RELATIVE DENSITY RATIO OPTIMIZATION FOR STABLE AND STATISTICALLY CONSISTENT MODEL ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

Aligning language models with human preferences is essential for ensuring their safety and reliability. Although most existing approaches assume specific human preference models such as the Bradley-Terry model, this assumption may fail to accurately capture true human preferences, and consequently, these methods lack statistical consistency, i.e., the guarantee that language models converge to the true human preference as the number of samples increases. In contrast, direct density ratio optimization (DDRO) achieves statistical consistency without assuming any human preference models. DDRO models the density ratio between preferred and non-preferred data distributions using the language model, and then optimizes it via density ratio estimation. However, this density ratio is unstable and often diverges, leading to training instability of DDRO. In this paper, we propose a novel alignment method that is both stable and statistically consistent. Our approach is based on the relative density ratio between the preferred data distribution and a mixture of the preferred and non-preferred data distributions. Our approach is stable since this relative density ratio is bounded above and does not diverge. Moreover, it is statistically consistent and yields significantly tighter convergence guarantees than DDRO. We experimentally show its effectiveness with Qwen 2.5 and Llama 3.

1 INTRODUCTION

Language models have demonstrated strong performance in text generation, translation, question answering, and information retrieval (Vaswani et al., 2017; Brown et al., 2020; Achiam et al., 2023). Nevertheless, they can produce inaccurate, harmful, or biased outputs (Ji et al., 2023; Bender et al., 2021; Sheng et al., 2019; Gehman et al., 2020), which may threaten their safety and reliability.

Alignment, which tunes language models to follow human preferences using a preference dataset containing preferred and non-preferred responses to various prompts, is essential for ensuring safety and reliability (Ouyang et al., 2022; Rafailov et al., 2023; Ethayarajh et al., 2024). To model human preferences, most existing approaches assume the Bradley-Terry model (Bradley & Terry, 1952), which is widely used in learning to rank. For example, reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022) first trains a reward model using the preference dataset under the Bradley-Terry assumption, and then tunes the language model via reinforcement learning to maximize the reward-model score. Direct preference optimization (DPO) (Rafailov et al., 2023) shows that the reward model can be represented by the language model, and aligns the language model through training the reward model without reinforcement learning. In contrast, Kahneman-Tversky optimization (KTO) (Ethayarajh et al., 2024) uses prospect theory (Tversky & Kahneman, 1992; Kahneman & Tversky, 2013), which captures how humans make decisions under uncertainty, to model human preferences. KTO can be regarded as a generalization of DPO.

Although these approaches are powerful, human preference models may not accurately capture the real complexities of human preference (Wu et al., 2022), and consequently these methods lack statistical consistency (Vapnik, 1999; Mohri et al., 2018), i.e., the guarantee that the language model converges to the true human preference as the number of samples increases (Higuchi & Suzuki, 2025). To address this, direct density ratio optimization (DDRO) (Higuchi & Suzuki, 2025) presents an alignment approach without assuming any human preference models. DDRO is based on density ratio estimation, which directly estimates the ratio between two distributions. DDRO models the density ratio between the preferred and non-preferred data distributions using the language model,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

and optimizes the language model through density ratio estimation. In contrast to other alignment approaches, DDRO satisfies statistical consistency. However, it has been empirically observed that training tends to be unstable (Higuchi & Suzuki, 2025).

Our purpose is to address the instability of DDRO and propose an alignment method that is both stable and statistically consistent. One reason for the instability of DDRO lies in the instability of the density ratio. Although DDRO models the density ratio between the preferred and non-preferred data distributions, it may diverge when the supports of these two distributions, the regions where their values are non-zero, are different (Sugiyama et al., 2012). Since preferred and non-preferred responses are likely to have different supports, this density ratio may diverge, leading to instability in DDRO. To address this, we employ a more stable approach called relative density ratio estimation (Yamada et al., 2013). The relative density ratio is defined as the ratio between the preferred data distribution and a mixture of the preferred and non-preferred data distributions. Unlike the ordinary density ratio, the relative density ratio is bounded above and does not diverge, enabling more stable estimation. We model the relative density ratio using the language model, and optimize the language model through relative density ratio estimation. We theoretically show that our approach is more stable and statistically consistent, and yields significantly tighter convergence guarantees than DDRO.

Our contributions can be summarized as follows:

- We propose a novel alignment method called relative density ratio optimization (RDRO), which is based on relative density ratio estimation (Yamada et al., 2013).
- We theoretically show that our RDRO is more stable and statistically consistent, and yields significantly tighter convergence guarantees than DDRO.
- We demonstrate the effectiveness of our RDRO with Qwen 2.5 (Team et al., 2024) and Llama 3 (Grattafiori et al., 2024).

2 Preliminaries

2.1 Problem Setup

Let x be the prompt and y be the response. Given a preferred dataset $\mathcal{D}^+ = (x_n^+, y_n^+)_{n=1}^N$ and a non-preferred dataset $\mathcal{D}^- = (x_m^-, y_m^-)_{m=1}^M$, we assume that the samples are independent and identically distributed within each dataset, with $\mathcal{D}^+$ following $p^+(x, y) = p^+(y|x)p(x)$ and $\mathcal{D}^-$ following $p^-(x, y) = p^-(y|x)p(x)$. Note that we assume the same prompt distribution: $p(x)$ for both datasets, while the response distributions given prompts: $p^+(y|x)$ and $p^-(y|x)$ differ. Following Higuchi & Suzuki (2025), we use "true human preference" to denote the preference data distribution p^+ defined by an annotator. Note that this term does not mean a universal preference shared by all humans. Let $p_\theta(y|x)$ be the policy of the language model to be aligned (e.g., a pre-trained or supervised fine-tuning model) with parameter θ, and $p_{\mathrm{ref}}(y|x)$ be the reference policy, which serves as a regularizer for $p_\theta(y|x)$ and is typically set to a frozen copy of the initial $p_\theta(y|x)$. Our goal is to align $p_\theta(y|x)$ with $p^+(y|x)$ using $\mathcal{D}^+$, $\mathcal{D}^-$, and $p_{\mathrm{ref}}(y|x)$.

2.2 Direct Density Ratio Optimization

Direct density ratio optimization (DDRO) (Higuchi & Suzuki, 2025) tries to align the policy $p_\theta(y|x)$ with the true preference $p^+(y|x)$ using density ratio estimation (Sugiyama et al., 2012). DDRO assumes that the reference policy $p_{\mathrm{ref}}(y|x)$ can be written as follows:

$$p_{\mathrm{ref}}(y|x) = \alpha p^+(y|x) + (1-\alpha)p^-(y|x), \tag{1}$$

where $\alpha \in (0, 1)$ is a hyperparameter. This assumption can be interpreted as labeling the response of the reference policy $p_{\mathrm{ref}}(y|x)$ as preferred with probability α and as non-preferred with probability $1-\alpha$. This assumption approximates $p(+|x)$ by a prompt-independent constant α, where $p(+|x)$ denotes the probability that a response sampled from $p_{\mathrm{ref}}(\cdot|x)$ is labeled as preferred (Higuchi & Suzuki, 2025). We discuss its limitations in Section 6. With this assumption, the following relation holds between the reference policy $p_{\mathrm{ref}}(y|x)$ and the true preference $p^+(y|x)$:

$$\frac{p_{\mathrm{ref}}(y|x)}{p^+(y|x)} = \frac{\alpha p^+(y|x) + (1-\alpha)p^-(y|x)}{p^+(y|x)} = \alpha + (1-\alpha)\frac{p^-(y|x)}{p^+(y|x)}. \tag{2}$$

Accordingly, the density ratio $p^-(y|x)/p^+(y|x)$ is written as follows:

$$\frac{p^-(y|x)}{p^+(y|x)} = \frac{1}{1-\alpha}\frac{p_{\mathrm{ref}}(y|x)}{p^+(y|x)} - \frac{\alpha}{1-\alpha} \equiv g^*(y|x). \tag{3}$$

By replacing the true preference $p^+(y|x)$ with the policy $p_\theta(y|x)$, this density ratio is modeled as follows:

$$g_\theta(y|x) \equiv \frac{1}{1-\alpha}\frac{p_{\mathrm{ref}}(y|x)}{p_\theta(y|x)} - \frac{\alpha}{1-\alpha}. \tag{4}$$

DDRO tries to align $p_\theta(y|x)$ with $p^+(y|x)$ by minimizing the Bregman divergence (Sugiyama et al., 2012) between $g^*(y|x)$ and $g_\theta(y|x)$:

$$\mathcal{L}_{\mathrm{DRE}}(\theta) = \mathbb{E}_{p^+(x,y)}\left[\mathrm{Breg}_f\left(g^*(y|x) \| g_\theta(y|x)\right)\right], \tag{5}$$

where

$$\mathrm{Breg}_f\left(u\|v\right) = f(u) - f(v) - f'(v)(u-v), \tag{6}$$

$\mathbb{E}[\cdot]$ is the expectation, $f : \mathbb{R} \to \mathbb{R}$ is a differentiable and strictly convex function, and f' is the derivative of f.

When $f(t) = t\log t - (1+t)\log(1+t)$, this optimization problem can be reduced to the logistic regression problem (Sugiyama et al., 2012):

$$\mathcal{L}_{\mathrm{DRE}}(\theta) = \mathbb{E}_{p^+(x,y)}\left[\log(1 + g_\theta(y|x))\right] + \mathbb{E}_{p^-(x,y)}\left[\log(1 + 1/g_\theta(y|x))\right]. \tag{7}$$

Using the datasets $\mathcal{D}^+$ and $\mathcal{D}^-$, $\mathcal{L}_{\mathrm{DRE}}(\theta)$ is approximated as follows:

$$\frac{1}{N}\sum_{n=1}^{N}\log(1 + g_\theta(y_n^+|x_n^+)) + \frac{1}{M}\sum_{m=1}^{M}\log(1 + 1/g_\theta(y_m^-|x_m^-)). \tag{8}$$

Unfortunately, this optimization is unstable in practice due to gradient spikes (Higuchi & Suzuki, 2025). To mitigate this, DDRO heuristically applies a monotonically increasing function $S(t) = \log\sigma(t)$, where $\sigma(t) = 1/(1+\exp(-t))$ is the sigmoid function, to this objective as follows:

$$\hat{\mathcal{L}}_{\mathrm{DRE}}(\theta) = \frac{1}{N}\sum_{n=1}^{N}S(\log(1 + g_\theta(y_n^+|x_n^+))) + \frac{1}{M}\sum_{m=1}^{M}S(\log(1 + 1/g_\theta(y_m^-|x_m^-))). \tag{9}$$

Since just minimizing this objective will lead to catastrophic forgetting (Kirkpatrick et al., 2017), DDRO introduces the regularizer based on the Kullback-Leibler (KL) divergence between $p_\theta(y|x)$ and $p_{\mathrm{ref}}(y|x)$ as follows:

$$\mathcal{L}_{\mathrm{DDRO}}(\theta) = \hat{\mathcal{L}}_{\mathrm{DRE}}(\theta) + \beta\mathrm{KL}\left(p_\theta\|p_{\mathrm{ref}}\right), \tag{10}$$

where β is a hyperparameter that controls the strength of the regularizer.

DDRO has the following property for the optimal parameter $\theta^* = \arg\min_\theta \mathcal{L}_{\mathrm{DRE}}(\theta)$:

$$\mathcal{L}_{\mathrm{DRE}}(\theta^*) = 0 \Leftrightarrow g_{\theta^*} = g^* \Leftrightarrow p_{\theta^*} = p^+. \tag{11}$$

Furthermore, as the number of samples increases, the policy $p_\theta(y|x)$ converges to the true preference $p^+(y|x)$ at the rate of $\mathcal{O}(1/\sqrt{N} + 1/\sqrt{M})$. Although DDRO has this ideal statistical consistency for the original objective $\mathcal{L}_{\mathrm{DRE}}$, training without the heuristic stabilization leads to gradient explosion. Conversely, the heuristic transformation S modifies the original objective $\mathcal{L}_{\mathrm{DRE}}$, causing the statistical consistency guarantee to no longer hold for $\hat{\mathcal{L}}_{\mathrm{DRE}}$ used in practice. This fundamental tension between stability and statistical consistency motivates our proposed method.

3 Proposed Method

One reason for the instability of DDRO lies in the instability of the density ratio $g_\theta(y|x)$. Although $g_\theta(y|x)$ models the true density ratio $p^-(y|x)/p^+(y|x)$, it may diverge when the supports of these two distributions, the regions where their values are non-zero, are different. Since the preferred response y^+ and the non-preferred response y^- are likely to have different supports, the density ratio may diverge, leading to instability in DDRO.

In this section, we propose an alignment method that is both stable and statistically consistent by addressing the instability of DDRO. To this end, we employ a more stable approach called relative density ratio estimation (Yamada et al., 2013), which generalizes density ratio estimation.

3.1 Relative Density Ratio Optimization

First, we define the relative density ratio $r^*(y|x)$ as follows:

$$r^*(y|x) \equiv \frac{p^+(y|x)}{p_{\mathrm{ref}}(y|x)} = \frac{p^+(y|x)}{\alpha p^+(y|x) + (1-\alpha)p^-(y|x)}. \tag{12}$$

This relative density ratio can be regarded as a generalization of the density ratio since it reduces to the standard density ratio when $\alpha = 0$. By definition, $r^*(y|x) \in [0, 1/\alpha]$, as also illustrated in Figure 3(a) (Appendix B). Hence, unlike the density ratio, the relative density ratio does not diverge and can be estimated in a stable manner.

By replacing the true preference $p^+(y|x)$ with the policy $p_\theta(y|x)$, we model this relative density ratio as follows:

$$r_\theta(y|x) \equiv \frac{p_\theta(y|x)}{p_{\mathrm{ref}}(y|x)}. \tag{13}$$

Analogous to DDRO, we align the policy $p_\theta(y|x)$ with the true preference $p^+(y|x)$ by minimizing the Bregman divergence between $r^*(y|x)$ and $r_\theta(y|x)$:

$$\mathcal{L}_{\mathrm{RDRE}}(\theta) = \mathbb{E}_{p_{\mathrm{ref}}(x,y)}\left[\mathrm{Breg}_f\left(r^*(y|x) \| r_\theta(y|x)\right)\right], \tag{14}$$

where $p_{\mathrm{ref}}(x, y) = p_{\mathrm{ref}}(y|x)p(x)$. When $f(t) = t\log t - (1+t)\log(1+t)$, this optimization problem can be reduced to the logistic regression problem (Sugiyama et al., 2012):

$$\mathcal{L}_{\mathrm{RDRE}}(\theta) = \mathbb{E}_{p_{\mathrm{ref}}(x,y)}\left[\log(1 + r_\theta(y|x))\right] + \mathbb{E}_{p^+(x,y)}\left[\log(1 + 1/r_\theta(y|x))\right]. \tag{15}$$

Defining $T_\theta(x, y) = \log p_\theta(y|x) - \log p_{\mathrm{ref}}(y|x)$, we have $r_\theta(y|x) = \exp(T_\theta(x, y))$. Substituting this into $\mathcal{L}_{\mathrm{RDRE}}(\theta)$, we obtain:

$$\begin{aligned}
\mathcal{L}_{\mathrm{RDRE}}(\theta) &= \mathbb{E}_{p_{\mathrm{ref}}(x,y)}\left[\log(1 + \exp(T_\theta(x, y)))\right] + \mathbb{E}_{p^+(x,y)}\left[\log(1 + 1/\exp(T_\theta(x, y)))\right] \\
&= \mathbb{E}_{p_{\mathrm{ref}}(x,y)}\left[\varphi(T_\theta(x, y))\right] + \mathbb{E}_{p^+(x,y)}\left[\varphi(T_\theta(x, y)) - T_\theta(x, y)\right] \\
&= \mathbb{E}_{p^+(x,y)}\left[(1+\alpha)\varphi(T_\theta(x, y)) - T_\theta(x, y)\right] + \mathbb{E}_{p^-(x,y)}\left[(1-\alpha)\varphi(T_\theta(x, y))\right],
\end{aligned} \tag{16}$$

where $\varphi(t) = \log(1 + \exp(t))$ is the softplus function. To obtain the last expression, we used $p_{\mathrm{ref}}(x, y) = \alpha p^+(x, y) + (1-\alpha)p^-(x, y)$. Figure 3(b) (Appendix B) shows our loss functions for preferred and non-preferred samples. Note that $S(t)$ used in Eq. (9) can be written as $S(t) = -\varphi(-t)$. This shows that the stabilizing transformation employed in DDRO arises naturally in our approach.

Using the datasets $\mathcal{D}^+$ and $\mathcal{D}^-$, $\mathcal{L}_{\mathrm{RDRE}}(\theta)$ is approximated as follows:

$$\mathcal{L}_{\mathrm{RDRO}}(\theta) \equiv \frac{1}{N}\sum_{n=1}^{N}\left\{(1+\alpha)\varphi(T_\theta(x_n^+, y_n^+)) - T_\theta(x_n^+, y_n^+)\right\} + \frac{1}{M}\sum_{m=1}^{M}(1-\alpha)\varphi(T_\theta(x_m^-, y_m^-)). \tag{17}$$

We optimize this objective using a stochastic optimization method such as AdamW (Loshchilov & Hutter, 2017). As we will show later, we can omit the KL regularizer used in DDRO, since α serves as a regularization parameter. We refer to our approach as relative density ratio optimization (RDRO). The PyTorch implementation and pseudocode are provided in Appendices A and C, respectively.

In the following section, we show that our approach is both stable and statistically consistent.

3.2 Stability Analysis

First, we analyze the stability of the proposed method by examining the gradient of the loss function. Since

$$\nabla_\theta \varphi(T_\theta(x, y)) = \sigma(T_\theta(x, y))\nabla_\theta T_\theta(x, y) \tag{18}$$

and

$$\nabla_\theta T_\theta(x, y) = \nabla_\theta \log p_\theta(y|x), \tag{19}$$

the gradient of $\mathcal{L}_{\mathrm{RDRE}}(\theta)$ can be written as follows:

$$\nabla_\theta \mathcal{L}_{\mathrm{RDRE}}(\theta) = \mathbb{E}_{p^+(x,y)}\left[c_\theta^+(x,y)\nabla_\theta \log p_\theta(y|x)\right] + \mathbb{E}_{p^-(x,y)}\left[c_\theta^-(x,y)\nabla_\theta \log p_\theta(y|x)\right], \quad (20)$$

where

$$c_\theta^+(x,y) = (1+\alpha)\sigma(T_\theta(x,y)) - 1, \qquad c_\theta^-(x,y) = (1-\alpha)\sigma(T_\theta(x,y)). \quad (21)$$

For the preferred dataset $\mathcal{D}^+$, we would like $p_\theta(y|x)$ to be large, whereas for the non-preferred dataset $\mathcal{D}^-$, we would like $p_\theta(y|x)$ to be small. Thus, $c_\theta^+(x,y) \leq 0$ and $c_\theta^-(x,y) \geq 0$ are required. By definition, $c_\theta^-(x,y)$ is always non-negative, and when $p_\theta(y|x)$ becomes sufficiently small for the non-preferred data, $c_\theta^-(x,y)$ becomes zero.

On the other hand, for $c_\theta^+(x,y)$, the following inequality holds:

$$c_\theta^+(x,y) \leq 0 \Leftrightarrow r_\theta(y|x) \equiv \frac{p_\theta(y|x)}{p_{\mathrm{ref}}(y|x)} \leq \frac{1}{\alpha}. \quad (22)$$

For the preferred data (x^+, y^+), the gradient is negative when $r_\theta(y^+|x^+) \leq 1/\alpha$. This enlarges $p_\theta(y^+|x^+)$ until $c_\theta^+(x^+,y^+)$ becomes zero at $r_\theta(y^+|x^+) = 1/\alpha$. Hence, the constraint on the relative density ratio, $r_\theta(y^+|x^+) \in [0, 1/\alpha]$, is automatically satisfied.

From another perspective, $p_\theta(y|x)$ is trained so that it does not exceed $p_{\mathrm{ref}}(y|x)/\alpha$. When α is large, $p_\theta(y|x)$ remains close to $p_{\mathrm{ref}}(y|x)$, whereas when α is small, $p_\theta(y|x)$ can deviate from $p_{\mathrm{ref}}(y|x)$. Hence, α can be interpreted as a regularization parameter.

Note that this bounding mechanism differs from gradient clipping. Gradient clipping rescales the aggregated gradient only after its direction has been determined. Thus, an unbounded density ratio may allow a few samples to dominate the aggregated gradient and produce a poorly conditioned update direction that clipping cannot correct. In contrast, RDRO bounds the coefficient for each sample before aggregating the gradients, since $c_\theta^+ \in [-1, \alpha]$ and $c_\theta^- \in [0, 1-\alpha]$. Gradient clipping remains useful for controlling large gradients of $\log p_\theta(y|x)$.

3.3 Statistical Consistency Analysis

Next, we analyze the statistical consistency of the proposed method. For the optimal parameter θ^* that minimizes the true risk $\mathcal{L}_{\mathrm{RDRE}}(\theta) = \mathbb{E}_{p_{\mathrm{ref}}}\left[\mathrm{Breg}_f\left(r^* \| r_\theta\right)\right]$, since f is strictly convex, the following consistency holds:

$$\mathcal{L}_{\mathrm{RDRE}}(\theta^*) = 0 \Leftrightarrow r_{\theta^*} = r^* \Leftrightarrow p_{\theta^*} = p^+. \quad (23)$$

Although this is an ideal property, since the number of samples is finite in practice, it is necessary to evaluate the convergence rate.

As a preliminary step, we evaluate the Lipschitz constant of the Bregman divergence, which can be rewritten as follows:

$$\mathrm{Breg}_f\left(u \| v\right) = \underbrace{-f(v) + f'(v)v}_{\psi_1(v)} + \underbrace{\left(-f'(v)\right)}_{\psi_2(v)} u + \underbrace{f(u)}_{Const.}. \quad (24)$$

Let L_1 and L_2 be the Lipschitz constants of ψ_1 and ψ_2, respectively. Then, the Lipschitz constant of the Bregman divergence can be written as follows:

$$\mathrm{Lip}\left(\mathrm{Breg}_f\left(u \| \cdot\right)\right) = L_1 + (\sup|u|)L_2. \quad (25)$$

For the proposed method, since $r^* \leq 1/\alpha$, this Lipschitz constant becomes:

$$C_{\mathrm{Lip}} = \mathrm{Lip}\left(\mathrm{Breg}_f\left(r^* \| \cdot\right)\right) = L_1 + (\sup|r^*|)L_2 = L_1 + \frac{1}{\alpha}L_2. \quad (26)$$

Let Θ be the parameter set, and $\mathcal{H} = \{p_\theta \mid \theta \in \Theta\}$ be the corresponding model class, following the convention of (Higuchi & Suzuki, 2025). For this $\mathcal{H}$, the Rademacher complexity (Mohri et al., 2018) is defined as $\mathcal{R}_K(\mathcal{H})$, where K is the number of samples. The Rademacher complexity measures the richness of a model class and is commonly used to bound generalization error. Based on this, we evaluate the estimation error of the policy with the true preference p^+.

Theorem 3.1. *We define $\hat{\theta} = \arg\min_\theta \hat{\mathcal{L}}_{\mathrm{RDRE}}(\theta)$, where $\hat{\mathcal{L}}_{\mathrm{RDRE}}(\theta)$ is the empirical approximation of $\mathcal{L}_{\mathrm{RDRE}}(\theta)$ by using the datasets $\mathcal{D}^{\pm} = \{\mathcal{D}^+, \mathcal{D}^-\}$. Then, the estimation error of $p_{\hat{\theta}}$ with p^+ is bounded above as follows:*

$$\mathbb{E}_{\mathcal{D}^{\pm}}\left[\mathbb{E}_{p^+}\left[\left(p_{\hat{\theta}} - p^+\right)^2\right]\right] \leq \frac{2}{\alpha\mu}\left[\inf_{\theta\in\Theta} \mathcal{L}_{\mathrm{RDRE}}(\theta) + 4C_{\mathrm{Lip}}\left\{\alpha\mathcal{R}_N(\mathcal{H}) + (1-\alpha)\mathcal{R}_M(\mathcal{H})\right\}\right], \tag{27}$$

where μ satisfies $f''(t) \geq \mu$ for all t.

Proof sketch. The key insight is the boundedness $r^* \leq 1/\alpha$, which (i) yields a smaller Lipschitz constant $C_{\mathrm{Lip}} = L_1 + L_2/\alpha$ than that of DDRO, and (ii) gives $\mathbb{E}_{p^+}[(p_\theta - p^+)^2] \leq \frac{2}{\alpha\mu}\mathcal{L}_{\mathrm{RDRE}}(\theta)$ via $p_\theta - p^+ = p_{\mathrm{ref}}(r_\theta - r^*)$. Combining these with a standard Rademacher complexity argument gives the result; see Appendix D for the complete proof. □

Since $\mathcal{R}_K(\mathcal{H}) = \mathcal{O}(1/\sqrt{K})$ usually holds (Mohri et al., 2018), the right-hand side of Eq. (27) can be evaluated as

$$\mathcal{O}\left(\inf_{\theta\in\Theta} \mathcal{L}_{\mathrm{RDRE}}(\theta) + 1/\sqrt{N} + 1/\sqrt{M}\right). \tag{28}$$

In addition, if the model class $\mathcal{H}$ contains the true preference p^+, $\inf_{\theta\in\Theta} \mathcal{L}_{\mathrm{RDRE}}(\theta)$ becomes zero. Therefore, $\mathbb{E}_{p^+}\left[\left(p_{\hat{\theta}} - p^+\right)^2\right] \to 0$ as $N, M \to \infty$.

3.4 Comparison of Convergence Rates

Finally, we compare the convergence rates of the proposed method and DDRO. According to (Higuchi & Suzuki, 2025), the upper bound of the estimation error for DDRO is:

$$\mathbb{E}_{\mathcal{D}^{\pm}}\left[\mathbb{E}_{p^+}\left[\left(p_{\hat{\theta}} - p^+\right)^2\right]\right] \leq \frac{2\left(1-\alpha\right)^2}{\alpha^2 m_+^2 \mu}\left[\inf_{\theta\in\Theta} \mathcal{L}_{\mathrm{DRE}}(\theta) + 4C'_{\mathrm{Lip}}\left\{\mathcal{R}_N(\mathcal{H}) + \mathcal{R}_M(\mathcal{H})\right\}\right], \tag{29}$$

where

$$m_+ = \min_{x\in\mathcal{X}, y\in\mathrm{supp}(p^+(\cdot|x))} p^+(y|x) > 0 \tag{30}$$

is the minimum value of $p^+(y|x)$ within its support $\mathrm{supp}(p^+(\cdot|x))$, $\mathcal{X}$ is the input space, and

$$C'_{\mathrm{Lip}} = \mathrm{Lip}\left(\mathrm{Breg}_f\left(g^* \| \cdot\right)\right) = L_1 + (\sup|g^*|)L_2. \tag{31}$$

When $\mathcal{H}$ contains p^+, both $\inf_{\theta\in\Theta} \mathcal{L}_{\mathrm{DRE}}(\theta)$ and $\inf_{\theta\in\Theta} \mathcal{L}_{\mathrm{RDRE}}(\theta)$ are zero. We therefore compare the remaining upper-bound terms. For $m_+ \ll 1$, the leading coefficients in the two bounds satisfy:

$$\frac{2}{\alpha\mu} \ll \frac{2\left(1-\alpha\right)^2}{\alpha^2 m_+^2 \mu}. \tag{32}$$

More precisely, this inequality holds whenever $\alpha < \{(\sqrt{m_+^2 + 4} - m_+)/2\}^2 = 1 - m_+ + O(m_+^2)$. Hence, for any fixed $\alpha \in (0, 1)$, Eq. (32) holds for sufficiently small m_+. Appendix E gives the full derivation.

Moreover, since $1/\alpha \ll \sup|g^*|$, the following inequality holds:

$$C_{\mathrm{Lip}} = L_1 + \frac{1}{\alpha}L_2 \ll L_1 + (\sup|g^*|)L_2 = C'_{\mathrm{Lip}}. \tag{33}$$

In addition, since $\alpha \in (0, 1)$, the following inequality holds:

$$\alpha\mathcal{R}_N(\mathcal{H}) + (1-\alpha)\mathcal{R}_M(\mathcal{H}) \leq \mathcal{R}_N(\mathcal{H}) + \mathcal{R}_M(\mathcal{H}). \tag{34}$$

From the above discussion, although both the proposed method and DDRO share the order of $\mathcal{O}(1/\sqrt{N}+1/\sqrt{M})$, our coefficient terms are significantly smaller than those for DDRO. In particular, in DDRO, the occurrence of $1/m_+$ and the potentially large value of $\sup|g^*|$ arise from the instability of the density ratio $g^* \in [0, \infty)$. The proposed method avoids this order deterioration by using the relative density ratio $r^* \in [0, 1/\alpha]$ instead of the density ratio g^*.

To summarize our contributions, our approach is more stable and statistically consistent, and yields significantly tighter convergence guarantees than DDRO.

4 Related Work

Many alignment methods assume specific human preference models, such as the Bradley-Terry model (Bradley & Terry, 1952) and prospect theory (Tversky & Kahneman, 1992; Kahneman & Tversky, 2013). The Bradley-Terry model is widely used in reinforcement learning approaches such as RLHF (Ouyang et al., 2022; Uesato et al., 2022; Ahmadian et al., 2024), as well as in DPO and its variants (Rafailov et al., 2023; Xu et al., 2024; Xie et al., 2024; Meng et al., 2024). KTO (Ethayarajh et al., 2024) is based on prospect theory and can be regarded as a generalization of DPO. Unlike RLHF and DPO, which require a preferred and a non-preferred response for each prompt, KTO can be applied to individual responses with binary labels indicating whether they are preferred. However, specific human preference models such as the Bradley-Terry model may fail to accurately capture true human preferences (Wu et al., 2022). Cyclic preference is a canonical example of this limitation (Higuchi & Suzuki, 2025); Appendix F gives an explicit construction and proof.

To address this limitation, several approaches have been presented, including those based on energy-based models (Hong et al., 2024), Nash equilibrium (Munos et al., 2023), binary classification (Jung et al., 2024), and density ratio estimation (Higuchi & Suzuki, 2025; Kim et al., 2025). In particular, DDRO (Higuchi & Suzuki, 2025) has statistical consistency as shown in Section 2.2. However, DDRO is also known to suffer from training instability. One cause lies in the instability of the density ratio, which may diverge when the supports of the two distributions differ (Sugiyama et al., 2012). Our approach avoids this instability by employing the bounded relative density ratio (Yamada et al., 2013). Consequently, our approach offers both training stability and statistical consistency, and yields significantly tighter convergence guarantees than DDRO. Note that Kim et al. (2025) is also based on density ratio estimation, but requires pairwise preference data and does not provide a guarantee of statistical consistency.

Note that using pointwise supervision does not by itself avoid the limitations of specific human preference models. For example, SPIN (Chen et al., 2024) constructs pairwise comparisons between supervised responses and responses from the current model, and iteratively trains a discriminator through self-play. Although SPIN does not explicitly assume the Bradley-Terry model, its final objective coincides with that of DPO and therefore cannot represent cyclic preferences. In contrast, our approach uses only pointwise labels without assuming any human preference models and satisfies statistical consistency.

5 Experiments

We primarily compared RDRO against KTO (Ethayarajh et al., 2024) and DDRO (Higuchi & Suzuki, 2025), which also learn from pointwise labels. Where pairwise labels are available, we additionally report DPO (Rafailov et al., 2023) and SimPO (Meng et al., 2024) to provide a broader comparison. We exclude these methods from the primary comparison because they require stronger pairwise supervision. Due to space constraints, additional results are deferred to Appendices G and H.

5.1 Base Models, Preference Datasets, and Metrics

We applied these alignment methods to the following base models: Qwen 2.5 Instruct (1.5B and 3B) (Team et al., 2024), and Llama 3 Instruct (3B and 8B) (Grattafiori et al., 2024). We refer to them as Qwen-1.5B, Qwen-3B, Llama-3B, and Llama-8B, respectively.

As the preference datasets, we used UltraFeedback GPT-3.5-Turbo Helpfulness (UF-G) and KTO Mix 14K (MIX-14K). UF-G is derived from the UltraFeedback (Cui et al., 2023) and filtered for helpfulness, containing preferred and non-preferred responses to various prompts labeled by GPT-3.5-turbo. MIX-14K mixes the following three datasets: Capybara, Orca, and UltraFeedback by sampling from each source in equal proportions. UF-G contains 6,088 preferred and 9,644 non-preferred samples, while MIX-14K contains 6,750 preferred and 6,750 non-preferred samples.

As the evaluation metric, we used AlpacaEval length-controlled (LC) win rates (Dubois et al., 2024). We first generated responses from the aligned models on the alpaca_eval prompts, then computed win rates against a baseline using an LLM-based auto-annotator, and finally reported LC win rates that regress out AlpacaEval's bias toward longer outputs. We used `gpt-4-turbo-2024-04-09` for the baseline and `weighted_alpaca_eval_gpt4_turbo` for the auto-annotator.

Table 1: Comparison of AlpacaEval LC Win Rates on UF-G.

	Qwen-1.5B	Qwen-3B	Llama-3B	Llama-8B
KTO	**5.220 ± 0.104**	7.920 ± 0.082	9.937 ± 0.024	9.344 ± 0.296
DDRO	**5.364 ± 0.162**	**7.888 ± 0.088**	12.069 ± 0.113	15.660 ± 0.343
RDRO	**5.599 ± 0.122**	**8.933 ± 0.318**	**12.634 ± 0.101**	**17.314 ± 0.078**

Table 2: Comparison of AlpacaEval LC Win Rates on MIX-14K. DPO (Rafailov et al., 2023) and SimPO (Meng et al., 2024) use pairwise labels (DPO-Mix-7K), while KTO, DDRO, and RDRO use pointwise labels (MIX-14K).

	Qwen-1.5B	Qwen-3B	Llama-3B	Llama-8B
KTO	5.320 ± 0.173	8.240 ± 0.124	**12.763 ± 0.190**	13.921 ± 0.041
DDRO	5.169 ± 0.166	8.458 ± 0.138	**12.487 ± 0.156**	**17.269 ± 0.675**
RDRO	5.297 ± 0.102	8.793 ± 0.238	**12.720 ± 0.205**	**17.102 ± 0.305**
DPO	**7.134 ± 0.145**	12.418 ± 0.085	10.056 ± 0.088	**16.039 ± 0.448**
SimPO	6.557 ± 0.059	**12.612 ± 0.142**	9.815 ± 0.231	**15.862 ± 0.242**

5.2 Setup

We implemented all methods using the Hugging Face TRL library. We used the KTO implementation in TRL and implemented DDRO and the proposed method by extending this KTO implementation. For training, we used AdamW (Loshchilov & Hutter, 2017) with a cosine learning rate scheduler (Loshchilov & Hutter, 2016). Following the TRL documentation, we trained all models for one epoch with a learning rate of 5×10^{-7} and a batch size of 128. We set the maximum gradient norm for gradient clipping to 1.0 and used a warmup ratio of 10% of the total training steps.

For DDRO and the proposed method, we set the hyperparameter α to the fraction of preferred samples in each dataset: 0.39 for UF-G and 0.5 for MIX-14K. This choice is motivated by Eq. (1): under the mixture model, α represents the proportion of preferred responses in the reference policy's output distribution, so setting it to the empirical fraction of preferred samples in the dataset is natural. For DDRO, we set the KL regularization coefficient β to 0.1, the default value in the KTO implementation of TRL. Following (Higuchi & Suzuki, 2025), which explicitly excludes the KL regularization term from gradient computation for training stability, β does not affect the gradient update in practice.

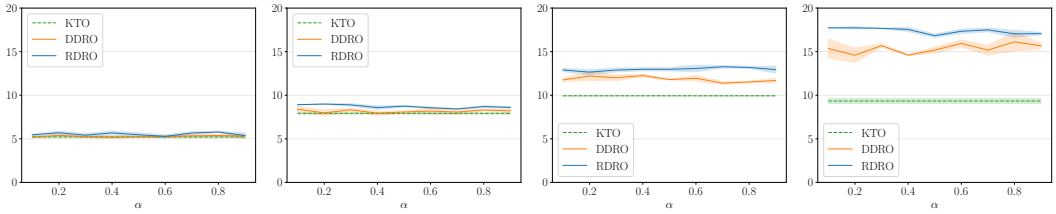
The experiments were conducted on machines equipped with eight NVIDIA H100 GPUs (80GB HBM3), Intel Xeon Platinum 8480+ CPUs, and 2.0TB of memory. We used Hugging Face Accelerate for distributed training and DeepSpeed ZeRO-3 (Rajbhandari et al., 2020) for memory efficiency. We repeated each experiment three times with different random seeds.

5.3 Results

Tables 1 and 2 show the AlpacaEval LC win rates on UF-G and MIX-14K, respectively. We used boldface to indicate the best result and any results not significantly different from it according to a pairwise t-test with a significance level of 5%.

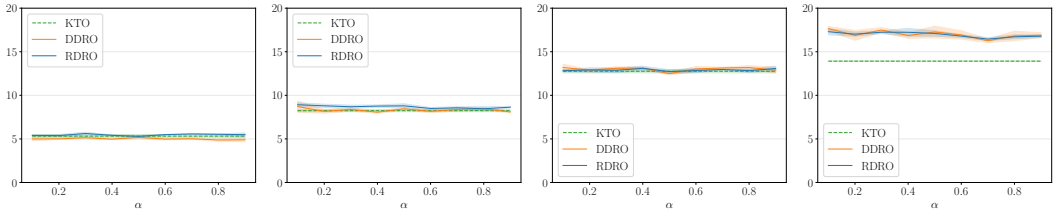
We first focus on the results on UF-G. Table 1 shows that, across all models, the proposed method achieves performance comparable to or better than both KTO and DDRO. In particular, for Llama-3B and Llama-8B, the proposed method yields a statistically significant improvement over KTO and DDRO. We next focus on the results on MIX-14K. Table 2 shows that, across all models, the proposed method achieves comparable performance to KTO and DDRO.

This difference across datasets may stem from structural differences between UF-G and MIX-14K. MIX-14K decomposes each triplet (x, y_w, y_l) from DPO-Mix-7K (Rafailov et al., 2023) into two prompt and response pairs, (x, y_w) and (x, y_l). Thus, both a preferred and a non-preferred response



(a) Qwen-1.5B. (b) Qwen-3B. (c) Llama-3B. (d) Llama-8B.

Figure 1: AlpacaEval LC win rates across α on UF-G, with shading indicating standard deviations.



(a) Qwen-1.5B. (b) Qwen-3B. (c) Llama-3B. (d) Llama-8B.

Figure 2: AlpacaEval LC win rates across α on MIX-14K, with shading indicating standard deviations.

are available for every prompt, which may make alignment easier and explain the competitive performance of all methods. In contrast, UF-G provides only one of y_w and y_l for each prompt, which may make alignment more challenging and produce larger performance differences. In practice, it is difficult to always provide both preferred and non-preferred responses for each prompt. Therefore, achieving strong performance on realistic datasets like UF-G is particularly important.

Table 2 also reports results for DPO and SimPO trained on DPO-Mix-7K, the pairwise version of MIX-14K. Note that this comparison is not fully comparable since DPO and SimPO have access to stronger (pairwise) supervision. Nevertheless, RDRO achieves performance comparable to or better than DPO and SimPO on larger models (Llama-3B and Llama-8B) using only pointwise labels.

Figures 1 and 2 show AlpacaEval LC win rates across α on UF-G and MIX-14K, respectively. Across both datasets, RDRO is robust to α: it outperforms DDRO at every tested value on UF-G and remains comparable on MIX-14K. These results show that RDRO does not require precise tuning of α across the evaluated base models and datasets.

Additional results on BIG-Bench Hard (Suzgun et al., 2023) are provided in Appendix G. Appendix H analyzes the training dynamics to empirically evaluate the stability of the proposed method. These results demonstrate the strong alignment performance and training stability of our approach.

6 Conclusion, Limitations, and Future Work

In this paper, we propose relative density ratio optimization (RDRO), a novel alignment method that resolves the inherent trade-off between training stability and statistical consistency in DDRO. Experiments across various models and datasets demonstrate that RDRO achieves performance comparable to or better than KTO and DDRO, and is competitive with pairwise methods such as DPO and SimPO despite using only pointwise labels. One limitation is that our approach requires fully labeled data; incorporating unlabeled data via positive-unlabeled learning (Kato et al., 2019; Kato & Teshima, 2021) is a promising direction. A further limitation concerns the mixture model assumption described in Section 2.2, which approximates $p(+|x)$ by a constant α across prompts. This assumption may be violated when $p(+|x)$ varies across prompts, as can occur in UF-G and MIX-14K; nevertheless, RDRO performs well on both datasets. Figures 1 and 2 further show that misspecifying α does not substantially degrade performance, and the bound in Theorem 3.1 remains finite for any fixed $\alpha \in (0, 1)$. Replacing α with a prompt-dependent $\alpha(x)$ estimated by a lightweight model (Devlin et al., 2019) would be a natural extension.

Ethics statement

This work does not collect new data from human participants. It studies model alignment with preference data intended to represent human preferences. Such data may reflect biases in the annotators and labeling protocols, and aligned models may reproduce or amplify these biases. Deployment should therefore include evaluation for the target population and application, together with appropriate safeguards and human oversight.

Reproducibility statement

Section 5 describes the experimental setup, datasets, evaluation protocol, and compute environment. Appendix A provides a PyTorch implementation of the proposed loss, Appendix D contains complete proofs, and anonymized source code is included in the supplementary material.

AI use statement

In this work, we used generative AI tools only to improve the English, clarity, and organization of our writing. We reviewed all suggested revisions and take responsibility for the final content of this work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Arash Ahmadian, Chris Cremer, Matthias Gallé, Marzieh Fadaee, Julia Kreutzer, Olivier Pietquin, Ahmet Üstün, and Sara Hooker. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*, 2024.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp. 610–623, 2021.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 6621–6642. PMLR, 2024.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Rei Higuchi and Taiji Suzuki. Direct density ratio optimization: A statistically consistent approach to aligning large language models. In *Forty-second International Conference on Machine Learning*, 2025.

Yuzhong Hong, Hanshan Zhang, Junwei Bao, Hongfei Jiang, and Yang Song. Energy-based preference model offers better offline alignment than the bradley-terry preference model. *arXiv preprint arXiv:2412.13862*, 2024.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38, 2023.

Seungjae Jung, Gunsoo Han, Daniel Wontae Nam, and Kyoung-Woon On. Binary classifier optimization for large language model alignment. *arXiv preprint arXiv:2404.04656*, 2024.

Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*, pp. 99–127. World Scientific, 2013.

Masahiro Kato and Takeshi Teshima. Non-negative bregman divergence minimization for deep direct density ratio estimation. In *International Conference on Machine Learning*, pp. 5320–5333. PMLR, 2021.

Masahiro Kato, Takeshi Teshima, and Junya Honda. Learning from positive and unlabeled data with a selection bias. In *International conference on learning representations*, 2019.

Yeongmin Kim, Heesun Bae, Byeonghu Na, and Il-Chul Moon. Preference optimization by estimating the ratio of the data distribution. *arXiv preprint arXiv:2505.19601*, 2025.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, et al. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114 (13):3521–3526, 2017.

Thomas Kwa, Drake Thomas, and Adrià Garriga-Alonso. Catastrophic goodhart: regularizing rlhf with kl divergence does not mitigate heavy-tailed reward misspecification. *Advances in Neural Information Processing Systems*, 37:14608–14633, 2024.

Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Andrea Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 18, 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.

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.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*, pp. 1–16. IEEE, 2020.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. *arXiv preprint arXiv:1909.01326*, 2019.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. Density-ratio matching under the bregman divergence: a unified framework of density-ratio estimation. *Annals of the Institute of Statistical Mathematics*, 64(5):1009–1044, 2012.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023.

Qwen Team et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3), 2024.

Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5(4):297–323, 1992.

Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. Solving math word problems with process-and outcome-based feedback. *arXiv preprint arXiv:2211.14275*, 2022.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Weichen Wu, Nynke Niezink, and Brian Junker. A diagnostic framework for the bradley-terry model. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 185:S461–S484, 2022.

Tengyang Xie, Dylan J Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf. *arXiv preprint arXiv:2405.21046*, 2024.

Haoran Xu, Amr Sharaf, Yunmo Chen, Weiting Tan, Lingfeng Shen, Benjamin Van Durme, Kenton Murray, and Young Jin Kim. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*, 2024.

Makoto Yamada, Taiji Suzuki, Takafumi Kanamori, Hirotaka Hachiya, and Masashi Sugiyama. Relative density-ratio estimation for robust distribution comparison. *Neural computation*, 25(5): 1324–1370, 2013.

A PyTorch Implementation

RDRO is easy to implement in practice; PyTorch code for the RDRO loss is provided below:

```
import torch
from torch.nn.functional import softplus as sp

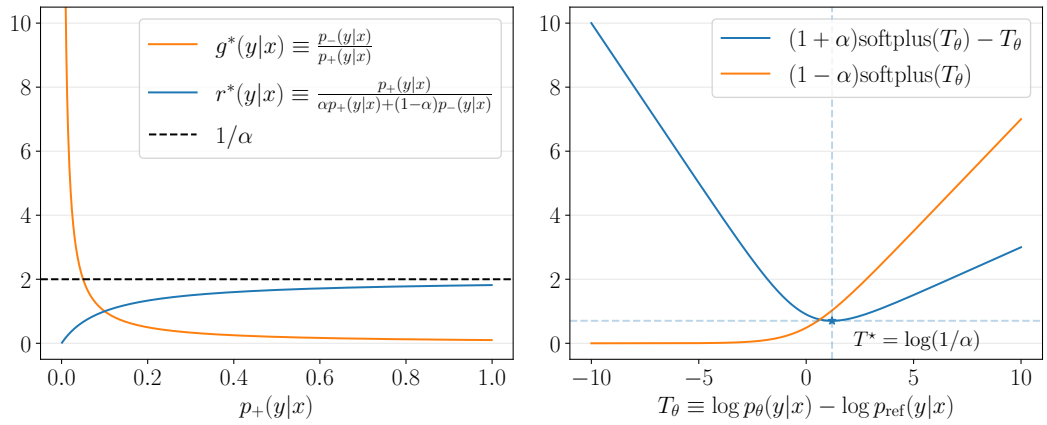

def rdro_loss(
    policy_chosen_logps,
    policy_rejected_logps,
    ref_chosen_logps,
    ref_rejected_logps,
    alpha,
):
    """
    policy_chosen_logps: policy logprobs for chosen responses
    policy_rejected_logps: policy logprobs for rejected responses
    ref_chosen_logps: reference policy logprobs for chosen responses
    ref_rejected_logps: reference policy logprobs for rejected responses
    alpha: hyperparameter in (0, 1)
    """

    chosen_logratios = policy_chosen_logps - ref_chosen_logps
    chosen_losses = (1 + alpha) * sp(chosen_logratios) - chosen_logratios

    rejected_logratios = policy_rejected_logps - ref_rejected_logps
    rejected_losses = (1 - alpha) * sp(rejected_logratios)

    return torch.cat([chosen_losses, rejected_losses], dim=0)
```

B Illustrations of Density Ratios and Loss Functions



(a) Comparing two ratios.

(b) Loss functions.

Figure 3: Density ratio comparison and loss functions.

Figure 3(a) compares the density ratio $g^*(y|x)$ and the relative density ratio $r^*(y|x)$, where the non-preferred data distribution $p^-(y|x) = 0.1$ and the hyperparameter $\alpha = 0.5$. Although $g^*(y|x)$ diverges as $p^+(y|x) \to 0$, $r^*(y|x)$ is bounded above by $1/\alpha$. Figure 3(b) shows the loss functions for preferred (blue) and non-preferred (orange) samples with $\alpha = 0.3$. The loss for preferred samples is minimized when $T_\theta \equiv \log p_\theta(y|x) - \log p_{\text{ref}}(y|x)$ attains $\log(1/\alpha)$.

C PSEUDOCODE

Algorithm 1 provides the pseudocode for RDRO.

Algorithm 1 Relative Density Ratio Optimization

Input: datasets $(\mathcal{D}^+, \mathcal{D}^-)$, reference policy $p_{\mathrm{ref}}(y|x)$, and hyperparameter $\alpha \in (0, 1)$
Output: target policy $p_\theta(y|x)$
Procedure:

```
while not converged do
    Sample mini-batch B from datasets (D+, D−)
    Compute gradient of Eq. (17) with B
    Update θ with the above gradient
end while
```

D PROOF OF THEOREM 3.1

In this section, we will prove Theorem 3.1 after providing four relevant lemmas. For simplicity, we denote the true risk $\mathcal{L}_{\mathrm{RDRE}}(\theta)$ as $\mathcal{L}(\theta)$, and the empirical risk $\hat{\mathcal{L}}_{\mathrm{RDRE}}(\theta)$ as $\hat{\mathcal{L}}(\theta)$.

Lemma D.1. *$\mathcal{L}(\theta)$ is bounded below as follows:*

$$\mathcal{L}(\theta) = \mathbb{E}_{p_{\mathrm{ref}}}\left[\mathrm{Breg}_f\left(r^* \| r_\theta\right)\right] \geq \frac{\mu}{2}\mathbb{E}_{p_{\mathrm{ref}}}\left[\left(r^* - r_\theta\right)^2\right], \tag{35}$$

where μ satisfies $f''(t) \geq \mu$ for all t.

Proof. Since f is a strongly convex function, the following inequality holds:

$$f(r^*) \geq f(r_\theta) + f'(r_\theta)(r^* - r_\theta) + \frac{\mu}{2}\left(r^* - r_\theta\right)^2, \tag{36}$$

and thus,

$$\mathrm{Breg}_f\left(r^* \| r_\theta\right) = f(r^*) - f(r_\theta) - f'(r_\theta)(r^* - r_\theta) \geq \frac{\mu}{2}\left(r^* - r_\theta\right)^2. \tag{37}$$

Taking expectations on both sides yields the claim. □

Lemma D.2. *The squared error between p_θ and p^+ can be bounded by $\mathcal{L}(\theta)$ as follows:*

$$\mathbb{E}_{p^+}\left[\left(p_\theta - p^+\right)^2\right] \leq \frac{2}{\alpha\mu}\mathcal{L}(\theta). \tag{38}$$

Proof. The squared error between p_θ and p^+ is bounded above as follows:

$$\begin{aligned}
\mathbb{E}_{p^+}\left[\left(p_\theta - p^+\right)^2\right] &= \mathbb{E}_{p^+}\left[p_{\mathrm{ref}}^2\left(r_\theta - r^*\right)^2\right] && (39)\\
&= \mathbb{E}_{p_{\mathrm{ref}}}\left[r^* p_{\mathrm{ref}}^2\left(r_\theta - r^*\right)^2\right] && (40)\\
&\leq \frac{1}{\alpha}\mathbb{E}_{p_{\mathrm{ref}}}\left[p_{\mathrm{ref}}^2\left(r_\theta - r^*\right)^2\right] && (41)\\
&\leq \frac{1}{\alpha}\mathbb{E}_{p_{\mathrm{ref}}}\left[\left(r_\theta - r^*\right)^2\right], && (42)
\end{aligned}$$

where we used $p_\theta - p^+ = p_{\mathrm{ref}}(r_\theta - r^*)$, $p^+ = r^* p_{\mathrm{ref}}$, $r^* \leq 1/\alpha$, and $p_{\mathrm{ref}} \leq 1$ since p_{ref} is a discrete probability distribution. Combining with Lemma D.1 yields the claim. □

Lemma D.3. *For $\hat{\theta}$ that satisfies $\hat{\theta} = \arg\min_\theta \hat{\mathcal{L}}(\theta)$, $\mathcal{L}(\hat{\theta})$ is bounded above as follows:*

$$\mathcal{L}(\hat{\theta}) \leq \inf_{\theta\in\Theta}\mathcal{L}(\theta) + \sup_{\theta\in\Theta}\left(\mathcal{L}(\theta) - \hat{\mathcal{L}}(\theta)\right) + \sup_{\theta\in\Theta}\left(\hat{\mathcal{L}}(\theta) - \mathcal{L}(\theta)\right). \tag{43}$$

Proof. $\mathcal{L}(\hat{\theta})$ is bounded above as follows:

$$\mathcal{L}(\hat{\theta}) = \mathcal{L}(\hat{\theta}) + \mathcal{L}(\theta) - \mathcal{L}(\theta) + \hat{\mathcal{L}}(\hat{\theta}) - \hat{\mathcal{L}}(\hat{\theta}) + \hat{\mathcal{L}}(\theta) - \hat{\mathcal{L}}(\theta) \tag{44}$$

$$= \mathcal{L}(\theta) + \left(\mathcal{L}(\hat{\theta}) - \hat{\mathcal{L}}(\hat{\theta})\right) + \left(\hat{\mathcal{L}}(\theta) - \mathcal{L}(\theta)\right) + \left(\hat{\mathcal{L}}(\hat{\theta}) - \hat{\mathcal{L}}(\theta)\right) \tag{45}$$

$$\leq \mathcal{L}(\theta) + \sup_{\theta\in\Theta}\left(\mathcal{L}(\theta) - \hat{\mathcal{L}}(\theta)\right) + \sup_{\theta\in\Theta}\left(\hat{\mathcal{L}}(\theta) - \mathcal{L}(\theta)\right), \tag{46}$$

where we used $\hat{\mathcal{L}}(\hat{\theta}) - \hat{\mathcal{L}}(\theta) \leq 0$. Taking the infimum over $\theta \in \Theta$ on both sides yields the claim. □

Lemma D.4. *The uniform deviation between the empirical risk $\hat{\mathcal{L}}(\theta)$ and the true risk $\mathcal{L}(\theta)$ is bounded above as follows:*

$$\mathbb{E}_{\mathcal{D}^\pm}\left[\sup_{\theta\in\Theta}\left(\hat{\mathcal{L}}(\theta) - \mathcal{L}(\theta)\right)\right] \leq 2C_{\mathrm{Lip}}\left\{\alpha\mathcal{R}_N(\mathcal{H}) + (1-\alpha)\mathcal{R}_M(\mathcal{H})\right\}, \tag{47}$$

where $\mathcal{D}^\pm = \{\mathcal{D}^+, \mathcal{D}^-\}$ is the dataset, C_{Lip} is the Lipschitz constant of the Bregman divergence, and $\mathcal{R}_N(\mathcal{H})$ and $\mathcal{R}_M(\mathcal{H})$ are Rademacher complexities of the model class $\mathcal{H} = \{p_\theta \mid \theta \in \Theta\}$ with sample sizes N and M, respectively.

Proof. Let $z = (x, y)$ and define $\ell_\theta(z) \equiv \mathrm{Breg}_f\left(r^*(z) \| r_\theta(z)\right)$. With this notation, we can rewrite $\mathcal{L}(\theta)$ and $\hat{\mathcal{L}}(\theta)$ as follows:

$$\mathcal{L}(\theta) = \alpha\mathbb{E}_{p^+}\left[\ell_\theta(z)\right] + (1-\alpha)\mathbb{E}_{p^-}\left[\ell_\theta(z)\right], \tag{48}$$

$$\hat{\mathcal{L}}(\theta) = \alpha\frac{1}{N}\sum_{n=1}^{N}\ell_\theta(z_n^+) + (1-\alpha)\frac{1}{M}\sum_{m=1}^{M}\ell_\theta(z_m^-), \tag{49}$$

where $(z_n^+)_{n=1}^N = (x_n^+, y_n^+)_{n=1}^N = \mathcal{D}^+$ and $(z_m^-)_{m=1}^M = (x_m^-, y_m^-)_{m=1}^M = \mathcal{D}^-$. Then, we obtain the following inequality:

$$\sup_{\theta\in\Theta}\left(\hat{\mathcal{L}}(\theta) - \mathcal{L}(\theta)\right) \leq \alpha\sup_{\theta\in\Theta}\left(\frac{1}{N}\sum_{n=1}^{N}\ell_\theta(z_n^+) - \mathbb{E}_{p^+}\left[\ell_\theta(z)\right]\right) + (1-\alpha)\sup_{\theta\in\Theta}\left(\frac{1}{M}\sum_{m=1}^{M}\ell_\theta(z_m^-) - \mathbb{E}_{p^-}\left[\ell_\theta(z)\right]\right). \tag{50}$$

By the symmetrization trick (Mohri et al., 2018) and the contraction inequality (Mohri et al., 2018), the uniform deviation is bounded above as follows:

$$\mathbb{E}_{\mathcal{D}^\pm}\left[\sup_{\theta\in\Theta}\left(\hat{\mathcal{L}}(\theta) - \mathcal{L}(\theta)\right)\right] \leq 2C_{\mathrm{Lip}}\left\{\alpha\mathcal{R}_N(\mathcal{H}) + (1-\alpha)\mathcal{R}_M(\mathcal{H})\right\}, \tag{51}$$

where

$$\mathcal{R}_K(\mathcal{H}) = \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_\sigma\left[\sup_{\theta\in\Theta}\frac{1}{K}\sum_{k=1}^{K}\sigma_k p_\theta(z_k)\right]\right] \tag{52}$$

is the Rademacher complexity of $\mathcal{H}$ with sample size $|\mathcal{D}| = K$, and $(\sigma_k)_{k=1}^K$ are independent uniform random variables taking values in $\{-1, +1\}$. □

In the same way, we can also obtain the following upper bound for the uniform deviation:

$$\mathbb{E}_{\mathcal{D}^\pm}\left[\sup_{\theta\in\Theta}\left(\mathcal{L}(\theta) - \hat{\mathcal{L}}(\theta)\right)\right] \leq 2C_{\mathrm{Lip}}\left\{\alpha\mathcal{R}_N(\mathcal{H}) + (1-\alpha)\mathcal{R}_M(\mathcal{H})\right\}. \tag{53}$$

By combining Lemma D.3 and Lemma D.4, we obtain the following upper bound:

$$\mathbb{E}_{\mathcal{D}^\pm}\left[\mathcal{L}(\hat{\theta})\right] \leq \mathbb{E}_{\mathcal{D}^\pm}\left[\inf_{\theta\in\Theta}\mathcal{L}(\theta)\right] + \mathbb{E}_{\mathcal{D}^\pm}\left[\sup_{\theta\in\Theta}\left(\hat{\mathcal{L}}(\theta) - \mathcal{L}(\theta)\right)\right] + \mathbb{E}_{\mathcal{D}^\pm}\left[\sup_{\theta\in\Theta}\left(\mathcal{L}(\theta) - \hat{\mathcal{L}}(\theta)\right)\right] \tag{54}$$

$$\leq \inf_{\theta\in\Theta}\mathcal{L}(\theta) + 4C_{\mathrm{Lip}}\left\{\alpha\mathcal{R}_N(\mathcal{H}) + (1-\alpha)\mathcal{R}_M(\mathcal{H})\right\}. \tag{55}$$

By combining this with Lemma D.2 evaluated at $\theta = \hat{\theta}$, we obtain Theorem 3.1:

$$\mathbb{E}_{\mathcal{D}^{\pm}}\left[\mathbb{E}_{p^{+}}\left[\left(p_{\hat{\theta}} - p^{+}\right)^2\right]\right] \leq \frac{2}{\alpha\mu}\mathbb{E}_{\mathcal{D}^{\pm}}\left[\mathcal{L}(\hat{\theta})\right] \tag{56}$$

$$\leq \frac{2}{\alpha\mu}\left[\inf_{\theta\in\Theta}\mathcal{L}(\theta) + 4C_{\mathrm{Lip}}\left\{\alpha\mathcal{R}_N(\mathcal{H}) + (1-\alpha)\mathcal{R}_M(\mathcal{H})\right\}\right]. \tag{57}$$

E Details of the Convergence-Rate Comparison

Here we derive the condition on α used in Eq. (32). Assuming $\alpha \in (0,1)$ and $\mu > 0$, we have

$$\frac{2}{\alpha\mu} < \frac{2\,(1-\alpha)^2}{\alpha^2 m_+^2 \mu} \Leftrightarrow \alpha m_+^2 < (1-\alpha)^2 \Leftrightarrow \alpha^2 - (2+m_+^2)\alpha + 1 > 0, \tag{58}$$

which yields

$$\alpha < \left(\frac{\sqrt{m_+^2+4} - m_+}{2}\right)^2. \tag{59}$$

Using a Taylor expansion around $m_+ \to 0$, we obtain

$$\left(\frac{\sqrt{m_+^2+4} - m_+}{2}\right)^2 = 1 - m_+ + O(m_+^2), \tag{60}$$

and hence

$$\alpha < 1 - m_+ + O(m_+^2). \tag{61}$$

For any fixed $\alpha \in (0,1)$, Eq. (61) holds for all sufficiently small m_+. Therefore, in the regime $m_+ \ll 1$, the right-hand side of Eq. (32) can be much larger than the left-hand side.

F Canonical Violation of Bradley-Terry Assumption

Under the Bradley-Terry assumption, for two responses y_1 and y_2 to a prompt x, the probability that y_1 is preferred over y_2 is defined as follows:

$$\Pr_\phi(y_1 \succ y_2|x) = \sigma(R_\phi(x,y_1) - R_\phi(x,y_2)), \tag{62}$$

where $R_\phi(x,y)$ denotes the reward model. Now consider a prompt x with three responses y_a, y_b, y_c and assume a cyclic preference pattern:

$$\Pr_\phi(y_a \succ y_b \mid x) = \Pr_\phi(y_b \succ y_c \mid x) = \Pr_\phi(y_c \succ y_a \mid x) = t \neq \tfrac{1}{2}. \tag{63}$$

Under the Bradley-Terry assumption, this would require

$$R_\phi(x,y_a) - R_\phi(x,y_b) = R_\phi(x,y_b) - R_\phi(x,y_c) = R_\phi(x,y_c) - R_\phi(x,y_a). \tag{64}$$

Taking the sum of these three equalities yields 0, and hence each difference must be zero. Therefore, $\Pr(y_a \succ y_b \mid x) = \sigma(0) = 1/2$, contradicting the assumption $t \neq 1/2$.

G Additional Results on BIG-Bench Hard

As additional experiments, we evaluate the aligned models on BIG-Bench Hard (BBH) (Suzgun et al., 2023), which is a benchmark for evaluating the reasoning capabilities of language models. We used the lm-evaluation-harness, where we set the number of fewshot examples to 3.

Tables 3 and 4 show the BBH results on UF-G and MIX-14K, respectively. We used boldface to indicate the best result and any results not significantly different from it according to a pairwise t-test with a significance level of 5%.

RDRO achieves the best or statistically comparable performance across all models and datasets, with the sole exception of Qwen-3B on MIX-14K. This indicates that RDRO performs equal to or better than DDRO across multiple benchmarks. Note that for Qwen-3B, scores are generally low on BBH across all methods and datasets, including the base model before alignment, suggesting that it is not well-suited to this benchmark.

Table 3: Comparison of BIG-Bench Hard (3 shot) on UF-G.

	Qwen-1.5B	Qwen-3B	Llama-3B	Llama-8B
KTO	0.252 ± 0.000	0.000 ± 0.000	0.438 ± 0.000	0.190 ± 0.012
DDRO	0.279 ± 0.001	0.004 ± 0.001	0.531 ± 0.001	0.615 ± 0.018
RDRO	$\mathbf{0.286 \pm 0.000}$	$\mathbf{0.010 \pm 0.000}$	$\mathbf{0.546 \pm 0.001}$	$\mathbf{0.701 \pm 0.000}$

Table 4: Comparison of BIG-Bench Hard (3 shot) on MIX-14K.

	Qwen-1.5B	Qwen-3B	Llama-3B	Llama-8B
KTO	0.243 ± 0.000	$\mathbf{0.017 \pm 0.000}$	0.538 ± 0.000	0.668 ± 0.001
DDRO	0.275 ± 0.001	0.015 ± 0.000	$\mathbf{0.552 \pm 0.001}$	$\mathbf{0.699 \pm 0.003}$
RDRO	$\mathbf{0.284 \pm 0.000}$	0.014 ± 0.000	$\mathbf{0.553 \pm 0.000}$	$\mathbf{0.694 \pm 0.001}$

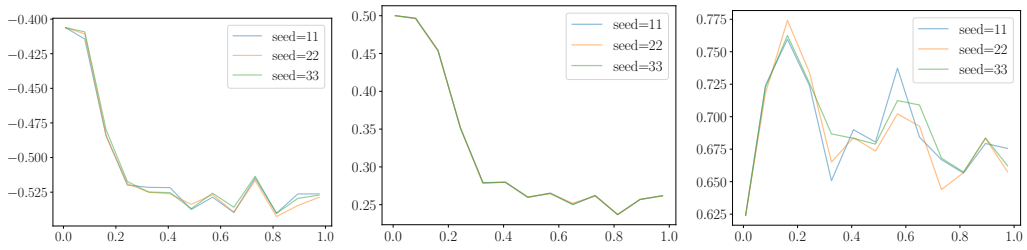
H Empirical Analysis of Training Dynamics

In this section, we present the training dynamics over steps (within an epoch), including training losses (Figure 4), gradient norms (Figure 5), and log-ratio trajectories (Figures 6 and 7). Here, log-ratios are defined as the difference between the log-probabilities of the target policy and the reference policy. Figure 6 shows the log-ratios for preferred data, and Figure 7 shows the log-ratios for non-preferred data. We use Llama-8B on UF-G with $\alpha = 0.39$.

DDRO introduces heuristic stabilization (e.g., the softplus transformation described in Eq. (9)) to prevent gradient explosion, as also observed in Figure 3 of (Higuchi & Suzuki, 2025). As a result, gradient norms appear small and training seems stable. However, these modifications change the original objective and may lead to a mismatch with its theoretical properties. Indeed, DDRO exhibits large changes in both preferred and non-preferred log-ratios, indicating overly large deviations from the reference policy. Prior work has shown that overly large deviations from the reference policy can degrade performance (Kwa et al., 2024), and our observations are consistent with this phenomenon. While one might expect that tuning the KL regularization parameter β could mitigate this issue, Appendix D of (Higuchi & Suzuki, 2025) explicitly states that the KL regularization is not included in the gradient calculation for training stability. This suggests that controlling such deviations within DDRO is non-trivial.

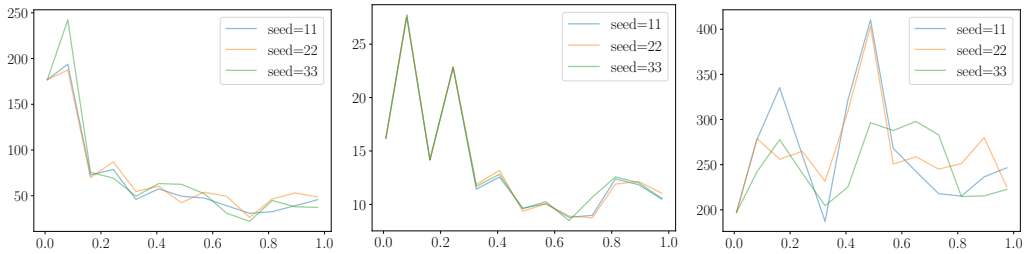
In contrast, RDRO does not rely on heuristic stabilization. The softplus transformation employed in DDRO arises naturally in our formulation, while preserving statistical consistency. As shown in Figure 5, although RDRO may exhibit larger gradient norms than heuristic-stabilized DDRO or KTO, it avoids the gradient explosion observed in the original DDRO formulation. Moreover, RDRO avoids overly aggressive updates in both preferred and non-preferred log-ratios, while consistently improving their separation (log-ratio margin), indicating controlled alignment behavior.

Overall, RDRO may appear less stable if stability is interpreted as having small gradient norms. However, we argue that stability should instead be understood as (i) avoiding gradient explosion and (ii) preventing overly large deviations from the reference policy. Under this notion, RDRO provides more principled and stable optimization behavior, consistent with both our theoretical guarantees and empirical observations.



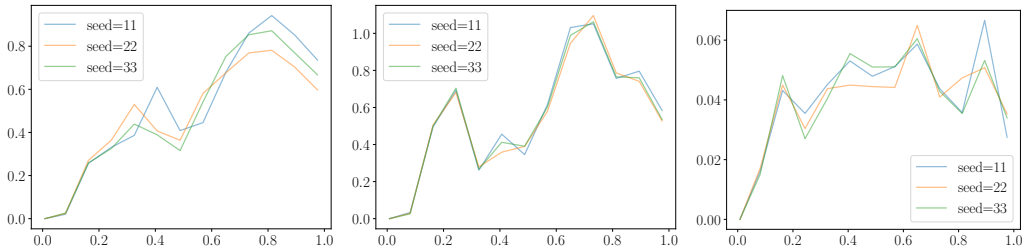
(a) DDRO. (b) KTO. (c) RDRO.

Figure 4: Training losses over steps for DDRO, KTO, and RDRO on Llama-8B with UF-G and $\alpha = 0.39$. Note that the loss values are not directly comparable across methods since the objectives differ.



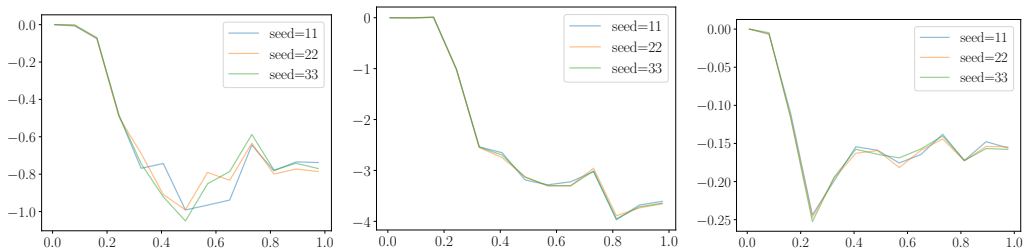
(a) DDRO. (b) KTO. (c) RDRO.

Figure 5: Gradient norms over steps for DDRO, KTO, and RDRO on Llama-8B with UF-G and $\alpha = 0.39$.



(a) DDRO. (b) KTO. (c) RDRO.

Figure 6: Preferred log-ratios over steps for DDRO, KTO, and RDRO on Llama-8B with UF-G and $\alpha = 0.39$.



(a) DDRO. (b) KTO. (c) RDRO.

Figure 7: Non-preferred log-ratios over steps for DDRO, KTO, and RDRO on Llama-8B with UF-G and $\alpha = 0.39$.