Term (i)} \leq \mathcal{O}\Bigg(\bigg[\frac{3+\exp(C)}{\sqrt{N}}\Big(\sqrt{d+\log(1/\delta)} + \frac{\alpha}{\sqrt{N}}\kappa_{\mathcal{D}}\Big) + \sqrt{\lambda C^2}\bigg]\cdot\Big\| \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi^\star(\cdot|x)}}[\phi(x,y)]\Big\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}}\Bigg). \tag{65}$$

Step 3: bounding term (ii). We can decompose and bound term (ii) by

$$\begin{aligned}J(\widehat{r},\widehat{\pi}) - J(r^\star,\widehat{\pi}) &= J(\widehat{r},\widehat{\pi}) + \frac{1}{\alpha}\ell(\widehat{r},\mathcal{D}) - \Big(J(r^\star,\widehat{\pi}) + \frac{1}{\alpha}\ell(r^\star,\mathcal{D})\Big) + \frac{1}{\alpha}(\ell(\widehat{r},\mathcal{D}) - \ell(r^\star,\mathcal{D}))\\ &\overset{(i)}{\leq} \frac{1}{\alpha}(\ell(\widehat{r},\mathcal{D}) - \ell(r^\star,\mathcal{D}))\\ &\leq \frac{1}{\alpha}(\ell(\widehat{r},\mathcal{D}) - \ell(r_{\mathsf{MLE}},\mathcal{D}) + \ell(r_{\mathsf{MLE}},\mathcal{D}) - \ell(r^\star,\mathcal{D})),\end{aligned}$$

where (i) follows from the fact that $(\widehat{r},\widehat{\pi})$ is a saddle point. Due to convexity of ℓ, we have

$$\begin{aligned}\ell(\widehat{r},\mathcal{D}) - \ell(r_{\mathsf{MLE}},\mathcal{D}) &\leq \big\langle \nabla_\theta \ell(\widehat{r},\mathcal{D}), \widehat{\theta}-\theta_{\mathsf{MLE}}\big\rangle\\ &= \Big\langle -\alpha \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\widehat{\pi}(\cdot|x)}}[\phi(x,y)] - \lambda_1 \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}}[\phi(x,y)], \widehat{\theta}-\theta_{\mathsf{MLE}}\Big\rangle\\ &= -\alpha\Big\langle \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\widehat{\pi}(\cdot|x)}}[\phi(x,y)] - \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}}[\phi(x,y)], \widehat{\theta}-\theta_{\mathsf{MLE}}\Big\rangle\end{aligned}$$

$$
\begin{aligned}
&\leq \alpha\kappa_{\mathcal{D}}\|\widehat{\theta} - \theta_{\mathsf{MLE}}\|_{\Sigma_{\mathcal{D}}+\lambda I} \\
&\leq \frac{\alpha^2(3+\exp(C))}{N}\kappa_{\mathcal{D}}^2 + 2\sqrt{\lambda C^2}\alpha\kappa_{\mathcal{D}},
\end{aligned}
$$

where the last step is due to (63). On the other hand, with probability $1-\delta$ we have (Zhan et al., 2023, Lemma 1):

$$
\ell(r_{\mathsf{MLE}}, \mathcal{D}) - \ell(r^\star, \mathcal{D}) \leq \widetilde{\mathcal{O}}(1).
$$

Putting pieces together,

$$
\textbf{Term (ii)} \leq \frac{\alpha(3+\exp(C))}{N}\kappa_{\mathcal{D}}^2 + 2\sqrt{\lambda C^2}\kappa_{\mathcal{D}} + \frac{1}{\alpha}. \tag{66}
$$

Step 4: putting things together. Combining (58) (65), (66), with probability $1-\delta$ we have

$$
\begin{aligned}
&J^\star(r^\star) - J(r^\star, \widehat{\pi}) \\
&\leq \mathcal{O}\Bigg(\frac{1}{\sqrt{N}}\Big[(3+\exp(C))\Big(\sqrt{d+\log(1/\delta)} + \kappa_{\mathcal{D}}\Big) + C\Big] \cdot \Big\| \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi^\star(\cdot|x)}}[\phi(x,y)]\Big\|_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} \\
&\qquad + \frac{1}{\sqrt{N}}\Big((3+\exp(C))\kappa_{\mathcal{D}}^2 + 2C\kappa_{\mathcal{D}} + 1\Big)\Bigg).
\end{aligned}
$$

Here we have set $\alpha = \sqrt{N}$ and $\lambda = 1/N$. We conclude by bounding $\kappa_{\mathcal{D}}$ as

$$
\begin{aligned}
\kappa_{\mathcal{D}}^2 &= \Big\| \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\widehat{\pi}(\cdot|x)}}[\phi(x,y)] - \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}}[\phi(x,y)]\Big\|^2_{(\Sigma_{\mathcal{D}}+\lambda I)^{-1}} \\
&\leq \Big\| \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\widehat{\pi}(\cdot|x)}}[\phi(x,y)] - \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\sim\pi_{\mathrm{cal}}(\cdot|x)}}[\phi(x,y)]\Big\|_2^2 \cdot \Big\|(\Sigma_{\mathcal{D}}+\lambda I)^{-1}\Big\|_2 \\
&\leq 4(\lambda_{\min}(\Sigma_{\mathcal{D}}) + \lambda)^{-1}.
\end{aligned}
$$

C Experimental details

C.1 Offline setting

For the offline setting experiments, we adopt instruction tuned models, LLAMA2-13B-CHAT and FLAN-T5-XL as the base models. To prompt these models, we prepend the question with

What is the choice to the following Question? Only provide the choice by providing a single letter.

and further append the question with

The answer is:.

The question is structured in a way that the multiple choices are shown as alphabets (letters) within parenthesis.

We set π_{cal} to the empirical distribution of the ground truth answer which is known to us. Based on preliminary experiments, we set β as 0.1 in DPO and τ as 1.0 in IPO. For VPO, we experiment with moving α from 0.01 to 10, choosing 1 for the reported results for FLAN-T5-XL results. For experiments on LLAMA2-13B-CHAT, we also set α to 1.

For both models, we train the base models with different algorithms DPO, VPO and IPO for 3000 steps and report the accuracy of the performance on the `ARC-challenge` test data set after every 500 steps. The training for LLAMA2-13B-CHAT model on 128 TPU-v4 takes around 2hrs and for FLAN-T5-XL on 64 TPU-v3 takes 1 hour.

C.2 Online setting

The prompt used for generating AI feedback (and rating) for `TL;DR` summarization is identical to (Guo et al., 2024). We set π_{cal} to the empirical distribution of the negative answer pairs (x, y_-) collected by the policy. We set β as 0.1 for the DPO term similar to (Guo et al., 2024). Additionally for VPO, we decrease the coefficient exponentially following $\frac{\alpha}{\sqrt{1+\text{training steps}}}$. We try different values of α and report the results for 0.1 and 0.01.

The training of the policy, PaLM2-XXS on 64 TPU-v3 for 5000 steps takes around 12 hours for both online DPO and VPO. We report the win rate percentage against the base SFT model for every 1000 steps using PaLM2-XS judge. We also further conduct side by side comparison of Online DPO and VPO at 5000 step.