Design Considerations in Offline Preference-based RL

Alekh Agarwal[1]   Christoph Dann[1]   Teodor V. Marinov[1]

Abstract

Offline algorithms for Reinforcement Learning from Human Preferences (RLHF), which use only a fixed dataset of sampled responses given an input, and preference feedback among these responses, have gained increasing prominence in the literature on aligning language models. In this paper, we study how the different design choices made in methods such as DPO, IPO, SLiC and many variants influence the quality of the learned policy, from a theoretical perspective. Our treatment yields insights into the choices of loss function, the policy which is used to normalize log-likelihoods, and also the role of the data sampling policy. Notably, our results do not rely on the standard reparameterization-style arguments used to motivate some of the algorithms in this family, which allows us to give a unified treatment to a broad class of methods. We also conduct a small empirical study to verify some of the theoretical findings on a standard summarization benchmark.

1. Introduction

The now substantial literature on Reinforcement Learning with Human Preferences (RLHF) can be broadly categorized into two families of methods. Given a dataset of human preferences, the first class of online algorithms are based on learning a reward function that assigns numerical scores to a response y given some input x, such that the high-scoring responses are preferred over the low-scoring ones in our preference dataset. These methods subsequently maximize this reward function using an *online RL* algorithm like PPO (Christiano et al., 2017; Stiennon et al., 2020; Ouyang et al., 2022) or Reinforce (Ahmadian et al., 2024). A different approach forgoes the reward learning step and uses a reparameterization trick to directly learn a

good policy from the preference dataset, with Direct Preference Optimization (DPO) (Rafailov et al., 2024b) being the pioneering approach in this line of work. These approaches are referred to as *offline or direct alignment*, since they do not draw any fresh samples from the learned policy during training, and only use the responses observed in the preference dataset. This paper focuses on this latter family of algorithms, and studies how the properties of the data and the learning objective affect the quality of the learned policy.

Offline methods for RLHF such as DPO (Rafailov et al., 2024b), IPO (Azar et al., 2024), SLiC (Zhao et al., 2022; 2023) and KTO (Ethayarajh et al., 2024), along with a growing number of variants have received a significant attention in the academic literature owing to a number of attractive properties. First, the removal of an explicit reward learning step simplifies the number of modeling choices and steps required for the RLHF pipeline, along with reducing the demand on computational resources. Furthermore, the requirement to only evaluate a policy's likelihood on a fixed set of responses in the preference dataset, as opposed to generating responses from the learned policy as in online RL, adds further resource efficiency. However, the growth of this literature has also spurred an equally large body of work now detailing the various deficiencies of these techniques, such as the tendency of the algorithms to shift the probability mass outside the support of the observed responses in the preference dataset, significant preference hacking behaviors and notorious collapses in the learning dynamics with continued training (Pal et al., 2024; Park et al., 2024; Rafailov et al., 2024a; Fisch et al., 2024).

While some of the issues raised above have received theoretical treatment for specific approaches, the literature still lacks a comprehensive theoretical foundation underlying these offline RLHF techniques. Most of these techniques have been motivated by some variant of the original reparameterization argument in the DPO paper, but the argument hinges on unformalized assumptions regarding the coverage of the data with respect to the learned policy, and does not capture all the algorithmic variants which have subsequently been developed in the literature.

In this paper, we instead adopt the perspective of offline RLHF as just solving a *loss minimization problem* on a preference dataset, and study when the optimal solution

---

*Equal contribution [1]Google Research. Correspondence to: Alekh Agarwal <alekhagarwal@google.com>, Christoph Dann <chrisdann@google.com>, Teodor V. Marinov <tvmarinov@google.com>.

Method	Base policy $\mu(y\|x)$	Loss ℓ	Constraint/Regularizer
DPO (Rafailov et al., 2024b)	$\pi_{\text{ref}}(y\|x)$	$\ell(z) = -\log\left(\frac{1}{1+\exp(-\beta z)}\right)$	-
IPO (Azar et al., 2024)	$\pi_{\text{ref}}(y\|x)$	$\ell(z) = (z-\tau)^2$	-
Slic-HF (Zhao et al., 2023)	1	$\ell(z) = \max\{\tau - z, 0\}$	$\text{CE}(\pi_{\text{ref}}, \pi)$
GPO (Tang et al., 2024)	$\pi_{\text{ref}}(y\|x)$	$\ell(z) = f(\beta z)$	-
CPO (Xu et al., 2024)	1	$\ell(z) = -\log\left(\frac{1}{1+\exp(-\beta z)}\right)$	$\text{CE}(\pi_w, \pi)$
R-DPO (Park et al., 2024)	$\pi_{\text{ref}}(y\|x)\exp\left(\frac{\alpha}{\beta}\|y\|\right)$	$\ell(z) = -\log\left(\frac{1}{1+\exp(-\beta z)}\right)$	-
ODPO[1] (Amini et al., 2024)	$\pi_{\text{ref}}(y\|x)$	$\ell(z) = -\log\left(\frac{1}{1+\exp(-\beta z+\tau)}\right)$	-
SimPO[2] (Meng et al., 2024)	1	$\ell(z) = -\log\left(\frac{1}{1+\exp(-\beta z)} - \gamma\right)$	-

Table 1. A collection of methods for offline RLHF from preference feedback, along with the instantiation of the different design choices that make them a special case of our general framework. For ODPO and SimPO, they are not included in the full generality in our framework as discussed in the footnotes.

of this loss minimization yields a desirable policy. A key challenge in undertaking such a study is the identification of a benchmark policy which is both desirable in terms of the responses it generates, and is attainable under reasonable assumptions using offine RLHF. With this background, our paper makes the following contributions.

1. We identify a benchmark policy to measure the performance of offline RLHF against. Prior work (Swamy et al., 2024) shows that the class of offline techniques considered here cannot attain the optimal policy for online RLHF in general, and we instead develop a weaker benchmark under reasonable assumptions on the data generating process and the learning setup. For DPO, this benchmark does correspond to the optimal softmax policy when the preference data follows the Bradley-Terry-Luce (Bradley & Terry, 1952; Luce, 2012) model.

2. We provide a bound on the sub-optimality of the learned policy to the yardstick identified above in a learning framework that encapsulates most existing offline RLHF variants. The bound depends on the curvature of the loss function, as well as the coverage of the offline dataset.

3. We corroborate some of our theoretical findings using an empirical study on a summarization task, and find that squared loss of IPO outperforms the logistic loss of DPO due to its nicer curvature properties, while dropping the normalization with reference policy likelihood causes a small, but consistent deterioration in the quality of the learned policy, as predicted by our theory.

2. Problem Setup

Preference-based learning. We consider the setting of learning from offline preference data. In this setting, we are typically given a dataset of samples where each sample consists of an input $x \in \mathcal{X} \sim D_x$, two responses $y, y' \in \mathcal{Y} \times \mathcal{Y} \stackrel{\text{i.i.d.}}{\sim} D_y(\cdot|x)$ and a binary label $\omega \in \{-1, 1\} \sim$

$D_\omega(\cdot|x, y, y')$. We use the shorthand $(x, y, y', \omega) \sim D_{xy\omega}$ to succinctly denote samples from this generative process. The label ω captures our preference for the response y over y', for the input x. We study offline preference-based RL methods, which are parameterized by some loss function $\ell : [-\mathfrak{R}, \mathfrak{R}] \to [0, \infty)$. In particular, we study methods which take as input a policy class $\Pi \subseteq \{\mathcal{X} \to \Delta(\mathcal{Y})\}$, and find a policy which minimizes the following loss, given n samples:

$$\widehat{\pi} = \operatorname*{argmin}_{\pi \in \Pi} \sum_{i=1}^{n} \ell\left(-\omega_i\underbrace{\left(\log\frac{\pi(y_i|x_i)}{\mu(y_i|x_i)} - \log\frac{\pi(y_i'|x_i)}{\mu(y_i'|x_i)}\right)}_{:=\omega_i^{\pi,\mu}}\right).$$

(1)

Here $\mu: \mathcal{X} \to \mathbb{R}_+^{\mathcal{Y}}$ refers to an arbitrary base policy, which does not need to be normalized. Typically ℓ is chosen so that it incentivizes $\omega_i\omega_i^{\widehat{\pi},\mu}$ to be positive. This means that $\widehat{\pi}$ tends to increase the probability of producing y_i compared with y_i', relative to the base probabilities under μ, when $\omega_i = 1$. In the particularly simple case when μ is just the uniform policy, we see that $\widehat{\pi}$ results in a larger probability of producing y_i over y_i', when $\omega_i = 1$, which captures the basic intuition behind offline preference-based RL.

Choice of policy class. A common policy parameterization is to use softmax policies $\pi_\theta(y|x) \propto \exp(f_\theta(y|x))$ with some parameter set $\theta \in \Theta$, where f is some fixed network architecture. While we could explicitly define the policy class Π this way, we intentionally leave Π general at this

---

[1] ODPO in the general case allows a margin of the form $f(\text{score}(x, y) - \text{score}(x, y'))$, which is not admissible in our setup as specified, though it can be incorporated in our analysis with some additional work. When f is the identity function, we can simply define $\mu(y|x) \propto \pi_{\text{ref}}(y|x)\text{score}(x, y)$, as a special case.

[2] SimPO in the general case has an additional length normalization, leading to the loss $\log(\omega\text{sigmoid}(\beta \log \pi(y|x)/|y| - \beta \log \pi(y'|x)/|y'|) - \gamma)$. This length normalization is currently not included in our theoretical setup and analysis.

point, to allow our framework to further include:

- Additional constraints (1): Several works (e.g. Zhao et al. (2023); Xu et al. (2024)) regularize or constrain the preference objective of Equation 1 with a cross-entropy term based on $CE(\pi_0, \pi) := -\mathbb{E}_{x \sim D_x, y \sim \pi_0(\cdot|x)} \ln \pi(y|x)$, where π_0 is some policy of a reasonable quality. This is done by adding $\alpha CE(\pi_0, \pi)$ to the objective or a constraint that $CE(\pi_0, \pi) \leq \lambda$. The addition keeps the loss optimization from degenerating when the distribution D underlying our data has a limited support, and the optimal policy $\widehat{\pi}$ might produce responses outside the support of D_{xy}, where we do not have preference feedback. We capture such additional constraints or regularizers as optimization with a reduced policy class Π.

- Early stopping: A different way to constrain the extent of optimization is directly in the parameter space. Most often the preferred optimizer is a first-order method such as Gradient Descent, AdaGrad, Adam or AdaFactor. Early stopping optimization with first-order methods directly corresponds to a constraint in the parameter space Θ (Yao et al., 2007; Raskutti et al., 2014; Neu & Rosasco, 2018; Suggala et al., 2018; Vaskevicius et al., 2019; Sonthalia et al., 2024). For example, if GD is used as the optimizer, early stopping would correspond to an ℓ_2 distance bound in parameter space between the parameters underlying π and the initial policy π_0 at the start of optimization. Early-stopping or limited training in typical fine-tuning setups is a common practice that we again capture through an appropriate choice for Π, e.g. with additional ℓ_2 constraints on the policy parameters.

We assume for simplicity that the policy class Π and the base policy μ are chosen such that $\log \frac{\pi(y|x)}{\mu(y|x)} - \log \frac{\pi(y'|x)}{\mu(y'|x)} \in [-\mathfrak{R}, \mathfrak{R}]$ for x, y, y' drawn according to the generative process described above, with probability 1. Consequently, the loss on each sample is also bounded by some B, and we expect that the empirical loss minimizer $\widehat{\pi}$ from (1) has also small population loss:

$$L_\mu(\widehat{\pi}; D_{xy\omega}) - \min_{\pi \in \Pi} L_\mu(\pi; D_{xy\omega}) \leq \epsilon_n, \quad \text{with} \quad (2)$$

$$L_\mu(\pi; D_{xy\omega}) = \mathbb{E}_{(x,y,y',\omega) \sim D_{xy\omega}}[\ell_\mu(\pi; \omega, x, y, y')],$$
$$\ell_\mu(\pi; \omega, x, y, y') = \ell(\omega \cdot \omega^{\pi,\mu}(x, y, y')).$$

The error bound ϵ_n typically scales as $O(\sqrt{Bd_\pi/n})$, with d_π being the statistical complexity of π, such as $\ln |\Pi|$ for finite classes, log-covering number for the infinite case, or other complexity measures like Rademacher or Gaussian complexities. We abstract such treatment into a general error term ϵ_n, as this analysis of the empirical loss minimizer is standard and not key focus of our study.

Existing offline RLHF methods. The formalization of preference-based learning above captures a wide range of existing offline RLHF methods through appropriate choices of the base policy μ and loss ℓ. Table 1 presents a selection of methods that fit the setup and which our formulation captures. We note that the methods predominantly vary along three axes: the loss function ℓ, the choice of the base policy μ and the choice of the constraint or regularizer to guide the policy optimization. In the next section, we present a theoretical framework and our main technical results on the quality of the learned policy $\widehat{\pi}$, with an emphasis on understanding the effects of these choices. We note here that the GPO paper of Tang et al. (2024) considers an almost identical set of design choices (other than the flexibility in μ and the choice of Π) as our work, but their emphasis is on empirical evaluation while we seek to understand the design space in theory.

While our setup is general and captures many existing RLHF methods, there are some approaches which are closely related but do not quite fit our framework. S-DPO, (Chen et al., 2024), samples multiple responses with a single response that is preferred to all others. Ramesh et al. (2024) study a version of the problem where multiple pairs of responses are sampled from multiple groups and the objective is to minimize the worst-case expected preference loss across groups. SimPO (Meng et al., 2024), nearly matches the DPO objective, however, the authors include additional length normalization, leading to the loss $\log(\omega \times \text{sigmoid}(\beta \log \pi(y|x)/|y| - \beta \log \pi(y'|x)/|y'|) - \gamma)$. There is no straightforward way to incorporate the length normalization in our analysis. One option is to incorporate the normalization into the reward class, however, this also affects the policy class definition. We note that the other main modification in SimPO, which is to remove the normalization by the log-likelihoods of the reference policy is indeed covered by our analysis through the choice of μ.

3. Analysis Framework and Main Results

In this section, we set up a framework for analyzing offline preference learning algorithms which optimize (1), and present our main results. We begin with a discussion to set up the performance criterion.

3.1. Analysis Framework

Performance criterion. How to measure the efficacy of a preference-based learning technique? As described above, based on our choices of ℓ, μ and Π, we get a guarantee on the expected loss of the resulting policy. But we want to measure how well the policy $\widehat{\pi}$ does in terms of producing highly preferred outputs y, given inputs x. It is not clear that a policy which has a small loss also produces good outputs. For instance, suppose that the learned $\widehat{\pi}$ is such that

$\mathbb{E}[\omega|x,y,y']\omega^{\widehat{\pi},\mu}(x,y,y') > 0$ for any x, y, y' in the support of our training distribution, and the base policy μ is uniform. In this case, we can conclude that $\widehat{\pi}(y|x) > \widehat{\pi}(y'|x)$ whenever $\mathbb{E}[\omega|x,y,y'] > 0$. But this does not preclude the two probabilities from being extremely close, though correctly ordered, and more generally $\widehat{\pi}$ might still place a non-trivial probability on the least desirable outputs y for many inputs x.

Ideally, we would like to say that $\widehat{\pi}$ places most of the mass on the most desirable outputs y. Since the conditional probabilities $D_\omega(\cdot|x,y,y')$ can be interpreted as a preference function $\mathbb{P}(y \succ y'|x)$, one notion of an optimal policy is provided by the Nash equilibrium policy for the two player-game encoded by this preference function, as considered in prior works (Wang et al., 2023; Munos et al., 2023; Swamy et al., 2024). However, as Swamy et al. (2024) show, this optimal solution cannot be attained by minimizers of the objective (1) in general, where they provide a lower bound for the special case of DPO. Consequently, we need a different yardstick to measure our performance for the setup of offline preference-based RL, which we do next. We begin with introducing some useful notation and our formal assumptions needed to define the benchmark policy.

Given the form of the loss in Equation 1, we have to reason about the log probabilities as the main object of interest. It is therefore convenient to denote for each policy π its log probabilities by R_π with $R_\pi(x,y) = \log \pi(y|x)$, and conversely by π_R the policy associated with R. We will also refer to such R as a *reward function* since it measures the quality of the output y for input x, and the policy π_R ascribes higher probability to outputs with high rewards under R. For the policy class Π we can then define the accompanying reward class as $\mathcal{R} = \{R_\pi : \pi \in \Pi\}$. Note that this is only nomenclature and is not a modeling assumption such as a reward-based Bradley-Terry-Luce (BTL) model of preferences in the data generating process.

Modeling assumptions. We start by assuming that the log-probabilities of all policies and the base policy μ are bounded, which ensures that the inputs of the loss function are in a bounded range. Additionally, we assume that also the loss outputs are bounded which holds for all practical loss functions, given the bounded domain $[-\mathfrak{R}, +\mathfrak{R}]$.

Assumption 3.1. For all x, y, y' and all $\pi \in \Pi$, we have $|R_\pi(x,y)| \leq \frac{\mathfrak{R}}{4}$, $|\log \mu(y|x)| \leq \frac{\mathfrak{R}}{4}$ and $\ell_\mu(\pi; \omega, x, y, y') \leq B$.

Given the policy class Π, we now make a realizability assumption on the data generating mechanism with respect to this class.

Assumption 3.2 (Realizability). There exists $\pi^\star \in \Pi$ such

that for all $x \in \mathcal{X}, y \in \mathcal{Y}, y' \in \mathcal{Y}$:

$$\omega^{\pi^\star,\mu}(x,y,y') = \underset{v \in [-\mathfrak{R}, \mathfrak{R}]}{\arg\min} \, \mathbb{E}_{D_\omega}[\ell(\omega \cdot v)|x,y,y']$$

We discuss in Appendix B.1 how this assumption can be relaxed to an approximate form of realizability, and still imply similar performance guarantees with a slightly different structural assumption on the loss function ℓ. We focus on exact realizability in the main text for the cleanest analysis. A necessary condition for Assumption 3.2 to hold is that there is a fixed policy $\pi^\star$ which minimizes the loss ℓ_μ in a pointwise manner for all x, y, y', when we take conditional expectation only over the preference labels ω. The assumption further requires that within the range of R_π which parameterizes $\pi \in \Pi$ we have that $\pi^\star$ is the pointwise minimizer of $\ell(\omega \cdot v)$. This makes the optimal policy $\pi^\star$ independent of the distributions D and D_y over x, y, y'. To further understand why it is helpful to have such an optimal policy $\pi^\star$, we make a standard calibration assumption on the loss function ℓ in Equation 1.

Assumption 3.3 (Proper loss). We assume that the loss function ℓ is a proper loss for class probability estimation (Reid & Williamson, 2010). That is, there is a function g_ℓ which depends only on ℓ, such that for all $\eta \in [0,1]$: $\arg\min_v \eta\ell(v) + (1-\eta)\ell(-v) = g_\ell(\eta)$.

That is, when we take conditional expectation over the binary label according to probability η, then the minimizer of the loss correctly recovers some loss-dependent function of η. This condition is satisfied by most commonly used differentiable losses for binary classification such as the logistic loss, squared loss, squared hinge loss etc. For instance, the function g_ℓ is given by $g_{\text{sq}}(\eta) = 2\eta - 1$ for squared loss and $g_{\log}(\eta) = \ln(\eta/(1-\eta))$ for the logistic loss.

Realizability, proper losses, and optimal policies. When we use proper losses, the realizability condition takes a particularly intuitive form when the data generating process and the g_ℓ function underlying the loss ℓ agree with each other. For instance, suppose we use the logistic loss and the preferences are generated according to a BTL model: $P(\omega = 1|x,y,y') = 1/(1 + \exp(R^\star(x,y') - R^\star(x,y)))$ for some $R^\star \in \mathcal{R}$. Then under Assumptions 3.2 and 3.3, we have that $\pi^\star = \pi_R$ for R such that for any x, y, y'

$$\begin{aligned} R(x,y) - R(x,y') = {} & R^\star(x,y) - R^\star(x,y') \\ & + \ln \mu(y|x) - \ln \mu(y'|x). \end{aligned} \quad (3)$$

That is, R is given by $R^\star + \ln \mu$, up to an x dependent offset, within the support of the data. A similar conclusion holds for the squared loss, and if $P(\omega = 1|x,y,y') = 0.5 + (R^\star(x,y) - R^\star(x,y'))/2$. A more detailed discussion of how modeling D_ω as part of the exponential family leads

to proper losses, ℓ, can be found in Appendix A. We see that in these cases, the policy underlying realizability learns a *ground-truth* reward function which underlies our data generation process. In such a scenario, where there is a reward function $R^\star$ underlying the observed preferences, a natural benchmark is the KL-regularized reward maximizing policy, $\pi(y|x) \propto \pi_0(y|x) \exp(\beta R^\star(x, y))$, where π_0 is some base policy with (such as the SFT policy), with respect to which the KL divergence is defined. When $\mu = \pi_0$, then we see that the policy $\pi^\star$ exactly corresponds to this optimal policy for an appropriately chosen loss function.

Performance criterion under realizability. Based on these insights, we adopt the policy $\pi^\star$ as our performance yardstick, and seek a policy π to minimize $\mathrm{KL}(\pi^\star||\pi)$. Approximately minimizing L_μ may not be sufficient to derive meaningful bounds on $\mathrm{KL}(\pi^\star||\pi)$, however, as there still needs to be alignment between the data-generating distribution D_y and $\pi^\star$. Otherwise, if there is no good coverage of the support of $\pi^\star(\cdot|x)$ by $D_y(\cdot|x)$ there is no guarantee that π will be able to distinguish good responses, y, according to $\pi^\star$ from highly sub-optimal ones. This necessitates making the following coverage assumption.

Assumption 3.4 (Coverage for optimal policy). Let $\bar{R}(x, y) = R(x, y) - \mathbb{E}_{D_y}[R(x, y)|x]$ denote the y-centered reward and let $R^\star$ denote the parameterization of $\pi^\star$. Further, let $\Delta\bar{R}(x, y) = \bar{R}(x, y) - \bar{R}^\star(x, y)$. We assume that there exists a constant C s.t. for $R \in \mathcal{R}$ it holds that

$$\mathbb{E}_{x, y \sim \pi^\star(\cdot|x)}[\Delta\bar{R}(x, y)^2] \leq C \, \mathbb{E}_{x, y \sim D}[\Delta\bar{R}(x, y)^2].$$

The coverage condition is akin to generalized coverage conditions used in the offline RL literature (Xie et al., 2021; Jiang & Xie, 2024). A sufficient condition to ensure this holds is to have $\sup_{x, y} \frac{\pi^\star(y|x)}{D_y(y|x)} \leq C$, but the generalized notion also holds when the class $\mathcal{R} = \{w^\top \phi(x, y) : \|w\|_2 \leq 1\}$ and we have that $\lambda_{\max}(\Sigma_y^{-1/2} \Sigma_{\pi_\mu^\star} \Sigma_y^{1/2}) \leq C$. Here we denote $\Sigma_\pi = \mathbb{E}_{(x \sim D, y \sim \pi(\cdot|x)}[\phi(x, y)\phi(x, y)^\top]$, and abbreviate $\Sigma_y = \Sigma_{D_y}$. Clearly, this second condition can be much weaker than the density ratio assumption, and indeed has underpinned several methods that effectively handle coverage issues in offline RL with large function spaces and high-dimensional data, motivating our definition here.

Before stating our main result we need a somewhat standard curvature assumption on the loss.

Assumption 3.5 (Curvature around optimum). For any policy $\pi \in \Pi$ and functions $R, R^\star \in \mathcal{R}$ such that $\pi = \pi_R, \pi^\star = \pi_{R^\star}$, there is a constant $c_\mu > 0$ such that

$$L_\mu(\pi; D_{xy\omega}) - L_\mu(\pi_\mu^\star; D_{xy\omega}) \geq \mathbb{E}\left[\omega \ell'(\omega \omega^{\pi^\star, \mu}(x, y, y')) \right.$$
$$\cdot (\omega^{\pi, \mu}(x, y, y') - \omega^{\pi^\star, \mu}(x, y, y'))\Big]$$
$$+ \frac{c_\mu}{2} \mathbb{E}\left[\left(\omega^{\pi^\star, \mu}(x, y, y') - \omega^{\pi, \mu}(x, y, y')\right)^2\right]$$

A sufficient condition for Assumption 3.5 is to instead have the stronger condition that for any $u, v \in [-\mathfrak{R}, \mathfrak{R}]$, we have

$$\ell(u) \geq \ell(v) + \ell'(v) \cdot (u - v) + \frac{c_\mu}{2}(u - v)^2.$$

This condition holds for $c_\mu = \frac{1}{2}$ for the squared loss: $\ell(u, v) = (u - v)^2$, and with an $\mathfrak{R}$-dependent constant for many other losses that are induced by log-likelihoods of exponential families, which includes the logistic loss and the probit loss. Assumption 3.5 weakens this condition by requiring curvature on the expected loss L only around the optimal policy $\pi^\star$, rather than pointwise on ℓ.

3.2. Main Results

With our main modeling assumptions set up, we now give the main theoretical result on the KL divergence between an approximate minimizer of the population loss $L_\mu(\pi; D_{xy\omega})$ and the benchmark $\pi^\star$.

Theorem 3.6. *For any $\pi \in \Pi$ such that $L_\mu(\pi; D_{xy\omega}) - L_\mu(\pi_\mu^\star; D_{xy\omega}) \leq \epsilon$, where the corresponding loss to L_μ, given by ℓ is proper, and under Assumptions 3.1-3.5, it holds that*

$$\mathbb{E}_x\left[\mathrm{KL}(\pi^\star(\cdot|x)||\pi(\cdot|x))\right] \leq \sqrt{\frac{\epsilon}{c_\mu}} + \sqrt{\frac{C\epsilon}{c_\mu}} + \frac{e^\mathfrak{R}}{2} \cdot \frac{C\epsilon}{c_\mu}.$$

In Section B we show a version of Theorem 3.6 which only uses approximate realizability, that is, the version utilizes a relaxed version of Assumption 3.2.

Remark 3.7 (Choice of loss function). Our bound scales inversely with the curvature constant, meaning that losses with high curvature will lead to more favourable bounds in Theorem 3.6. The squared loss satisfies Assumption 3.5 with $c_\mu = 1$ for any range of $\omega^{\pi^\star, \mu}$ as it is strongly convex, while the squared hinge loss only satisfies Assumption 3.5 in $(-\infty, 1)$, finally the logistic loss satisfies Assumption 3.5 with a range dependent c_μ as the loss becomes less curved as the $\omega^{\pi^\star, \mu}$ approaches ∞. Our theorem suggests that optimizing the squared loss is ideal in terms of the final bound as c_μ is constant and bounded away from 0 across the full range of the loss. Squared loss is also a proper loss, so that the main assumption which might fail is realizability. We do note that the probabilistic model corresponding to squared loss naturally is less realistic than say, the BTL model, corresponding to the logistic loss. Nevertheless, the benefits of squared loss are verified by our experiments in Section 5. We note that a related discussion of the curvature properties comparing the squared and logistic cases can also be found in Azar et al. (2024), who correctly identify similar issues with the curvature of the logistic loss, and use this to motivate the IPO algorithm. However, there is no

explicit quantitative analysis of the role of the link function beyond a simple example, since the focus is more on obtaining a practical algorithm for the identity link function case. We instead derive precise convergence guarantees, and also highlight the role of offline data coverage which is not captured in the IPO paper.

Remark 3.8 (Choice of base policy). While the choice of base policy μ does not appear directly in the KL bound, it influences the realizability assumption. As already discussed, under a proper loss such as the logistic loss and a corresponding reward model such as BTL, realizability becomes equivalent to having the reward model plus a $\log \mu$ term be part of the reward space $\mathcal{R}$. Further, the choice of μ can change $\pi^\star$, and as we discuss in Section 3.1, the choice of $\mu = \pi_{\text{ref}}$ naturally yields a desirable $\pi^\star$. In our experiments we use two commonly studied choices of μ, the uniform policy which puts equal probability on all responses, and a SFT policy π_{ref}.

Remark 3.9 (Effect of constraints). Recall that we constrain the optimization problem to a policy class Π which captures any constraints that we incorporate such as CE to a SFT policy, π_0, or the implicit constraints induced by the choice of optimizer and early stopping. We note that the choice of Π determines if the benchmark, $\pi^\star$, which satisfies the realizability and curvature assumptions has to be part of Π. In this context, a cross-entropy regularization to π_{ref} essentially makes an assumption that $\pi^\star$ lies in the vicinity of π_{ref}. Biasing the reference policy in cross-entropy towards preferred responses such as in CPO (Xu et al., 2024) can be further beneficial in ensuring the feasibility of $\pi^\star$.

Remark 3.10 (Connections with prior results). As mentioned earlier, there is now a substantial literature on the degeneracies of DPO in particular, due to its popularity, with primarily empirical (Park et al., 2024; Rafailov et al., 2024a; Fisch et al., 2024), but also some theoretical results in these works that demonstrate that DPO tends to shift mass away from the support of the preference data, with probabilities of both preferred and dispreferred responses in the data rapidly degrading to zero. The reader might wonder how to reconcile these negative observations with our positive result on the loss minimizers of DPO-style losses. However, there are a few caveats which apply to the specific case of DPO. First, as we remarked earlier, the curvature constant c_μ for DPO degrades exponentially fast with $\mathfrak{R}$. While this scaling is not ideal we note that it is typical and appears in prior and concurrent works as well (Zhao et al., 2024; Zhu et al., 2023; Xiong et al., 2023). Further, since DPO does not control the log-likelihood ratios through regularization terms, the quantity $\mathfrak{R}$ can rapidly grow large empirically, as pointed out in multiple papers, and as we corroborate in the next section. In fact, some works on incorporating pessimistic reasoning in DPO (Fisch et al., 2024; Liu et al., 2024; Cen et al., 2024; Huang et al., 2024) result in adding regularization terms which partly mitigate some of these degeneracies.

4. Analysis

The high-level reasoning to prove Theorem 3.6 is the following. We first use Assumption 3.5 to establish that any ϵ-minimizer of $L(\pi_R; D_{xy\omega})$ admits a bound on the expected error in the centered rewards $\mathbb{E}_{(x,y)\sim D}[\Delta \bar{R}(x,y)^2]$. We then invoke the coverage condition of Assumption 3.4 to translate this error bound to be under the benchmark policy $\pi^\star$. Subsequently, we relate the KL divergence between π and $\pi^\star$ in terms of expectation of $\Delta \bar{R}(x,y)^2$ under $\pi^\star$, using a careful anaylsis of the log-partition function.

We start with the first step in the sketch above.

Lemma 4.1. *Under Assumptions 3.2 and 3.5, any policy* $\pi \in \Pi$ *with* $L_\mu(\pi; D_{xy\omega}) - L_\mu(\pi^\star; D_{xy\omega}) \leq \epsilon$, *satisfies*

$$\mathbb{E}_{x,y\sim D}\left[\Delta \bar{R}(x,y)^2\right]$$
$$= \mathbb{E}_{x,y,y'\sim D}[(\Delta R(x,y) - \Delta R(x,y'))^2] \leq \frac{2\epsilon}{c_\mu}.$$

Proof. The first condition of the lemma is equivalent to

$$\mathbb{E}[\ell\left(\omega\omega^{\pi,\mu}(x,y,y')\right)] \geq \mathbb{E}[\ell\left(\omega\omega^{\pi^\star,\mu}(x,y,y')\right)]$$
$$+ \mathbb{E}[\omega\ell'\left(\omega\omega^{\pi^\star,\mu}(x,y,y')\right)\left((\omega^{\pi,\mu}(x,y,y') - \omega^{\pi^\star,\mu}(x,y,y'))\right)]$$
$$+ \mathbb{E}[\frac{c_\mu}{2}\left(\omega(R(x,y) - R(x,y') - R^\star(x,y) + R^\star(x,y'))\right)^2]$$

Now, Assumption 3.2 implies that for any $\pi \in \Pi$,

$$\mathbb{E}_{\omega|x,y,y'}[\ell'\left(\omega\omega^{\pi^\star,\mu}(x,y,y')\right)$$
$$\cdot \left(\omega(\omega^{\pi,\mu}(x,y,y') - \omega^{\pi^\star,\mu}(x,y,y'))\right)] \geq 0,$$

where we have used the fact that $v^\star = \omega^{\pi^\star,\mu}(x,y,y')$ is the minimizer of $\mathbb{E}_{\omega|x,y,y'}[\ell(\omega \cdot v^\star)]$ together with first order optimality so that $\mathbb{E}_{\omega|x,y,y'}[\ell'(\omega v^\star)\omega(v - v^\star)] \geq 0$ for any v and in particular for any, π, such that $v = \omega^{\pi,\mu}(x,y,y')$. This, together with the second assumption of the lemma imply that

$$\epsilon \geq \mathbb{E}[\ell\left(\omega\omega^{\pi,\mu}(x,y,y')\right)] - \mathbb{E}[\ell\left(\omega\omega^{\pi^\star,\mu}(x,y,y')\right)]$$
$$\geq \frac{c_\mu}{2}\mathbb{E}[(\omega(R(x,y) - R(x,y') - R^\star(x,y) + R^\star(x,y')))^2],$$

where we have used the parametrization of π and $\pi^\star$. $\square$

Next we show how to bound the KL by $\Delta \bar{R}$ and the fraction of log-partition functions $\bar{Z}_\pi$ and $\bar{Z}_\star$, where $\bar{Z}_\pi(x) = \sum_y \exp(\bar{R}_\pi(x,y))$.

Lemma 4.2. *Under Assumption 3.4, for any* $\pi \in \Pi$ *such that* $\pi \propto \exp(\bar{R}(x,y))$, *the expected KL divergence* $\mathbb{E}_x\left[\text{KL}(\pi^\star(\cdot|x)||\pi(\cdot|x))\right]$ *is bounded by*

$$\sqrt{2C\,\mathbb{E}_{x,y}[\Delta \bar{R}(x,y)^2]} + \mathbb{E}_x\left|\log \frac{\bar{Z}_\pi(x)}{\bar{Z}_\star(x)}\right|.$$

Proof.

$$\mathbb{E}_x[\mathrm{KL}(\pi^\star(\cdot|x)||\pi(\cdot|x))] \le \mathbb{E}_x\left[\left|\mathbb{E}_{y\sim\pi^\star(\cdot|x)}\log\frac{\pi^\star(y|x)}{\pi(y|x)}\right|\right]$$

$$\le \mathbb{E}_x\sqrt{2\,\mathbb{E}_{y\sim\pi^\star(\cdot|x)}\left(\log\frac{\pi^\star(y|x)}{\pi(y|x)}-\log\frac{\bar{Z}_\star(x)}{\bar{Z}_\pi(x)}\right)^2}$$

$$+\ \mathbb{E}_x\sqrt{2\,\mathbb{E}_{y\sim\pi^\star(\cdot|x)}\left(\log\bar{Z}_\pi(x)-\log\bar{Z}_\star(x)\right)^2}$$

$$\le\ \sqrt{2C\,\mathbb{E}_{x,y}[\Delta\bar{R}(x,y)^2]}+\mathbb{E}_x\left|\log\frac{\bar{Z}_\pi(x)}{\bar{Z}_\star(x)}\right|,$$

where in second inequality follows from Jensen's inequality, and the third inequality follows from a combination of $(x+y)^2 \le 2x^2 + 2y^2$ and $\sqrt{x+y} \le \sqrt{x}+\sqrt{y}, \forall x,y \ge 0$, and the third inequality uses Jensen to push the expectation inside the square root, along with Assumption 3.4. $\quad\square$

The next lemma bounds the ratio of log-partitions.

Lemma 4.3. *For any π such that $L(\pi; D_{xy\omega}) - L(\pi^\star; D_{xy\omega}) \le \epsilon$, it holds that*

$$\mathbb{E}_x\left|\log\frac{\bar{Z}_\pi(x)}{\bar{Z}_\star(x)}\right| \le \sqrt{\frac{C\epsilon}{c_\mu}}+\frac{e^{\mathfrak{R}}}{2}\cdot\frac{C\epsilon}{c_\mu}$$

Proof. Using the definition of $\bar{Z}_\pi$ we have

$$\mathbb{E}_x\left|\log\frac{\bar{Z}_\pi(x)}{\bar{Z}_\star(x)}\right| = \mathbb{E}_x\left|\log\sum_y\frac{\exp(\bar{R}_\pi(x,y))}{\bar{Z}_\star(x)}\right|$$

$$= \mathbb{E}_x\left|\log\sum_y\pi^\star(y|x)\frac{\exp(\bar{R}_\pi(x,y))}{\exp(\bar{R}_\star(x,y))}\right|.$$

Here the second equation rearranges the definition $\pi^\star(y|x) = \exp(\bar{R}_\star(x,y))/\bar{Z}_\star(x)$. Proceeding further

$$\mathbb{E}_x\left|\log\frac{\bar{Z}_\pi(x)}{\bar{Z}_\star(x)}\right| = \mathbb{E}_x\left|\ln\mathbb{E}_{y\sim\pi^\star(\cdot|x)}\exp(\Delta\bar{R}(x,y))\right|$$

$$\le \mathbb{E}_x\left|\ln\left(1+\mathbb{E}_{y\sim\pi_\star(\cdot|x)}\Delta\bar{R}(x,y)\right.\right.$$

$$\left.\left.+\frac{e^{\mathfrak{R}}}{2}\mathbb{E}_{y\sim\pi^\star(\cdot|x)}\Delta\bar{R}(x,y)^2\right)\right|$$

$$\le \mathbb{E}_x\left|\mathbb{E}_{y\sim\pi^\star(\cdot|x)}\Delta\bar{R}(x,y)\right.$$

$$\left.+\frac{e^{\mathfrak{R}}}{2}\mathbb{E}_x\mathbb{E}_{y\sim\pi^\star(\cdot|x)}\Delta\bar{R}(x,y)^2\right|.$$

Here the first inequality uses that $e^x \le 1 + x + e^A x^2/2$, $\forall x \le A$. The second inequality uses $\ln(1+x) \le x$. Con-

tinuing with our simplification, we get the bound

$$\mathbb{E}_x\left|\log\frac{\bar{Z}_\pi(x)}{\bar{Z}_\star(x)}\right| \le \sqrt{\mathbb{E}_x\mathbb{E}_{y\sim\pi^\star(\cdot|x)}\Delta\bar{R}(x,y)^2}$$

$$+\frac{e^{\mathfrak{R}}}{2}\mathbb{E}_x\mathbb{E}_{y\sim\pi^\star(\cdot|x)}\Delta\bar{R}(x,y)^2$$

$$\le \sqrt{\frac{C\epsilon}{c_\mu}}+\frac{e^{\mathfrak{R}}}{2}\cdot\frac{C\epsilon}{c_\mu}.$$

The first inequality above follows by triangle inequality to the sum inside absolute value, followed by Jensen's inequality on the first term. Finally, we invoke Assumption 3.4 and Lemma 4.1 to control the last term, which completes the proof of the lemma. $\quad\square$

Combining Lemma 4.2 and Lemma 4.3 finishes the proof of Theorem 3.6.

5. Experiments

We evaluate the impact of different design choices for offline RLHF methods on the standard TL;DR summarization task (Völske et al., 2017; Stiennon et al., 2020), where the task is to provide short summaries of articles. For our policy, we use a large T5 model (Raffel et al., 2020) with 770M parameters, which has been fine-tuned to maximize the log-likelihood of the human responses in the TL;DR dataset. This ensures that the policy is initialized with parameters that have reasonable likelihoods for the responses observed in our data. We experiment with two different losses, ℓ:

$$\ell(x) = \log(1 + \exp(-\beta x)) \qquad \text{(Logistic loss)}$$
$$\ell(x) = (\beta x - 1)^2, \qquad \text{(Squared loss)}$$

where $\beta > 0$ is a hyper-parameter, governing the strength of how much each preference in the dataset should in the policy. We pair each of losses with two possible choices for the base policy μ: The uniform base policy $\mu = 1$, which results in the following optimization

$$\min_{\pi\in\Pi}\sum_{i=1}^n\ell\big(-\omega_i\big(\log\pi(y_i|x_i)-\log\pi(y_i'|x_i)\big)\big),$$

and the second choice uses $\mu = \pi_{\text{ref}}$, the policy obtained after fine-tuning on TL;DR which our optimization is initialized with. This corresponds to the objective is:

$$\min_{\pi\in\Pi}\sum_{i=1}^n\ell\big(-\omega_i\big(\log\tfrac{\pi(y_i|x_i)}{\pi_{\text{ref}}(y_i|x_i)}-\log\tfrac{\pi(y_i'|x_i)}{\pi_{\text{ref}}(y_i'|x_i)}\big)\big).$$

We recall that when $\mu = \pi_{\text{ref}}$, using the logistic loss from (Logistic loss) corresponds to the DPO algorithm (Rafailov et al., 2024b) and the squared loss from (Squared loss) corresponds to IPO (Azar et al., 2024).
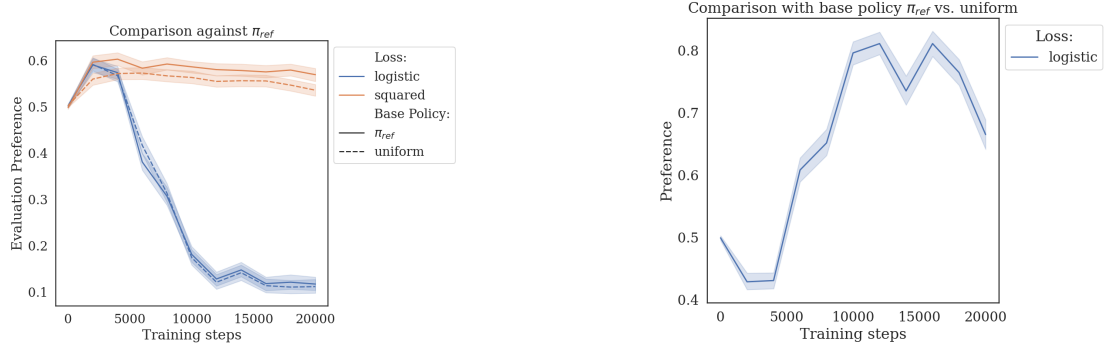
Figure 1. **Left** panel shows the preference of the learned policy's summaries against those from the initial policy π_{ref}, as evaluated by a prompted Gemini 1.0 Ultra model. Shaded regions represent 95% error bands. Both the logistic loss variants quickly improve in terms of the preference scores initially, but then suffer a catastrophic collapse. Squared loss improves at a similar rate initially, and remains stable throughout the training regime. **Right** panel shows a direct comparison between the variants of logistic loss using $\mu = $ uniform and $\mu = \pi_{\text{ref}}$ (DPO) at regular intervals in the training process. Interestingly, the uniform variant is preferred in the early stages of training, but as the training collapses around the training step 5K, the π_{ref} variant starts to improve. Nevertheless, the absolute performance of both variants reaches its peak earlier in the training and rapidly worsens after 5K steps, suggesting that the preference for π_{ref} over uniform in this region might not be particularly significant. See text for a more nuanced discussion.

For each variant, we tuned the β parameter in the loss in the interval $\{0.1, 0.5, 1.0\}$. For logistic loss, the best results are obtained at $\beta = 0.1$, while we did not see a significant difference across these choices for the squared loss, and show the results at $\beta = 0.5$. See Appendix C for details.

We evaluate the different variants in two ways. First we compare the policies generated by each method at regular intervals in the training process, with the initial policy π_{ref}, by comparing the generated summaries in terms of their quality and conciseness by a prompted Gemini 1.0 Ultra model (Team et al., 2023). Figure 1 (left) shows that squared loss variants perform significantly better than those with logistic loss, which reach a peak preference at 2k steps and suffer from a dramatic collapse afterwards. In the case of $\mu = \pi_{\text{ref}}$, this collapse is consistent with prior findings on the DPO algorithm in the literature (Fisch et al., 2024; Rafailov et al., 2024a). In comparison, both squared loss variants reach peak performance after similar number of steps, but maintain that performance more stably afterwards.

Notably, using $\mu = \pi_{\text{ref}}$ is consistently better than $\mu = 1$, when combined with the squared loss. In case of logistic loss, the comparison against π_{ref} in Figure 1 (left) suggests that the base policy choice does not affect performance. However, when we compare the policies produced by the two variants with $\mu = 1$ and $\mu = \pi_{\text{ref}}$ directly in Figure 1 (left), we do observe a strong impact. Initially, the uniform variant is slightly preferred with a strong preference afterwards in the other direction, after roughly 5K steps. However, at this point the preference of each variants against π_{ref} has already collapsed, indicating a perhaps less meaningful comparison between two sets of bad responses in this region.

To further understand the training dynamics, we plot the log-probabilities of the preferred and dispreferred responses from the preference data for all the variants in Figure 2. For the logistic loss, these plots demonstrate that while both the likelihoods rapidly decrease to zero, the rate is faster for the dispreferred responses. In terms of our analysis setup, this corresponds to $\omega^{\pi,\mu}$ being large, which in turn leads to a small value of the curvature constant c_μ and a large value of $\mathfrak{R}$. This makes the bound in Theorem 3.6 drastically worse, confirming our theoretical findings empirically. In the case of $\mu = \pi_{\text{ref}}$, corresponding to DPO, a similar collapse of log probabilities was also observed in prior works (Fisch et al., 2024; Rafailov et al., 2024b). While the log-likelihoods also decline for the squared loss, the decrease is milder, which in turn means that the magnitude of $\omega^{\pi,\mu}$ and the necessary value of $\mathfrak{R}$ remain adequately bounded. This, in addition with the better curvature constant of the squared loss makes the findings to be consistent with the theory.

In Table 2 we report approximations to c_μ and $\mathfrak{R}$ for all the different loss variants, as measured empirically. We approximate $\mathfrak{R}$ by the value of $\omega^{\pi,\mu}$ averaged over the last mini-batch of training. Further, we approximate c_μ by the curvature of ℓ_μ for the values of $\mathfrak{R}$ and $\omega^{\pi,\mu}$ obtained this way. We note that $\omega^{\pi,\mu}$ decreases monotonically during training and so the approximation to $\mathfrak{R}$ and c_μ is tightest at the end of training. The reported values of c_μ and $\mathfrak{R}$ are consistent with our theory and the empirical observations in Figures 1 and 2.
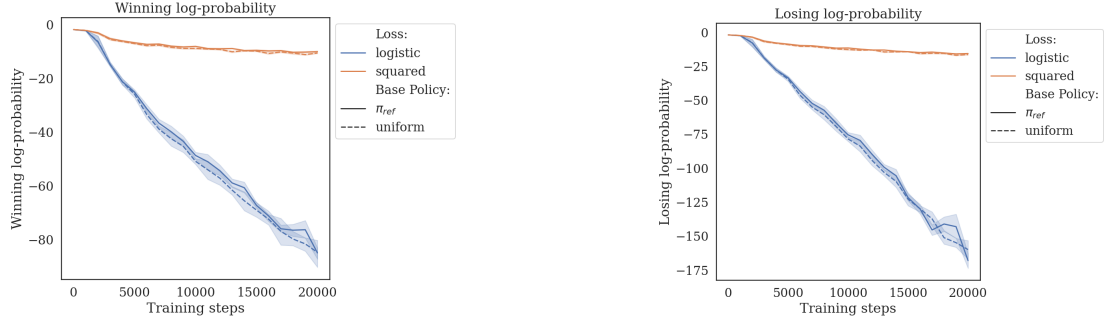
Figure 2. Evolution of the log-likelihoods of the preferred response (**left**) and dispreferred response (**right**) from the preference dataset across the training process. Both variants of the squared loss decrease the log-likelihoods of both the responses during training, but the decrease is relatively mild. The logistic loss, on the other hand, sends these log-likelihoods crashing sharply, even though the dispreferred responses have significantly lower values, so the difference of log-likelihoods remains highly negative, driving the loss to zero. We suspect that this degeneration of log-likelihoods is responsible for the eventual collapse observed for the logistic loss in Figure 1.

Loss	Base policy	$\mathfrak{R}$	c_μ
$\ell(z) = -\log\left(\frac{1}{1+\exp(-\beta z)}\right)$	$\mu = \pi_{\text{ref}}$	73.203	0.00066
$\ell(z) = -\log\left(\frac{1}{1+\exp(-\beta z)}\right)$	$\mu = 1$	72.949	0.00068
$\ell(z) = (\beta z - 1)^2$	$\mu = \pi_{\text{ref}}$	4.678	2.0
$\ell(z) = (\beta z - 1)^2$	$\mu = 1$	4.924	2.0

Table 2. We approximate $\mathfrak{R}$ by $\omega^{\pi,\mu}$ averaged over the last mini-batch and approximate c_μ by computing the curvature of ℓ_μ at $\omega^{\pi,\mu}$.

6. Discussion

This works adopts a different line of reasoning than the typical reparameterization arguments to motivate the correspondence between offline and online RLHF techniques, with a goal of understanding the impact of design choices in the offline methods, many of which do not fit cleanly in the arguments for equivalence of online and offline methods. Our theory and experiments collectively indicate that perhaps the limited coverage of offline data, and the propensity of log-likelihoods of the preference data to precipitously drop in certain methods are the key obstacles to reliable learning in this setup. Our findings suggest that using losses which do not decay to zero at a slow rate, like the logistic loss, and using experimental design techniques for data collection before offline RLHF can be fruitful avenues for addressing the concerns uncovered here.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Ahmadian, A., Cremer, C., Gallé, M., Fadaee, M., Kreutzer, J., Pietquin, O., Üstün, A., and Hooker, S. Back to basics: Revisiting reinforce style optimization for learning from human feedback in llms. *arXiv preprint arXiv:2402.14740*, 2024.

Amini, A., Vieira, T., and Cotterell, R. Direct preference optimization with an offset. *arXiv preprint arXiv:2402.10571*, 2024.

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

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

Cen, S., Mei, J., Goshvadi, K., Dai, H., Yang, T., Yang, S., Schuurmans, D., Chi, Y., and Dai, B. Value-incentivized preference optimization: A unified approach to online and offline rlhf. *arXiv preprint arXiv:2405.19320*, 2024.

Chen, Y., Tan, J., Zhang, A., Yang, Z., Sheng, L., Zhang, E., Wang, X., and Chua, T.-S. On softmax direct preference optimization for recommendation. *arXiv preprint arXiv:2406.09215*, 2024.

Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

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

Fisch, A., Eisenstein, J., Zayats, V., Agarwal, A., Beirami, A., Nagpal, C., Shaw, P., and Berant, J. Robust preference optimization through reward model distillation. *arXiv preprint arXiv:2405.19316*, 2024.

Huang, A., Zhan, W., Xie, T., Lee, J. D., Sun, W., Krishnamurthy, A., and Foster, D. J. Correcting the mythos of kl-regularization: Direct alignment without overoptimization via chi-squared preference optimization. *arXiv preprint arXiv:2407.13399*, 2024.

Jiang, N. and Xie, T. Offline reinforcement learning in large state spaces: Algorithms and guarantees. *Statistical Science*, 2024.

Liu, Z., Lu, M., Zhang, S., Liu, B., Guo, H., Yang, Y., Blanchet, J., and Wang, Z. Provably mitigating overoptimization in rlhf: Your sft loss is implicitly an adversarial regularizer. *arXiv preprint arXiv:2405.16436*, 2024.

Luce, R. *Individual Choice Behavior: A Theoretical Analysis*. Dover Books on Mathematics. Dover Publications, 2012. ISBN 9780486153391. URL https://books.google.com/books?id=ERQsKkPiKkkC.

Meng, Y., Xia, M., and Chen, D. SimPO: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024.

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

Neu, G. and Rosasco, L. Iterate averaging as regularization for stochastic gradient descent. In *Conference On Learning Theory*, pp. 3222–3242. PMLR, 2018.

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

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

Park, R., Rafailov, R., Ermon, S., and Finn, C. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*, 2024.

Rafailov, R., Chittepu, Y., Park, R., Sikchi, H., Hejna, J., Knox, B., Finn, C., and Niekum, S. Scaling laws for reward model overoptimization in direct alignment algorithms. *arXiv preprint arXiv:2406.02900*, 2024a.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024b.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21 (140):1–67, 2020.

Ramesh, S. S., Hu, Y., Chaimalas, I., Mehta, V., Sessa, P. G., Bou Ammar, H., and Bogunovic, I. Group robust preference optimization in reward-free rlhf. *Advances in Neural Information Processing Systems*, 37:37100–37137, 2024.

Raskutti, G., Wainwright, M. J., and Yu, B. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *The Journal of Machine Learning Research*, 15(1):335–366, 2014.

Reid, M. D. and Williamson, R. C. Composite binary losses. *The Journal of Machine Learning Research*, 11:2387–2422, 2010.

Sonthalia, R., Lok, J., and Rebrova, E. On regularization via early stopping for least squares regression. *arXiv preprint arXiv:2406.04425*, 2024.

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

Suggala, A., Prasad, A., and Ravikumar, P. K. Connecting optimization and regularization paths. *Advances in Neural Information Processing Systems*, 31, 2018.

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

Tang, Y., Guo, Z. D., Zheng, Z., Calandriello, D., Munos, R., Rowland, M., Richemond, P. H., Valko, M., Pires, B. A., and Piot, B. Generalized preference optimization:

A unified approach to offline alignment. In *Forty-first International Conference on Machine Learning*, 2024.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A., et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Vaskevicius, T., Kanade, V., and Rebeschini, P. Implicit regularization for optimal sparse recovery. *Advances in Neural Information Processing Systems*, 32, 2019.

Völske, M., Potthast, M., Syed, S., and Stein, B. Tl; dr: Mining reddit to learn automatic summarization. In *Proceedings of the Workshop on New Frontiers in Summarization*, pp. 59–63, 2017.

Wang, Y., Liu, Q., and Jin, C. Is rlhf more difficult than standard rl? *arXiv preprint arXiv:2306.14111*, 2023.

Xie, T., Cheng, C.-A., Jiang, N., Mineiro, P., and Agarwal, A. Bellman-consistent pessimism for offline reinforcement learning. *Advances in neural information processing systems*, 34:6683–6694, 2021.

Xiong, W., Dong, H., Ye, C., Wang, Z., Zhong, H., Ji, H., Jiang, N., and Zhang, T. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. *arXiv preprint arXiv:2312.11456*, 2023.

Xu, H., Sharaf, A., Chen, Y., Tan, W., Shen, L., Van Durme, B., Murray, K., and Kim, Y. J. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*, 2024.

Yao, Y., Rosasco, L., and Caponnetto, A. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

Zhao, H., Ye, C., Gu, Q., and Zhang, T. Sharp analysis for kl-regularized contextual bandits and rlhf. *arXiv preprint arXiv:2411.04625*, 2024.

Zhao, Y., Khalman, M., Joshi, R., Narayan, S., Saleh, M., and Liu, P. J. Calibrating sequence likelihood improves conditional language generation. *arXiv preprint arXiv:2210.00045*, 2022.

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

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

A. From a preference model to a proper loss

A natural family for the reward generating distributions is the exponential family. In particular we are going to assume that D_ω is in the exponential family parametrized by some $v^\star := v^\star(x, y, y')$, that is $\mathbb{P}(\omega|x, y, y') = \exp(\omega v^\star - \phi(v^\star))$, where ϕ is some strictly convex function. Considering the exponential family naturally leads to an objective function for learning the unknown parameter v, that is to maximize the log-likelihood which will recover $v^\star$ in the following way. First we take the derivative of the log-likelihood with respect to v

$$\nabla_v \mathbb{E}_\omega[\log \exp(\omega v - \phi(v))] = \mathbb{E}_\omega[\omega - \nabla\phi(v)] = \nabla\phi(v^\star) - \nabla\phi(v).$$

Setting the derivative to 0 and using the strict convexity to invert $\nabla\phi(v)$ shows that $v^\star = \nabla\phi^{-1}(\mathbb{E}_\omega[\omega])$. To summarize the above, a preference model which follows exponential family distribution gives rise to a natural loss function given by

$$\min_{v \in [-\mathfrak{R}, \mathfrak{R}]} E_\omega[\phi(v) - \omega v],$$

and has a closed form solution $v^\star = \nabla\phi^{-1}(\mathbb{E}_\omega[\omega])$.

To make this discussion concrete we focus on a BTL model, where we set $v(x, y, y') = R(x, y) - R(x, y')$. Then $\phi(v(x, y, y')) = \log(\exp(R(x, y) - R(x, y')) + \exp(R(x, y') - R(x, y)))$ and the corresponding loss function is then

$$\begin{aligned}
\phi(v(x, y, y')) - \omega v(x, y, y') &= \log(\exp(R(x, y) - R(x, y')) + \exp(R(x, y') - R(x, y))) \\
&\quad - \log(\exp(\omega(R(x, y) - R(x, y')))) \\
&= \log(1 + \exp(-\omega(R(x, y) - R(x, y')))),
\end{aligned}$$

which is precisely the logistic loss. The above derivation already shows that any loss derived from the exponential family with link function $\nabla\phi$ is going to be proper as ϕ is strictly convex and hence $\nabla\phi$ is invertible with g_ℓ in Assumption 3.3 satisfying $g_\ell \equiv \nabla\phi^{-1}$. We can further check that for a minimizer, v, of the logistic loss we must have

$$\nabla_v \left(\eta \log(1 + \exp(-v)) + (1 - \eta) \log(1 + \exp(v))\right) = -\frac{\eta \exp(-v)}{1 + \exp(-v)} + \frac{(1 - \eta) \exp(v)}{1 + \exp(v)} = 0,$$

where $\eta = \mathbb{P}(\omega = 1|x, y, y')$. The above implies $v = \log \frac{\eta}{1-\eta}$. Equivalently, we could have computed the derivative of the convex conjugate of ϕ, $\nabla\phi^\star$, which is precisely $\nabla\phi^{-1}$. Finally, to establish the claimed connection between the parametrization of Π to the BTL parametrization with $v^\star$ we have the following. The fact that the logistic loss is proper implies

$$R(x, y) - R(x, y') - \log \frac{\mu(y|x)}{\mu(y'|x)} = \omega^{\pi, \mu}(x, y, y')$$

$$= \log \frac{\eta}{1 - \eta} = \log \frac{\exp(R^\star(x, y) - R^\star(x, y'))}{Z(x) - (R^\star(x, y) - R^\star(x, y'))},$$

which further simplifying gives Equation 3.

A similar line of reasoning shows that when $\ell(v) = (1 - v)^2$, the link function satisfies $\eta = \frac{1 + v^\star}{2}$ and so the resulting reward model is $\mathbb{P}(\omega|x, y, y') = \frac{1 + R^\star(x, y) - R^\star(x, y')}{2}$.

B. Approximate realizability

We now relax Assumption 3.2 to the following.

Assumption B.1 (Approximate Realizability). There exists $\pi^\star \in \Pi$ such that for all $x \in \mathcal{X}, y \in \mathcal{Y}, y' \in \mathcal{Y}$:

$$\mathbb{E}_{D_\omega}[\ell(\omega \cdot \omega^{\pi^\star, \mu}(x, y, y'))|x, y, y'] - \min_{v \in [-\mathfrak{R}, \mathfrak{R}]} \mathbb{E}_{D_\omega}[\ell(\omega \cdot v)|x, y, y'] \le \epsilon'$$

To recover a version of Theorem 3.6 we need to modify Lemma 4.1 and its proof as that is the only place where realizability is utilized. We will also need to use the stronger version of Assumption 3.5.

Assumption B.2. For any $u, v \in [-\mathfrak{R}, \mathfrak{R}]$, we have

$$\ell(u) \geq \ell(v) + \ell'(v) \cdot (u - v) + \frac{c_\mu}{2}(u - v)^2.$$

Assumption B.2 posits that ℓ is c_μ-strongly convex.

Lemma B.3. *Under Assumptions B.1 and B.2, any policy $\pi \in \Pi$ with $L_\mu(\pi; D_{xy\omega}) - L_\mu(\pi^\star; D_{xy\omega}) \leq \epsilon$, satisfies*

$$\mathbb{E}_{x,y\sim D}\left[\Delta\bar{R}(x,y)^2\right]$$
$$= \mathbb{E}_{x,y,y'\sim D}[(\Delta R(x,y) - \Delta R(x,y'))^2] \leq \frac{4(2\epsilon' + \epsilon)}{c_\mu}.$$

Proof. Using the strong convexity part of Assumption B.2 we have that for π and $\pi^\star$

$$\mathbb{E}[\frac{c_\mu}{2}(\omega(R(x,y) - R(x,y') - v^\star))^2] \leq E[\ell(\omega\omega^{\pi,\mu}(x,y,y'))] - \mathbb{E}[\ell(\omega v^\star)]$$
$$= L_\mu(\pi; D_{xy\omega}) - L_\mu(\pi^\star; D_{xy\omega}) + \mathbb{E}[\ell(\omega \cdot \omega^{\pi^\star,\mu}(x,y,y'))] - \mathbb{E}[\ell(\omega v^\star)]$$
$$\leq \epsilon + \epsilon'$$
$$\mathbb{E}[\frac{c_\mu}{2}(\omega(R^\star(x,y) - R^\star(x,y') - v^\star))^2]$$
$$\leq E[\ell(\omega\omega^{\pi,\mu}(x,y,y'))] - \mathbb{E}[\ell(\omega v^\star)] \leq \epsilon',$$

where $v^\star = \arg\min_{v\in[-\mathfrak{R},\mathfrak{R}]} \mathbb{E}_{D_\omega}[\ell(\omega \cdot v)|x,y,y']$ and we have used first order optimality, together with the strong convexity of ℓ. Using the above two inequalities we have

$$\frac{c_\mu}{4} \mathbb{E}[(\omega(R(x,y) - R(x,y') - R^\star(x,y) + R^\star(x,y')))^2]$$
$$= \frac{c_\mu}{4} \mathbb{E}[(\omega(R(x,y) - R(x,y') - v^\star + v^\star R^\star(x,y) + R^\star(x,y')))^2]$$
$$\leq \mathbb{E}[\frac{c_\mu}{2}(\omega(R(x,y) - R(x,y') - v^\star))^2] + E[\frac{c_\mu}{2}(\omega(R^\star(x,y) - R^\star(x,y') - v^\star))^2] \leq 2\epsilon' + \epsilon.$$

$\square$

The remainder of the proof of Theorem 3.6 remains unchanged except for replacing ϵ by $2\epsilon' + \epsilon$. And so we have

Theorem B.4. *For any $\pi \in \Pi$ such that $L_\mu(\pi; D_{xy\omega}) - L_\mu(\pi_\mu^\star; D_{xy\omega}) \leq \epsilon$, where the corresponding loss to L_μ, given by ℓ is proper, and under Assumption 3.1, Assumption 3.4, Assumption B.1 and Assumption B.2, it holds that*

$$\mathbb{E}_x\left[\mathrm{KL}(\pi^\star(\cdot|x)||\pi(\cdot|x))\right] \leq \sqrt{\frac{(\epsilon + 2\epsilon'))}{c_\mu}} + \sqrt{\frac{C(\epsilon + 2\epsilon')}{c_\mu}} + \frac{e^{\mathfrak{R}}}{2} \cdot \frac{C(\epsilon + 2\epsilon')}{c_\mu}.$$

C. Experiment details

We evaluate the different variants on the TL;DR dataset (Völske et al., 2017), where the task is to summarize posts on reddit forums. The dataset consists of an original reddit posts, along with a pair of responses which are rated by human judges to provide the groumd-truth preference annotations (Stiennon et al., 2020). Our experiments use a T5 large model (Raffel et al., 2020) with 770M parameters, which is further fine-tuned to maximize the log-likelihood of the winning responses in the TL;DR dataset. We train for 20000 iterations, with a batch size of 32. A KL regularizer is used to the reference π_{ref} checkpoint with coefficient equal to 0.005. The optimizer used is Adafactor with learning rate that is constant with a linear warm-up for 2000 steps and a base rate of $1e - 4$.

We used the following text to prompt the Gemini evaluator used for our experiments:

```
 You are an expert summary rater who prefers very short and high quality
summaries.  Given a document and two candidate summaries, say 1 if SUMMARY1
is the better summary, or 2 if SUMMARY2 is the better summary.  Give a short
```

```
reasoning for your answer.
ARTICLE: <article-here >
SUMMARY1:  <summary-by-π>
SUMMARY2:  <summary-by-πref>.
```