DISTORTION OF AI ALIGNMENT REVISITED: RLHF IS A DECENT UTILITARIAN ALIGNER

Kazusato Oko[1,2], **Annie Ulichney**[1], **Nika Haghtalab**[1], **Han Bao**[2,3,4]
[1]University of California, Berkeley
[2]Center for Advanced Intelligence Project, RIKEN
[3]The Institute of Statistical Mathematics and the Graduate University for Advanced Studies
[4]Tohoku University
{oko,annie_ulichney,nika}@berkeley.edu, bao.han@ism.ac.jp

ABSTRACT

While Reinforcement Learning from Human Feedback (RLHF) is the standard paradigm for aligning large language models with human preferences, its effectiveness in pluralistic settings has been called into question. Notably, recent work by Gölz et al. (2025) demonstrated that the *distortion*—defined as the multiplicative gap between the average user utility of the RLHF policy and the optimal average utility—can scale exponentially with the Bradley–Terry temperature parameter β when users have heterogeneous preferences. In this work, we present a fine-grained analysis of the distortion of RLHF with reward clipping and demonstrate that such exponential degradation is not a fundamental property of the algorithm but rather a consequence of distribution mismatch between the distribution generating preference data (μ) and the KL reference policy (π_{ref}). To this end, we establish tight upper and lower bounds on the distortion of RLHF across multiple regimes of the KL regularization strength. We show that in a representative regime, under the Bradley–Terry model, the distortion is $\tilde{\Theta}(\beta B + \beta)$, where B is an upper bound on the log density ratio between μ and π_{ref}. In particular, when there is no distribution mismatch (i.e., $\mu = \pi_{\mathrm{ref}}$), RLHF achieves the optimal distortion of $O(\beta)$ up to a constant. Our results suggest that, to reasonably maximize average utility with RLHF, it is preferable to use on-policy sampled preference data or to fine-tune models on samples from μ prior to RLHF.

1 INTRODUCTION

In the post-training of large language models (LLMs), the alignment problem of matching model outputs to human preferences and ethical values is essential for safe and effective deployment (Bai et al., 2022a; Wang et al., 2023). A representative approach is preference-based, which assumes the existence of an underlying reward function behind human-labeled preference data, estimates this reward, and then fine-tunes the model's policy to maximize it, as exemplified by Reinforcement Learning from Human Feedback (RLHF) (Christiano et al., 2017; Ziegler et al., 2019; Ouyang et al., 2022). This theoretical assumption of an underlying reward prevails even in preference-based methods without an explicit reward model (Ethayarajh et al., 2024; Meng et al., 2024; Huang et al., 2025), including direct preference optimization (DPO) (Rafailov et al., 2023).

Such preference-based policy optimization methods have been criticized for optimizing a single reward function, which may not represent diverse user populations well (Bai et al., 2022b; Chakraborty et al., 2024; Conitzer et al., 2024). Alternatively, the community has begun to acknowledge the existence of diverse users and to model such diversity explicitly (Sorensen et al., 2024). A *utilitarian* framework is one such attempt to assess how much an alignment method satisfies multiple users in terms of their average utility (Siththaranjan et al., 2023; Gölz et al., 2025). To assess the utilitarian performance of alignment methods, the notion called *distortion* is introduced by following social choice theory (Procaccia & Rosenschein, 2006; Boutilier et al., 2012; Anshelevich et al., 2021): the distortion measures how far a given mechanism's average utility falls short of the highest achievable average utility. This framework can be applied to analyzing the utilitarian performance of alignment methods by identifying a utility and a mechanism with a reward and a distribution optimization

Table 1: Comparison of RLHF distortion bounds.

	AI alignment	Social choice
Gölz et al.	$e^{\Omega(\beta)}$ $(B \to \infty)$	$O(\beta^2)$ $\Omega(\beta)$
Ours.	$O(B\beta)$ (Thm. 2) $\Omega(B\beta)$ (Thm. 4)	$O(\beta)$ (Thm. 1)
Boutilier et al.	N/A	$\Omega(m^{\frac{1}{2}})$ $(\beta = \infty)$

$*$ We assume a constant KL budget $\tau = \Theta(1)$ in the AI alignment setting. β is the temperature in the Bradley–Terry model, B is the distribution mismatch, and m is the number of alternatives.

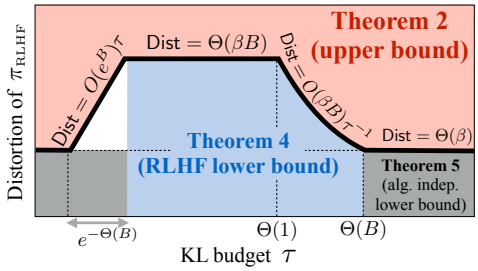

Figure 1: Overview of the upper and lower bounds in the analysis in the AI alignment setting (with the KL constraint), shown for a given distribution mismatch B.

problem under a KL constraint, respectively. Under this setup, Gölz et al. (2025) prove that some distortion is unavoidable due to the nonlinearity of the Bradley–Terry reward model. Specifically, an algorithm-independent lower bound of the distortion is $\Omega(\beta)$ with $\beta > 0$ denoting the Bradley–Terry temperature, which can be achieved by Nash Learning from Human Feedback (Munos et al., 2024). In stark contrast, RLHF suffers from an exponential lower bound of the distortion $e^{\Omega(\beta)}$. Since this is a consequence of the nonlinearity of the BT model and practical reward models often operate in this nonlinear regime as shown later in Section 6, RLHF can significantly amplify the curse of nonlinearity in theory.

Although the distortion theory suggests a potentially catastrophic failure mode of RLHF, it has been widely applied in post-training of many LLMs such as GPT-4 (Achiam et al., 2023), where RLHF does not exhibit extremely poor empirical performance (Touvron et al., 2023; Georgiev et al., 2024). To reconcile this gap between theory and practice, this paper refines the distortion theory by raising the following questions:

Under what conditions can the distortion of RLHF be reasonably controlled? Conversely, is the remaining distortion fundamentally unavoidable?

1.1 Our Contributions

We show that distortion can be controlled by the mismatch between the reference policy π_{ref} of the KL constraint and the distribution μ from which preference data are sampled. Concretely, defining the maximum log density ratio $B = \left\| \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi_{\mathrm{ref}}} \right\|_\infty$ and the temperature parameter of the Bradley–Terry model β, we show that the worst-case distortion scales as $\tilde{\Theta}(B\beta + \beta)$. This suggests that, unless there is extreme distribution mismatch, RLHF can reasonably solve the problem of maximizing the average underlying utility. We summarize our contributions below, compare bounds across settings and prior work in Table 1, and discuss practical implications in Section 7. See Appendix A for literature review.

- We begin our analysis of RLHF by considering a special case where no KL constraint is imposed on the policy space. This corresponds to the traditional social choice setting, where the mechanism selects the alternative that maximizes the estimated reward. We show that the distortion of RLHF is bounded by $O(\beta)$. This improves upon the $O(\beta^2)$ bound of Gölz et al. (2025). The proof introduces *effective utility*, which retains only the components of the utility that provide informative signals to the RLHF rewards, and it serves as the foundation for the general case.
- Section 4 generalizes the analysis to the AI alignment setting, where the policy is subject to a KL constraint $\mathrm{KL}(\pi \| \pi_{\mathrm{ref}}) \leq \tau$. This reflects that the policy is usually regularized to remain close to the base model, whose distribution we denote by π_{ref}, and this constraint is known to result in an interesting increase in distortion (Gölz et al., 2025). We show that RLHF with reward clipping yields distortion that varies from $O(\beta)$ to $O(B\beta)$ depending on the interplay between the KL budget τ and the distribution mismatch B; see Figure 1. Also, we show that fine-tuning base models with samples from μ prior to RLHF, so that π_{ref} is close to μ, mitigates the effect of distribution mismatch.

- In Section 5, we establish matching lower bounds, up to logarithmic factors, across all ranges of the KL budget τ. These results decompose into an RLHF-specific (Theorem 4) and an algorithm-independent lower bound (Theorem 5); see the blue and gray shades, respectively, in Figure 1. For the former, a distortion of $\tilde{\Omega}(B\beta)$ persists even when the KL budget τ is exponentially small, i.e., $\tau = e^{-\Theta(B)}$, demonstrating the fundamental prevalence of RLHF distortion caused by distribution mismatch.
- To help interpret the theoretical results, we present two experiments in Section 6: one examining the practical impact of the BT model's nonlinear regime (Section 6.1), and another illustrating how mismatch between π_{ref} and μ induces distortion (Section 6.2).

2 Problem Setting

Our problem setting follows that of Gölz et al. (2025), except that we define distortion with respect to individual distributions, and we apply reward clipping in Section 4. Let $A = \{1, \dots, m\}$ denote a finite set of alternatives, which corresponds in alignment scenarios to LLM candidate completions, or groups of semantically equivalent completions. They also correspond to candidate LLMs in AI leaderboard scenarios, as explained in Section 3. Each user is associated with a utility vector $u = (u(1), \dots, u(m))$, whose entries satisfy $0 \leq u(x) \leq 1$ and represent the user's utility for alternative x, and we denote the distribution of the utility vectors by $\mathcal{D}$. Our goal is to find a distribution $\pi \in \Delta(A)$ that maximizes the average utility $\mathbb{E}_{u\sim\mathcal{D},\, x\sim\pi}[u(x)]$ over users, with or without a KL constraint. When a KL constraint is (is not, resp.) imposed, we call the problem *AI alignment (social choice, resp.)* setting.

Generation of preference data. Let $\mu \in \Delta(A)$ be a distribution which selects alternatives to be compared and satisfies $\mu(x) > 0$ for all $x \in A$. We sample n users' utility vectors $u^1, \dots, u^n \sim_{\text{i.i.d.}} \mathcal{D}$. For each user i, we draw a pair of alternatives $x^i, y^i \sim_{\text{i.i.d.}} \mu$. Then, user i compares x^i and y^i via the Bradley–Terry (BT) model. The probability that the user i prefers x^i over y^i (denoted by $x^i \succ y^i$) is:

$$p(x^i \succ y^i) = \sigma\big(\beta(u^i(x^i) - u^i(y^i))\big),$$

where $\sigma(t) = 1/(1+e^{-t})$ is the sigmoid function and $\beta \geq 1$ is the temperature parameter controlling preference sharpness; otherwise, i prefers y^i over x^i.

Reward estimation. The standard RLHF proceeds in two stages (Ouyang et al., 2022). First, a reward model is trained using human-labeled preference data $\{x^i \cdot^i y^i\}_{i=1}^n$ ($\cdot^i \in \{\succ, \prec\}$). Specifically, the preference data are assumed to be generated by a *single* user, and the reward model $\bar{r}$ is learned via maximum likelihood estimation of a single BT model:

$$\max_{\tilde{r}\in\mathbb{R}^m} \sum_{i=1}^{n} \Big[\mathbb{1}[x^i \succ y^i] \log(\sigma(\tilde{r}(x^i) - \tilde{r}(y^i))) + \mathbb{1}[x^i \prec y^i] \log(\sigma(\tilde{r}(y^i) - \tilde{r}(x^i)))\Big].$$

We consider the regime where n is large such that the empirical loss in (1) converges to the population loss. Specifically,

$$\begin{aligned}\bar{r} &:= \arg\max_{\tilde{r}\in\mathbb{R}^m} \mathbb{E}_{u\sim\mathcal{D},x,y\sim\mu}\Big[\mathbb{1}[x \succ y] \log(\sigma(\tilde{r}(x) - \tilde{r}(y))) + \mathbb{1}[x \prec y] \log(\sigma(\tilde{r}(y) - \tilde{r}(x)))\Big] \\ &= \arg\max_{\tilde{r}\in\mathbb{R}^m} \mathbb{E}_{u\sim\mathcal{D},x,y\sim\mu}\big[\sigma(u(x) - u(y)) \log(\sigma(\tilde{r}(x) - \tilde{r}(y)))\big]. \qquad (1)\end{aligned}$$

If utilities are constant across users, $\bar{r}$ recovers βu up to an additive constant as $n \to \infty$. However, when u is stochastic, $\bar{r}$ cannot exactly recover the average utility across users $\mathbb{E}_{u\sim\mathcal{D}}[u]$ (Gölz et al., 2025).

Equation (1) admits an additive constant shift in $\bar{r}(x)$. We fix the translational degree of freedom, for convenience, by imposing the constraint that a virtual alternative x with utility $u(x) \equiv 0$ and infinitesimal sampling probability $\mu(x)$ satisfies $\bar{r}(x) = 0$. Formally, we impose the following constraint:

$$\mathbb{E}_{x\sim\mu}\big[\sigma(-\bar{r}(x))\big] = \mathbb{E}_{x\sim\mu,u\sim\mathcal{D}}\big[\sigma(-\beta u(x))\big]. \qquad (2)$$

Appendix B.7.1 shows that this is indeed equivalent to considering such a virtual alternative.

Reward clipping. To concentrate the reward on a range of values that is informative as a learning signal, we introduce reward clipping only in the AI alignment setting. Fix constants $r_{\min}$ and $r_{\max}$ (specified in Section 4), and truncate the learned reward by

$$r(x) = \max\{\min\{\bar{r}(x), r_{\max}\}, r_{\min}\}. \tag{3}$$

While this is a technical device for obtaining the optimal rate, in Appendix C we remove reward clipping and show an $O(\beta^2)$ distortion when $\mu = \pi_{\mathrm{ref}}$.

Proximal policy optimization (PPO). In the second stage of RLHF, it optimizes a policy with respect to the learned reward. We analyze two distinct settings:

(i) AI alignment setting. LLMs initialized from pre-trained models are fine-tuned to maximize reward. During optimization, the generation policy is typically regularized to stay close to the pre-trained model, which serves as the reference policy (Schulman et al., 2017). Prior work (Gölz et al., 2025) shows that this KL constraint may lead to an increase in distortion. Formally, given a reference policy $\pi_{\mathrm{ref}} \in \Delta(A)$ and a KL budget $\tau > 0$, define

$$\pi_{\mathrm{RLHF}} = \arg \max_{\pi\colon \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau} \mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[r(x)]. \tag{4}$$

If the maximizer is not unique, we may select any distribution among the maximizers.

In the following analysis, we show that large distortion arises from a mismatch between π_{ref} and μ, through its interaction with the KL budget τ. Among several ways to define distribution mismatch, we use the maximum log probability ratio. Although this has the drawback of being sensitive to perturbations, we choose it for its simplicity, which facilitates analysis and helps clarify the effect of distribution mismatch on distortion.

Assumption 1 (Distribution mismatch)**.** *The log-likelihood ratio between π_{ref} and μ is uniformly bounded:* $\max_{x\in A} \log \frac{\mu(x)}{\pi_{\mathrm{ref}}(x)} = B < \infty$.

We note that the distortion upper bounds require only the upper bound on $\log \frac{\mu(x)}{\pi_{\mathrm{ref}}(x)}$, whereas assuming a bound on its absolute value does not affect the lower bounds.

(ii) Social choice setting. In this setting, we may output any distribution maximizing $\mathbb{E}_{u\sim\mathcal{D},\,x\sim\pi}[r(x)]$ (with no KL constraint). This corresponds to a traditional formulation in social choice theory and corresponds to the limit $\tau \to \infty$ in (4). We do not need reward clipping in this setting, so we set $r_{\min} = -\infty$ and $r_{\max} = \infty$.

Distortion. In the **(i) AI alignment setting**, for any $\pi \in \Delta(A)$ we define distortion as the ratio between the best achievable average utility within the KL ball and the average utility achieved by π:

$$\mathsf{Dist}(\pi) = \frac{\max_{\pi'\colon \mathrm{KL}(\pi'\|\pi_{\mathrm{ref}})\le\tau} \mathbb{E}_{u\sim\mathcal{D},x\sim\pi'}[u(x)]}{\mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[u(x)]}.$$

In the **(ii) social choice setting**, which corresponds to the case $\tau = \infty$, the numerator is replaced with $\max_{x\in A} \mathbb{E}_{u\sim\mathcal{D}}[u(x)]$ (Procaccia & Rosenschein, 2006).

3 Warm-up: Social Choice Setting

We begin by analyzing RLHF in the special case where no KL constraint is imposed. This setting is equivalent to analyzing the Borda winner in social choice theory and has implications for the reliability of AI leaderboards. Moreover, our proof strategy here forms the basis for the AI alignment setting with a KL constraint.

In this setting, the policy π_{RLHF} concentrates on the alternatives with the maximum rewards. Given that the reward ranking is consistent with the ordering of the Borda scores $\mathsf{Borda}(x) = \mathbb{E}_{u\sim\mathcal{D},y\sim\mu}[p(x \succ y)]$ under the Bradley–Terry model (see Lemma 18 borrowed from Siththaranjan et al. (2023, Theorem 3.1)), the alternatives with the highest rewards can also be interpreted as the Borda winners, i.e., the maximizers of $\mathsf{Borda}(x)$. The distortion of π_{RLHF} is also the distortion of the Borda winners, i.e., the maximizers of $\mathsf{Borda}(x)$. In deference to the long history of the Borda voting rule in social choice theory (Procaccia & Rosenschein, 2006; Boutilier et al., 2012; Anshelevich et al., 2021), we present π_{RLHF} as π_{Borda} in the social choice setting.

Theorem 1. *In the social choice setting, no reward clipping is imposed, and we may choose any policy maximizing $\mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[\bar{r}(x)]$ as π_{Borda}. Then, the distortion of the policy π_{RLHF} is bounded by*

$$\mathsf{Dist}(\pi_{\mathrm{Borda}}) \leq C_1\beta + 4,$$

where C_1 is an absolute constant.

Before proceeding to the proof overview, we discuss an implication of this result for AI leaderboards:

AI leaderboards are not too distorted. Our results have implications for AI leaderboards such as Chatbot Arena (Chiang et al., 2024), which present users with responses from two anonymized models and ask them to select the preferred one. Based on the collected data, these leaderboards assume that each model has a deterministic utility and fit a single BT model to rank the models. This setting can be naturally viewed through the lens of social choice, with models viewed as alternatives. In particular, our results characterize how suboptimal the top-ranked model on a leaderboard can be relative to the true maximizer of average utility. Our improved bound in Theorem 1 shows that this distortion is at most a constant factor larger than the algorithm-independent lower bound (Theorem 5).

3.1 Proof Overview of Theorem 1

3.1.1 Effective Utility

We outline the proof of Theorem 1. First, we introduce the *effective utility*, which extracts the utilities that actually contribute to rewards. Formally, we define the effective utility $\hat{u} : A \to [0, c\beta^{-1}]$ as

$$\hat{u}(x) = \begin{cases} 0 & \text{(i) if } \mathbb{P}_{y\sim\mu}[u(y) - u(x) > c\beta^{-1}] \geq \frac{1}{2}, \\ c\beta^{-1} & \text{(ii) else if } u(x) > c\beta^{-1}, \\ u(x) & \text{(iii) otherwise} \end{cases}$$

for a constant c with $0 < c \leq \frac{1}{316}$. The reduction is applied to the true utility u under the two cases (i) and (ii), where the preference data become uninformative to recover the true utility u. We describe these two cases below.

(i) Signal degradation in pairwise comparisons. Even when an alternative x has high utility $u(x)$, it may still lose to a majority of opponents under μ (e.g., if μ concentrates on alternatives with higher utility). In this case, pairwise outcomes provide little evidence that x has high utility, so a reward model fit to comparisons will accidentally treat such an x as low utility. Thus, we set $\hat{u}(x) = 0$.

(ii) Underweighting of large utilities. When $u(x) - u(y)$ is large, the win probability $\sigma(\beta(u(x) - u(y)))$ is close to 1. However, further increases in $u(x)$ have negligible impact on the observed comparisons and thus on the rewards. This prevents the reward from distinguishing well between large utilities. This motivates capping the utility at a threshold $c\beta^{-1}$ for a fixed constant $c > 0$.

Then, we have the following relationship between the true utility and the effective utility. The proof is found in Appendix B.1.

Lemma 1. *The maximal average utility can be bounded by the maximal average effective utility as follows:*

$$\max_{\pi\in\Delta(A)} \mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[u(x)] \lesssim \beta \max_{\pi} \mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[\hat{u}(x)].$$

3.1.2 Reward Sandwich

Next, we "sandwich" the learned reward $\bar{r}$ with the effective utility and true utility. Specifically, we can lower bound the reward with the effective utility and upper bound the reward with the true utility as follows. The proofs for the following lemmas can be found in Appendices B.2 and B.3, respectively.

Lemma 2. *For any $x \in A$, the expectation of the effective utility can lower bound the upper-clipped reward as*

$$\mathbb{E}_{u\sim\mathcal{D}}[\hat{u}(x)] \lesssim \beta^{-1} \min\{\bar{r}(x), R\}, \tag{5}$$

where R is some problem dependent constant.

Lemma 3. *Suppose that $\mathbb{P}_{u\sim\mathcal{D},x\sim\mu}[\beta u(x) > c] \le c^2$ and let R be the same constant as in Lemma 5. Then, for any $x \in A$, we have that*

$$\min\{\bar{r}(x), R\} \lesssim \beta\mathbb{E}_{u\sim\mathcal{D}}[u(x)].$$

Putting the above together, we obtain Theorem 1 if $\mathbb{P}_{u\sim\mathcal{D},x\sim\mu}[\beta u(x) > c] \le c^2$. According to Lemmas 1 and 2,

$$\max_{\pi\in\Delta(A)} \mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[u(x)] \lesssim \max_{\pi\in\Delta(A)} \mathbb{E}_{x\sim\pi}[\min\{\bar{r}(x), R\}]. \tag{6}$$

When $\pi_{\mathrm{RLHF}} = \max_{\pi\in\Delta(A)} \mathbb{E}_{x\sim\pi}[\bar{r}(x)]$, π_{RLHF} is a maximizer of RHS of (6). By using this fact together with Lemma 3 to (6), we obtain that

$$\max_{\pi\in\Delta(A)} \mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[u(x)] \lesssim \beta\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{Borda}}}[u(x)],$$

which proves $\mathsf{Dist}(\pi_{\mathrm{Borda}}) \lesssim \beta$. See Appendix B for the details of this argument.

The case $\mathbb{P}_{u\sim\mathcal{D},x\sim\mu}[\beta u(x) > c] > c^2$ requires a separate argument. In this case, we prove the result by observing that μ itself achieves an average utility of $\Omega(\beta^{-1})$. See Appendix B.5 for the details.

4 AI Alignment Setting

We now turn to the AI alignment setting, where we incorporate the KL constraint into the reward estimation (4). The main theorem in this section is the following.

Theorem 2. *Suppose that the mismatch between μ and π_{ref} satisfies Assumption 1, and let the KL budget be $\tau > 0$. To define reward clipping (3), we take $r_{\min}$ to be the solution to the following equation*

$$\mathbb{E}_{y\sim\mu}\big[\sigma(r_{\min} - \bar{r}(y))\big] = \frac{1}{2} - \frac{c^3}{16}, \tag{7}$$

and set $r_{\max} = r_{\min} + 2c$, where c is an arbitrary constant satisfying $0 < c \le \frac{1}{316}$. Then, the distortion of π_{RLHF} satisfies

$$\mathsf{Dist}(\pi_{\mathrm{RLHF}}) \le C_2\big(\min\big\{e^B\tau, B, B\tau^{-1}\big\} + 1\big)\beta + 4.$$

Here, $C_2 > 0$ is a constant polynomially depending on c^{-1}.

In the following, we take the constant c in reward clipping to be the same as that in the definition of the effective utility in Section 3.1.1. In Appendix C, we consider removing reward clipping and in particular present bounds for the case $\mu = \pi_{\mathrm{ref}}$. We make the following remarks about this theorem.

Optimality of RLHF for utility maximization. This result addresses an open problem raised by Gölz et al. (2025), who introduced distortion into the AI alignment setting. They did not establish an upper bound in the AI alignment setting, and in particular left obtaining tighter bounds under the assumption $\mu = \pi_{\mathrm{ref}}$ as future work. Theorem 2 shows that when $\mu = \pi_{\mathrm{ref}} \Leftrightarrow B = 0$, the distortion upper bound is $O(\beta)$, which matches the algorithm-independent lower bound up to a constant factor. Consequently, our results provide an answer to this open question by establishing the RLHF optimality under the distribution matching scenario. It is surprising that RLHF implicitly achieves optimal average utility maximization, despite the fact that reward estimation with a single BT model (1) does not account for the stochasticity in utilities.

On the source of RLHF distortion. Under the distribution mismatching scenario ($B > 0$), the distortion can grow up to $O(B\beta)$, which is tight up to logarithmic factors (see Theorem 4 and Figure 1). That is, RLHF distortion is caused multiplicatively by nonlinearity of the BT model (β) and the distribution mismatch (B). While Gölz et al. (2025, Theorem 6) show that RLHF distortion can scale as $e^{\Omega(\beta)}$, this behavior arises in the limit of distributions for which $B \to \infty$. In contrast, we quantify distribution mismatch explicitly and show that distortion is controlled linearly in the log density ratio B. In particular, exponential distortion requires the density ratio to be doubly exponential in β. Overall, our results indicate that distribution mismatch drives RLHF distortion, but its effect is comparatively mild unless mismatch is significant.

A benefit of fine-tuning on samples from μ. To mitigate the effect of distribution mismatch for off-policy sampled preference data, one might consider bringing the reference policy closer to the preference data distribution. While this reduces distortion, it also changes the maximum average utility in the definition of distortion, and thus it is unclear whether the average utility is improved compared to the original π_{RLHF}. Nevertheless, the following corollary, which can be obtained with one additional step beyond Theorem 2, shows that the worst-case ratio relative to the maximum average utility under the original reference policy is in fact improved by that. See Appendix B.6 for the proof.

Corollary 3. *Let a base model π_{base} be given, and suppose that $\max_{x\in A}\log\frac{\mu(x)}{\pi_{\mathrm{base}}(x)}\le B$. Define $\pi_{\mathrm{ref}}=(1-e^{-\lambda})\pi_{\mathrm{base}}+e^{-\lambda}\mu$ ($\lambda>0$), and perform RLHF as in Theorem 2. Then, the resulting policy π_{RLHF} satisfies*

$$\frac{\max_{\pi\colon \mathrm{KL}(\pi\|\pi_{\mathrm{base}})\le\tau}\mathbb{E}_{u,x\sim\pi}[u(x)]}{\mathbb{E}_{u,x\sim\pi_{\mathrm{RLHF}}}[u(x)]}\le\frac{\beta C_2(\lambda+1)+4}{1-e^{-\lambda}}.$$

According to the distortion bound in Theorem 2, the ratio of the average utility of the original π_{RLHF} relative to $\max_{\pi\colon \mathrm{KL}(\pi\|\pi_{\mathrm{base}})\le\tau}\mathbb{E}_{u,x\sim\pi}[u(x)]$ is $O(\beta B)$ when $Be^{-B}\le\tau\le 1$. Therefore, making the reference policy closer to the preference data distribution so that $\lambda\lesssim B$ improves the ratio relative to $\max_{\pi\colon \mathrm{KL}(\pi\|\pi_{\mathrm{base}})\le\tau}\mathbb{E}_{u,x\sim\pi}[u(x)]$. This implies that fine-tuning the model based on the samples from the preference data distribution prior to RLHF can mitigate the effect of distribution mismatch.

4.1 Proof Overview of Theorem 2

We prove Theorem 2 by extending the proof in the social choice setting. Focusing on the case where $\mathbb{P}_{u\sim\mathcal{D},x\sim\mu}[\beta u(x)>c]\le c^2$, the proof in the social choice setting was obtained by combining Lemmas 1, 2, and 3. Among these, Lemmas 2 and 3 do not depend on π_{ref}, and therefore can be used without modification, so the main difficulty lies in Lemma 1. When the KL divergence constraint prevents π from ranging over the entire simplex $\Delta(A)$, the lemma is updated as follows.

Lemma 4. *Under Assumption 1, the maximal average utility can be bounded by the maximal average effective utility as follows:*

$$\begin{aligned}&\max_{\pi\colon \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau}\mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[u(x)]\\&\lesssim\beta\Big(\min\Big\{e^{B}\tau,B,\frac{B}{\tau}\Big\}+1\Big)\max_{\pi\colon \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau}\mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[\hat{u}(x)]+\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{RLHF}}}[u(x)].\end{aligned}\tag{8}$$

The proof is found in Appendix B.1. Intuitively, the additional factor of $\min\{e^B\tau,B,\frac{B}{\tau}\}$ appears because the KL constraint can restrict a mass to alternatives where the gap between the effective utility and the true utility is large.

More precisely, we need to consider two scenarios in which reduction is applied: **(i) Signal degradation in pairwise comparisons**, corresponding to $\mathbb{P}_{y\sim\mu}\big[u(y)-u(x)\ge c\beta^{-1}\big]\ge\frac{1}{2}$, i.e., when x loses to at least half of the alternatives drawn from μ, so that x is treated as a low-utility alternative; and **(ii) Underweighting of large utilities**, corresponding to $u(x)>c\beta^{-1}$, where the sigmoid function $\sigma(\beta(u(x)-u(y)))$ may saturate, making large utilities difficult to distinguish.

Accounting for the second effect is straightforward. Since reduction in the second case scales the utility $u(x)$ by at most $c\beta^{-1}$, as far as this effect alone is concerned, the maximum average utility is reduced by at most a factor of $c\beta^{-1}$.

The first effect is, however, more subtle, since reducing $u(x)>0$ to $\hat{u}(x)=0$ results in an unbounded ratio $\frac{u(x)}{\hat{u}(x)}$. Instead, we must compare the maximum expectations under policy distributions. Letting $I_1(u,x)$ be the indicator of the event $\mathbb{P}_{y\sim\mu}\big[u(y)-u(x)\ge c\beta^{-1}\big]\ge\frac{1}{2}$, we would like to account for the maximum loss caused by this reduction, that is, $\max_{\pi\colon \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau}\mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[I_1(u,x)u(x)]$. If $\mu=\pi_{\mathrm{ref}}$, we have the following bound:

$$\begin{aligned}\max_{\pi\colon \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau}\mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[I_1(u,x)u(x)]&\le 2c^{-1}\beta\mathbb{E}_{\hat{u},x\sim\mu}[\hat{u}(x)]&(9)\\&\le 2c^{-1}\beta\max_{\pi\colon \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau}\mathbb{E}_{\hat{u},x\sim\pi}[\hat{u}(x)].&(10)\end{aligned}$$

The first inequality shows that the data distribution μ is a "good" policy that absorbs the loss caused by the reduction, and directly follows from combining (14) and (15) in the proof of Lemma 5. Then the second inequality is obtained simply by choosing $\pi = \mu$ in (10), as $\mu = \pi_{\mathrm{ref}}$ trivially satisfies the KL constraint.

However, under distribution mismatch ($\pi_{\mathrm{ref}} \neq \mu$), μ may not satisfy the KL constraint with respect to π_{ref}, and the second inequality may break down. To overcome this issue, we consider an interpolating distribution between μ and π_{ref}, defined by $\pi' = \lambda\mu + (1-\lambda)\pi_{\mathrm{ref}}$. By taking $\lambda = B\tau^{-1}$, the distribution π' lies in the KL ball, as shown in Lemma 6. With this π', the inequality (10) is valid under the general case $\pi_{\mathrm{ref}} \neq \mu$. Specifically, since $\pi'(x) \geq \lambda\mu(x)$ holds for all $x \in A$, it follows that $\mathbb{E}_{\hat{u},y\sim\mu}[\hat{u}(y)] \leq \lambda^{-1}\mathbb{E}_{\hat{u},y\sim\pi'}[\hat{u}(y)]$, and therefore

$$(9) \leq \lambda^{-1} \times 2c^{-1}\beta\mathbb{E}_{\hat{u},y\sim\pi'}[\hat{u}(y)] \leq \lambda^{-1} \times 2c^{-1}\beta \max_{\pi\colon \mathrm{KL}(\pi\|\mu)\leq\tau} \mathbb{E}_{\hat{u},x\sim\pi}[\hat{u}(x)].$$

This is given as (17) in the proof of Lemma 5. Here, we have an additional factor of $\lambda^{-1} = B\tau^{-1}$ compared to the case $\pi_{\mathrm{ref}} = \mu$.

Still, for small $\tau \leq 1$, any feasible distribution π may lie far from μ, and the above argument leads to a blow-up of distortion. To handle this regime, we exploit the fact that, when τ is small, any policy π satisfying $\mathrm{KL}(\pi\|\pi_{\mathrm{ref}}) \leq \tau$ must lie close to the reference policy π_{ref} in the total variation distance. However, applying the standard Pinsker's inequality only yields an $O(\sqrt{\tau})$ bound on the total variation distance, which is insufficient to obtain the tight upper bound. Instead, we use Lemma 7, which provides a linear bound on the total variation restricted to regions where the density ratio is large. Using this lemma, we obtain bounds involving the factors B and $e^B\tau$. The effect from the regions where the density ratio is close to 1 is handled by the additive term $\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{RLHF}}}[u(x)]$ in (8).

5 Lower Bounds

The upper bounds of Theorem 2 are tight up to logarithmic factors. Specifically, we present two lower bounds corresponding to different ranges of τ. The RLHF-specific lower bounds for $e^{-\Theta(B)} \leq \tau \leq B$ are given in Section 5.1, while an algorithm-independent lower bound for the remaining cases is presented in Section 5.2.

5.1 Lower Bound for RLHF

We first show that the increase in RLHF distortion due to distribution mismatch is tight up to logarithmic factors. This bound holds regardless of the choice of $r_{\min}$ and $r_{\max}$.

Theorem 4. *Assume that the KL budget $\tau > 0$, the Bradley–Terry model temperature parameter $\beta \geq 3$, and the maximum log-likelihood ratio $B = \max_{x\in A} \left|\log \frac{\mu(x)}{\pi_{\mathrm{ref}}(x)}\right|$ satisfy*

$$\max\{e^{-\frac{B}{3}}, e^{-\frac{\beta}{2}}\} \leq \tau \leq B, \quad 3 \leq \beta \leq e^{\frac{B}{3}} - 1, \quad \textit{and} \quad 3 \leq B \leq \min\{e^{\frac{B}{3}}, e^{\frac{\beta}{2}}\}.$$

Then, there exists a pair of an instance $\mathcal{D}$, a data distribution μ, and a reference policy π_{ref} such that, regardless of how $r_{\min}$ and $r_{\max}$ are chosen in reward clipping, the distortion of π_{RLHF} is lower bounded by

$$\mathsf{Dist}(\pi_{\mathrm{RLHF}}) \gtrsim \min\left\{\frac{B}{\log B\beta}, \frac{B}{\tau}\right\}\beta.$$

The proof is found in Appendix D.1. An implication from this lower bound is that a distortion of $\tilde{\Theta}(B\beta)$ is unavoidable even under an exponentially small KL budget $\tau = e^{-\Theta(B)}$. This shows that distortion arising from distribution mismatch is persistent even when LLM updates are restricted to small-scale fine-tuning. Also, we note that this lower bound holds regardless of how $r_{\min}$ and $r_{\max}$ are chosen in reward clipping.

5.2 Algorithm Independent Lower Bound

To cover all regimes of the KL budget τ, we establish an $\Omega(\beta)$ lower bound for all τ in the case $\pi_{\mathrm{ref}} = \mu$, i.e., $B = 0$. We note that, while Gölz et al. (2025, Theorem 3) also prove a similar lower

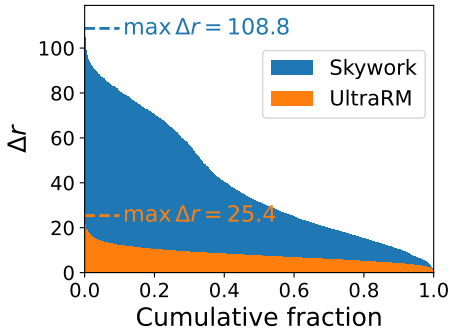

Figure 2: Empirical distribution of pairwise reward differences.

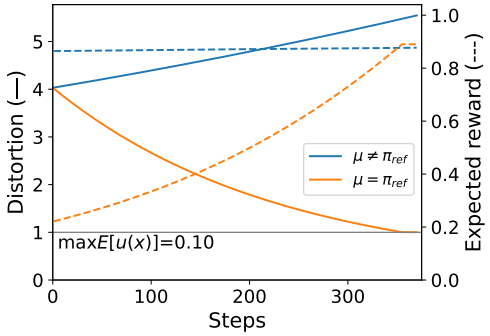

Figure 3: Distortion with different reference policies.

bound, it applies to the social choice setting in which the alternative with the maximum reward is selected, and thus extends to the AI alignment setting ($\tau < \infty$) only for sufficiently large τ, rather than for all τ.[1] The proof of this theorem can be found in Appendix D.2.

Theorem 5. *When $\mu = \pi_{\mathrm{ref}}$, for any KL budget $\tau > 0$ and any Bradley–Terry temperature parameter $\beta > 0$, there exist a collection of instances $\{\mathcal{D}_i\}_{i=1}^N$ and a data distribution μ such that, when an instance is drawn uniformly at random from $\{\mathcal{D}_i\}_{i=1}^N$, the output of any algorithm incurs expected distortion at least*

$$\frac{\beta}{2}\frac{1+e^{-\beta}}{1-e^{-\beta}} - \epsilon,$$

where $\epsilon > 0$ is an arbitrarily small constant.

6 Experiments

6.1 Reward Scale in Practical Reward Models

In the theoretical analysis, larger values of the temperature parameter β induce stronger nonlinearity in the BT model, leading to larger distortion. While the effect of this nonlinearity cannot be observed directly, since β upper bounds the maximum difference between rewards $\max_{x\in A} \bar{r}(x) - \min_{x\in A} \bar{r}(x)$ by Lemma 20, we visualize the reward scale of open-weight reward models to obtain rough estimates of β and to see how much the nonlinear regime of the BT model matters in practice.

From RewardBench (Lambert et al., 2025), we selected two open-weight reward models: Skywork-Reward-V2-Llama-3.1-8B (Liu et al., 2024) and UltraRM-13B (Cui et al., 2023). The former model was trained with the original MLE loss (1), while the latter added a regularization m with $|m| \le 1$ to $r(x) - r(y)$. Using their training datasets, Skywork-Reward-Preference-80K-v0.1 (Liu et al., 2024) and UltraFeedback (Cui et al., 2023), we sampled 5000 preference instances and plotted the values of $\Delta r \coloneqq |r(x \mid z) - r(y \mid z)|$, where z denotes the prompt and x, y denote completions. See Appendix E.1 for details.

The results are shown in Figure 2. The maximum Δr is 108.8 for Skywork and 25.4 for UltraRM. While they do not exactly correspond to the effective value of β due to optimization errors, they nevertheless mean that reward models in practice operate in a regime where the MLE loss (1) exhibits substantial nonlinearity. This implies that practical reward models can have non-negligible distortion.

[1] We remark that Gölz et al. (2025) report a $(\frac{1}{2} + o(1))\beta$ lower bound for both the social choice setting and the AI alignment setting in their Table 1. However, both of these refer to Gölz et al. (2025, Theorem 3), which considers the setting in which a single alternative is selected, namely the social choice setting. Indeed, in their construction, as $\tau \to 0$, any π satisfying $\mathrm{KL}(\pi \| \pi_{\mathrm{ref}}) \le \tau$ can no longer allocate mass to the alternative a with exceptionally high utility, and since the remaining alternatives also have positive average utility, the distortion converges to 1.

6.2 SYNTHETIC EXPERIMENT

Next, we present a toy example in which distortion arises from distribution mismatch between μ and π_{ref}. The setup is detailed in Appendix E.2, with $\beta = 10$ and $B = 11.5$ in particular. For reward computation, we consider two cases for generating preference data: one using μ, and another using the reference policy π_{ref} in place of μ, which results in $B = 1$. In both cases, we optimized a distribution initialized at π_{ref}, using mirror descent (Beck & Teboulle, 2003) with step size $\eta = 10^{-3}$.

Figure 3 shows the evolution of the distortion and the average reward (the latter differs between the two cases). When π_{ref} is used for preference data generation (denoted by $\mu = \pi_{\mathrm{ref}}$), the distortion converges to 1, whereas when μ is used ($\mu \neq \pi_{\mathrm{ref}}$), the distortion grows as optimization proceeds, corresponding to a decrease in average utility. This succinctly illustrates that, in RLHF, distribution mismatch between μ and π_{ref} is a key driver of increased distortion.

7 DISCUSSION AND PRACTICAL IMPLICATIONS

In this work, we revisited the question of whether RLHF is well-suited to aggregating diverse human preferences. By isolating the effect of distribution mismatch B, we proved that the RLHF distortion varies from $\Theta(\beta)$ to $\tilde{\Theta}(B\beta)$ depending on the KL budget τ, with the matching upper and lower bounds. This decomposes the sources of RLHF distortion into two components: the nonlinearity of the preference data generation model (β), which is unavoidable for any algorithm, and the distribution mismatch between the reference policy and the data distribution (B). This shows that misspecification in reward estimation, namely the use of a single Bradley–Terry model, does not by itself worsen distortion as a fundamental property of the algorithm. While the effect of distribution mismatch can persist even under a small KL budget, its dependence on the mismatch is upper bounded by the logarithm of the density ratio, making it comparatively moderate and ruling out the pessimistic exponential bound from Gölz et al. (2025) unless the mismatch is extreme.

Our results suggest several practical implications. First, on-policy data collection (Ouyang et al., 2022) or online variants of RLHF (Xiong et al., 2024; Zhang et al., 2024; Guo et al., 2024; Xie et al., 2025) may offer advantages for reducing distortion over offline methods. However, in practice, we are often faced with preference data that is not collected on-policy. For example, the community has recently been making efforts to construct more heterogeneous public datasets that better reflect a broader range of users (Zhang et al., 2026), and it is common to use data generated by previous models when training newer ones (Ettinger et al., 2025). In such cases, if sufficiently many samples from μ are available, Corollary 3 suggests that fine-tuning on samples from the same distribution prior to RLHF can mitigate distribution mismatch and improve the expected utility. Moreover, in addition to fine-tuning on samples from μ, our results naturally point to two research directions on how to handle off-policy preference data:

- **How should off-policy preference data be preconditioned?** An alternative to fine-tuning for bringing μ and π_{ref} closer is to filter, reweight, or otherwise precondition the preference data, which adjusts μ instead of π_{ref}. However, performing this efficiently at modern scale remains future work. This is because per-completion log-likelihood differences can exhibit unnecessarily high variance, and a potential remedy is to treat clusters of completions as alternatives.

- **When should RLHF be used versus more explicit pluralistic alignment methods?** RLHF may perform well in some regimes, while in others it may be preferable to resort to more computationally expensive but heterogeneity-robust methods such as NLHF (Munos et al., 2024). An important practical question is therefore to understand when the extra robustness is worth the additional optimization cost, in which distribution mismatch emerges as one of the key variables.

In a broader sense, amid the community's emerging attention to user heterogeneity (Sorensen et al., 2024; Poddar et al., 2024; Zhang et al., 2026), our results suggest that the impact of heterogeneity depends on the relationship between π_{ref} and μ, i.e., how base models are trained and preference data is collected. This insight encourages moving beyond evaluating whether a particular algorithm is robust to heterogeneity, and calls for a more holistic perspective on the entire training pipeline to account for the effects of user heterogeneity.

ACKNOWLEDGMENTS

We thank Song Mei for helpful feedback. KO was partially supported by JST, ACT-X Grant Number JPMJAX23C4, ONR Grant Number N00014-24-S-B001, and DARPA AIQ Grant Number HR001124S0029-AIQ-FP-003. AU is supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 2146752. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the National Science Foundation. HB is supported by JST, BOOST Grant Number JPMJBY24E8. This work was supported in part by the National Science Foundation under grant CCF-2145898, by the Office of Naval Research under grant N00014-24-1-2159, a Google Research Scholar Award, an Alfred P. Sloan fellowship, and a Schmidt Science AI2050 fellowship.

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.

Hamidreza Alipour and Mohak Goyal. Utilitarian distortion under probabilistic voting. *arXiv preprint arXiv:2602.11152*, 2026.

Elliot Anshelevich and John Postl. Randomized social choice functions under metric preferences. *Journal of Artificial Intelligence Research*, 58:797–827, 2017.

Elliot Anshelevich, Onkar Bhardwaj, Edith Elkind, John Postl, and Piotr Skowron. Approximating optimal social choice under metric preferences. *Artificial Intelligence*, 264:27–51, 2018.

Elliot Anshelevich, Aris Filos-Ratsikas, Nisarg Shah, and Alexandros A Voudouris. Distortion in social choice problems: The first 15 years and beyond. In *30th International Joint Conference on Artificial Intelligence*, pp. 4294–4301, 2021.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

Craig Boutilier, Ioannis Caragiannis, Simi Haber, Tyler Lu, Ariel D Procaccia, and Or Sheffet. Optimal social choice functions: A utilitarian view. In *Proceedings of the 13th ACM Conference on Electronic Commerce*, pp. 197–214, 2012.

Ioannis Caragiannis and Ariel D Procaccia. Voting almost maximizes social welfare despite limited communication. *Artificial Intelligence*, 175(9-10):1655–1671, 2011.

Souradip Chakraborty, Jiahao Qiu, Hui Yuan, Alec Koppel, Dinesh Manocha, Furong Huang, Amrit Bedi, and Mengdi Wang. MaxMin-RLHF: Alignment with diverse human preferences. In *International Conference on Machine Learning*, pp. 6116–6135. PMLR, 2024.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

Flavio Chierichetti, Ravi Kumar, and Andrew Tomkins. Learning a mixture of two multinomial logits. In *International Conference on Machine Learning*, pp. 961–969. PMLR, 2018.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 2017.

Vincent Conitzer, Rachel Freedman, Jobst Heitzig, Wesley H Holliday, Bob M Jacobs, Nathan Lambert, Milan Mossé, Eric Pacuit, Stuart Russell, Hailey Schoelkopf, et al. Position: social choice should guide AI alignment in dealing with diverse human feedback. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 9346–9360, 2024.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, et al. Ultrafeedback: Boosting language models with scaled ai feedback. *arXiv preprint arXiv:2310.01377*, 2023.

Soroush Ebadian, Anson Kahng, Dominik Peters, and Nisarg Shah. Optimized distortion and proportional fairness in voting. *ACM Transactions on Economics and Computation*, 12(1):1–39, 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.

Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, et al. Olmo 3. *arXiv preprint arXiv:2512.13961*, 2025.

Luise Ge, Daniel Halpern, Evi Micha, Ariel D Procaccia, Itai Shapira, Yevgeniy Vorobeychik, and Junlin Wu. Axioms for AI alignment from human feedback. *Advances in Neural Information Processing Systems*, 37:80439–80465, 2024.

Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Paul Gölz, Nika Haghtalab, and Kunhe Yang. Distortion of AI alignment: Does preference optimization optimize for preferences? *arXiv preprint arXiv:2505.23749*, 2025.

Mohak Goyal and Sahasrajit Sarmasarkar. Metric distortion under probabilistic voting. *arXiv preprint arXiv:2405.14223*, 2024.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. Direct language model alignment from online AI feedback. *arXiv preprint arXiv:2402.04792*, 2024.

Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason D Lee, Wen Sun, Akshay Krishnamurthy, and Dylan J Foster. Correcting the mythos of KL-regularization: Direct alignment without overoptimization via chi-squared preference optimization. In *The Thirteenth International Conference on Learning Representations*, 2025.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, Lester James Validad Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, et al. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, 2025.

Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. Skywork-reward: Bag of tricks for reward modeling in LLMs. *arXiv preprint arXiv:2410.18451*, 2024.

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.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, et al. Nash learning from human feedback. In *Forty-first International Conference on Machine Learning*, 2024.

Sewoong Oh and Devavrat Shah. Learning mixed multinomial logit model from ordinal data. *Advances in Neural Information Processing Systems*, 27, 2014.

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.

Sriyash Poddar, Yanming Wan, Hamish Ivison, Abhishek Gupta, and Natasha Jaques. Personalizing reinforcement learning from human feedback with variational preference learning. *Advances in Neural Information Processing Systems*, 37:52516–52544, 2024.

Ariel D Procaccia and Jeffrey S Rosenschein. The distortion of cardinal preferences in voting. In *International Workshop on Cooperative Information Agents*, pp. 317–331. Springer, 2006.

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

Arun Rajkumar and Shivani Agarwal. A statistical convergence perspective of algorithms for rank aggregation from pairwise data. In *International Conference on Machine Learning*, pp. 118–126. PMLR, 2014.

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

Anand Siththaranjan, Cassidy Laidlaw, and Dylan Hadfield-Menell. Distributional preference learning: Understanding and accounting for hidden context in RLHF. *arXiv preprint arXiv:2312.08358*, 2023.

Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, et al. A roadmap to pluralistic alignment. *arXiv preprint arXiv:2402.05070*, 2024.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*, 2023.

Rui Wu, Jiaming Xu, Rayadurgam Srikant, Laurent Massoulié, Marc Lelarge, and Bruce Hajek. Clustering and inference from pairwise comparisons. In *Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pp. 449–450, 2015.

Tengyang Xie, Dylan J Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Hassan Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit Q*-approximation for sample-efficient RLHF. In *The Thirteenth International Conference on Learning Representations*, 2025.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: bridging theory and practice for RLHF under KL-constraint. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 54715–54754, 2024.

Lily H Zhang, Smitha Milli, Karen Long Jusko, Jonathan Smith, Brandon Amos, Wassim Bouaziz, Jack Kussman, Manon Revel, Lisa Titus, Bhaktipriya Radharapu, et al. Cultivating pluralism in algorithmic monoculture: The community alignment dataset. In *The Fourteenth International Conference on Learning Representations*, 2026.

Shenao Zhang, Donghan Yu, Hiteshi Sharma, Han Zhong, Zhihan Liu, Ziyi Yang, Shuohang Wang, Hany Hassan, and Zhaoran Wang. Self-exploring language models: Active preference elicitation for online alignment. *arXiv preprint arXiv:2405.19332*, 2024.

Xiaomin Zhang, Xucheng Zhang, Po-Ling Loh, and Yingyu Liang. On the identifiability of mixtures of ranking models. *arXiv preprint arXiv:2201.13132*, 2022.

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

Appendix

TABLE OF CONTENTS

A Related Work

Distortion in social choice theory. The notion of distortion, defined as the ratio between the average utility achieved by a selected distribution and the maximum achievable average utility, originates in social choice theory. In the classical setting, each user has a deterministic preference ordering that is consistent with their underlying utility values (Procaccia & Rosenschein, 2006). In this setting, for m alternatives, the distortion is $\Omega(m^2)$ for deterministic voting rules (Caragiannis & Procaccia, 2011), and $\Omega(m^{1/2})$ when randomized voting rules are allowed (Boutilier et al., 2012; Ebadian et al., 2024).

Recent work has studied metric distortion as a setting that yields distortion independent of m by imposing structure on utilities (Anshelevich & Postl, 2017; Anshelevich et al., 2018). In this setting, Goyal & Sarmasarkar (2024) show that distortion is reduced as preferences become less deterministic. In this context, following Gölz et al. (2025), our work serves as the non-metric counterpart to Goyal & Sarmasarkar (2024) and derives distortion bounds that depend on the strength of randomness β but not on m.

Utilitarian analysis of alignment methods. Our closest prior work is Gölz et al. (2025), which introduced distortion for the analysis of AI alignment methods and incorporated a KL-divergence constraint. Their analysis of RLHF builds on results from Siththaranjan et al. (2023), which establish an ordering equivalence between rewards fitted under a single Bradley–Terry model and the Borda score; related results can be traced back to Rajkumar & Agarwal (2014). Also, Ge et al. (2024) analyze RLHF under a linear utility model and prove that it fails to satisfy several desired properties.

Concurrent with and independent of our work, Alipour & Goyal (2026) consider the same data generation process and analyze the distortion of various mechanisms in the traditional social choice setting. Their results on the Borda rule imply that, in this setting, the distortion of RLHF is $\beta(1+o(1))$. Their analysis focuses on the alternative that maximizes the Borda score. In contrast, our analysis based on effective utility can evaluate the magnitude of differences in the estimated rewards across alternatives, which allows our approach to generalize to the AI alignment setting.

Relation between Theorem 5 and identifiability results. The problem of identifying the underlying parameters of mixtures of ranking models is closely related to this lower bound. For the BT model, Wu et al. (2015) establish identifiability under the assumption that comparisons for all pairs of alternatives are observed from the same user, while Oh & Shah (2014) and Chierichetti et al. (2018) assume that additional comparison information is available for each user. Subsequently, Zhang et al. (2022) show that when there are two distinct users, the parameters are identifiable up to a measure-zero set.

Our proof considers two users, and thus falls within the scope of the positive result of Zhang et al. (2022). Our lower bound is based on the fact that, even when the model parameters are identifiable, a degree of freedom corresponding to a permutation of alternatives remains, which prevents distinguishing differences in average utility on the order of $\Theta(\beta)$.

B Proof of Upper Bounds

In this section, we prove Theorem 1 and Theorem 2. Throughout this section, we use reward clipping (3) in the definition of the AI alignment setting, and denote the unclipped rewards by $\bar{r}$.

Both Theorem 1 and Theorem 2 require a case analysis depending on the value of $\mathbb{P}_{u\sim\mathcal{D},x\sim\mu}[\beta u(x) > c]$, and, only in the case of Theorem 2, on $r_{\min}$. Among these cases, the technically most involved ones arise when $\mathbb{P}_{u\sim\mathcal{D},\,x\sim\mu}[\beta u(x) > c] \leq c^2$ in Theorem 1, and when $\mathbb{P}_{u\sim\mathcal{D},\,x\sim\mu}[\beta u(x) > c] \leq c^2$ and $r_{\min} \leq 0$ in Theorem 2. We therefore state the theorems restricted to these cases as Theorems 6 and 7, respectively, and their proofs are found in Appendices B.1–B.4.

Specifically, the proof of Theorem 6 proceeds in the following steps, as sketched in Section 3.1. First, in Appendix B.1, we prove Lemma 1, which introduces a hypothetical effective utility $\hat{u}(x)$, and gives the upper bound of the true utility $u(x)$ by $\hat{u}(x)$. Next, in Appendix B.2, we prove Lemma 2, showing that the expectation of the effective utility, $\mathbb{E}_{u\sim\mathcal{D}}[\hat{u}(x)]$, lower bounds the clipped reward. We then prove Lemma 3 in Appendix B.3, which establishes that the clipped reward in turn lower bounds the

average true utility $\mathbb{E}_{u\sim\mathcal{D}}[u(x)]$. Finally, these components are combined in Appendix B.4 to obtain Theorem 6. The proof of Theorem 7 mainly reduces to a generalization of Lemma 1 to Lemma 4, which is discussed in Appendix B.1.

In the assumptions for Theorems 6 and 7, the condition $\mathbb{P}_{u\sim\mathcal{D},\,x\sim\mu}[\beta u(x) > c] \le c^2$ informally ensures that not too many alternatives have large utility. Otherwise, the presence of many high-utility alternatives may lead to inaccurate reward estimation for alternatives with small average utility. Also, if $r_{\min} > 0$, the clipping may ignore signals from small utilities. Therefore, when these assumptions do not hold, separate arguments are required. For Theorem 1, the case $\mathbb{P}_{u\sim\mathcal{D},\,x\sim\mu}[\beta u(x) > c] > c^2$ is handled in Theorem 8. For Theorem 2, Theorem 9 covers the case $\mathbb{P}_{u\sim\mathcal{D},\,x\sim\mu}[\beta u(x) > c] > c^2$, while Theorem 10 treats the case $\mathbb{P}_{u\sim\mathcal{D},\,x\sim\mu}[\beta u(x) > c] \le c^2$ with $r_{\min} > 0$. The proofs of these results are given in Appendix B.5.

B.1 Analysis of the Reduction Effect on u

We begin by comparing the maximum expected value of the true utility with that of the effective utility. This quantifies how much the signal to identify the true utility is lost due to pairwise comparisons and the nonlinearity of the Bradley–Terry model.

The following lemma is a formal version of Lemma 4. Lemma 1 can also be obtained by taking the limit $\tau \to \infty$.

Lemma 5. *For $u \in \mathbb{R}^m$ drawn from $\mathcal{D}$, we define $\hat{u} \in \mathbb{R}^m$ by the following mapping, where $c > 0$ is an arbitrary constant.*

$$\hat{u}(x) = \begin{cases} 0 & (\text{if } \mathbb{P}_{y\sim\mu}[u(y) - u(x) \ge c\beta^{-1}] \ge \frac{1}{2}) \\ c\beta^{-1} & (\text{else if } u(x) \ge c\beta^{-1}) \\ u(x) & (\text{otherwise}) \end{cases}.$$

Assume $\pi_{\mathrm{ref}} \ll \mu$ and define $B = \max_{x\in A} \log \frac{\mu(x)}{\pi_{\mathrm{ref}}(x)}$. Then we have that

$$\begin{aligned} &\max_{\pi\colon \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau} \mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[u(x)] \\ &\le c^{-1}\beta\big(40\min\left\{e^B\tau, B, B\tau^{-1}\right\} + 3\big) \max_{\pi\colon \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau} \mathbb{E}_{\hat{u},x\sim\pi}[\hat{u}(x)] + 4\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{RLHF}}}[u(x)]. \end{aligned} \tag{11}$$

As explained in Section 4.1, the proof of this lemma uses the following two auxiliary lemmas. The proofs of these lemmas are deferred to Section B.7.2.

Lemma 6. *Let $B, \tau > 0$ be constants. Let π_1, π_2 be probability measures with $\pi_2 \ll \pi_1$ and $\log \frac{\mathrm{d}\pi_2}{\mathrm{d}\pi_1} \le B$. Then, the probability measure π_λ defined by $\pi_\lambda := (1-\lambda)\pi_1 + \lambda\pi_2$ with $\lambda := \min\{1, B^{-1}\tau\}$ satisfies $\mathrm{KL}(\pi_\lambda\|\pi_1) \le \tau$.*

Lemma 7. *Let π and π' be probability measures such that $\pi' \ll \pi$ and define the tail regions $A_- := \{x\colon \frac{\mathrm{d}\pi'}{\mathrm{d}\pi}(x) \le \frac{1}{2}\}$ and $A_+ := \{x\colon \frac{\mathrm{d}\pi'}{\mathrm{d}\pi}(x) \ge 2\}$. Then, the total variation restricted to A_- is bounded by*

$$\frac{1}{2}\int_{A_-} |\mathrm{d}\pi' - \mathrm{d}\pi| \le (\log(e/2))^{-1}\mathrm{KL}(\pi'\|\pi). \tag{12}$$

Also, the total variation restricted to A_+ is bounded by

$$\frac{1}{2}\int_{A_+} |\mathrm{d}\pi' - \mathrm{d}\pi| \le (\log(4/e))^{-1}\mathrm{KL}(\pi'\|\pi). \tag{13}$$

Proof of Lemma 5. **(i) When $\tau \ge 1$.** Let $I_1(u,x)$ be the indicator for the event $\mathbb{P}_{y\sim\mu}\big[u(y) - u(x) \ge c\beta^{-1}\big] \ge \frac{1}{2}$. For a fixed u, the set of x with $I_1(u,x) = 1$ has μ-measure at most $1/2$, and for such x we have $u(x) < u(y)$ for every y with $I_1(u,y) = 0$. This implies that

$$\begin{aligned} &u(x)I_1(u,x) \le 2\mathbb{E}_{y\sim\mu}[u(y)(1 - I_1(u,y))] \\ &\Rightarrow \max_{\pi\colon \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau} \mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[u(x)I_1(u,x)] \le 2\mathbb{E}_{u\sim\mathcal{D},y\sim\mu}[u(y)(1 - I_1(u,y))]. \end{aligned} \tag{14}$$

When $I_1(u,x)=0$, we have $u(x)\le c^{-1}\beta\hat{u}(x)$ by the definition of $\hat{u}$. Thus, we can further bound (14) by

$$\mathbb{E}_{u\sim\mathcal{D},y\sim\mu}[u(y)(1-I_1(u,y))]\le c^{-1}\beta\mathbb{E}_{\hat{u},y\sim\mu}[\hat{u}(y)(1-I_1(u,y))]\le c^{-1}\beta\mathbb{E}_{\hat{u},y\sim\mu}[\hat{u}(y)]. \quad (15)$$

Now define the mixture $\pi' := (1-\lambda)\pi_{\mathrm{ref}}+\lambda\mu$ with $\lambda := \min\{B^{-1}\tau,1\}$. Since $\mathrm{d}\pi'\ge\lambda\mathrm{d}\mu$,

$$\mathbb{E}_{\hat{u},y\sim\mu}[\hat{u}(y)]\le\lambda^{-1}\mathbb{E}_{\hat{u},y\sim\pi'}[\hat{u}(y)]. \quad (16)$$

By combining (14), (15), and (16), we have that

$$\begin{aligned}\max_{\pi:\ \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau}\mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[u(x)I_1(u,x)] &\le 2c^{-1}\beta\lambda^{-1}\mathbb{E}_{\hat{u},y\sim\pi'}[\hat{u}(y)]\\ &\le 2c^{-1}\beta\lambda^{-1}\max_{\pi:\ \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau}\mathbb{E}_{\hat{u},x\sim\pi}[\hat{u}(x)], \quad (17)\end{aligned}$$

In the final inequality, we used the fact that $\mathrm{KL}(\pi'\|\pi_{\mathrm{ref}})\le\tau$ from Lemma 6 (with $\pi_1=\pi_{\mathrm{ref}}$, $\pi_2=\mu$).

For the case where $I_1(u,x)=0$, we have $u(x)\le c^{-1}\beta\hat{u}(x)$, thus

$$\begin{aligned}\max_{\pi:\ \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau}\mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[u(x)(1-I_1(u,x))] &\le c^{-1}\beta\max_{\pi:\ \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau}\mathbb{E}_{\hat{u},x\sim\pi}[\hat{u}(x)(1-I_1(u,x))]\\ &\le c^{-1}\beta\max_{\pi:\ \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau}\mathbb{E}_{\hat{u},x\sim\pi}[\hat{u}(x)]. \quad (18)\end{aligned}$$

Combining (17) and (18) and using $\lambda=\min\{B^{-1}\tau,1\}$, we obtain

$$\max_{\pi:\ \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau}\mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[u(x)]\le c^{-1}\beta(2B\tau^{-1}+3)\max_{\pi:\ \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau}\mathbb{E}_{\hat{u},x\sim\pi}[\hat{u}(x)].$$

(ii) When $Be^{-B}\le\tau<1$. Let π^* be a maximizer of $\mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[u(x)]$, and let π_{RLHF} be a maximizer of $\mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[r(x)]$, both within the KL ball $\mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau$. For a fixed u,

$$\begin{aligned}&\mathbb{E}_{x\sim\pi^*}[u(x)I_1(u,x)]-4\mathbb{E}_{x\sim\pi_{\mathrm{RLHF}}}[u(x)I_1(u,x)]\\ &=\int u(x)I_1(u,x)(\mathrm{d}\pi^*-4\mathrm{d}\pi_{\mathrm{RLHF}})\\ &=\int u(x)I_1(u,x)(\mathrm{d}\pi^*-2\mathrm{d}\pi_{\mathrm{ref}})+2\int u(x)I_1(u,x)(\mathrm{d}\pi_{\mathrm{ref}}-2\mathrm{d}\pi_{\mathrm{RLHF}})\\ &\le\int u(x)I_1(u,x)\mathbb{1}\Big[\mathrm{d}\pi^*>2\mathrm{d}\pi_{\mathrm{ref}}\Big](\mathrm{d}\pi^*-2\mathrm{d}\pi_{\mathrm{ref}})\\ &\quad+2\int u(x)I_1(u,x)\mathbb{1}\left[\mathrm{d}\pi_{\mathrm{RLHF}}<\frac{1}{2}\mathrm{d}\pi_{\mathrm{ref}}\right](\mathrm{d}\pi_{\mathrm{ref}}-2\mathrm{d}\pi_{\mathrm{RLHF}})\\ &\le\int\mathbb{1}\Big[\mathrm{d}\pi^*>2\mathrm{d}\pi_{\mathrm{ref}}\Big]|\mathrm{d}\pi^*-\mathrm{d}\pi_{\mathrm{ref}}|+2\int\mathbb{1}\left[\mathrm{d}\pi_{\mathrm{RLHF}}<\frac{1}{2}\mathrm{d}\pi_{\mathrm{ref}}\right]|\mathrm{d}\pi_{\mathrm{ref}}-\mathrm{d}\pi_{\mathrm{RLHF}}| \quad (19)\\ &\le 2(\log(4/e))^{-1}\mathrm{KL}(\pi^*\|\pi_{\mathrm{ref}})+4(\log(e/2))^{-1}\mathrm{KL}(\pi_{\mathrm{RLHF}}\|\pi_{\mathrm{ref}}) \quad (20)\\ &\le\Big(2(\log(4/e))^{-1}+4(\log(e/2))^{-1}\Big)\tau \quad (21)\\ &\le 20\tau, \quad (22)\end{aligned}$$

where we applied Lemma 7 in passing from (19) to (20) (to the first integral on $A_+=\{\mathrm{d}\pi^*>2\mathrm{d}\pi_{\mathrm{ref}}\}$ and to the second integral on $A_-=\{\mathrm{d}\pi_{\mathrm{RLHF}}<\frac{1}{2}\mathrm{d}\pi_{\mathrm{ref}}\}$), and in passing from (20) to (21) we used that $\mathrm{KL}(\pi^*\|\pi_{\mathrm{ref}}),\mathrm{KL}(\pi_{\mathrm{RLHF}}\|\pi_{\mathrm{ref}})\le\tau$.

Suppose that there exists some x such that $\mathbb{P}_{y\sim\mu}\big[u(y)-u(x)\ge c\beta^{-1}\big]\ge\frac{1}{2}$. Then $u(y)\ge u(x)+c\beta^{-1}$ holds on a μ-measure of at least $\frac{1}{2}$. For such y, we have $\mathbb{P}_{z\sim\mu}\big[u(z)-u(y)\ge c\beta^{-1}\big]<\frac{1}{2}$, which implies that y is not reduced to 0 by condition (i), but instead to $c\beta^{-1}$ by condition (ii). Therefore,

$$\mathbb{E}_{y\sim\mu}[\hat{u}(y)]\ge\frac{1}{2}c\beta^{-1}. \quad (23)$$

Consider $\pi' = (1-\lambda)\pi_{\mathrm{ref}} + \lambda\mu$ with $\lambda = B^{-1}\tau \le 1$. Then, $\mathrm{d}\pi' \ge \lambda \mathrm{d}\mu$ holds, so (23) implies that

$$\mathbb{E}_{y\sim\pi'}[\hat{u}(y)] \ge \lambda \mathbb{E}_{y\sim\mu}[\hat{u}(y)] \ge \frac{1}{2}\lambda c \beta^{-1}. \tag{24}$$

By combining (22) and (24), we have that

$$\mathbb{E}_{x\sim\pi^*}[u(x)I_1(u,x)] - 4\mathbb{E}_{x\sim\pi_{\mathrm{RLHF}}}[u(x)I_1(u,x)] \le 40\lambda^{-1}c^{-1}\beta\tau\mathbb{E}_{x\sim\pi'}[\hat{u}(x)]. \tag{25}$$

Note that, if no such x exists, the LHS of (25) is 0 and the bound holds trivially.

By Lemma 6, π' lies in the KL ball $\mathrm{KL}(\pi\|\pi_{\mathrm{ref}}) \le \tau$. Taking the expectation of $u \sim \mathcal{D}$ in (25), we have that

$$\begin{aligned}
&\mathbb{E}_{u\sim\mathcal{D},x\sim\pi^*}[u(x)I_1(u,x)] - 4\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{RLHF}}}[u(x)I_1(u,x)] \\
&\le 40\lambda^{-1}c^{-1}\beta\tau\mathbb{E}_{\hat{u},x\sim\pi'}[\hat{u}(x)] \\
&\le 40\lambda^{-1}c^{-1}\beta\tau \max_{\pi:\,\mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau} \mathbb{E}_{\hat{u},x\sim\pi}[\hat{u}(x)].
\end{aligned} \tag{26}$$

For the case where $I_1(u,x) = 0$, we again use $u(x) \le c^{-1}\beta\hat{u}(x)$ to obtain

$$\mathbb{E}_{u\sim\mathcal{D},x\sim\pi^*}[u(x)(1-I_1(u,x))] \le c^{-1}\beta \max_{\pi:\,\mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau} \mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[\hat{u}(x)]. \tag{27}$$

Summing (26) and (27) and substituting $\lambda = B^{-1}\tau$,

$$\begin{aligned}
&\max_{\pi:\,\mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau} \mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[u(x)] - 4\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{RLHF}}}[u(x)] \\
&\le c^{-1}\beta\big(40\lambda^{-1}\tau + 1\big) \max_{\pi:\,\mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau} \mathbb{E}_{\hat{u},x\sim\pi}[\hat{u}(x)] \\
&= c^{-1}\beta\big(40B + 1\big) \max_{\pi:\,\mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau} \mathbb{E}_{\hat{u},x\sim\pi}[\hat{u}(x)].
\end{aligned} \tag{28}$$

(iii) When $\tau < Be^{-B}$. For the same mixture $\pi' = (1-\lambda)\pi_{\mathrm{ref}} + \lambda\mu$ as in (ii), we have $\mathrm{d}\pi' \ge e^{-B}\mathrm{d}\mu$. Therefore,

$$\mathbb{E}_{y\sim\pi'}[\hat{u}(y)] \ge e^{-B}\mathbb{E}_{y\sim\pi_{\mathrm{ref}}}[\hat{u}(y)] \ge \frac{1}{2}e^{-B}c\beta^{-1},$$

and λ in (24) can be replaced by e^{-B}. The rest of the proof is identical to (ii) until (28), and we have that

$$\begin{aligned}
&\max_{\pi:\,\mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau} \mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[u(x)] - 4\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{RLHF}}}[u(x)] \\
&\le (28) \le c^{-1}\beta\big(40e^{B}\tau + 1\big) \max_{\pi:\,\mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau} \mathbb{E}_{\hat{u},x\sim\pi}[\hat{u}(x)].
\end{aligned}$$

Combining (i), (ii), and (iii), we obtain (11). □

B.2 Upper Bounding the Effective Utility by the Reward

Next, we show that the upper-clipped reward is lower bounded by the expectation of the effective utility.

Lemma 8 (Lemma 2, restated)**.** *Let $\bar{r}(x)$ be the maximizer of the population log-likelihood (1) satisfying the constraint (3), and assume that $c \le 1$.*

Then, for any $x \in A$, the expectation of the effective utility can lower bound the upper-clipped reward as

$$\mathbb{E}_{u\sim\mathcal{D}}[\hat{u}(x)] \le \big(2\beta\sigma'(2c)\big)^{-1} \min\{\bar{r}(x), r_{\max}\}.$$

Lemma 2 is obtained by simply setting $R = r_{\max}$. In the proof, we use the following auxiliary lemmas.

Lemma 9. *Recall that $r_{\min}$ is defined as the solution to*

$$\mathbb{E}_{y\sim\mu}\big[\sigma(r_{\min}-\bar{r}(y))\big]=\frac{1}{2}-\frac{c^3}{16}, \tag{29}$$

and $r_{\max}=r_{\min}+2c$. If $c\le 1$, then $r_{\min}$ and $r_{\max}$ satisfy $r_{\min}\ge -c$ and $r_{\max}\ge c$, respectively.

Proof. Since $\bar{r}(y)\ge 0$ and $\sigma(\cdot)$ is increasing, we have that

$$\frac{1}{2}-\frac{c^3}{16}=\mathbb{E}_{y\sim\mu}\big[\sigma(r_{\min}-\bar{r}(y))\big]\le\sigma(r_{\min}).$$

We consider the case where $r_{\min}<0$ (otherwise the assertion follows immediately). We apply Lemma 11 in the next subsection with $s=-r_{\min}$, $t=0$ and $a=\frac{1}{2}$ to obtain the bound

$$\frac{1}{2}-\frac{c^3}{16}\le\frac{1}{2}-\frac{1}{8}\min\Big\{-r_{\min},\frac{1}{2}\Big\}.$$

Since $c<1$, we have $\frac{1}{2}-\frac{c^3}{16}>\frac{1}{2}-\frac{1}{8}\cdot\frac{1}{2}$. Therefore, $\frac{1}{2}-\frac{c^3}{16}\le\frac{1}{2}+\frac{1}{8}r_{\min}$ which implies $r_{\min}\ge-\frac{c^3}{2}\ge-c$, completing the proof. Also, $r_{\max}\ge c$ follows from $r_{\max}=r_{\min}+2c\ge -c+2c=c$. □

Lemma 10. *Let $a,b>0$ be arbitrary constants, and assume that $s,t>0$ satisfy $t-s\ge -a$. Then the sigmoid function $\sigma(t)=1/(1+e^{-t})$ satisfies*

$$\sigma(t-s)-\sigma(-s)\ge\sigma'(a+b)\min\{t,b\}.$$

The proof of this lemma is deferred to Section B.7.2.

Proof of Lemma 8. Fix $x\in A$. We begin with the optimality condition of $\bar{r}$. By subtracting (2) from the LHS of (57) in Lemma 17,

$$\mathbb{E}_{y\sim\mu}\big[\sigma(\bar{r}(x)-\bar{r}(y))\big]-\mathbb{E}_{y\sim\mu}\big[\sigma(-\bar{r}(y))\big]=\mathbb{E}_{y\sim\mu}\big[\sigma(\bar{r}(x)-\bar{r}(y))-\sigma(-\bar{r}(y))\big]\le\frac{1}{4}\bar{r}(x), \tag{30}$$

where we used $\sigma'(t)\le\frac{1}{4}$.

On the other hand, subtracting (2) from the RHS of (57),

$$\begin{aligned}(\text{LHS of (30)})&=\mathbb{E}_{u\sim\mathcal{D},y\sim\mu}\big[\sigma(\beta(u(x)-u(y)))\big]-\mathbb{E}_{u\sim\mathcal{D},y\sim\mu}\big[\sigma(-\beta u(y))\big]\\&=\mathbb{E}_{u\sim\mathcal{D},y\sim\mu}\big[\sigma(\beta(u(x)-u(y)))-\sigma(-\beta u(y))\big].\end{aligned} \tag{31}$$

To evaluate (31), recall the indicator function $I_1(u,x)$, introduced in Lemma 5, which indicates the event $\mathbb{P}_{y\sim\mu}\big[u(y)-u(x)\ge c\beta^{-1}\big]\ge\frac{1}{2}$. Further, let $I_2(u,x,y)$ denote the indicator function of the event $u(y)-u(x)\ge c\beta^{-1}$. Then we have that

$$\mathbb{E}_{y\sim\mu}\big[\sigma(\beta(u(x)-u(y)))-\sigma(-\beta u(y))\big]\ge\mathbb{E}_{y\sim\mu}\big[(1-I_2(u,x,y))(\sigma(\beta(u(x)-u(y)))-\sigma(-\beta u(y)))\big].$$

Applying Lemma 10 with $a=b=c$, $t=\beta u(x)$, and $s=\beta u(y)$, on the event $u(y)-u(x)\le c\beta^{-1}$, it holds that

$$\sigma(\beta(u(x)-u(y)))-\sigma(-\beta u(y))\ge\sigma'(2c)\min\{\beta u(x),c\}=\beta\sigma'(2c)\hat{u}(x).$$

Therefore, when $I_1(u,x)=0$,

$$\mathbb{E}_{y\sim\mu}\big[\sigma(\beta(u(x)-u(y)))-\sigma(-\beta u(y))\big]\ge\mathbb{E}_{y\sim\mu}\big[(1-I_2(u,x,y))\beta\sigma'(2c)\hat{u}(x)\big]\ge\frac{\beta\sigma'(2c)}{2}\hat{u}(x), \tag{32}$$

where we used $\mathbb{E}_{y\sim\mu}[1-I_2(u,x,y)]\ge\frac{1}{2}$ for the second inequality.

If instead $I_1(u,x)=1$, (32) also holds because, by definition, $\hat{u}(x)=0$. Therefore, taking expectation of both sides of (32) over $u\sim\mathcal{D}$, eq. (31) is lower bounded by $\frac{\beta\sigma'(2c)}{2}\mathbb{E}_{u\sim\mathcal{D}}[\hat{u}(x)]$. Combining this result with (30) and rearranging, we have that

$$\mathbb{E}_{u\sim\mathcal{D}}[\hat{u}(x)]\le\big(2\beta\sigma'(2c)\big)^{-1}\bar{r}(x). \tag{33}$$

Finally, we consider the reward clipping. Because $r_{\min} \geq -c$ by Lemma 9, we have $r_{\max} = r_{\min} + 2c \geq c$. Also, $\big(2\beta\sigma'(2c)\big)^{-1} \geq 2\beta^{-1}$ holds from $\sigma'(2c) \leq \frac{1}{4}$, and $\hat{u}(x) \leq c\beta^{-1}$ from the definition of $\hat{u}$. Combining them, we have that

$$\mathbb{E}_{u\sim\mathcal{D}}[\hat{u}(x)] \leq c\beta^{-1} \leq \big(2\beta\sigma'(2c)\big)^{-1} c \leq \big(2\beta\sigma'(2c)\big)^{-1} r_{\max}.$$

Therefore, $\bar{r}(x)$ in the RHS (33) can be replaced by $\min\{\bar{r}(x), r_{\max}\}$, and we obtain the desired bound. □

B.3 Upper Bounding the Reward by the True Utility

Finally, we relate the learned reward to the welfare objective by upper bounding the reward in terms of the average true utility, under the assumption that $\mathbb{P}_{u\sim\mathcal{D},x\sim\mu}[\beta u(x) > c] \leq c^2$. The remaining case is discussed in Section B.5.

Our proof depends on a refined linearization of the sigmoid function. This result draws on Gölz et al. (2025, Lemma 1), but differs in that we linearize only in a neighborhood of the origin. This localization allows us to obtain a more accurate evaluation in the regime of small utilities. As a result, the error introduced by this linearization contributes with only a constant factor to the final distortion bound. This contrasts with Gölz et al. (2025), where each application of linearization incurs a $O(\beta)$ loss.

Lemma 11. *For any $a \geq 0$ and $s, t \geq 0$, the sigmoid function $\sigma(t) = 1/(1+e^{-t})$ satisfies*

$$\frac{1}{2} + \frac{1-a}{4}\min\{t, a\} - \frac{1}{4}\min\{s, 4\} \leq \sigma(t-s) \leq \frac{1}{2} + \frac{1}{4}\min\{t, 4\} - \frac{1-a}{4}\min\{s, a\}. \quad (34)$$

The proof of this lemma is deferred to Section B.7.2.

Based on this lemma, we show that if only a small fraction of alternatives have large utility, then $\bar{r}(x)$ is small for the rest of alternatives x.

Lemma 12. *Suppose that $\mathbb{P}_{u\sim\mathcal{D},x\sim\mu}[\beta u(x) > c] \leq c^2$ with $0 < c \leq \frac{1}{148}$. Then, for all x with $\mathbb{P}_{u\sim\mathcal{D}}[\beta u(x) > c] \leq c$, we have that*

$$\bar{r}(x) \leq 74c.$$

Moreover, such x exist with μ-measure at least $1 - c$.

Proof. Under the assumption, it is immediate that there exists a c-fraction of x satisfying $\mathbb{P}_{u\sim\mathcal{D}}\big[\beta u(x) > c\big] \leq c$ with respect to μ. Therefore, it remains to prove the first part of the lemma, for x such that $\mathbb{P}_{u\sim\mathcal{D}}\big[\beta u(x) > c\big] \leq c$.

Applying Lemma 11 with $a = \frac{1}{2}$ to each side of (57) of Lemma 17, we have that

$$\mathbb{E}_{y\sim\mu}\Big[\frac{1}{2} + \frac{1}{4}\min\{\bar{r}(x), 4\} - \frac{1}{8}\min\Big\{\bar{r}(y), \frac{1}{2}\Big\}\Big] \geq \mathbb{E}_{u\sim\mathcal{D},y\sim\mu}\Big[\frac{1}{2} + \frac{1}{8}\min\Big\{\beta u(x), \frac{1}{2}\Big\} - \frac{1}{4}\min\{\beta u(y), 4\}\Big],$$

for $x = 1, 2, \ldots, m$. By considering an alternative x with infinitesimal sampling probability $\mu(x)$ such that $u(x) \equiv 0$ and $r(x) = 0$, and rearranging the terms, we obtain

$$\mathbb{E}_{y\sim\mu}\Big[\min\Big\{\bar{r}(y), \frac{1}{2}\Big\}\Big] \leq 2\mathbb{E}_{u\sim\mathcal{D},y\sim\mu}[\min\{\beta u(y), 4\}] \leq 4c, \quad (35)$$

where we used the assumption that $\mathbb{P}_{u\sim\mathcal{D},x\sim\mu}[\beta u(x) > c] \leq c^2$ and $c \leq \frac{1}{148}$ to obtain $\mathbb{E}_{u\sim\mathcal{D}}[\min\{\beta u(y), 4\}] \leq c + 4c^2 \leq 2c$.

We then apply Lemma 11 with $a = \frac{1}{2}$ to (57) of Lemma 17 in the reverse direction to obtain that

$$\mathbb{E}_{y\sim\mu}\Big[\frac{1}{2} + \frac{1}{8}\min\Big\{\bar{r}(x), \frac{1}{2}\Big\} - \frac{1}{4}\min\{\bar{r}(y), 4\}\Big] \leq \mathbb{E}_{u\sim\mathcal{D},y\sim\mu}\Big[\frac{1}{2} + \frac{1}{4}\min\{\beta u(x), 4\} - \frac{1}{8}\min\Big\{\beta u(x), \frac{1}{2}\Big\}\Big].$$

Rearranging the terms yields that

$$\begin{aligned}
\min\Big\{\bar{r}(x), \frac{1}{2}\Big\} &\leq 2\mathbb{E}_{y\sim\mu}[\min\{\bar{r}(y), 4\}] + 2\mathbb{E}_{u\sim\mathcal{D}}[\min\{\beta u(x), 4\}] \\
&\leq 16\mathbb{E}_{y\sim\mu}\Big[\min\Big\{\bar{r}(y), \frac{1}{2}\Big\}\Big] + 2(c + 4\cdot\mathbb{P}_{u\sim\mathcal{D}}\big[\beta u(x) > c\big]) \\
&\leq 16\cdot 4c + 2(c + 4c) = 74c
\end{aligned}$$

where we used $\min\{t, 4\} \le 8\min\{t, \frac{1}{2}\}$ and $\mathbb{E}_{u\sim\mathcal{D}}[\min\{\beta u(x), 4\}] \le c + 4\cdot\mathbb{P}_{u\sim\mathcal{D}}\big[\beta u(x) > c\big]$ for the second inequality, and we used (35) and $\mathbb{P}_{u\sim\mathcal{D}}\big[\beta u(x) > c\big] \le c$ for the third inequality. Finally, since $c \le \frac{1}{148}$ implies $74c \le \frac{1}{2}$, $\bar{r}(x) \le 74c$ as claimed. □

The preceding lemma shows that, under our tail assumption $\mathbb{P}_{u\sim\mathcal{D},x\sim\mu}[\beta u(x) > c] \le c^2$, $\bar{r}(y)$ is small for a $1-c$ fraction under $y \sim \mu$. The following lemma uses this to upper bound $r_{\max}$.

Lemma 13. *Suppose that* $\mathbb{P}_{u\sim\mathcal{D},x\sim\mu}[\beta u(x) > c] \le c^2$ *and* $0 < c \le \frac{1}{316}$. *Then,* $r_{\max} \le 159c$.

Proof. According to Lemma 12, $\bar{r}(y) \le 74c$ with probability $1-c$ with respect to μ, which implies that

$$\frac{1}{2} - \frac{c^3}{16} = \mathbb{E}_{y\sim\mu}\big[\sigma(r_{\min} - \bar{r}(y))\big] \ge (1-c)\sigma(r_{\min} - 74c).$$

Applying Lemma 11 with $s = 75c$, $t = r_{\min} + c$ and $a = \frac{1}{2}$, this is further bounded by

$$\frac{1}{2} - \frac{c^3}{16} \ge (1-c)\Big(\frac{1}{2} + \frac{1}{8}\min\Big\{r_{\min} + c, \frac{1}{2}\Big\} - \frac{1}{4}\min\{75c, 4\}\Big). \tag{36}$$

When $c \le \frac{1}{148}$, we have that $\frac{1}{1-c} \le 1 + 2c$ and $75c \le 4$. By using this, rearranging (36) yields that

$$\min\Big\{r_{\min} + c, \frac{1}{2}\Big\} \le 158c.$$

When $c \le \frac{1}{316}$, $158c \le \frac{1}{2}$, which implies that $r_{\min} + c \le 158c \Leftrightarrow r_{\min} \le 157c \Leftrightarrow r_{\max} \le 159c$. □

Using the above results, we show that $r(x)$ is upper bounded by the average utility $\mathbb{E}_{u\sim\mathcal{D}}[u(x)]$. We can obtain Lemma 3 by taking $R = r_{\max}$.

Lemma 14 (Lemma 3, restated)**.** *Suppose that* $\mathbb{P}_{u\sim\mathcal{D},x\sim\mu}[\beta u(x) > c] \le c^2$ *with* $0 < c \le \frac{1}{316}$. *Then, for all* x, *we have that*

$$\min\{\bar{r}(x), r_{\max}\} \le \frac{1}{2}\beta(\sigma'(233c))^{-1}\mathbb{E}_{u\sim\mathcal{D}}[u(x)],$$

Proof. First, consider x such that $\mathbb{P}_{u\sim\mathcal{D}}[\beta u(x) > c] \le c$. By subtracting (2) from the LHS of (57) in Lemma 17,

$$\begin{aligned}
&\mathbb{E}_{y\sim\mu}\big[\sigma(\bar{r}(x) - \bar{r}(y))\big] - \mathbb{E}_{y\sim\mu}\big[\sigma(-\bar{r}(y))\big] && (37)\\
&\ge \mathbb{E}_{y\sim\mu}\big[\mathbb{1}[\bar{r}(y) \le 74c](\sigma(\bar{r}(x) - \bar{r}(y)) - \sigma(-\bar{r}(y)))\big] \\
&\ge \mathbb{E}_{y\sim\mu}\big[\mathbb{1}[\bar{r}(y) \le 74c]\sigma'(233c)\min\{\bar{r}(x), 159c\}\big] && (38)\\
&\ge \frac{1}{2}\sigma'(233c)\min\{\bar{r}(x), 159c\} && (39)\\
&\ge \frac{1}{2}\sigma'(233c)\min\{\bar{r}(x), r_{\max}\}, && (40)
\end{aligned}$$

where we have used Lemma 10 with $a = 74c$ and $b = 159c$ for (38), the fact that at least $\frac{1}{2} \le 1 - c$ fraction of y satisfies $\bar{r}(y) \le 74c$ according to Lemma 12 for (39), and Lemma 13 for (40).

On the other hand, by subtracting (2) from the RHS of (57) in Lemma 17,

$$\begin{aligned}
(37) &= \mathbb{E}_{u\sim\mathcal{D},y\sim\mu}\big[\sigma(\beta(u(x) - u(y)))\big] - \mathbb{E}_{u\sim\mathcal{D},y\sim\mu}\big[\sigma(-\beta u(y))\big] \\
&= \mathbb{E}_{u\sim\mathcal{D},y\sim\mu}\big[\sigma(\beta(u(x) - u(y))) - \sigma(-\beta u(y))\big] \\
&\le \frac{\beta}{4}\mathbb{E}_{u\sim\mathcal{D}}[u(x)], && (41)
\end{aligned}$$

because $\sigma'(t) \le \frac{1}{4}$.

Comparing (40) and (41), we have that

$$\min\{\bar{r}(x), r_{\max}\} \le \frac{1}{2}\beta(\sigma'(233c))^{-1}\mathbb{E}_{u\sim\mathcal{D}}[u(x)]$$

as desired. □

B.4 PUTTING IT ALL TOGETHER

Putting the above arguments together, we obtain an upper bound on the distortion in the case where $\mathbb{P}_{u\sim\mathcal{D},x\sim\mu}[\beta u(x) > c] \le c^2$ and $r_{\min} \le 0$ (required only in the AI alignment setting).

We first present a linear distortion bound for the social choice setting.

Theorem 6. *Suppose that $\mathbb{P}_{u\sim\mathcal{D},x\sim\mu}[\beta u(x) > c] \le c^2$ with $0 < c \le \frac{1}{316}$. In the social choice setting the distortion of policy π_{Borda} is bounded by*

$$\mathsf{Dist}(\pi_{\mathrm{Borda}}) \le \frac{3\beta}{4c(\sigma'(233c))^2} + 4.$$

Proof. Since the social choice setting can be viewed as the limit $\tau \to \infty$, Lemma 5 implies that

$$\max_{x\in A} \mathbb{E}_{u\sim\mathcal{D}}[u(x)] \le 3c^{-1}\beta \max_{\pi} \mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[\hat{u}(x)] + 4\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{Borda}}}[u(x)].$$

From Lemma 8, we have that

$$\max_{\pi} \mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[\hat{u}(x)] \le \big(2\beta\sigma'(2c)\big)^{-1} \max_{\pi} \mathbb{E}_{x\sim\pi}[\min\{\bar{r}(x), r_{\max}\}].$$

Because π_{RLHF} is the maximizer of the average reward without reward clipping, and it is also the maximizer of the expectation of $\min\{\bar{r}(x), r_{\max}\}$:

$$\max_{\pi} \mathbb{E}_{x\sim\pi}\big[\min\{\bar{r}(x), r_{\max}\}\big] = \mathbb{E}_{x\sim\pi_{\mathrm{Borda}}}\big[\min\{\bar{r}(x), r_{\max}\}\big]$$

Finally, according to Lemma 14

$$\mathbb{E}_{x\sim\pi_{\mathrm{Borda}}}\big[\min\{\bar{r}(x), r_{\max}\}\big] \le \frac{1}{2}\beta(\sigma'(233c))^{-1}\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{Borda}}}[u(x)].$$

Putting it all together, we have that

$$\max_{x\in A} \mathbb{E}_{u\sim\mathcal{D}}[u(x)] \le \Big(\frac{3\beta}{4c(\sigma'(233c))^2} + 4\Big)\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{Borda}}}[u(x)].$$

Therefore, the distortion is bounded as desired. □

Next, we present the result for the AI alignment setting.

Theorem 7. *Suppose that $\mathbb{P}_{u\sim\mathcal{D},x\sim\mu}[\beta u(x) > c] \le c^2$ with $0 < c \le \frac{1}{316}$ and Assumption 1 holds. Let $\bar{r}(x)$ be the maximizer of the population log-likelihood*

$$\mathcal{L}(\tilde{r}) = \mathbb{E}_{u\sim\mathcal{D},x,y\sim\mu}\big[\sigma(u(x) - u(y))\log(\sigma(\tilde{r}(x) - \tilde{r}(y)))\big],$$

and the clipped reward $r(x) = \max\{\min\{\bar{r}(x), r_{\max}\}, r_{\min}\}$ be as defined in Theorem 2. Based on this reward $r(x)$, the RLHF distribution is obtained as the maximizer of

$$\max_{\pi:\ \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau} \mathbb{E}_{x\sim\pi}[r(x)].$$

Then, the distortion of π_{RLHF} is bounded by

$$\mathsf{Dist}(\pi_{\mathrm{RLHF}}) \le c^{-1}(2\sigma'(233c))^{-2} \times \big(40\min\big\{e^B\tau, B, B\tau^{-1}\big\} + 3\big)\beta + 4.$$

Proof. According to Lemma 5,

$$\begin{aligned}&\max_{\pi:\ \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau} \mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[u(x)] \\ &\le c^{-1}\beta\big(40\min\big\{e^B\tau, B, B\tau^{-1}\big\} + 3\big) \max_{\pi:\ \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau} \mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[\hat{u}(x)] + 4\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{RLHF}}}[u(x)].\end{aligned} \tag{42}$$

By using Lemma 8, $\mathbb{E}_{u\sim\mathcal{D}}[\hat{u}(x)] \le \big(2\beta\sigma'(2c)\big)^{-1}\min\{\bar{r}(x), r_{\max}\} = \big(2\beta\sigma'(2c)\big)^{-1} r(x)$ holds for each x, and thus

$$\max_{\pi:\ \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau} \mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[\hat{u}(x)] \le \big(2\beta\sigma'(2c)\big)^{-1} \max_{\pi:\ \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau} \mathbb{E}_{x\sim\pi}[r(x)]. \tag{43}$$

Furthermore, Lemma 14 implies that, for each x, $r(x) = \min\{\bar{r}(x), r_{\max}\} \le \beta(2\sigma'(233c))^{-1}\mathbb{E}_{u\sim\mathcal{D}}[u(x)]$, which yields

$$\max_{\pi:\ \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau} \mathbb{E}_{x\sim\pi}[r(x)] = \mathbb{E}_{x\sim\pi_{\mathrm{RLHF}}}[r(x)] \le \beta(2\sigma'(233c))^{-1}\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{RLHF}}}[u(x)]. \tag{44}$$

By combining (42), (43), and (44), we obtain the desired bound. □

B.5 When the Sampling Distribution Has High Average Utility

In this subsection, we derive the distortion in the case where $\mathbb{P}_{u\sim\mathcal{D},x\sim\mu}[\beta u(x) > c] \geq c^2$ or $r_{\min} > 0$. Theorem 8 establishes a linear bound using the unclipped reward in the case of $\pi_{\mathrm{ref}} = \mu$. This Theorem 8 also serves as the upper bound for the social choice setting, as the KL constraint is effective in the social choice setting. For the AI alignment setting, we prove Theorem 9 for the case $\mathbb{P}_{u\sim\mathcal{D},x\sim\mu}[\beta u(x) > c] \geq c^2$, and Theorem 10 for the case $r_{\min} > 0$.

We begin by presenting two technical lemmas.

Lemma 15. *Define the Borda score of $x \in A$ by $\mathsf{Borda}(x) = \mathbb{E}_{u\sim\mathcal{D},y\sim\mu}[\sigma(\beta(u(x)-u(y)))]$ $(x = 1,\dots,m)$, and let $\mathcal{B}$ be a set of x such that $\mathsf{Borda}(x) \geq \frac{1}{2} - \frac{c^3}{32}$. Assume that $c \leq 1$. Then, we have that*

$$r(x) \geq r_{\min} + \frac{c^3}{8}, \tag{45}$$

and

$$\mu(\mathcal{B}) = \mathbb{P}_{x\sim\mu}[\mathbb{1}[x \in \mathcal{B}]] \geq \frac{c^3}{c^3+16}. \tag{46}$$

Proof. For $x \in \mathcal{B}$, we have that

$$\begin{aligned}
\frac{1}{4}(\bar{r}(x) - r_{\min}) &\geq \mathbb{E}_{y\sim\mu}\big[\sigma(\bar{r}(x) - \bar{r}(y))\big] - \mathbb{E}_{y\sim\mu}\big[\sigma(r_{\min} - \bar{r}(y))\big] \quad \Big(\because \sigma'(t) \leq \frac{1}{4}\Big) \\
&= \mathsf{Borda}(x) - \Big(\frac{1}{2} - \frac{c^3}{16}\Big) \quad (\because \text{ Lemma 17 and the definition of } r_{\min}\ (7)) \\
&\geq \frac{1}{2} - \frac{c^3}{32} - \Big(\frac{1}{2} - \frac{c^3}{16}\Big) = \frac{c^3}{32},
\end{aligned}$$

which implies that $\bar{r}(x) \geq r_{\min} + \frac{c^3}{8}$. When $c \leq 1$, it holds that $r_{\max} = r_{\min} + 2c \geq r_{\min} + \frac{c^3}{8}$. Therefore, we obtain the first claim

$$r(x) \geq r_{\min} + \frac{c^3}{8}$$

for $x \in \mathcal{B}$.

We then prove the second claim. Note that $\mathbb{E}_{x\sim\mu}[\mathsf{Borda}(x)] = \frac{1}{2}$ from the symmetry, and $\mathsf{Borda}(x) \leq 1$. Given this, to lower bound $\mu(\mathcal{B})$, it suffices to consider the case where $\bar{r}(x) = 1$ for all $x \in \mathcal{B}$ and $\bar{r}(x) = \frac{1}{2} - \frac{c^3}{32}$ for all $x \notin \mathcal{B}$. Therefore, we have that

$$\mu(\mathcal{B}) \geq \frac{\frac{c^3}{32}}{\frac{1}{2} + \frac{c^3}{32}} = \frac{c^3}{c^3+16}.$$

□

Lemma 16. *Suppose that $\mathbb{P}_{u\sim\mathcal{D},x\sim\mu}[\beta u(x) > c] \geq c^2$ with $c \leq 1$. Let $\mathcal{B}$ the set defined in Lemma 15, and $\mathcal{B}'$ be a set of x such that $\mathbb{E}_{u\sim\mathcal{D}}[\min\{\beta u(x), c\}] \geq \frac{c^4}{16}$. Then $\mathcal{B} \subseteq \mathcal{B}'$, and $r(x) = r_{\min}$ for all $x \notin \mathcal{B}'$.*

Proof. For $x \notin \mathcal{B}'$, we have

$$\begin{aligned}
\mathsf{Borda}(x) - \frac{1}{2} &= \mathbb{E}_{u\sim\mathcal{D},y\sim\mu}\Big[\sigma(\beta(u(x) - u(y)))\Big] - \frac{1}{2} \quad (\because \text{ Lemma 17}) \\
&\leq \mathbb{E}_{u\sim\mathcal{D},y\sim\mu}\Big[\frac{1}{4}\min\{\beta u(x), 4\} - \frac{1}{8}\min\Big\{\beta u(y), \frac{1}{2}\Big\}\Big] \quad \Big(\because \text{ Lemma 11 with } a = \frac{1}{2}\Big) \\
&\leq c^{-1}\mathbb{E}_{u\sim\mathcal{D}}[\min\{\beta u(x), c\}] - \frac{1}{8}\mathbb{E}_{u\sim\mathcal{D},y\sim\mu}[\min\{\beta u(y), c\}] \\
&\leq c^{-1}\cdot\frac{c^4}{16} - \frac{1}{8}\cdot c^3 = -\frac{c^3}{16},
\end{aligned} \tag{47}$$

For the final inequality, we used $\mathbb{E}_{u\sim\mathcal{D}}[\min\{\beta u(x), c\}] < \frac{c^4}{16}$ for $x \notin \mathcal{B}'$ from the definition of $\mathcal{B}'$, and $\mathbb{E}_{u\sim\mathcal{D},y\sim\mu}[\min\{\beta u(y), c\}] \geq c^3$ from the assumption that $\mathbb{P}_{u\sim\mathcal{D},x\sim\mu}[\beta u(x) > c] \geq c^2$. Because $\mathcal{B}$ is the set of x satisfying $\mathsf{Borda}(x) \geq \frac{1}{2} - \frac{c^3}{32}$, we have $\mathcal{B} \subseteq \mathcal{B}'$.

Furthermore,

$$\begin{aligned}\mathbb{E}_{y\sim\mu}[\sigma(\bar{r}(x) - \bar{r}(y))] &= \mathsf{Borda}(x) \quad (\because \text{ Lemma 17}) \\ &\leq \frac{1}{2} - \frac{c^3}{16} \quad (\because \text{ (47)}) \\ &= \mathbb{E}_{y\sim\mu}[\sigma(r_{\min} - \bar{r}(y))]. \quad (\because \text{ the definition of } r_{\min} \text{ (7)})\end{aligned}$$

The monotonicity of σ implies that $\bar{r}(x) \leq r_{\min}$. Since $\bar{r}(x)$ smaller than $r_{\min}$ is clipped, we have $r(x) = r_{\min}$ for all $x \notin \mathcal{B}'$. □

Based on the above technical lemmas, we establish a linear distortion bound for the case $\pi_{\mathrm{ref}} = \mu$ with unclipped rewards, under $\mathbb{P}_{u\sim\mathcal{D},x\sim\mu}[\beta u(x) > c] \geq c^2$. Since the social choice setting does not impose a KL constraint, the corresponding result for the social choice setting is obtained by taking the limit $\tau \to \infty$ in this bound.

Theorem 8. *Suppose that $\mathbb{P}_{u\sim\mathcal{D},x\sim\mu}[\beta u(x) > c] \geq c^2$ with $0 < c \leq 1$ and $\pi_{\mathrm{ref}} = \mu$. Let $\bar{r}(x)$ be the maximizer of the population log-likelihood (1), and use $\bar{r}$ as the reward without reward clipping (3). Thus, the RLHF distribution under the unclipped reward is defined as*

$$\pi_{\mathrm{RLHF}} = \arg\max_{\pi:\ \mathrm{KL}(\pi\|\mu)\leq\tau} \mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[\bar{r}(x)].$$

Then, the distortion of this distribution π_{RLHF} is bounded by

$$\frac{16(c^3+16)\beta}{c^7}.$$

Proof. According to Lemma 16, for $x \notin \mathcal{B}$, we have $r_{\min} = r(x)$. This implies that $\pi_{\mathrm{RLHF}}(x) \leq \pi_{\mathrm{ref}}(x)$ for all $x \notin \mathcal{B}$, as π_{RLHF} maximizes $\mathbb{E}[\bar{r}(x)]$. From this, we obtain $1 - \pi_{\mathrm{RLHF}}(\mathcal{B}) \leq 1 - \pi_{\mathrm{ref}}(\mathcal{B})$, implying $\pi_{\mathrm{RLHF}}(\mathcal{B}) \geq \pi_{\mathrm{ref}}(\mathcal{B})$. By using this, Lemma 15, and $\mathcal{B} \subseteq \mathcal{B}'$ from Lemma 16, we have that

$$\mathbb{P}_{x\sim\pi_{\mathrm{RLHF}}}[x \in \mathcal{B}'] \geq \mathbb{P}_{x\sim\pi_{\mathrm{ref}}}[x \in \mathcal{B}'] \geq \mathbb{P}_{x\sim\pi_{\mathrm{ref}}}[x \in \mathcal{B}] \geq \frac{c^3}{c^3+16}.$$

Also, $x \in \mathcal{B}'$ satisfies $\mathbb{E}_{u\sim\mathcal{D}}[u(x)] \geq \beta^{-1}\mathbb{E}_{u\sim\mathcal{D}}[\min\{\beta u(x), c\}] \geq \frac{c^4}{16\beta}$. Therefore,

$$\mathbb{E}_{x\sim\pi_{\mathrm{RLHF}}}[u(x)] \geq \frac{c^4}{16\beta}\mathbb{P}_{x\sim\pi_{\mathrm{RLHF}}}[x \in \mathcal{B}'] \geq \frac{c^4}{16\beta} \cdot \frac{c^3}{c^3+16} = \frac{c^7}{16(c^3+16)\beta}.$$

Because $\mathbb{E}_{x\sim\pi}[u(x)]$ is at most 1, the distortion is bounded by $\frac{16(c^3+16)\beta}{c^7}$. □

Next, as complements to Theorem 7, we present Theorem 9 and Theorem 10. As explained at the beginning of this subsection, Theorem 9 covers the case $\mathbb{P}_{u\sim\mathcal{D},\,x\sim\mu}[\beta u(x) > c] > c^2$, while Theorem 10 covers the case $\mathbb{P}_{u\sim\mathcal{D},\,x\sim\mu}[\beta u(x) > c] \leq c^2$ with $r_{\min} > 0$.

Theorem 9. *Suppose that $\mathbb{P}_{u\sim\mathcal{D},\,x\sim\mu}[\beta u(x) > c] \geq c^2$ with $0 < c \leq \frac{1}{316}$, and suppose also that Assumption 1 holds. Let π_{RLHF} be defined as in Theorem 7. Then, the distortion of π_{RLHF} is bounded by*

$$\frac{5120(c^3+16)}{c^9}\min\left\{e^B\tau, B, \frac{B}{\tau}\right\}\beta + 4.$$

Proof. The proof proceeds by a case analysis on the value of τ, following the proof of Lemma 5.

(i) When $\tau \geq 1$. Consider the probability measure $\pi' = (1-\lambda)\pi_{\mathrm{ref}} + \lambda\mu$ with $\lambda = \min\{B^{-1}\tau, 1\}$. Because $\mathrm{d}\pi' \geq \lambda \mathrm{d}\mu$ holds, we have that

$$\begin{aligned}\mathbb{E}_{x\sim\pi'}[r(x)] &\geq \lambda \cdot \pi'(\mathcal{B})\big(\min_{x\in\mathcal{B}} r(x) - r_{\min}\big) + r_{\min} \\ &\geq \lambda \cdot \mu(\mathcal{B})\big(\min_{x\in\mathcal{B}} r(x) - r_{\min}\big) + r_{\min} \\ &\geq \lambda \cdot \frac{c^3}{c^3+16} \cdot \frac{c^3}{8} + r_{\min},\end{aligned}$$

where we used (45) and (46) from Lemma 15 for the final inequality. Therefore,

$$\max_{\mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\leq\tau} \mathbb{E}_{x\sim\pi}[r(x)] \geq \frac{c^6\lambda}{8(c^3+16)} + r_{\min}. \tag{48}$$

Because $\mathbb{E}_{x\sim\pi}[r(x)] \leq r_{\max} = r_{\min} + 2c$ and $r(x) = r_{\min}$ for $x \notin \mathcal{B}$ from Lemma 16, (48) implies that $\pi_{\mathrm{RLHF}}(\mathcal{B}) \geq \frac{c^5\lambda}{16(c^3+16)}$.

For $x \in \mathcal{B}$, since $\mathcal{B} \subseteq \mathcal{B}'$ from Lemma 16, the average utility is at least

$$\mathbb{E}_{u\sim\mathcal{D}}[u(x)] \geq \beta^{-1}\mathbb{E}_{u\sim\mathcal{D}}[\min\{\beta u(x), c\}] \geq \frac{c^4}{16}\beta^{-1}.$$

By using this, the average utility with respect to π_{RLHF} is

$$\begin{aligned}\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{RLHF}}}[u(x)] &\geq \pi_{\mathrm{RLHF}}(\mathcal{B}) \times \frac{c^4}{16}\beta^{-1} \geq \frac{c^9}{256(c^3+16)}\lambda\beta^{-1} \\ &= \frac{c^9}{256(c^3+16)} \min\Big\{\frac{\tau}{B}, 1\Big\}\beta^{-1}.\end{aligned} \tag{49}$$

Because the maximum average utility is at most 1, the distortion is bounded by $\frac{256(c^3+16)}{c^9}\max\big\{\frac{B}{\tau}, 1\big\}\beta$.

(ii) When $Be^{-B} \leq \tau < 1$. Similarly to (22) in the proof of Lemma 5, Lemma 7 implies that

$$\mathbb{E}_{u\sim\mathcal{D},x\sim\pi^*}[u(x)] - 4\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{RLHF}}}[u(x)] \leq 20\tau. \tag{50}$$

Here π^* is the distribution that maximizes the average utility.

Consider the probability measure $\pi' = (1-\lambda)\pi_{\mathrm{ref}} + \lambda\mu$ with $\lambda = B^{-1}\tau < 1$. By following the argument for (i) $\tau \geq 1$ until (49), we have that

$$\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{RLHF}}}[u(x)] \geq \frac{c^9}{256(c^3+16)}\lambda\beta^{-1} = \frac{c^9}{256(c^3+16)} \times \frac{\tau}{B}\beta^{-1}. \tag{51}$$

Combining (50) and (51), we obtain that

$$\mathbb{E}_{u\sim\mathcal{D},x\sim\pi^*}[u(x)] \leq \Big(\frac{5120(c^3+16)}{c^9}B\beta + 4\Big)\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{RLHF}}}[u(x)]$$

Therefore, the distortion is bounded by $\frac{5120(c^3+16)}{c^9}B\beta + 4$.

(iii) When $\tau < Be^{-B}$. For a probability measure $\pi' = (1-\lambda)\pi_{\mathrm{ref}} + \lambda\mu$ with $0 \leq \lambda \leq 1$, we also have $\mathrm{d}\pi' \geq e^{-B}\mathrm{d}\pi_{\mathrm{ref}}$ from the assumption. Therefore, (48) holds with λ replaced by e^{-B}, and

$$\max_{\pi:\ \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\leq\tau} \mathbb{E}_{x\sim\pi}[r(x)] \geq \frac{c^6 e^{-B}}{8(c^3+16)} + r_{\min}. \tag{52}$$

By following the subsequent argument from (48), (49) is also modified to

$$\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{RLHF}}}[u(x)] \geq \frac{c^9}{256(c^3+16)}e^{-B}\beta^{-1}. \tag{53}$$

By combining (50) and (53), we have that

$$\mathbb{E}_{u\sim\mathcal{D},x\sim\pi^*}[u(x)] \leq \Big(\frac{5120(c^3+16)}{c^9}e^B\beta\tau + 4\Big)\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{RLHF}}}[u(x)] \leq 10\tau.$$

Therefore, the distortion is bounded by $\frac{5120(c^3+16)}{c^9}e^B\beta\tau + 4$.

Summarizing the three cases, the distortion is bounded by $\frac{5120(c^3+16)}{c^9}\min\big\{e^B\tau, B, \frac{B}{\tau}\big\}\beta + 4$ as desired. □

Theorem 10. *Suppose that* $\mathbb{P}_{u\sim\mathcal{D},\,x\sim\mu}[\beta u(x) > c] \le c^2$ *with* $0 < c \le \frac{1}{316}$, $r_{\min} > 0$, *and Assumption 1 holds. Let* π_{RLHF} *be defined as in Theorem 7. Then, the distortion of* π_{RLHF} *is bounded by*

$$\frac{80(c^3+16)}{c^6\sigma'(233c)}\min\Big\{e^B\tau, B, \frac{B}{\tau}\Big\}\beta + 4.$$

Proof. We follow the proof of Theorem 9, reusing its intermediate results as needed.

(i) When $\tau \ge 1$**.** Consider the intermediate distribution $\pi' = (1-\lambda)\pi_{\mathrm{ref}} + \lambda\mu$ with $\lambda = \min\big\{\frac{\tau}{B}, 1\big\}$. From (48),

$$\mathbb{E}_{x\sim\pi_{\mathrm{RLHF}}}[r(x)] = \max_{\pi:\,\mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau} \mathbb{E}_{x\sim\pi}[r(x)] \ge \frac{c^6\lambda}{8(c^3+16)} + r_{\min}.$$

According to Lemma 14, we have that

$$\frac{1}{2}\beta(\sigma'(233c))^{-1}\mathbb{E}_{u\sim\mathcal{D}}[u(x)] \ge \mathbb{E}_{x\sim\pi_{\mathrm{RLHF}}}[\min\{\bar{r}(x), r_{\max}\}] \ge \mathbb{E}_{x\sim\pi_{\mathrm{RLHF}}}[r(x)] - r_{\min}. \quad (54)$$

By combining these two, we obtain that

$$\frac{c^6\lambda}{8(c^3+16)} \le \frac{1}{2}\beta(\sigma'(233c))^{-1}\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{RLHF}}}[u(x)]. \quad (55)$$

Therefore, the distortion is bounded by

$$\frac{8(c^3+16)}{c^6\lambda}\cdot\frac{1}{2}\beta(\sigma'(233c))^{-1} = \frac{4(c^3+16)}{c^6\sigma'(233c)}\max\Big\{\frac{B}{\tau}, 1\Big\}\beta.$$

(ii) When $Be^{-B} \le \tau < 1$**.** Let $\lambda = B^{-1}\tau < 1$ instead of $\lambda = \min\big\{\frac{\tau}{B}, 1\big\}$. Eq. (55) still holds despite this modification. Combining that with (50), we have

$$\mathbb{E}_{u\sim\mathcal{D},x\sim\pi^*}[u(x)] - 4\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{RLHF}}}[u(x)] \le 20\tau \le \frac{80(c^3+16)}{c^6\sigma'(233c)}\beta B \times \mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{RLHF}}}[u(x)],$$

which implies that the distortion is bounded by

$$\frac{80(c^3+16)}{c^6\sigma'(233c)}\beta B + 4.$$

(iii) When $\tau < Be^{-B}$**.** Recalling (52), we have

$$\max_{\pi:\,\mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau} \mathbb{E}_{x\sim\pi}[r(x)] \ge \frac{c^6e^{-B}}{8(c^3+16)} + r_{\min}.$$

Combining this with the result (54) of Lemma 14,

$$\frac{1}{2}\beta(\sigma'(233c))^{-1}\mathbb{E}_{u\sim\mathcal{D}}[u(x)] \ge \frac{c^6e^{-B}}{8(c^3+16)}.$$

Combining this with (50), we have that

$$\mathbb{E}_{u\sim\mathcal{D},x\sim\pi^*}[u(x)] - 4\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{RLHF}}}[u(x)] \le 20\tau \le \frac{80(c^3+16)}{c^6\sigma'(233c)}\beta e^B\tau\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{RLHF}}}[u(x)],$$

which implies that the distortion is bounded by

$$\frac{80(c^3+16)}{c^6\sigma'(233c)}\beta e^B\tau + 4.$$

□

B.6 Proof of Corollary 3

Since $\log \frac{\mu(x)}{\pi_{\mathrm{ref}}(x)} = \log \frac{\mu(x)}{(1-e^{-\lambda})\pi_{\mathrm{base}}(x)+e^{-\lambda}\mu(x)} \leq \log \lambda$ for all $x \in A$, according to Theorem 2 and the definition of distortion, we have

$$\frac{\max_{\pi\colon \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\leq\tau} \mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[u(x)]}{\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{RLHF}}}[u(x)]} \leq \beta C_2(\lambda+1)+4.$$

Let π^* denote a maximizer of $\max_{\pi\colon \mathrm{KL}(\pi\|\pi_{\mathrm{base}})\leq\tau} \mathbb{E}_{u,x\sim\pi}[u(x)]$. To replace $\max_{\pi\colon \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\leq\tau} \mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[u(x)]$ by $\max_{\pi\colon \mathrm{KL}(\pi\|\pi_{\mathrm{base}})\leq\tau} \mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[u(x)]$, we consider $\pi' = (1-e^{-\lambda})\pi^* + e^{-\lambda}\mu$, and we have that

$$\max_{\pi\colon \mathrm{KL}(\pi\|\pi_{\mathrm{base}})\leq\tau} \mathbb{E}_{u,x\sim\pi}[u(x)] = \mathbb{E}_{u\sim\mathcal{D},x\sim\pi^*}[u(x)] \leq \frac{1}{1-e^{-\lambda}}\mathbb{E}_{u\sim\mathcal{D},x\sim\pi'}[u(x)].$$

Also, according to Lemma 19 ($P=\pi^*, Q=\pi_{\mathrm{base}}, R=\mu, \lambda=e^{-\lambda}$), π' satisfies $\mathrm{KL}(\pi'\|\pi_{\mathrm{ref}}) \leq (1-e^{-\lambda})\tau \leq \tau$, which implies that

$$\mathbb{E}_{u\sim\mathcal{D},x\sim\pi'}[u(x)] \leq \max_{\pi\colon \mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\leq\tau} \mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[u(x)].$$

Combining these three bounds above yields

$$\frac{\max_{\pi\colon \mathrm{KL}(\pi\|\pi_{\mathrm{base}})\leq\tau} \mathbb{E}_{u,x\sim\pi}[u(x)]}{\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{RLHF}}}[u(x)]} \leq \frac{\beta C_2(\lambda+1)+4}{1-e^{-\lambda}},$$

which concludes the proof. □

B.7 Auxiliary Lemmas

B.7.1 First-order Optimality

In our upper-bound analysis, we frequently use the following first-order condition to connect the reward and the average utility. The right-hand side of (57) corresponds to the so-called (expected) Borda score. While a similar relationship between the reward and the Borda score has already appeared in prior work (Siththaranjan et al., 2023; Rajkumar & Agarwal, 2014), we formalize it here as a lemma in a form convenient for our analysis and provide a complete proof.

Lemma 17. *Let $\bar{r}(x)$ be the maximizer of the population log-likelihood*

$$\mathcal{L}(\tilde{r}) = \mathbb{E}_{u\sim\mathcal{D},x,y\sim\mu}\big[\sigma(u(x)-u(y))\log(\sigma(\tilde{r}(x)-\tilde{r}(y)))\big] \tag{56}$$

such that $\bar{r}(x)=0$ for an alternative x with utility $u(x)\equiv 0$ and infinitesimal sampling probability $\mu(x)$. Then, for each $x=1,\dots,m$, it holds that

$$\mathbb{E}_{y\sim\mu}\big[\sigma(\bar{r}(x)-\bar{r}(y))\big] = \mathbb{E}_{u\sim\mathcal{D},y\sim\mu}\big[\sigma(\beta(u(x)-u(y)))\big]. \tag{57}$$

Moreover,

$$\mathbb{E}_{x\sim\mu}\big[\sigma(-\bar{r}(x))\big] = \mathbb{E}_{u\sim\mathcal{D},x\sim\mu}\big[\sigma(\beta(-u(x)))\big]. \tag{58}$$

Proof. To account for boundary conditions, we add a new alternative $x=m+1$ such that $u(m+1)\equiv 0$ and $\mu(m+1)$ is negligibly small. The effect of this modification on the rewards of the original alternatives vanishes as $\mu(m+1)\to 0$.

Let us prove (57) for $x=1,\dots,m$. Note that $\frac{\mathrm{d}}{\mathrm{d}t}\log\sigma(t) = 1-\sigma(t)$ and $1-\sigma(t)=\sigma(-t)$. By differentiating (56) with $r(z)$ ($z=2,\dots,m$), we have that

$$\begin{aligned}
&\frac{\mathrm{d}}{\mathrm{d}r(z)}\mathbb{E}_{u\sim\mathcal{D},x,y\sim\mu}\Big[\sigma(\beta(u(x)-u(y)))\log(\sigma(r(x)-r(y)))\Big] \\
&= \mathbb{E}_{u\sim\mathcal{D},x,y\sim\mu}\big[\sigma(\beta(u(x)-u(y)))\sigma(r(y)-r(x))(\mathbb{1}[x=z]-\mathbb{1}[y=z])\big] \\
&= \mu(z)\mathbb{E}_{u\sim\mathcal{D},y\sim\mu}\big[\sigma(\beta(u(z)-u(y)))\sigma(r(y)-r(z))\big] \\
&\quad - \mu(z)\mathbb{E}_{u\sim\mathcal{D},x\sim\mu}\big[\sigma(\beta(u(x)-u(z)))\sigma(r(z)-r(x))\big] \\
&= \mu(z)\mathbb{E}_{u\sim\mathcal{D},y\sim\mu}\big[\sigma(\beta(u(z)-u(y)))\big(1-\sigma(r(z)-r(y))\big)\big] \\
&\quad - \mu(z)\mathbb{E}_{u\sim\mathcal{D},y\sim\mu}\big[\big(1-\sigma(\beta(u(z)-u(y)))\big)\sigma(r(z)-r(y))\big] \\
&= \mu(z)\Big(\mathbb{E}_{u\sim\mathcal{D},y\sim\mu}\big[\sigma(\beta(u(z)-u(y)))\big] - \mathbb{E}_{u\sim\mathcal{D},y\sim\mu}\big[\sigma(r(z)-r(y))\big]\Big).
\end{aligned}$$

Because $\mu(z) > 0$, $\frac{\mathrm{d}\mathcal{L}(r)}{\mathrm{d}r(z)}\big|_{r=\bar{r}} = 0$ implies (57) for all $x = 1, \ldots, m$.

Finally, because of the symmetry,

$$\mathbb{E}_{x\sim\mu}\big[\mathbb{E}_{y\sim\mu}\big[\sigma(\bar{r}(x) - \bar{r}(y))\big]\big] = \mathbb{E}_{x\sim\mu}\mathbb{E}_{u\sim\mathcal{D},y\sim\mu}\big[\sigma(\beta(u(x) - u(y)))\big]\big] = \frac{1}{2}.$$

From this, when (57) holds for $x = 1, \ldots, m$, (57) must also be true for $x = m + 1$. Equation (57) for the alternative $m + 1$, which has $\bar{r}(m + 1) = 0$ and $u(m + 1) \equiv 0$, is written as

$$\mathbb{E}_{x\sim\mu}\big[\sigma(-\bar{r}(x))\big] = \mathbb{E}_{u\sim\mathcal{D},x\sim\mu}\big[\sigma(\beta(-u(x)))\big].$$

Therefore, we obtain (58). □

The first-order optimality condition implies that the ordering of the Borda scores coincides with that of the estimated rewards.

Lemma 18 (Theorem 3.1 of Siththaranjan et al. (2023))**.** *For any $x, x' \in A$, we have*

$$\mathbb{E}_{u\sim\mathcal{D},y\sim\mu}\big[\sigma(\beta(u(x) - u(y)))\big] > \mathbb{E}_{u\sim\mathcal{D},y\sim\mu}\big[\sigma(\beta(u(x') - u(y)))\big],$$

if and only if $\bar{r}(x) > \bar{r}(x')$.

Proof. According to the optimality condition (57),

$$\mathbb{E}_{u\sim\mathcal{D},y\sim\mu}\big[\sigma(\beta(u(x) - u(y)))\big] > \mathbb{E}_{u\sim\mathcal{D},y\sim\mu}\big[\sigma(\beta(u(x') - u(y)))\big]$$

holds if and only if

$$\mathbb{E}_{y\sim\mu}\big[\sigma(\bar{r}(x) - \bar{r}(y))\big] > \mathbb{E}_{y\sim\mu}\big[\sigma(\bar{r}(x') - \bar{r}(y))\big].$$

Since the sigmoid function is strictly increasing, the latter inequality holds if and only if $\bar{r}(x) > \bar{r}(x')$. □

B.7.2 Proof of Technical Lemmas

Below we provide the proofs of the technical lemmas. For clarity, we restate each lemma before presenting its proof.

Lemma 6. *Let $B, \tau > 0$ be constants. Let π_1, π_2 be probability measures with $\pi_2 \ll \pi_1$ and $\log \frac{\mathrm{d}\pi_2}{\mathrm{d}\pi_1} \le B$. Then, the probability measure π_λ defined by $\pi_\lambda := (1 - \lambda)\pi_1 + \lambda\pi_2$ with $\lambda := \min\{1, B^{-1}\tau\}$ satisfies $\mathrm{KL}(\pi_\lambda\|\pi_1) \le \tau$.*

Proof. Let $f = \frac{\mathrm{d}\pi_2}{\mathrm{d}\pi_1}$ so $\log f \le B$ and

$$\mathrm{KL}(\pi_2\|\pi_1) = \int f \log f \mathrm{d}\pi_1 = \int \log f \mathrm{d}\pi_2 \le B.$$

Then, since $\frac{\mathrm{d}\pi_\lambda}{\mathrm{d}\pi_1} = (1 - \lambda) + \lambda f$ and $\phi(t) = t \log t$ is convex for $t > 0$ with $\phi(1) = 0$,

$$\mathrm{KL}(\pi_\lambda\|\pi_1) \le (1 - \lambda)\phi(1) + \lambda \int \phi(f)\mathrm{d}\pi_1 = \lambda\mathrm{KL}(\pi_2\|\pi_1) \le \lambda B \le \tau.$$

□

Lemma 7. *Let π and π' be probability measures such that $\pi' \ll \pi$ and define the tail regions $A_- := \{x\colon \frac{\mathrm{d}\pi'}{\mathrm{d}\pi}(x) \le \frac{1}{2}\}$ and $A_+ := \{x\colon \frac{\mathrm{d}\pi'}{\mathrm{d}\pi}(x) \ge 2\}$. Then, the total variation restricted to A_- is bounded by*

$$\frac{1}{2}\int_{A_-} |\mathrm{d}\pi' - \mathrm{d}\pi| \le (\log(e/2))^{-1}\mathrm{KL}(\pi'\|\pi). \tag{12}$$

Also, the total variation restricted to A_+ is bounded by

$$\frac{1}{2}\int_{A_+} |\mathrm{d}\pi' - \mathrm{d}\pi| \le (\log(4/e))^{-1}\mathrm{KL}(\pi'\|\pi). \tag{13}$$

Proof. Rewrite the KL divergence as

$$\mathrm{KL}(\pi'\|\pi) = \int f \log f \mathrm{d}\pi = \int \phi(f)\mathrm{d}\pi,$$

where $f = \frac{\mathrm{d}\pi'}{\mathrm{d}\pi}$ and $\phi(f) = f\log f - f + 1$.

(i) Restriction to A_-. The function $\phi(f)$ is nonnegative and decreasing on $(0, 1]$. Hence, on A_-, $\phi(f) \geq \phi(\frac{1}{2})$, and

$$\mathrm{KL}(\pi'\|\pi) \geq \int_{A_-} \phi(f)\mathrm{d}\pi \geq \phi\Big(\frac{1}{2}\Big)\pi(A_-). \tag{59}$$

Moreover, since $f \leq \frac{1}{2}$ on A_-, we have that

$$\int_{A_-} |\mathrm{d}\pi' - \mathrm{d}\pi| = \int_{A_-} |f-1|\mathrm{d}\pi = \int_{A_-} (1-f)\mathrm{d}\pi \leq \pi(A_-). \tag{60}$$

Combining (59) and (60) gives

$$\int_{A_-} |\mathrm{d}\pi' - \mathrm{d}\pi| \leq \phi\Big(\frac{1}{2}\Big)^{-1}\mathrm{KL}(\pi'\|\pi) = 2(\log(e/2))^{-1}\mathrm{KL}(\pi'\|\pi),$$

which yields (12).

(ii) Restriction to A_+. Since $\phi(f)$ is nonnegative and $\phi(f) \geq \frac{\phi(2)}{2}f = \frac{\log(4/e)}{2}f$ holds on $[2, \infty)$, we have that

$$\mathrm{KL}(\pi'\|\pi) \geq \int_{A_+} \phi(f)\mathrm{d}\pi \geq \int_{A_+} \frac{\log(4/e)}{2} f\mathrm{d}\pi \geq \frac{\log(4/e)}{2}\int_{A_+} (f-1)\mathrm{d}\pi. \tag{61}$$

Moreover, since $f \geq 2$ on A_+, we obtain

$$\int_{A_+} |\mathrm{d}\pi' - \mathrm{d}\pi| = \int_{A_+} |f-1|\mathrm{d}\pi = \int_{A_+} (f-1)\mathrm{d}\pi. \tag{62}$$

Combining (61) and (62) and dividing by 2 gives

$$\frac{1}{2}\int_{A_+} |\mathrm{d}\pi' - \mathrm{d}\pi| \leq \frac{1}{\log(4/e)}\mathrm{KL}(\pi'\|\pi),$$

which yields (13) and thus completes the proof. □

Lemma 10. *Let $a, b > 0$ be arbitrary constants, and assume that $s, t > 0$ satisfy $t - s \geq -a$. Then the sigmoid function $\sigma(t) = 1/(1+e^{-t})$ satisfies*

$$\sigma(t-s) - \sigma(-s) \geq \sigma'(a+b)\min\{t, b\}.$$

Proof. **(i) When $t \leq b$.** By the mean value theorem, there exists some $p \in [-s, t-s]$ such that

$$\sigma(t-s) - \sigma(-s) = \sigma'(p)t = \sigma'(p)\min\{t, b\}.$$

Since $t \leq b$ and $t - s \geq -a$, we have $-a-b \leq -a \leq t-s \leq b \leq a+b$, and hence $p \in [-a-b, a+b]$. Moreover, since $\sigma'(t)$ is decreasing on $t \geq 0$, $\sigma'(p) \geq \sigma'(a+b)$, and the claim follows.

(ii) When $t > b$. By assumption, the length of the interval $[-s, t-s]$ has length $t > b$. We show that it contains a subinterval of length at least b lying inside $[-a-b, a+b]$. Since $t - s \geq -a$, its right endpoint is larger than $-a-b$ by at least b. Since the left endpoint $-s < 0$, the left endpoint is smaller than $a+b$ by at least b. Therefore, the intersection of the intervals $[-s, t-s]$ and $[-a-b, a+b]$, which is $[\max\{-s, -a-b\}, \min\{t-s, a+b\}]$, has length at least b. Therefore, by the mean value theorem, there exists some $p \in [-a-b, a+b]$ such that

$$\begin{aligned}\sigma(t-s) - \sigma(-s) &\geq \sigma(\min\{t-s, a+b\}) - \sigma(\max\{-s, -a-b\}) \\ &= \sigma'(p)(\min\{t-s, a+b\} - \max\{-s, -a-b\}) \\ &\geq \sigma'(p)b.\end{aligned}$$

Because $p \in [-a-b, a+b]$, $\sigma'(p) \geq \sigma'(a+b)$, which completes the proof. □

Lemma 11. *For any $a \geq 0$ and $s, t \geq 0$, the sigmoid function $\sigma(t) = 1/(1+e^{-t})$ satisfies*

$$\frac{1}{2} + \frac{1-a}{4}\min\{t, a\} - \frac{1}{4}\min\{s, 4\} \leq \sigma(t-s) \leq \frac{1}{2} + \frac{1}{4}\min\{t, 4\} - \frac{1-a}{4}\min\{s, a\}. \quad (34)$$

Proof. We first prove the left-hand inequality. When $s > 4$, the left-hand side of (34) can be bounded as $\frac{1}{2} + \frac{1-a}{4}\min\{t,a\} - \frac{1}{4}\min\{s,4\} < \frac{1}{2} + \frac{(1-a)a}{4} - \frac{1}{4}\cdot 4 \leq \frac{1}{2} + \frac{1}{16} - 1 < 0$, while $\sigma(t-s)$ is always positive, so the left-hand inequality of (34) is true. Therefore, we focus on the case $s \leq 4$ in the following.

Recall that $\sigma(0) = \frac{1}{2}$ and $\sigma'(w) = \frac{e^{-w}}{(1+e^{-w})^2}$. For any $w \geq 0$, since $1 - w \leq e^{-w}$, we have $1 - w \leq (1+w)e^{-w}$, and hence $\frac{1-e^{-w}}{1+e^{-w}} \leq w$. Combining this with $\frac{1-e^{-w}}{1+e^{-w}} \in [0,1]$, we further obtain $(\frac{1-e^{-w}}{1+e^{-w}})^2 \leq w$. Substituting this bound, we obtain

$$\sigma'(w) = \frac{e^{-w}}{(1+e^{-w})^2} = \frac{1}{4}\left(1 - \left(\frac{1-e^{-w}}{1+e^{-w}}\right)^2\right) \geq \frac{1}{4}(1-w). \quad (63)$$

(i) When $t - s \geq 0$. Letting $p = \min\{t-s, a\} \geq 0$, we have

$$\sigma(p) = \sigma(0) + \int_0^p \sigma'(z)\mathrm{d}z = \frac{1}{2} + \int_0^p \sigma'(z)\mathrm{d}z \geq \frac{1}{2} + p\sigma'(p) \geq \frac{1}{2} + p\sigma'(a),$$

where we used the fact that $\sigma'(z)$ is decreasing for $z \geq 0$, which implies that $\sigma'(z) \geq \sigma'(p) \geq \sigma'(a)$ when $z \leq p \leq a$. By using (63) with $w = a$, we continue as follows:

$$\begin{aligned}
\sigma(p) &\geq \frac{1}{2} + \sigma'(a)\min\{t-s, a\} \\
&\geq \frac{1}{2} + \frac{1-a}{4}\min\{t-s, a\} \\
&\geq \frac{1}{2} + \frac{1-a}{4}\min\{t, a\} - \frac{1-a}{4}s \\
&\geq \frac{1}{2} + \frac{1-a}{4}\min\{t, a\} - \frac{s}{4}, \quad (64)
\end{aligned}$$

where we used $\min\{t-s, a\} \geq \min\{t, a\} - s$ in the third inequality. When $s \leq 4$, we have $\frac{s}{4} = \frac{1}{4}\min\{s, 4\}$ and (64) immediately yields the left-hand inequality of (34).

(ii) When $t - s < 0$. As $\sigma'(w) \leq \frac{1}{4}$ for all $w \in \mathbb{R}$,

$$\sigma(t-s) \geq \sigma(0) + \frac{1}{4}(t-s) = \frac{1}{2} + \frac{1}{4}(t-s) \geq \frac{1}{2} + \frac{1-a}{4}\min\{t, a\} - \frac{1}{4}s, \quad (65)$$

where we used $t \geq (1-a)\min\{t, a\}$ for the second inequality. As above, when $s \leq 4$, we have $\frac{s}{4} = \frac{1}{4}\min\{s, 4\}$ and (65) immediately yields the left-hand inequality of (34).

Finally, the right-hand inequality of (34) is obtained by applying the left-hand inequality of (34) to $\sigma(s-t)$ and using the identity $\sigma(t-s) = 1 - \sigma(s-t)$:

$$\begin{aligned}
\sigma(t-s) &= 1 - \sigma(s-t) \\
&\leq 1 - \Big(\frac{1}{2} + \frac{1-a}{4}\min\{s, a\} - \frac{1}{4}\min\{t, 4\}\Big) \\
&= \frac{1}{2} - \frac{1}{4}\min\{t, 4\} + \frac{1-a}{4}\min\{s, a\}.
\end{aligned}$$

□

Lemma 19. *Let P, Q, R be probability measures. For $\lambda \in [0,1]$, define*

$$P_\lambda := (1-\lambda)P + \lambda R, \qquad Q_\lambda := (1-\lambda)Q + \lambda R.$$

Then

$$\mathrm{KL}(P_\lambda \| Q_\lambda) \;\leq\; (1-\lambda)\,\mathrm{KL}(P\|Q).$$

Proof. Introduce an auxiliary Bernoulli random variable $Z \in \{0, 1\}$ with

$$\mathbb{P}[Z = 0] = 1 - \lambda, \qquad \mathbb{P}[Z = 1] = \lambda.$$

Now define two probability measures $\tilde{P}$ and $\tilde{Q}$ by

$$\tilde{P}(\mathrm{d}z, \mathrm{d}x) := (1 - \lambda)\delta_0(\mathrm{d}z)P(\mathrm{d}x) + \lambda\delta_1(\mathrm{d}z)R(\mathrm{d}x),$$

and

$$\tilde{Q}(\mathrm{d}z, \mathrm{d}x) := (1 - \lambda)\delta_0(\mathrm{d}z)Q(\mathrm{d}x) + \lambda\delta_1(\mathrm{d}z)R(\mathrm{d}x).$$

Since the two measures differ only on the event $\{Z = 0\}$, we obtain

$$\mathrm{KL}(\tilde{P}\|\tilde{Q}) = \int \log \frac{\mathrm{d}\tilde{P}}{\mathrm{d}\tilde{Q}}(z, x)\mathrm{d}\tilde{P}(z, x) = (1 - \lambda) \int \log \frac{\mathrm{d}P}{\mathrm{d}Q}(x)\mathrm{d}P(x) = (1 - \lambda)\,\mathrm{KL}(P\|Q).$$

Next, let π be the projection map $\pi(z, x) = x$. The pushforwards of $\tilde{P}$ and $\tilde{Q}$ under π are precisely

$$\pi_{\#}\tilde{P} = P_\lambda, \qquad \pi_{\#}\tilde{Q} = Q_\lambda.$$

Therefore, by the data-processing inequality for KL divergence,

$$\mathrm{KL}(P_\lambda\|Q_\lambda) = \mathrm{KL}(\pi_{\#}\tilde{P}\|\pi_{\#}\tilde{Q}) \le \mathrm{KL}(\tilde{P}\|\tilde{Q}) = (1 - \lambda)\,\mathrm{KL}(P\|Q).$$

This proves the claim. □

C Removal of Reward Clipping

The purpose of reward clipping (3) is to focus optimization on the informative regime of the reward. The interval $r_{\min} \le \bar{r}(x) \le r_{\max}$ corresponds to the region where "majority" alternatives preferred with probability around $\frac{1}{2}$ are concentrated. Outside this interval, reward estimation can incur substantial errors. Reward clipping is therefore designed to prevent the KL budget from being spent on optimizing such uninformative regions.

However, we do not believe that the absence of such clipping leads to a catastrophic deterioration in distortion. Indeed, even without imposing reward clipping, we can still derive a polynomial distortion bound, at least in the case $\mu = \pi_{\mathrm{ref}}$.

Theorem 11. *Suppose that $\pi_{\mathrm{ref}} = \mu$. Let $\bar{r}(x)$ be the maximizer of the population log-likelihood (1), and use this $\bar{r}$ as the reward without clipping. Thus, the RLHF distribution is defined as*

$$\pi_{\mathrm{RLHF}} = \arg \max_{\pi:\ \mathrm{KL}(\pi\|\mu) \le \tau} \mathbb{E}_{u\sim\mathcal{D}, x\sim\pi}[\bar{r}(x)].$$

Then, the distortion of this distribution π_{RLHF} is bounded by

$$\max\left\{\frac{16(c^3 + 16)\beta}{c^7}, \frac{3\max\{\beta, c^{-1}\beta^2\} + 4}{c(2\sigma'(233c))^{-2}}\right\}, \tag{66}$$

for any $0 < c < \frac{1}{316}$.

The proof is divided into two cases depending on whether $\mathbb{P}_{u\sim\mathcal{D},\, x\sim\mu}[\beta u(x) > c] > c^2$ holds. When this condition holds, an $O(\beta)$ bound has already been established in Theorem 8 for the case $\pi_{\mathrm{ref}} = \mu$.

On the other hand, when $\mathbb{P}_{u\sim\mathcal{D},\, x\sim\mu}[\beta u(x) > c] \le c^2$, removing reward clipping requires bounding $\bar{r}(x)$ itself in Lemma 3, rather than $\min\{\bar{r}(x), R\}$. Therefore, if $\bar{r}(x) \ge R$, the following modification is required:

$$\bar{r}(x) \le \big(R^{-1} \max_{x'\in A} \bar{r}(x')\big) \times \underbrace{\min\{\bar{r}(x), R\}}_{=R} \underbrace{\lesssim}_{\text{Lemma 3}} \big(R^{-1} \max_{x'\in A} \bar{r}(x')\big) \times \beta\mathbb{E}_{u\sim\mathcal{D}}[u(x)].$$

Because Lemma 14 implies that $R = \Theta(1)$, it suffices to bound $\max_x \bar{r}(x)$. However, even if $\beta u \in [0, \beta]^m$, it is not immediate that the estimate $\bar{r}$ obtained from the mixture of preferences generated by heterogeneous utilities also lies in $[0, \beta]^m$. This non-expansiveness of the maximum likelihood estimator for mixtures of Bradley–Terry models can be summarized as follows.

Lemma 20. *Let μ be a distribution on $\{1, \dots, m\}$, and let $\mathcal{D}$ be an arbitrary distribution on $[0, \beta]^m$. Define*

$$\mathcal{L}(\tilde{r}) = \mathbb{E}_{u\sim\mathcal{D}, x, y\sim\mu}\big[\sigma(u(x) - u(y)) \log \sigma(\tilde{r}(x) - \tilde{r}(y))\big].$$

for $\tilde{r} \in \mathbb{R}^m$. Then the maximizer $\bar{r}$ of $\mathcal{L}$ satisfies

$$\max_x \bar{r}(x) - \min_x \bar{r}(x) \le \beta.$$

Proof. Let c be a constant satisfying $0 < c < \frac{1}{316}$. If $\mathbb{P}_{u\sim\mathcal{D}, x\sim\mu}[\beta u(x) > c] \ge c^2$, Theorem 8 bounds the distortion by $\frac{16(c^3+16)\beta}{c^7}$. Therefore, it suffices to show that the latter term in the distortion bound (66) arises when $\mathbb{P}_{u\sim\mathcal{D},\, x\sim\mu}[\beta u(x) > c] \le c^2$.

When $\mathbb{P}_{u\sim\mathcal{D}, x\sim\mu}[\beta u(x) > c] \le c^2$, we consider how Lemma 14 is modified. If $\bar{r}(x) \le c$, it implies that $\min\{\bar{r}(x), r_{\max}\} = \bar{r}(x)$ from Lemma 9. This implies that

$$\bar{r}(x) \le 2\beta(\sigma'(233c))^{-1}\mathbb{E}_{u\sim\mathcal{D}}[u(x)] \tag{67}$$

On the other hand, if $\bar{r}(x) \ge c$, then $\min\{\bar{r}(x), r_{\max}\} \ge c$ [2] by Lemma 9, and thus Lemma 14 implies that

$$c \le 2\beta(\sigma'(233c))^{-1}\mathbb{E}_{u\sim\mathcal{D}}[u(x)] \Leftrightarrow 1 \le \frac{2\beta}{c\sigma'(233c)}\mathbb{E}_{u\sim\mathcal{D}}[u(x)].$$

By letting β' be the uniform upper bound on $\bar{r}(x)$, we have

$$\bar{r}(x) \le \beta' \le \frac{2\beta\beta'}{c\sigma'(233c)}\mathbb{E}_{u\sim\mathcal{D}}[u(x)]. \tag{68}$$

By combining (67) and (68), Lemma 14 is modified to

$$\bar{r}(x) \le \frac{2\max\{\beta, c^{-1}\beta'\beta\}}{\sigma'(233c)}\mathbb{E}_{u\sim\mathcal{D}}[u(x)].$$

Therefore, the key step is to bound the quantity β'. Lemma 20 shows that β' is bounded by β, and hence

$$\bar{r}(x) \le \frac{2\max\{\beta, c^{-1}\beta^2\}}{\sigma'(233c)}\mathbb{E}_{u\sim\mathcal{D}}[u(x)]. \tag{69}$$

Using this bound in place of Lemma 14, the proof of Theorem 7 is unaffected by removing clipping, except that β is replaced by $\max\{\beta, c^{-1}\beta^2\}$. We therefore obtain the desired bound by making this substitution in Theorem 7 with $B = 0$. □

Remark 1. *Before proceeding to the proof of Lemma 20, we discuss the current limitations in removing reward clipping in Theorem 2 for a general reference policy $\pi_{\text{ref}} \neq \mu$, without incurring the additional factor of β. First, for the case $\mathbb{P}_{u\sim\mathcal{D},\, x\sim\mu}[\beta u(x) > c] \le c^2$, the proof of Theorem 2 is valid for a general π_{ref} without reward clipping, except that we need to use $\bar{r}(x) \lesssim \beta^2\mathbb{E}_{u\sim\mathcal{D}}[u(x)]$ (69) in place of Lemma 14, which incurs the additional factor of β. Regarding this, the following counterexample shows that the bound $\bar{r}(x) \lesssim \beta^2\mathbb{E}_{u\sim\mathcal{D}}[u(x)]$ is tight: for $m = 3$, let $u = (0, 1, 1)$ with probability $\frac{1}{\beta}$ and $u = (0, 1, 0)$ with probability $1 - \frac{1}{\beta}$, and take $\mu = (1 - c^2,\, c^2 - \epsilon,\, \epsilon)$ with $\epsilon \to 0$. Then, one can see that $\bar{r}(1) = 0$, $\bar{r}(2) \xrightarrow{\epsilon\to 0} \beta$, and $\bar{r}(3) \xrightarrow{\epsilon\to 0} \beta(1 - o_\beta(1))$, as well as $\mathbb{E}_{u\sim\mathcal{D}}[u(3)] \simeq \beta^{-1}$, which implies that $\bar{r}(3) \lesssim \beta^2\mathbb{E}_{u\sim\mathcal{D}}[u(3)]$. Consequently, a linear bound without the additional factor of β is not obtained at least from the point-wise bound of $\bar{r}(x)$ by $\mathbb{E}_{u\sim\mathcal{D}}[u(x)]$; in other words, if it were to be obtained, it would require an argument that accounts for the interaction among alternatives.*

On the other hand, when $\mathbb{P}_{u\sim\mathcal{D},\, x\sim\mu}[\beta u(x) > c] < c^2$, the reference policy already achieves the optimal $O(\beta)$ distortion when $\pi_{\text{ref}} = \mu$, which yields the optimal distortion of π_{RLHF}. However, once a general reference policy $\pi_{\text{ref}} \neq \mu$ is allowed, this property of π_{ref} no longer holds, which prevents the current proof from carrying over directly.

[2] While we are not using reward clipping here, we treat $r_{\max}$ as the quantity defined in (29), with $r_{\max} = r_{\min} + 2c$, to utilize the previous lemmas.

C.1 Non-expansiveness of Mixture of Bradley–Terry Models

To prove Lemma 20, we introduce an auxiliary functional $\Phi(v; v')$ for $v, v' \in \mathbb{R}^m$. Specifically, let $g_v = \nabla \mathcal{L}(v)$ and $D_{v'} = \nabla^2 \mathcal{L}(v')$, and consider a vector $h_{v'} \in \mathbb{R}^m$ satisfying the following condition.

$$h_{v'}(1) = 0,\ h_{v'}(m) = 1,\ \text{and } (D_{v'} h_{v'})(x) = 0 \text{ for all } x = 2, \dots, m-1.$$

Using this vector $h_{v'}$, the functional $\Phi(v; v')$ is defined as

$$\Phi(v; v') = h_{v'}^\top g_v.$$

We prove Lemma 20 by contradiction by supposing that $\max_x \bar{r}(x) - \min_x \bar{r}(x) = \beta' > \beta$ for the maximizer $\bar{r}$ of $\mathcal{L}$, and defining $\bar{r}' = \frac{\beta}{\beta'} \bar{r}$. Then, by Lemma 21, the functional $\Phi(v; v')$ attains its maximum at $v = v'$ for $v, v' \in [0, \beta]^m$. Moreover, by the optimality condition of $\bar{r}$ in Lemma 17, we have $\mathbb{E}_{u \sim \mathcal{D}}[g_{\beta u}] = g_{\bar{r}}$. Hence,

$$\Phi(r'; r') \geq \mathbb{E}_{u \sim \mathcal{D}}[\Phi(\beta u; r')] = h_{r'}^\top \mathbb{E}_{u \sim \mathcal{D}}[g_{\beta u}] = h_{r'}^\top g_{\bar{r}} = \Phi(\bar{r}; r').$$

On the other hand, we can show that $\Phi(r'; r') < \Phi(\bar{r}; r')$ by comparing the two quantities term by term. Therefore, we obtain a contradiction, which implies that $\max_x \bar{r}(x) - \min_x \bar{r}(x) \leq \beta$.

Proof of Lemma 20. If we redefine the alternatives by splitting each alternative into multiple ones, the value of $\bar{r}(x)$ remains unchanged between the original and the resulting alternatives. Therefore, the quantity $\max_x \bar{r}(x) - \min_x \bar{r}(x)$ is invariant under such transformations. Building on this, by splitting the alternatives so that each has (approximately) equal sampling probability, the general case can be approximated arbitrarily well by the case where μ is uniform. Therefore, in the following we assume that μ is the uniform distribution. When μ is uniform, relabeling the coordinates does not change the problem, and we may assume without loss of generality that

$$r(1) \leq r(2) \leq \cdots \leq r(m),$$

in the following.

For $v \in \mathbb{R}^m$, define $g_v \in \mathbb{R}^m$ as

$$g_v(x) = \mathbb{E}_{y \sim \mu}[\sigma(v(x) - v(y))] = \frac{1}{m} \sum_{y=1}^{m} \sigma(v(x) - v(y)). \tag{70}$$

According to Lemma 17, the maximizer $\bar{r}$ is characterized by

$$g_r(x) = \mathbb{E}_{u \sim \mathcal{D}}[g_{\beta u}(x)] \quad (x = 1, \dots, m). \tag{71}$$

For later use, we denote the Jacobian of g_v with respect to $(v(1), \dots, v(m))$ by D_v, i.e.,

$$D_v(x, y) = \frac{\mathrm{d}}{\mathrm{d}v(y)} g_v(x) = \begin{cases} -\frac{1}{m} \sigma'(v(x) - v(y)) & (y \neq x) \\ \frac{1}{m} \sum_{y' \neq x} \sigma'(v(x) - v(y')) & (y = x). \end{cases} \tag{72}$$

Suppose, for contradiction, that $\beta' := \max_x \bar{r}(x) - \min_x \bar{r}(x) = \bar{r}(m) - \bar{r}(1) > \beta$. Define

$$r'(x) := \frac{\beta}{\beta'} \big(\bar{r}(x) - \bar{r}(1)\big) \quad (x = 1, \dots, m).$$

Then

$$0 = r'(1) \leq r'(2) \leq \cdots \leq r'(m) = \beta,$$

and in particular $r' \in [0, \beta]^m$.

Also, for $v' \in \mathbb{R}^m$, we introduce an auxiliary vector $h_{v'} \in \mathbb{R}^m$ such that

$$h_{v'}(1) = 0,\ h_{v'}(m) = 1,\ \text{and } (D_{v'} h_{v'})(x) = 0 \text{ for all } x = 2, \dots, m-1. \tag{73}$$

(Note that, because $(D_{v'} h_{v'})(x) = 0 \Leftrightarrow \sum_{y \neq x} \sigma'(v'(x) - v'(y)) h_{v'}(x) = \sum_{y \neq x} \sigma'(v'(x) - v'(y)) h_{v'}(y)$ from (72), we can regard the definition of $h_{v'}$ as the Dirichlet problem, and the existence and uniqueness of the solution follow from this perspective.) Then, we define

$$\Phi(v; v') = h_{v'}^\top g_v.$$

By Lemma 21, for every $v, v' \in [0,\beta]^m$ with $v'(1) = 0 \le v'(2) \le \cdots \le v'(m) = \beta$, we have

$$\Phi(v; v') \le \Phi(v'; v').$$

Since βu takes values in $[0,\beta]^m$ when $u \sim \mathcal{D}$ and $0 = r'(1) \le r'(2) \le \cdots \le r'(m) = \beta$, it follows that

$$h_{r'}^\top \mathbb{E}_{u\sim\mathcal{D}}[g_{\beta u}] = \mathbb{E}_{u\sim\mathcal{D}}[\Phi(\beta u; r')] \le \Phi(r'; r') = h_{r'}^\top g_{r'}.$$

Using (71), we obtain

$$h_{r'}^\top g_{\bar{r}} \le h_{r'}^\top g_{r'} \Leftrightarrow \Phi(\bar{r}; r') \le \Phi(r'; r'). \tag{74}$$

On the other hand,

$$\begin{aligned} h_{r'}^\top g_{\bar{r}} - h_{r'}^\top g_{r'} &= \frac{1}{m} \sum_{x,y=1}^{m} h_{r'}(x)\big(\sigma(\bar{r}(x) - \bar{r}(y)) - \sigma(r'(x) - r'(y))\big) \\ &= \frac{1}{m} \sum_{1\le x<y\le m} (h_{r'}(x) - h_{r'}(y))\big(\sigma(\bar{r}(x) - \bar{r}(y)) - \sigma(r'(x) - r'(y))\big). \end{aligned} \tag{75}$$

By Lemma 22, we have $h_{r'}(x) \le h_{r'}(y)$ whenever $x < y$. Moreover,

$$\bar{r}(x) - \bar{r}(y) = \frac{\beta'}{\beta}(r'(x) - r'(y)) \le r'(x) - r'(y) \le 0 \quad (x < y),$$

and since σ is increasing,

$$\sigma(\bar{r}(x) - \bar{r}(y)) \le \sigma(r'(x) - r'(y)) \quad (x < y).$$

Therefore every term in (75) is nonnegative. Furthermore, we have $h_{r'}(m) - h_{r'}(1) = 1$ and $\bar{r}(m) - \bar{r}(1) = \beta' > \beta = r'(m) - r'(1)$ implies that $\sigma(\bar{r}(1) - \bar{r}(m)) < \sigma(r'(1) - r'(m))$. Hence one of the terms in (75) is strictly positive, and consequently

$$h_{r'}^\top g_{\bar{r}} - h_{r'}^\top g_{r'} > 0 \Leftrightarrow \Phi(\bar{r}; r') > \Phi(r'; r').$$

This contradicts (74).

Therefore $\beta' > \beta$ is impossible, and we conclude that

$$\max_x \bar{r}(x) - \min_x \bar{r}(x) \le \beta.$$

□

Lemma 21. *Consider $v' \in [0,\beta]^m$ such that $0 = v'(1) \le v'(2) \le \cdots \le v'(m-1) \le v'(m) = \beta$. Let $D_{v'} \in \mathbb{R}^{m\times m}$ be as defined in Lemma 20, i.e.,*

$$D_{v'}(x,y) = \frac{\mathrm{d}}{\mathrm{d}v'(y)} g_{v'}(x) = \begin{cases} -\frac{1}{m}\sigma'(v'(x) - v'(y)) & (y \ne x) \\ \frac{1}{m}\sum_{y'\ne x} \sigma'(v'(x) - v'(y')) & (y = x), \end{cases} \tag{76}$$

and define $h_{v'} \in \mathbb{R}^m$ by

$$h_{v'}(1) = 0,\ h_{v'}(m) = 1,\ \textit{and}\ (D_{v'}h_{v'})(x) = 0\ \textit{for all}\ x = 2, \ldots, m,$$

following (73). Moreover, for $v \in [0,\beta]^m$, define $\Phi(v; v') = h_{v'}^\top g_v$, where $g_v \in \mathbb{R}^m$ is defined as

$$g_v(x) = \frac{1}{m}\sum_{y=1}^{m} \sigma(v(x) - v(y)),$$

as in (70).

Then, when v runs over $[0,\beta]^m$, this $\Phi(v; v')$ is maximized at $v = v'$.

Proof. Consider an arbitrary $v \in [0,\beta]^m$. If there exist $x, x' \in \{1,\dots,m\}$ with $x < x'$ and $v(x) > v(x')$, swapping the values of $v(x)$ and $v(x')$ does not decrease the value of $\Phi(v;v')$. Let v_{swap} denote the vector obtained by this swap. Then,

$$\Phi(v_{\mathsf{swap}};v') - \Phi(v;v') = (h_{v'}(x') - h_{v'}(x))(g_v(x) - g_v(x')). \tag{77}$$

According to Lemma 22, $h_{v'}(x') - h_{v'}(x) \geq 0$. Also, the monotonicity of σ implies that $\sigma(v(x) - v(y)) - \sigma(v(x') - v(y)) \geq 0$ for any $y \in A$, and

$$g_v(x) - g_v(x') = \frac{1}{m}\sum_{y=1}^{m}(\sigma(v(x)-v(y)) - \sigma(v(x')-v(y))) \geq 0.$$

By combining them, it holds that (77) is always nonnegative. Therefore, the maximum of $\Phi(v;v')$ is achieved in $C = \{v \in [0,\beta]^m \mid v(1) \leq v(2) \leq \cdots \leq v(m)\}$. Therefore, we will restrict our attention to C in the following.

From the definition of $\Phi(v;v')$, we have that

$$\begin{aligned}
\Phi(v;v') &= h_{v'}^\top g_v \\
&= \sum_{x=1}^{m} h_{v'}(x)\frac{1}{m}\sum_{y=1}^{m}\sigma(v(x)-v(y)) \\
&= \frac{1}{m}\sum_{1\leq x<y\leq m} h_{v'}(x)\sigma(v(x)-v(y)) + \frac{1}{m}\sum_{1\leq y<x\leq m} h_{v'}(x)(1-\sigma(v(x)-v(y))) + \frac{1}{2m}\sum_{x=1}^{m} h_{v'}(x) \\
&= \frac{1}{m}\sum_{1\leq x<y\leq m}(h_{v'}(x) - h_{v'}(y))\sigma(v(x)-v(y)) + (\text{a constant independent of } v).
\end{aligned}$$

Note that $v(x) - v(y) \leq 0$ for all $x < y$ for $v \in C$, and $h_{v'}(x) - h_{v'}(y) \leq 0$ according to Lemma 22. Because σ is convex if the domain is restricted to $t \geq 0$, and thus $\Phi(v;v')$ is concave in $v \in C$. Therefore,

$$\begin{aligned}
\Phi(v;v') &\leq \Phi(v';v') + \nabla\Phi(v';v')(v-v') \\
&\leq \Phi(v';v') + h_{v'}^\top D_{v'}(v-v').
\end{aligned} \tag{78}$$

Because $(D_{v'}h_{v'})(x) = 0$ $(x = 2,\dots,m-1)$ and because the column sum of $D_{v'}$ is 0 from (76), we have that

$$D_{v'}h_{v'} = \kappa(\mathbb{1}(m) - \mathbb{1}(1)) \quad \text{with some } \kappa > 0,$$

where $\mathbb{1}(x)$ is the one-hot vector with 1 at x. The positivity of κ follows from the fact that $h_{v'}^\top D_{v'}h_{v'} = \kappa = \frac{1}{m}\sum_{x,y=1}^{m}\sigma'(v'(x)-v'(y))(h_{v'}(x)-h_{v'}(y))^2 > 0$. Also, $D_{v'}$ is symmetric from its definition (76) and the fact that σ' is even. Therefore, (78) is further bounded by

$$\begin{aligned}
\Phi(v;v') &\leq \Phi(v';v') + \kappa(\mathbb{1}(m) - \mathbb{1}(1))^\top(v-v') \\
&= \Phi(v';v') + \kappa(v(m) - v(1) - (v'(m) - v'(1))) \\
&= \Phi(v';v') + \kappa(v(m) - v(1) - \beta).
\end{aligned}$$

As $v(m) - v(1) \leq \beta$ holds from $v \in C$, we obtain that $\Phi(v;v') \leq \Phi(v';v')$, which concludes the proof. □

Lemma 22. *Consider $v' \in [0,\beta]^m$ such that $0 = v'(1) \leq v'(2) \leq \cdots \leq v'(m-1) \leq v'(m) = \beta$. Let $D_{v'} \in \mathbb{R}^{m\times m}$ be as defined in Lemma 20, i.e.,*

$$D_{v'}(x,y) = \frac{\mathrm{d}}{\mathrm{d}v'(y)}g_{v'}(x) = \begin{cases} -\frac{1}{m}\sigma'(v'(x)-v'(y)) & (y \neq x) \\ \frac{1}{m}\sum_{y'\neq x}\sigma'(v'(x)-v'(y')) & (y = x), \end{cases}$$

and define $h_{v'} \in \mathbb{R}^m$ by

$$h_{v'}(1) = 0,\ h_{v'}(m) = 1,\ \textit{and}\ (D_{v'}h_{v'})(x) = 0\ \textit{for all}\ x = 2,\dots,m, \tag{79}$$

following (73).

Then, $h_{v'}(x)$ satisfies that

$$0 = h_{v'}(1) \leq h_{v'}(2) \leq \cdots \leq h_{v'}(m) = 1.$$

Proof. To show that this $h_{v'}(x)$ is monotonically increasing in x, we consider a stochastic process $\{X_t\}_{t=0}^{\infty}$ corresponding to a random walk over $1, \dots, m$. Specifically, define

$$\Lambda = \max_x \sum_{y=1}^{m} \sigma'(v'(x) - v'(y)), \tag{80}$$

and, for each interior state $x \in \{2, \dots, m-1\}$, define the transition probability $p(x \to y)$ to y by

$$p(x \to y) = \frac{\sigma'(v'(x) - v'(y))}{\Lambda} \ (y \neq x), \quad p(x \to x) = 1 - \frac{\sum_{y \neq x} \sigma'(v'(x) - v'(y))}{\Lambda}. \tag{81}$$

Also, let 1 and m be absorbing states. In Lemma 23, we will show that, for $x, x' \in \{2, \dots, m-1\}$ with $x < x'$, there exists some $x \leq x'' \leq x'$ such that

$$p(x \to y) \geq p(x' \to y) \text{ if } y \leq x'', \text{ and } p(x \to y) \leq p(x' \to y) \text{ if } y > x''. \tag{82}$$

Let $\tau(x)$ be the hitting time when the process $\{X_t\}_{t=0}^{\infty}$ arrives at x for the first time. If there is no such t, we let $\tau(x) = \infty$. We consider the probability $\mathbb{P}[\tau(m) < \tau(1) \mid X_0 = x] =: q(x)$, i.e., the probability where it arrives at m before visiting 1, when the process starts from x.

Then, q satisfies $q(1) = 0$ and $q(m) = 1$. Also, for $x \in \{2, \dots, m-1\}$, because the probability to move from x to y is $p(x \to y)$ and the probability of arriving at m before 1 starting from y is $q(y) = \mathbb{P}[\tau(m) < \tau(1) \mid X_0 = y]$, we have that

$$\begin{aligned}
&q(x) = \sum_{y=1}^{m} p(x \to y) q(y) \\
&\Leftrightarrow q(x) = \Big(1 - \frac{\sum_{y \neq x} \sigma'(v'(x) - v'(y))}{\Lambda}\Big) q(x) + \sum_{y \neq x} \frac{\sigma'(v'(x) - v'(y))}{\Lambda} q(y) \\
&\Leftrightarrow \sum_{y \neq x} \sigma'(v'(x) - v'(y)) q(x) = \sum_{y \neq x} \sigma'(v'(x) - v'(y)) q(y). \\
&\Leftrightarrow q(x) = (D_{v'} q)(x).
\end{aligned}$$

By looking at (79), these properties of q are the same as the definition of $h_{v'}$, and thus $q = h_{v'}$ holds. Therefore, what we need to show is that $q(x) \leq q(x') \Leftrightarrow \mathbb{P}[\tau(m) < \tau(1) \mid X_0 = x] \leq \mathbb{P}[\tau(m) < \tau(1) \mid X_0 = x']$ for all x, x' with $x \leq x'$.

To prove that, we consider an equivalent definition of $\{X_t\}_{t=0}^{\infty}$. For $x \in \{2, \dots, m-1\}$, define the cumulative mass function of the next step by

$$L(y; x) := \sum_{1 \leq z \leq y} p(x \to z).$$

For convenience, let $L(0; x) = 0$. We introduce $\{\alpha_t\}_{t=0}^{\infty}$, where $\alpha_t \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}([0, 1])$. If the current state is $X_t = x$, define $X_{t+1} = y$, where

$$L(y-1; x) < \alpha_t \leq L(y; x). \tag{83}$$

Because $L(y; x) - L(y-1; x) = p(x \to y)$, this is an equivalent definition of $\{X_t\}_{t=0}^{\infty}$.

We then consider two instances of this process with different initial states, denoted by $\{X_t^{(x)}\}_{t=0}^{\infty}$ and $\{X_t^{(x')}\}_{t=0}^{\infty}$ with $x \leq x'$, both driven by the same randomness $\{\alpha_t\}_{t=0}^{\infty}$. For each t, we show that $X_{t+1}^{(x)} \leq X_{t+1}^{(x')}$ by induction.

For $x, x' \in \{2, \dots, m-1\}$ with $x < x'$, $L(y; x) - L(y; x') = 0$ at $y = 1, m$. Also, as y goes from 1 to $m-1$, the sign of $\Delta(y) = (L(y+1; x) - L(y+1; x')) - (L(y; x) - L(y; x')) = p(x \to y+1) - p(x' \to y+1)$ changes only once from $-$ to $+$ according to (82). Therefore, for x, x' with $x < x'$, $L(y; x) \geq L(y; x')$ always holds.

Therefore, when $X_t^{(x)} \leq X_t^{(x')}$, we have that $L(X_{t+1}^{(x)} - 1; X_t^{(x)}) \geq L(X_{t+1}^{(x)} - 1; X_t^{(x')})$, and because of the definition of $X_{t+1}^{(x)}$ (83)

$$\alpha_t > L(X_{t+1}^{(x)} - 1; X_t^{(x)}) \geq L(X_{t+1}^{(x)} - 1; X_t^{(x')}).$$

As $X_{t+1}^{(x')}$ is defined as y such that $L(y-1; X_t^{(x')}) < \alpha_t \le L(y; X_t^{(x')})$, it follows that $X_{t+1}^{(x')} \ge X_{t+1}^{(x)}$.

Therefore, the two processes with different initial points are coupled so that $X_t^{(x)} \le X_t^{(x')}$ always holds, which implies that $\mathbb{P}[\tau(m) < \tau(1) \mid X_0 = x] \le \mathbb{P}[\tau(m) < \tau(1) \mid X_0 = x'] \Leftrightarrow q(x) \le q(x') \Leftrightarrow h_{v'}(x) \le h_{v'}(x')$ for all x, x' with $x \le x'$. □

Lemma 23. *Consider $v' \in [0, \beta]^m$ satisfying $0 = v'(1) \le v'(2) \le \cdots \le v'(m-1) \le v'(m) = \beta$. Let $p(x \to y)$ be defined as in (80) and (81) in Lemma 22. For any $x, x' \in \{2, \ldots, m-1\}$ with $x < x'$, there exists some $x \le x'' \le x'$ such that the following holds:*

for all $y \in \{1, \ldots, x''\}$, $p(x \to y) \ge p(x' \to y)$, and for all $y \in \{x''+1, \ldots, m\}$, $p(x \to y) \le p(x' \to y)$.

Proof. Remember that $\sigma'(t)$ is even and monotonically decreasing in $t \ge 0$. Because $v'(1) \le v'(2) \le \cdots \le v'(m-1) \le v'(m)$, the sign of $\sigma'(v'(x) - v'(y)) - \sigma'(v'(x') - v'(y))$ only changes as y increases from $y = 1$ to $y = m$. Also, because $v'(x) \le v'(x')$ from $x < x'$, $\sigma'(v'(x) - v'(y)) - \sigma'(v'(x') - v'(y)) \ge 0$ when $y \le x$, and $\sigma'(v'(x) - v'(y)) - \sigma'(v'(x') - v'(y)) \le 0$ when $y \ge x'$. Therefore, there exists some x'' with $x \le x'' \le x'$ such that $\sigma'(v'(x) - v'(y)) - \sigma'(v'(x') - v'(y)) \ge 0$ when $y \le x''$ and $\sigma'(v'(x) - v'(y)) - \sigma'(v'(x') - v'(y)) \le 0$ when $y > x''$.

For $y \ne x, x'$, $p(x \to y) = \Lambda^{-1}\sigma'(v'(x) - v'(y))$ and $p(x' \to y) = \Lambda^{-1}\sigma'(v'(x') - v'(y))$. Therefore, the assertion directly follows.

For $y = x$,

$$p(x \to x) = 1 - \frac{\sum_{y' \ne x} \sigma'(v'(x) - v'(y))}{\Lambda} = 1 - \frac{\sum_{y=1}^{m} \sigma'(v'(x) - v'(y))}{\Lambda} + \frac{\sigma'(v'(x) - v'(x))}{\Lambda}$$
$$\ge \frac{\sigma'(v'(x) - v'(x))}{\Lambda} = \frac{\sigma'(0)}{\Lambda},$$

as $\sum_{y=1} \sigma'(v'(x) - v'(y)) \le \Lambda$ follows from the definition of Λ. Also, $p(x' \to y) = \frac{\sigma'(v'(x') - v'(x))}{\Lambda}$. Therefore, according to $\sigma'(0) \ge \sigma'(v'(x') - v'(x))$, we have $p(x \to x) \le p(x' \to x)$. Because $x \le x''$, this is precisely the inequality that we need to establish.

Similarly, for $y = x'$, it holds that $p(x' \to x') \ge \frac{\sigma'(0)}{\Lambda} \ge p(x' \to x) = \frac{\sigma'(v'(x) - v'(x'))}{\Lambda}$. Because $x'' \le x'$, this is also the inequality that we need to establish. Combining the above, the claim is established for all y. □

D Proof of Lower Bounds

In this section, we present the proofs of the lower bounds. The proof of Theorem 4 is given in Appendix D.1, and the proof of Theorem 5 is provided in Appendix D.2.

D.1 A Lower Bound Dependent on the KL Constraint

We first show below that, for $e^{-\Theta(B)} \le \tau \le B$, the distortion is lower bounded by $\tilde{\Theta}(\min\{B\beta, B\beta/\tau\})$.

Theorem 4. *Assume that the KL budget $\tau > 0$, the Bradley–Terry model temperature parameter $\beta \ge 3$, and the maximum log-likelihood ratio $B = \max_{x \in A} \left| \log \frac{\mu(x)}{\pi_{\mathrm{ref}}(x)} \right|$ satisfy*

$$\max\{e^{-\frac{B}{3}}, e^{-\frac{\beta}{2}}\} \le \tau \le B, \quad 3 \le \beta \le e^{\frac{B}{3}} - 1, \quad \textit{and} \quad 3 \le B \le \min\{e^{\frac{B}{3}}, e^{\frac{\beta}{2}}\}.$$

Then, there exists a pair of an instance $\mathcal{D}$, a data distribution μ, and a reference policy π_{ref} such that, regardless of how $r_{\min}$ and $r_{\max}$ are chosen in reward clipping, the distortion of π_{RLHF} is lower bounded by

$$\mathsf{Dist}(\pi_{\mathrm{RLHF}}) \gtrsim \min\left\{\frac{B}{\log B\beta}, \frac{B}{\tau}\right\}\beta.$$

Proof. We consider four alternatives and set $(\mu(1), \mu(2), \mu(3), \mu(4)) = (e^{-B}, e^{-B}, 1 - 3e^{-B}, e^{-B})$. We define the utility u as follows.

(i) With probability p_1: $(u(1), u(2), u(3), u(4)) = (\frac{1}{\beta}, 0, 0, 0)$.

(ii) With probability p_2: $(u(1), u(2), u(3), u(4)) = (0, \frac{1}{2}, 1, 0)$.

(iii) With probability $p_3 = 1 - p_1 - p_2$: $(u(1), u(2), u(3), u(4)) = (0, 0, 0, \frac{1}{\beta})$.

By Lemma 18, the ordering of the Borda scores $\mathsf{Borda}(x) = \mathbb{E}_{y\sim\mu}[p(x \succ y)]$ coincides with that of the estimated rewards $\bar{r}$. We compute the expected Borda scores and define p_1, p_2, and p_3 so that $\mathsf{Borda}(1) = \mathsf{Borda}(2) \le \mathsf{Borda}(3) = \mathsf{Borda}(4)$, which implies that $\bar{r}(1) = \bar{r}(2) \le \bar{r}(3) = \bar{r}(4)$. Regardless of the lower and upper bounds in the reward clipping, this implies that $r(1) = r(2) \le r(3) = r(4)$.

Alternative 1:

$$\begin{aligned}\mathsf{Borda}(1) =& p_1\Big(e^{-B}\cdot\frac{1}{2} + (1-e^{-B})\cdot\sigma(1)\Big) + p_2\Big(2e^{-B}\cdot\frac{1}{2} + (1-3e^{-B})\cdot\sigma(-\beta) + e^{-B}\sigma\Big(-\frac{\beta}{2}\Big)\Big)\\ &+ p_3\Big((1-e^{-B})\cdot\frac{1}{2} + e^{-B}\cdot\sigma(-1)\Big).\end{aligned}$$

Alternative 2:

$$\begin{aligned}\mathsf{Borda}(2) =& p_1\Big(e^{-B}\cdot\sigma(-1) + (1-e^{-B})\cdot\frac{1}{2}\Big) + p_2\Big(2e^{-B}\cdot\sigma\Big(\frac{\beta}{2}\Big) + e^{-B}\cdot\frac{1}{2} + (1-3e^{-B})\cdot\sigma\Big(-\frac{\beta}{2}\Big)\Big)\\ &+ p_3\Big((1-e^{-B})\cdot\frac{1}{2} + e^{-B}\cdot\sigma(-1)\Big).\end{aligned}$$

Alternative 3:

$$\begin{aligned}\mathsf{Borda}(3) =& p_1\Big(e^{-B}\cdot\sigma(-1) + (1-e^{-B})\cdot\frac{1}{2}\Big) + p_2\Big(2e^{-B}\cdot\sigma(\beta) + e^{-B}\cdot\sigma\Big(\frac{\beta}{2}\Big) + (1-3e^{-B})\cdot\frac{1}{2}\Big)\\ &+ p_3\Big((1-e^{-B})\cdot\frac{1}{2} + e^{-B}\cdot\sigma(-1)\Big).\end{aligned}$$

Alternative 4:

$$\begin{aligned}\mathsf{Borda}(4) =& p_1\Big(e^{-B}\cdot\sigma(-1) + (1-e^{-B})\cdot\frac{1}{2}\Big) + p_2\Big(2e^{-B}\cdot\frac{1}{2} + (1-3e^{-B})\cdot\sigma(-\beta) + e^{-B}\sigma\Big(-\frac{\beta}{2}\Big)\Big)\\ &+ p_3\Big((1-e^{-B})\cdot\sigma(1) + e^{-B}\cdot\frac{1}{2}\Big).\end{aligned}$$

Using the above calculations, the condition where $\mathsf{Borda}(1) = \mathsf{Borda}(2)$ holds is

$$p_1 = \underbrace{\frac{e^{-B}\big(2\sigma\big(\frac{\beta}{2}\big) - \sigma\big(-\frac{\beta}{2}\big) - \frac{1}{2}\big) + (1-3e^{-B})\big(\sigma\big(-\frac{\beta}{2}\big) - \sigma(-\beta)\big)}{e^{-B}\big(\frac{1}{2} - \sigma(-1)\big) + (1-e^{-B})\big(\sigma(1) - \frac{1}{2}\big)}}_{=c_1'} p_2,$$

and the condition where $\mathsf{Borda}(3) = \mathsf{Borda}(4)$ holds is

$$p_3 = \underbrace{\frac{e^{-B}\big(2\sigma(\beta) + \sigma\big(\frac{\beta}{2}\big) - \sigma\big(-\frac{\beta}{2}\big) - 1\big) + (1-3e^{-B})\big(\frac{1}{2} - \sigma(-\beta)\big)}{e^{-B}\big(\frac{1}{2} - \sigma(-1)\big) + (1-e^{-B})\big(\sigma(1) - \frac{1}{2}\big)}}_{=c_2'} p_2.$$

Here, $c_1' = \Theta(e^{-B} + \sigma(-\frac{\beta}{2})) = \Theta(e^{-B} + e^{-\frac{\beta}{2}})$ and $c_2' = \Theta(1)$. So we rewrite $p_1 = c_1(e^{-B} + e^{-\frac{\beta}{2}})p_3$ and $p_2 = c_2 p_3$ with $c_2 = (c_2')^{-1} = \Theta(1)$, $c_1 = (e^{-B} + e^{-\frac{\beta}{2}})^{-1}c_1'c_2 = \Theta(1)$, and $p_3 = \big(1 + c_1(e^{-B} + e^{-\frac{\beta}{2}}) + c_2\big)^{-1} = \Theta(1)$. Also, because $u(3) \ge u(2)$ is always true, $\mathsf{Borda}(2) \le \mathsf{Borda}(3)$.

Next, we specify π_{ref} and analyze π_{RLHF}. When $\tau \le B$, $\beta \ge 3$, and $B \ge \log 3(1+\beta)$, we define

$$(\pi_{\mathrm{ref}}(1), \pi_{\mathrm{ref}}(2), \pi_{\mathrm{ref}}(3), \pi_{\mathrm{ref}}(4)) = \Big(1 - \frac{\tau}{B\beta} - (1+\beta)e^{-B}, \frac{\tau}{B\beta}, e^{-B}, \beta e^{-B}\Big)$$

and we can see that $\left\| \log \frac{\mathrm{d}\mu}{\mathrm{d}\pi_{\mathrm{ref}}} \right\|_\infty = B$. Because π_{RLHF} is the maximizer of $\mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[r(x)]$ within $\mathrm{KL}(\pi\|\pi_{\mathrm{ref}}) \le \tau$ (if the maximizer is not unique, we choose the one with the smallest KL divergence), $r(1) = r(2) \le r(3) = r(4)$ implies that $\frac{\pi_{\mathrm{RLHF}}(1)}{\pi_{\mathrm{ref}}(1)} = \frac{\pi_{\mathrm{RLHF}}(2)}{\pi_{\mathrm{ref}}(2)} \le 1$ while $\frac{\pi_{\mathrm{RLHF}}(3)}{\pi_{\mathrm{ref}}(3)} = \frac{\pi_{\mathrm{RLHF}}(4)}{\pi_{\mathrm{ref}}(4)} \ge 1$. We denote $\pi_{\mathrm{RLHF}}(3)$ by q in the following.

Then,

$$\mathrm{KL}(\pi\|\pi_{\mathrm{ref}}) = (1-(1+\beta)q)\log\frac{1-(1+\beta)q}{1-(1+\beta)e^{-B}} + (1+\beta)q\log\frac{(1+\beta)q}{(1+\beta)e^{-B}} \le \tau.$$

By using $\log(1-x) \ge -x$, we have that

$$(1+\beta)q\Big(\log q + B - 1\Big) \le \tau.$$

When $3\log\frac{B\beta}{3\tau} \le B$ and $B \ge 3$, this implies that $q \le \frac{3\tau}{B\beta}$. Therefore, the average utility of π_{RLHF} is

$$\begin{aligned}
\mathbb{E}_{u\sim\mathcal{D},x\sim\pi_{\mathrm{RLHF}}}[u(x)] &\le \pi_{\mathrm{ref}}(1)\cdot\frac{1}{\beta}\cdot p_1 + \pi_{\mathrm{ref}}(2)\cdot\frac{1}{2}\cdot p_2 + q\cdot 1\cdot p_2 + \beta q\cdot\frac{1}{\beta}\cdot p_3 \\
&\le \frac{c_1(e^{-B}+e^{-\frac{\beta}{2}})p_3}{\beta^2} + \frac{c_2 p_3\tau}{2B\beta} + \frac{3c_2p_3\tau}{B\beta} + \frac{3p_3\tau}{B\beta} \\
&\le \Big(\frac{27c_1\tau^2}{\beta^3B^2} + \frac{c_1}{\beta} + \frac{7c_2}{2} + 3\Big)\frac{\tau p_3}{B\beta}, \qquad (84)
\end{aligned}$$

where we have used $3\log\frac{B\beta}{3\tau} \le B$ and $\beta \ge 2\log\frac{B}{\tau}$ to evaluate e^{-B} and $e^{-\frac{\beta}{2}}$, respectively, in the final inequality.

On the other hand, when $B\beta \ge e$, consider a distribution $\pi' = \min\{\frac{\tau}{\log B\beta}, 1\}\mathbb{1}_2 + \big(1 - \min\{\frac{\tau}{\log B\beta}, 1\}\big)\mu$. Then, we have that

$$\mathrm{KL}(\pi'\|\pi_{\mathrm{ref}}) \le \min\Big\{\frac{\tau}{\log B\beta}, 1\Big\} \times \log\frac{\min\big\{\frac{\tau}{\log B\beta}, 1\big\}}{\frac{\tau}{B\beta}} \le \frac{\tau}{\log B\beta}\times\log B\beta \le \tau.$$

Also, the average utility is

$$\max_{\mathrm{KL}(\pi\|\pi_{\mathrm{ref}})\le\tau} \mathbb{E}_{u\sim\mathcal{D},x\sim\pi}[u(x)] \ge \mathbb{E}_{u\sim\mathcal{D},x\sim\pi'}[u(x)] \ge \min\Big\{\frac{\tau}{\log B\beta}, 1\Big\}\cdot\frac{1}{2}\cdot c_2 p_3. \qquad (85)$$

From (84) and (85), the distortion is lower bounded by

$$\mathsf{Dist}(\pi_{\mathrm{RLHF}}) \ge \frac{c_2}{2}\Big(\frac{27c_1\tau^2}{\beta^3B^2} + \frac{c_1}{\beta} + \frac{7c_2}{2} + 3\Big)^{-1}\min\Big\{\frac{B}{\log B\beta}, \frac{B}{\tau}\Big\}\beta,$$

as desired. □

D.2 A Lower Bound Independent of the KL Constraint

Theorem 4 provided a lower bound that depends on the KL budget τ and the density ratio B between μ and π_{ref}. Here, we complement Theorem 4 by showing that an $\Omega(\beta)$ distortion arises for any value of B, including the case $B = 0$, and for any choice of τ.

Theorem 5. *When $\mu = \pi_{\mathrm{ref}}$, for any KL budget $\tau > 0$ and any Bradley–Terry temperature parameter $\beta > 0$, there exist a collection of instances $\{\mathcal{D}_i\}_{i=1}^N$ and a data distribution μ such that, when an instance is drawn uniformly at random from $\{\mathcal{D}_i\}_{i=1}^N$, the output of any algorithm incurs expected distortion at least*

$$\frac{\beta}{2}\frac{1+e^{-\beta}}{1-e^{-\beta}} - \epsilon,$$

where $\epsilon > 0$ is an arbitrarily small constant.

Proof. Take δ, ε to be sufficiently small and $K \in \mathbb{N}$ to be sufficiently large. We consider $K+2$ alternatives and set $(\mu(1), \dots, \mu(K+1), \mu(K+2)) = (\delta, \dots, \delta, 1-(K+1)\delta)$. We define an instance $\mathcal{D}$ by specifying the utility u as follows. The collection $\{\mathcal{D}_i\}_{i=1}^N$ is then obtained by uniformly randomly permuting the alternatives $1, \dots, K+1$.

(i) With probability p_1: Define the utility vector $(u(1), u(2), \dots, u(K+2)) = (1, 0, \dots, 0)$.

(ii) With probability $p_2 = 1 - p_1$: Choose y uniformly at random from $\{2, \dots, K+1\}$, and set $u(x) = \varepsilon\beta^{-1}$ (if $x = y$), 0 (otherwise).

Now, let us compute the expected win rates between alternatives in $\mathcal{D}$:

When $x, y \in \{2, \dots, K+1\}$: Because of the symmetry within each set of alternatives, $p(x \succ y) = \frac{1}{2}$.

When $x = 1$ and $y \in \{2, \dots, K+1\}$: By considering the deviation from the win rate of $\frac{1}{2}$, we have that

$$p(x \succ y) - \frac{1}{2} = p_1 \cdot \delta\Big(\sigma(\beta) - \frac{1}{2}\Big) - \frac{p_2}{K} \cdot \delta\Big(\frac{1}{2} - \sigma(-\varepsilon)\Big). \tag{86}$$

When $x = 1$ and $y = K+2$:

$$p(x \succ y) - \frac{1}{2} = p_1 \cdot \delta\Big(\sigma(\beta) - \frac{1}{2}\Big). \tag{87}$$

When $x \in \{2, \dots, K+1\}$ and $y = K+2$:

$$p(x \succ y) - \frac{1}{2} = \frac{p_2}{K} \cdot \delta\Big(\frac{1}{2} - \sigma(-\varepsilon)\Big). \tag{88}$$

The condition where $(86) = 0$ and $(87) = (88)$ hold is

$$p_1 = \underbrace{\frac{\sigma(\varepsilon) - \frac{1}{2}}{\sigma(\beta) - 1}}_{=c_p} \cdot K^{-1} p_2 \quad (> 0).$$

Under this condition, $\hat{\pi}$ cannot identify the permutation of the alternatives $1, \dots, K+1$.

Now, consider the output $\hat{\pi}$ of the algorithm that achieves the optimal distortion. For δ chosen sufficiently small, it is not possible to allocate all probability mass to alternatives 1 through $1+K$, which implies that $\mathrm{KL}(\hat{\pi} \| \pi_{\mathrm{ref}}) = \tau$. Letting P be a random permutation of $\{1, \dots, K+2\}$ fixing $K+2$, using $P_{\#}\hat{\pi}$ instead of $\hat{\pi}$ decreases the distortion and the KL constraint unless $P_{\#}\hat{\pi}$ and $\hat{\pi}$ are equal as distributions. Therefore, $\hat{\pi}$ must satisfy

$$\hat{\pi}(1) = \cdots = \hat{\pi}(K+1) = q \geq \delta.$$

Let us consider the KL constraint.

$$\mathrm{KL}(\hat{\pi} \| \pi_{\mathrm{ref}}) = \tau \Leftrightarrow q \log \frac{q}{\delta} + Kq \log \frac{q}{\delta} + (1 - (1+K)q) \log \frac{1 - (1+K)q}{1 - (1+K)\delta} = \tau.$$

By using $\log(1-x) \geq -x$, we have that

$$(1+K)q\Big(\log \frac{q}{\delta} - 1\Big) \leq \tau, \tag{89}$$

and

$$(1+K)q \log \frac{q}{\delta} \geq \tau \tag{90}$$

By taking $\delta \leq (1+K)^{-1}(K^{K+1} \log K)^{-1}\tau$, we can see that $\log \frac{q}{\delta} \geq K \log K$ from (90). Under this, (89) implies that

$$\begin{aligned} &\Big(1 - \frac{1}{K \log K}\Big)(1+K)q \log \frac{q}{\delta} \leq (1+K)q\Big(\log \frac{q}{\delta} - 1\Big) \leq \tau \\ &\Rightarrow \Big(1 - \frac{1}{K \log K}\Big)(1+K)q \log \frac{q}{\delta} \leq \tau. \end{aligned} \tag{91}$$

Let us consider a distribution $\pi' = \big(1 - \frac{1}{K \log K}\big)\big(Kq - \delta\big)\mu' + \big(1 - \big(1 - \frac{1}{K \log K}\big)(Kq - \delta)\big)\mu$, where μ' is the uniform distribution over $1, \dots, a$. Then, by using $\log \frac{q}{\delta} \geq K \log K$,

$$\begin{aligned}\mathrm{KL}(\pi' \| \pi_{\mathrm{ref}}) &\leq \Big(1 - \frac{1}{K \log K}\Big) Kq \Big(\log \frac{q}{\delta} + \log K \Big) \\ &\leq \Big(1 - \frac{1}{K \log K}\Big)(1 + K) q \log \frac{q}{\delta},\end{aligned}$$

where the RHS is bounded by τ according to (91). Using this distribution π', the maximum average utility is lower bounded by

$$\max_{\mathrm{KL}(\pi \| \pi_{\mathrm{ref}}) \leq \tau} \mathbb{E}_{u \sim \mathcal{D}, x \sim \pi}[u(x)] \geq \mathbb{E}_{u \sim \mathcal{D}, x \sim \pi'}[u(x)] \geq \Big(1 - \frac{1}{K \log K}\Big) Kq p_1.$$

On the other hand,

$$\mathbb{E}_{u \sim \mathcal{D}, x \sim \hat{\pi}}[u(x)] = q \cdot p_1 + q \cdot \varepsilon \beta^{-1} p_2 = q p_1 + q \varepsilon \beta^{-1} \beta' c_p^{-1} p_1.$$

Therefore, the distortion of $\hat{\pi}$ is lower bounded by

$$\mathsf{Dist}(\hat{\pi}) \geq \Big(1 - \frac{1}{K \log K}\Big) \frac{K}{1 + \varepsilon \beta^{-1} K c_p^{-1}}. \tag{92}$$

We let $K \to \infty$ and $\varepsilon \to 0$. Note that $\frac{c_p}{\varepsilon} \to \frac{\varepsilon}{4(\sigma(\beta) - 1/2)}$ as $\varepsilon \to 0$. Therefore, the RHS of (92) converges to

$$\frac{\beta}{4(\sigma(\beta) - 1/2)} = \frac{\beta}{2} \frac{1 + e^{-\beta}}{1 - e^{-\beta}},$$

as desired. □

E Details of the Experiments

E.1 Details of Reward Scale Evaluation

Skywork-Reward-V2-Llama-3.1-8B (Liu et al., 2024) and UltraRM-13B (Cui et al., 2023) are trained on preference data by maximizing the likelihood under a single Bradley–Terry model. For a dataset of prompt z, chosen completion x, and rejected completion y (i.e., $x \prec y$), consisting of pairs $\{(x^i, y^i, z^i)\}_{i=1}^n$, the log-likelihood conditioned by the prompt z is defined as follows. (Eq. (1) is the unconditioned version.)

$$\max_{\theta \in \mathbb{R}^d} \sum_{i=1}^n \log \big(\sigma(r_\theta(x^i \mid z^i) - r_\theta(y^i \mid z^i))\big).$$

While this objective was used in Skywork (Liu et al., 2024), UltraRM-13B made a slight modification to incorporate a guide of the annotated rewards for a subset of their datasets. Specifically, when the annotated reward difference $\Delta r_{\mathsf{annotated}}$ between a chosen completion x and a rejected completion y is available, the objective is modified to

$$\max_{\theta \in \mathbb{R}^d} \sum_{i=1}^n \log \big(\sigma(r_\theta(x^i \mid z^i) - r_\theta(y^i \mid z^i) - \Delta r_{\mathsf{annotated}})\big).$$

Here, $\Delta r_{\mathsf{annotated}}$ is the absolute difference between the annotated reward of two texts. It is normalized so that $\Delta r_{\mathsf{annotated}} \in (0, 1]$. When using datasets with only preference rankings, they simply set $\Delta r_{\mathsf{annotated}} = 0$. We denote the resulting reward model by $r_{\hat{\theta}}$.

For Skywork and UltraRM, we evaluate the reward scale using 5000 samples drawn from their training datasets, Skywork-Reward-Preference-80K-v0.1 (Liu et al., 2024) and UltraFeedback (Cui et al., 2023), respectively. Specifically, for each pair of prompt z^i, completions x^i and y^i, we calculated the reward difference $\Delta r_{\hat{\theta}}(x^i, y^i; z^i)$ defined by

$$\Delta r_{\hat{\theta}}(x^i, y^i; z^i) = \big| r_{\hat{\theta}}(x^i \mid z^i) - r_{\hat{\theta}}(y^i \mid z^i) \big|.$$

Because UltraFeedback has four completions for each prompt, we selected a completion with the highest reward as x^i and a completion with the lowest reward as y^i for each.

E.2 Details of the Synthetic Experiment

We consider three alternatives. The utility distribution is defined as $u = (0, 0.5, 1)$ with probability 0.99, and $u = (0.01, 0, 0)$ with probability 0.01. We set $\beta = 10$, $\mu = (10^{-4}, 10^{-4}, 1 - 2 \cdot 10^{-4})$, and $\pi_{\mathrm{ref}} = (1 - 0.05 - 10^{-5},\ 0.05,\ 10^{-5})$, which yields $B = 11.5$. Also, we tuned the KL budget $\tau = 0.143$ so that the maximum average utility $\max_{\pi:\ \mathrm{KL}(\pi \| \pi_{\mathrm{ref}}) \le \tau} \mathbb{E}_{u \sim \mathcal{D}, x \sim \pi}[u(x)]$, which is numerically computed, is 0.1.

Rewards are computed by first evaluating the objective in (1) exactly in the limit $n \to \infty$, and then performing second-order optimization using the L-BFGS algorithm by exploiting the concavity of the problem. In Figure 3, the case $\mu \neq \pi_{\mathrm{ref}}$ corresponds to rewards computed under the above setting, whereas the case $\mu = \pi_{\mathrm{ref}}$ corresponds to replacing μ with π_{ref} as the data distribution in the same setup.

In both cases, we optimize the distribution using a mirror descent-style update (Beck & Teboulle, 2003) with step size $\eta = 10^{-3}$. Specifically, starting from $\pi^0 = \pi_{\mathrm{ref}}$, we gradually update π^t until π^t violates the KL constraint by repeating the following update:

$$\pi_{t+1} = \arg\max_{\pi} \left\{ \langle \pi, r \rangle - \frac{1}{\eta} \mathrm{KL}(\pi \| \pi_t) \right\} \quad \Longrightarrow \quad \pi_{t+1}(x) = \frac{\pi_t(x) \exp(\eta r(x))}{\sum_{x'} \pi_t(x') \exp(\eta r(x'))}.$$