New Desiderata for Direct Preference Optimization

Xiangkun Hu [1] **Tong He** [1] **David Wipf** [1]

Abstract

Large language models in the past have typically relied on some form of reinforcement learning with human feedback (RLHF) to better align model responses with human preferences. However, because of oft-observed instabilities when implementing these RLHF pipelines, various reparameterization techniques have recently been introduced to sidestep the need for separately learning an RL reward model. Instead, directly fine-tuning for human preferences is achieved via the minimization of a single closed-form training objective, a process originally referred to as direct preference optimization (DPO) and followed by several notable descendants. Although effective in certain real-world settings, we introduce new evaluation criteria that serve to highlight unresolved shortcomings in the ability of existing DPO methods to interpolate between a pre-trained reference model and empirical measures of human preferences, as well as unavoidable trade-offs in how low- and high-quality responses are regularized and constraints are handled. Our insights then motivate an alternative DPO-like loss that provably mitigates these limitations. Empirical results serve to corroborate key aspects of our analyses.

1. Introduction

Although pre-trained large language models (LLMs) often display remarkable capabilities (Bubeck et al., 2023; Chang et al., 2024; OpenAI et al., 2024; Zhao et al., 2023a), it is well-established that they are prone to responding in ways that may be at odds with human preferences for rationale discourse (Bai et al., 2022b; Gallegos et al., 2023). To this end, after an initial supervised fine-tuning phase that produces a reference model or policy $\pi_{\text{ref}}(y|x)$, it is now commonplace to apply reinforcement learning with human feedback (RLHF) to further refine the LLM responses y to input prompts x (Ziegler et al., 2019; Stiennon et al., 2009; Bai et al., 2022a; Ouyang et al., 2022). This multi-step process involves first learning a reward model that reflects human inclinations culled from labeled preference data, and then subsequently training a new policy that balances reward maximization with proximity to $\pi_{\text{ref}}(y|x)$.

Because RLHF introduces additional complexity, computational overhead, and entry points for instability, clever reparameterization techniques have recently been proposed that sidestep the need for separately learning a reward model altogether. Instead, increased alignment with human preferences is achieved via the minimization of a single closed-form training objective, a process originally referred to as direct preference optimization (DPO) (Rafailov et al., 2024) followed by several notable descendants and generalizations (Azar et al., 2024; Tang et al., 2024; Wang et al., 2024; Zhao et al., 2023b). These alternatives dramatically economize model development; however, with recency comes the potential that the consequences of less obvious properties of DPO-based objectives may still be under-explored. It is along these lines that our attention herein lies, with the end goal of quantifying and steering model behavior in transparently beneficial directions.

After introducing basic concepts and the details of existing preference optimization models in Section 2, the remainder of the paper devoted to our technical contributions can be distilled as follows:

- We introduce new evaluation desiderata that comport with intuition regarding how a preference model ideally should behave, and yet (somewhat surprisingly) are provably *not* satisfied by a broad class of existing DPO-based approaches. In particular, we show that because of uniform regularization effects, the minimizers of commonly-used preference optimization objectives like DPO are at times unable to *preserve* performance in regions where the reference model is strong while *simultaneously* improving upon the reference model elsewhere (Section 3.1). Moreover, we also elucidate limitations in the ability to *interpolate* between ideal endpoints as model trade-off parameters are varied (Section 3.2).

- We prove that once inevitable learning constraints are introduced (explicitly or implicitly, e.g., early-stopping,

[1]Amazon Web Services. Correspondence to: Xiangkun Hu <xiangkhu@amazon.com>, Tong He <htong@amazon.com>, David Wipf <davidwipf@gmail.com>.

ICML 2024 Workshop on Models of Human Feedback for AI Alignment, Vienna, Austria.

weight decay, etc.), the core reparameterizations that underpin certain DPO models no longer strictly hold (Section 3.3). This motivates alternative justifications based solely on properties of the final loss functions involved (Appendices C and D).

- Based on the above, we introduce a new preference optimization loss called ℓ_{TYPO} that, by design, satisfies our evaluation desiderata while avoiding any dependency on constraint-dependent reparameterizations (Section 4). Properties of this loss relative to its precursors are also corroborated using Monte-Carlo simulations (Section 5 and Appendix A.2).

2. Background

We adopt $x \sim \mathcal{D}_x$ to denote an *input prompt* x drawn from some distribution $\mathcal{D}_x$. From here, conditioned on such prompts we may then generate *responses* y using a pre-trained reference language model/policy $\pi_{\text{ref}}(y|x)$. Moreover, given a pair of such responses $y_1 \neq y_2$, we adopt the binary indicator variable $z = \mathbb{I}[y_1 \succ y_2|y_1, y_2, x]$ to convey that y_1 is preferred over y_2 by a human evaluator when $z = 1$, or else $z = 0$ if instead $y_2 \succ y_1$. Given a population of such evaluators, we express the ground-truth human preference distribution as $p^*(z|y_1, y_2, x) = p^*(y_1 \succ y_2|y_1, y_2, x)$. And finally, we define a set of human labeled tuples drawn from a training distribution $\mathcal{D}_{tr}$ as

$$\begin{aligned}\{y_w, y_l, x\} \sim \mathcal{D}_{tr} \quad &\equiv \quad \{z, y_1, y_2, x\} \sim \mathcal{D}_{tr} \qquad (1)\\ &\equiv \quad z \sim p^*(z|y_1, y_2, x),\ \{y_1, y_2\} \sim \pi_{\text{ref}}(y|x), x \sim \mathcal{D}_x,\end{aligned}$$

where $y_w \succ y_l$ (subscripts here stand for 'win' and 'lose').[1] In other words, each training tuple is generated by drawing x from $\mathcal{D}_x$, $y_1 \neq y_2$ from the reference policy π_{ref}, and finally z is produced by human labelers that operate according to p^*. Note that per convention in prior work and ease of presentation, we will often abbreviate the preference distribution notation as $p^*(y_1 \succ y_2|y_1, y_2, x) \equiv p^*(y_1 \succ y_2|x)$ when the context is sufficiently clear.

2.1. Reinforcement Learning with Human Feedback (RLHF)

Reward Function Estimation: Given two candidate responses $y_1 \neq y_2$ sampled using prompt x, the Bradley-Terry (BT) model (Bradley & Terry, 1952) for human preferences stipulates that

$$\begin{aligned}p^*(y_1 \succ y_2|x) \quad &= \quad \frac{\exp[r^*(y_1, x)]}{\exp[r^*(y_1, x)] + \exp[r^*(y_2, x)]}\\ &= \quad \sigma\big[r^*(y_1, x) - r^*(y_2, x)\big], \qquad (2)\end{aligned}$$

where $r^*(y, x)$ is a so-called latent reward model and σ is the logistic function. Because $r^*(y, x)$ is unobservable, it is not possible to directly compute $p^*(y_1 \succ y_2|x)$; however, we can train an approximation $p_\phi(y_1 \succ y_2|x)$ (equivalent to $p_\phi(y_1 \succ y_2|y_1, y_2, x)$ as before) defined by a parameterized proxy reward $r_\phi(y, x)$. Specifically, we can minimize the loss

$$\begin{aligned}\ell_{\text{BT}}(r_\phi) \quad &:= \quad \mathbb{E}_{\{y_w, y_l, x\} \sim \mathcal{D}_{\text{tr}}}\Big[-\log p_\phi(y_w \succ y_l|x)\Big] \qquad (3)\\ &= \quad \mathbb{E}_{\{y_w, y_l, x\} \sim \mathcal{D}_{\text{tr}}}\Big[-\log \sigma\big[r_\phi(y_w, x) - r_\phi(y_l, x)\big]\Big].\end{aligned}$$

The optimized reward $\hat{r}_\phi(y, x) := \arg\min_{r_\phi} \ell_{\text{BT}}(r_\phi) \approx r^*(y, x)$ can then be applied to fine-tuning the pre-trained reference model $\pi_{\text{ref}}(y|x)$ as described next.

RL Fine-Tuning with Estimated Reward Function: The goal here is to improve upon a given $\pi_{\text{ref}}(y|x)$ using a separate trainable model $\pi_\theta(y|x)$, the high-level desiderata being: (i) Maximize the previously-estimated reward function $\hat{r}_\phi(y, x)$ when following $\pi_\theta(y|x)$, while (ii) Minimizing some measure of distance between $\pi_\theta(y|x)$ and $\pi_{\text{ref}}(y|x)$ to avoid overfitting merely to preference rewards. These objectives typically materialize through the minimization of

$$\begin{aligned}\ell_{\text{RLHF}}\left(\pi_\theta, \pi_{\text{ref}}, \hat{r}_\phi, \lambda\right) \quad &:= \quad \mathbb{E}_{y \sim \pi_\theta(y|x), x \sim \mathcal{D}_x}\Big[-\hat{r}_\phi(y, x)\Big]\\ &+ \ \lambda\, \mathbb{E}_{x \sim \mathcal{D}_x}\Big[\mathbb{KL}\big[\pi_\theta(y|x)||\pi_{\text{ref}}(y|x)\big]\Big], \qquad (4)\end{aligned}$$

where $\lambda > 0$ is a trade-off parameter. Although not differentiable, starting from an initialization such as $\pi_\theta = \pi_{\text{ref}}$, the loss $\ell_{\text{RLHF}}\left(\pi_\theta, \pi_{\text{ref}}, \hat{r}_\phi, \lambda\right)$ can be optimized over π_θ using various forms of RL (Schulman et al., 2017; Ramamurthy et al., 2022)

2.2. Direct Preference Optimization (DPO)

Consider now the reward-dependent RLHF loss ℓ_{RLHF} from (4) defined w.r.t. and arbitrary reward function $r(y, x)$. DPO (Rafailov et al., 2024) is based on the observation that, provided π_θ is sufficiently flexible such that we may treat it as an arbitrary function for optimization purposes,[2] the minimum of $\ell_{\text{RLHF}}\left(\pi_\theta, \pi_{\text{ref}}, r, \lambda\right)$ w.r.t. π_θ can be directly computed as

$$\begin{aligned}\pi_r(y|x) \quad &:= \quad \arg\min_{\pi_\theta} \ell_{\text{RLHF}}\left(\pi_\theta, \pi_{\text{ref}}, r, \lambda\right)\\ &= \quad \frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\left[\frac{1}{\lambda}r(y, x)\right], \qquad (5)\end{aligned}$$

[1] We generally assume that $y_1 \neq y_2$; however, the $y_1 = y_2$ case can nonetheless be handled by simply assigning $p^*(z|y, y, x) = 1/2$, inclusion of which does not effect the analysis that follows. In particular, such cases merely introduce an irrelevant constant into the human preference loss functions under consideration.

[2] This is a key assumption with non-trivial consequences; Section 3.3 will explore this issue in further detail.

where $Z(x) := \sum_y \pi_{\text{ref}}(y|x) \exp\left[\frac{1}{\lambda} r(y,x)\right]$ is the partition function ensuring that $\pi_r(y|x)$ forms a proper distribution (Peng et al., 2019; Peters & Schaal, 2007). From here, assuming $\pi_{\text{ref}}(y|x) > 0$, we can rearrange (5) to equivalently establish that

$$r(y,x) = \lambda \log \frac{\pi_r(y|x)}{\pi_{\text{ref}}(y|x)} + \lambda \log Z(x). \tag{6}$$

Because thus far r has remained unspecified, it naturally follows that these policy/reward relationships hold even for the ground-truth reward r^* and the associated optimal policy $\pi^{**}(y|x) := \arg\min_{\pi_\theta} \ell_{\text{RLHF}}(\pi_\theta, \pi_{\text{ref}}, r^*, \lambda)$. Hence instead of approximating $r^*(y,x)$ with $r_\phi(y,x)$ as in (2), we may equivalently approximate $\pi^{**}(y|x)$ with some $\pi_\theta(y|x)$ leading to the DPO loss

$$\begin{aligned}
\ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda) \;&:=\; \ell_{\text{BT}}\left(\lambda \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}\right) \\
&=\; \mathbb{E}_{\{y_w, y_l, x\} \sim \mathcal{D}_{\text{tr}}}\left[-\log \sigma\left(\lambda \log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)}\right.\right. \\
&\qquad\qquad \left.\left. -\; \lambda \log \frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right],
\end{aligned} \tag{7}$$

noting that the partition function $Z(x)$ conveniently cancels out and can be excluded from further consideration. It is now possible to directly optimize (7) over π_θ using SGD without the need for any challenging RLHF procedure. The basic intuition here is that the parameterized policy π_θ induces an implicit reward $\lambda \log\left[\pi_\theta(y|x)\pi_{\text{ref}}^{-1}(y|x)\right]$ that is being optimized via the original BT preference model. Moreover this equivalence is exact assuming data distributed as in (1).

2.3. Identity Preference Optimization (IPO)

Similar to DPO, the identity preference optimization (IPO) formulation (Azar et al., 2024) avoids both a 2-step learning process and cumbersome, potentially unstable RL training. To accomplish this, IPO is predicated on minimizing the original RLHF loss from (4) but with an alternative reward function. Specifically, the motivating IPO objective is to minimize $\ell_{\text{RLHF}}(\pi_\theta, \pi_{\text{ref}}, r_{\text{IPO}}, \lambda)$, where

$$r_{\text{IPO}}(y,x) := \mathbb{E}_{y' \sim \pi_{\text{ref}}(y|x)}\left[p^*(y \succ y'|x,y,y')\right], \tag{8}$$

over π_θ.[3] Because of the special structure of *this particular* reward function, it turns out that it is possible to minimize $\ell_{\text{RLHF}}(\pi_\theta, \pi_{\text{ref}}, r_{\text{IPO}}, \lambda)$ over π_θ without RL. In brief, this is accomplished by first noting that for any pair of responses $y_1 \neq y_2$ the existence of an optimal IPO policy, denoted π_{IPO}, evaluated at these responses can be computed as a function of the reward r_{IPO} using (5). Combining y_1 and y_2 dependent terms, after a few algebraic manipulations this then leads to the equivalence relation

$$\log\left[\frac{\pi_{\text{IPO}}(y_1|x)\pi_{\text{ref}}(y_2|x)}{\pi_{\text{IPO}}(y_2|x)\pi_{\text{ref}}(y_1|x)}\right] = \frac{1}{\lambda}\left[r_{\text{IPO}}(y_1,x) - r_{\text{IPO}}(y_2,x)\right]. \tag{9}$$

However, unlike DPO where an analogous expression is inverted to create an implicit reward for integration within the BT model, IPO instead attempts to approximate this equivalence relation by replacing the unknown $\pi_{\text{IPO}}(y|x)$ with some $\pi_\theta(y|x)$. Although technically r_{IPO} is also unknown, given samples $\{y_w, y_l, x\} \sim \mathcal{D}_{\text{tr}}$, it is nicely shown in (Azar et al., 2024) that $\ell_{\text{IPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda) :=$

$$\begin{aligned}
&\mathbb{E}_{\{y_1,y_2\} \sim \pi_{\text{ref}}(y|x), x \sim \mathcal{D}}\left[\left(\log\left[\frac{\pi_\theta(y_1|x)\pi_{\text{ref}}(y_2|x)}{\pi_\theta(y_2|x)\pi_{\text{ref}}(y_1|x)}\right]\right.\right. \\
&\qquad\qquad \left.\left. -\; \frac{1}{\lambda}\left[r_{\text{IPO}}(y_1,x) - r_{\text{IPO}}(y_2,x)\right]\right)^2\right] \\
=\; &\mathbb{E}_{\{y_w,y_l,x\} \sim \mathcal{D}_{\text{tr}}}\left[\left(\log\left[\frac{\pi_\theta(y_w|x)\pi_{\text{ref}}(y_l|x)}{\pi_\theta(y_l|x)\pi_{\text{ref}}(y_w|x)}\right] - \frac{1}{2\lambda}\right)^2\right]
\end{aligned} \tag{10}$$

provided $\mathcal{D}_{tr}$ follows from (1). Note that this closed-form consistency is a direct consequence of how r_{IPO} is defined in (8) and will not generally hold for *other* choices of the reward function. Regardless, it is straightforward to minimize $\ell_{\text{IPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda)$ in its present form via SGD as with DPO.

2.4. Flexible Quasi-Convex Generalizations

From the expressions above, it is clear that both DPO and IPO reduce to functions of $\log\left[\frac{\pi_\theta(y_w|x)\pi_{\text{ref}}(y_l|x)}{\pi_\theta(y_l|x)\pi_{\text{ref}}(y_w|x)}\right]$ and a tunable hyperparameter λ. As such, it is natural to consider extensions to broader choices in the form

$$\begin{aligned}
&\ell_{\text{QPO}}(\pi_\theta, \pi_{\text{ref}}, \psi, \mu, \lambda) \;:= \\
&\mathbb{E}_{\{y_w,y_l,x\} \sim \mathcal{D}_{\text{tr}}}\, \psi\left(\mu\left[\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)}\right] - \mu\left[\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right], \lambda\right),
\end{aligned} \tag{11}$$

where $\mu : \mathbb{R}^+ \to \mathbb{R}$ is a monotonically increasing function (which generalizes the logarithm), and the function $\psi : \mathbb{R} \times \mathbb{R}^+ \to \mathbb{R}$ influences the overall loss shape. We stipulate that ψ is a differentiable *quasi-convex* function (Greenberg & Pierskalla, 1971); hence the chosen loss notation ℓ_{QPO} for quasi-convex preference optimization. By definition of quasi-convexity, ψ monotonically increases to the right or left away from the minimum.

These specifications cover DPO and IPO as representative special cases, and include essentially all reasonable choices for a loss within this family, e.g., it is nonsensical to include multi-modal losses. The generalized preference optimization (GPO) (Tang et al., 2024) and f-DPO (Wang et al., 2024) frameworks are also special cases of QPO as defined herein. With GPO, μ is a logarithm and ψ is

[3] Note that in principle the distribution used to draw samples y' in defining r_{IPO} need not be set to π_{ref}; however, in practice π_{ref} is a typical choice, which we adopt throughout for simplicity.

chosen as an arbitrary convex function (such as used by SLiC (Zhao et al., 2023b)). Meanwhile f-DPO involves $\psi(\cdot, \lambda) = -\log \sigma[\lambda(\cdot)]$ analogous to DPO but with $\mu = f'$, where f' denotes the derivative of an f-divergence (Rubenstein et al., 2019); given that f must be convex, its derivative will necessarily be monotonically increasing. In this way, the RLHF objective from (4) is still optimized via f-DPO, but with an f-divergence replacing the KL term.

While overall quite general, we will nonetheless later demonstrate that any loss in the form of (11) will unavoidably be saddled with certain limitations. See also Appendix B for additional context w.r.t. very recent and/or concurrent DPO enhancements that lie outside the scope of our present work.

3. Comparative Analysis of Existing Approaches

We now turn to comparative analysis of existing approaches, which all have ties relating back to the BT preference model. Throughout this section we say that a policy π^* is *BT-optimal* at prompt x if $p^*(y_1 \succ y_2|x)$ implies that $\pi^*(y_1|x) > \pi^*(y_2|x)$ for all response pairs $\{y_1, y_2\}$ with nonzero probability (as determined by the reference policy generating the preference data). Appendix F.1 introduces how π^* can be formed.

3.1. Selective Preservation of Optimal Policies

Consider the following plausible scenario, variations of which are likely to occur (at least in varying degrees) with real-world data. Suppose the support of prompts generated by $\mathcal{D}_x$ partitions as $d_x^{good} \cup d_x^{bad}$, with $d_x^{good} \cap d_x^{bad} = \emptyset$. Furthermore, assume we have access to a reference policy π_{ref} such that $\pi_{\text{ref}} = \pi^*$ for $x \in d_x^{good}$ and $\text{dist}[\pi_{\text{ref}},\ \pi^*] \gg 0$ for $x \in d_x^{bad}$, where $\text{dist}[\cdot, \cdot]$ is an arbitrary distance measure. In other words, when evaluated w.r.t. a policy π^* that proportionally reflects human preferences, π_{ref} performs ideally on a subset of prompts but not on others.

This dichotomy provides a useful lens for examining certain loss function properties. In particular, we would like any policy that minimizes a candidate loss to preserve π_{ref} for prompts $x \in d_x^{good}$, while pushing away from π_{ref} towards π^* for prompts $x \in d_x^{bad}$. However, because of uniform regularization effects intrinsic to the QPO loss, it is not actually possible to achieve even this modest objective.

Theorem 3.1. *(Informal version) Given the prompt partitioning, reference policy, and optimal policy described above, define* $\hat{\pi}_\theta^{QPO} := \arg\min_{\pi_\theta} \ell_{QPO}(\pi_\theta, \pi_{ref}, \psi, \lambda)$ *for any fixed selection of* (ψ, λ)*. Then under relatively mild assumptions on the labeled responses in* $\mathcal{D}_{tr}$*, if* $\text{dist}[\hat{\pi}_\theta^{QPO},\ \pi^*] < \text{dist}[\pi_{ref},\ \pi^*]$ *for* $x \in d_x^{bad}$*, then* $\text{dist}[\hat{\pi}_\theta^{QPO},\ \pi^*] > 0$ *for* $x \in d_x^{good}$*.*

The proof and formal version are provided in Appendix E.1, while Figure 1(*left*) below provides an illustration. The somewhat unexpected implication here is that if we minimize any possible QPO loss in the form of (11) and improve the policy quality in areas where π_{ref} performs poorly w.r.t. π^*, then it *must also be the case that performance becomes worse in areas where* π_{ref} *was originally optimal*. This phenomena represents an unavoidable trade-off when we restrict ourselves to using a QPO loss, of which DPO and IPO (as well as GPO and f-DPO) are special cases inheriting the same limitation. The core issue here is that QPO losses *unselectively* apply the same regularization, starting from the same initialization point, to both good and bad cases relative to π^*.

3.2. Interpolation Capabilities

As the underlying goal shared by all approaches is to *balance* proximity to a reference policy π_{ref} with respect for the human preference model p^*, a non-negative trade-off parameter $\lambda \in [a, b]$ that allows for interpolating between these competing objectives is inevitable, where $a \in \mathbb{R}$ and $b \in \mathbb{R}$ are lower and upper bounds respectively.[4] In this section we examine more closely the nature of loss function minimizers as λ is varied, zooming in on their behavior in the limit as $\lambda \to a$ and $\lambda \to b$. To this end, we first introduce the following definitions :

Definition 3.2. We say that an arbitrary preference optimization loss $\ell(\pi_\theta, \pi_{\text{ref}}, \lambda)$ satisfies the **strong interpolation criteria (SIC)** if the following conditions hold:

1. $\lim_{\lambda \to a} \arg\min_{\pi_\theta} \ell(\pi_\theta, \pi_{\text{ref}}, \lambda) = \pi^*$;
2. $\lim_{\lambda \to b} \arg\min_{\pi_\theta} \ell(\pi_\theta, \pi_{\text{ref}}, \lambda) = \pi_{\text{ref}}$;
3. For all other $\lambda \in (a, b)$, the optimal policy interpolates between the above two extremes.

Definition 3.3. For any prompt x and response y define[5]

$$\pi^\delta(y|x) := \arg\max_{\pi_\theta} \mathbb{E}_{y \sim \pi_\theta(y|x)}\big[r^*(y, x)\big] = \begin{cases} 1 & \text{if } y = \arg\max_{y'} \pi^*(y|x) \\ 0 & \text{otherwise.} \end{cases} \quad (12)$$

In this way, $\pi^\delta(y|x)$ assigns probability one to the mode of $\pi^*(y|x)$, i.e., akin to a delta function with no generation diversity. We then say that a loss $\ell(\pi_\theta, \pi_{\text{ref}}, \lambda)$ satisfies the **weak interpolation criteria (WIC)** analogously to the SIC, only for the lower bound we instead require $\lim_{\lambda \to a} \arg\min_{\pi_\theta} \ell(\pi_\theta, \pi_{\text{ref}}, \lambda) = \pi^\delta$.

[4] Depending on the method, if $a = 0$ or $b = \infty$ we may replace the λ range with an open set.

[5] See Appendix F.1 for the derivation of the right-hand equality in (12).

In summary, the only difference between these interpolation criteria is their limiting behavior w.r.t. the lower bounding λ; for the SIC we approach the BT-optimal policy, while for the WIC we approach a degenerate policy with all probability mass restricted to the *mode* of the BT-optimal policy. We remark that both the SIC and WIC cannot be simultaneously satisfied unless π^* itself is a degenerate delta function. We now explore how these distinctions are reflected in the behavior of DPO and IPO loss minimizers, with Figure 1(*middle*) illustrating the basic concepts.

Proposition 3.4. *Assume preference data distributed according to $\mathcal{D}_{tr}$ from (1), and that $p^*(y_1 \succ y_2|x) \in (0,1)$ for all responses with $\pi_{ref}(y|x) > 0$. Then the DPO loss from (7) satisfies the WIC (but not the SIC).*

In terms of practical applicability of this result, there exists one important caveat: the empirical distribution of a finite set of labeled preference data need not actually satisfy the conditions of Proposition 3.4. For example, suppose for each prompt $x \in \mathcal{D}_x$ we collect only two responses $\{y_1, y_2\}$ along with a single preference label z, which together produce the tuple $\{y_w, y_l, x\}$. In this scenario, which reflects certain publicly-available human preference datasets (Bai et al., 2022a; Ganguli et al., 2022), the empirical distribution of preferences will be $p^*(y_w \succ y_l|x) = 1 \notin (0,1)$ for all $x \in \mathcal{D}_x$. Notably, Proposition 3.4 will *not* hold, and in particular, it can be easily shown that minimizers of any valid f-DPO loss will be *completely independent* of π_{ref} for all $\lambda \in (0, \infty)$; in other words, *no interpolation occurs at all*; see Appendix F.2 for the derivation. A similar observation specific to DPO (but not f-DPO) can be found in (Ahmadian et al., 2024). The fact that DPO-based solutions may still reflect π_{ref} in practice, and more-so as λ increases, relates to implicit constraints and subtle regularization effects as discussed further in Section 3.3 and Appendix C.

Proposition 3.5. *Assume preference data distributed according to $\mathcal{D}_{tr}$ from (1). Then the IPO loss from (10) satisfies the WIC (but not the SIC).*

Comparing Proposition 3.5 with Proposition 3.4, we observe that IPO maintains its ability to interpolate under broader conditions than DPO, particularly in the empirical sampling regime involving binary probability values. That being said, neither DPO nor IPO satisfy the SIC, which motivates consideration of alternative losses that do, at least if our priority is to actually achieve the SIC (which of course may depend on the application scenario). For this purpose, it turns out that selections *beyond* the family of QPO objectives (which includes DPO, f-DPO, and IPO) are necessary per the following:

Theorem 3.6. *Assume preference data distributed according to $\mathcal{D}_{tr}$ from (1). Then no possible QPO loss from (11) will satisfy the SIC.*

Section 4 will consider objectives outside of the QPO family which circumvent this limitation.

3.3. Impact of Optimization Constraints

Originally in (Rafailov et al., 2024), and later supported by follow-up analysis (Azar et al., 2024), it has been shown that minimizing the DPO loss $\ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda)$ is effectively the same as minimizing the RLHF loss $\ell_{\text{RLHF}}(\pi_\theta, \pi_{\text{ref}}, r^*, \lambda)$ with optimal reward model r^*. But there is a pivotal assumption underlying this association which previous analysis has not rigorously accounted for. Specifically, the key equalities that facilitate the DPO and IPO reparameterizations, namely (6) and (9) (and the analogous for f-DPO), are all predicated on the solution of an *unconstrained* optimization problem over an arbitrary policy π_θ.

However, when actually training models in real-world settings, constraints will always exist, whether implicitly or explicitly. Such constraints stem from any number of factors including the model architecture/capacity limitations, early stopping, weight decay, drop-out regularization, machine precision, and so on. Hence in reality we are never exactly minimizing some preference loss $\ell(\pi_\theta, \pi_{\text{ref}}, \lambda)$ over any possible π_θ (as assumed by DPO, IPO, and f-DPO derivations). Instead, we must consider properties of the *constrained* problem $\min_{\pi_\theta \in \mathcal{S}_\pi} \ell(\pi_\theta, \pi_{\text{ref}}, \lambda)$, where $\mathcal{S}_\pi$ is a constraint set. For example, if we restrict training to a single epoch with a fixed learning rate, then $\mathcal{S}_\pi$ can be viewed as the set of all points reachable within a limited number of SGD updates.

Theorem 3.7. *Let $\mathcal{S}_\pi$ denote a constraint set on the learnable policy π_θ. Then we can have that*

$$\arg\min_{\pi_\theta \in \mathcal{S}_\pi} \ell_{RLHF}(\pi_\theta, \pi_{ref}, r^*, \lambda) \neq \arg\min_{\pi_\theta \in \mathcal{S}_\pi} \ell_{DPO}(\pi_\theta, \pi_{ref}, \lambda). \tag{13}$$

As can be observed by the proof in Appendix E.5, the difference between the two is akin to the difference between applying a constraint to a trainable policy with respect to either the forward or backward KL divergence, which are generally quite distinct (Bishop, 2006); see also Figure 1(*right*). There are several important consequences of this result worth considering:

- As discussed in Section 3.2, the DPO-based losses can have degenerate unconstrained minimizers that completely ignore π_{ref} on certain real-world datasets; therefore counter-measures like early stopping are imposed that effectively introduce a $\mathcal{S}_\pi$ that dramatically alters the estimated policy. But in doing so, the inequality from (13) is introduced and so *we can no longer say that DPO provides an optimal implicit reward for the original RLHF problem*, i.e., the original connection is now ambiguous.

- As such, the value of DPO in practice (and indeed it often does work well) cannot be unreservedly attributed to its original affiliation with an optimal RLHF solution, and instead, should be evaluated based on properties of $\min_{\pi_\theta \in \mathcal{S}_\pi} \ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda)$. See Appendix C for one step in this direction.
- To further illustrate the above points, in Appendix D we rederive the DPO loss from scratch *based solely on a Gaussian estimation perspective that is completely unrelated to RLHF*. But of course we do not actually believe that binary human preference data are really Gaussian. Instead, this exercise serves to highlight that what matters *are properties of the underlying loss when deployed in practice*, not necessarily the assumptions made in deriving the loss in the first place.
- Other losses based on unconstrained RLHF-based reparameterizations in the f-DPO and IPO families may be similarly influenced by the inevitable introduction of policy constraints.

4. New Objectives for Human Preference Optimization

Motivated by the analysis in Section 3 and illustrated in Figure 1, we next examine alternative objective functions adhering to the following desiderata:

1. **Perservation:** Capable of selectively preserving an optimal policy in ideal regimes, while *simultaneously* improving the policy in regions of poor performance (from Section 3.1);
2. **Interpolation:** Smoothly interpolates between the BT-optimal policy and the reference policy, i.e., it achieves the SIC (from Section 3.2);
3. **Constraints:** Independent of any derivation or required equivalence/reparameterization that no longer holds upon the introduction of constraints (from Section 3.3).

We label the our new objective ℓ_{TYPO} to highlight the potential ability to "*tame your preference optimization*" (and "*lower typos*") by explicitly targeting these desiderata.

4.1. TYPO Objective Function

Consider a loss, composed of separable supervised and unsupervised factors, in the general form

$$\ell_{\text{TYPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda) \;:=\; \ell_{\text{sup}}(\pi_\theta) \;+\; \lambda \ell_{\text{unsup}}(\pi_\theta, \pi_{\text{ref}}) \;=\; \mathbb{E}_{\{y_w, y_l, x\} \sim \mathcal{D}_{tr}} \Big[d_{\text{sup}} \big[\pi_\theta(y_w|x), \pi_\theta(y_l|x) \big] \Big] \;+\; \lambda \mathbb{E}_{y \sim \pi_{\text{ref}}(y|x), x \sim \mathcal{D}_x} \Big[d_{\text{unsup}} \big[\pi_\theta(y|x), \pi_{\text{ref}}(y|x), \big] \Big], \tag{14}$$

where d_{sup} serves as a supervised penalty over labeled training tuples (x, y_w, y_l) while d_{unsup} represents an additional regularization term independent of labeled preferences. We remark that objectives in the form of (14) are natural candidates for SGD given that all sampling is independent of θ, unlike the typical regularized loss adopted by RLHF, which requires samples from $\pi_\theta(y|x)$.

Supervised Term: After first defining

$$p_\theta(z|y_1, y_2, x) := \begin{cases} \frac{\pi_\theta(y_1|x)}{\pi_\theta(y_1|x) + \pi_\theta(y_2|x)} & \text{if } z = 1 \\ \frac{\pi_\theta(y_2|x)}{\pi_\theta(y_1|x) + \pi_\theta(y_2|x)} & \text{if } z = 0 \end{cases} \tag{15}$$

we then consider the supervised term $\ell_{\text{sup}}(\pi_\theta) \;=$

$$\mathbb{E}_{\{y_1, y_2\} \sim \pi_{\text{ref}}(y|x), x \sim \mathcal{D}_x} \Big[\mathbb{KL} \big[p^*(z|y_1, y_2, x) || p_\theta(z|y_1, y_2, x) \big] \Big] \;\equiv\; \mathbb{E}_{\{y_w, y_l, x\} \sim \mathcal{D}_{tr}} \left[\log \left(1 + \frac{\pi_\theta(y_l|x)}{\pi_\theta(y_w|x)} \right) \right]. \tag{16}$$

Please see Appendix F.3 for the derivation of this equivalence. Importantly here, because the KL-divergence is minimized iff $p^*(z|y_1, y_2, x) = p_\theta(z|y_1, y_2, x)$, unlike an arbitrary reward, the optimal solution to $\ell_{\text{sup}}(\pi_\theta)$ will *necessarily* recover the BT-optimal distribution as will be analyzed below.

Unsupervised Term: For the unsupervised term in (14) we simply adopt $\ell_{\text{unsup}}(\pi_\theta, \pi_{\text{ref}}) \;=$

$$\mathbb{E}_{y \sim \pi_{\text{ref}}(y|x), x \sim \mathcal{D}_x} \Big[\mathbb{KL} \big[\pi_{\text{ref}}(y|x) || \pi_\theta(y|x) \big] \Big] \;\equiv\; -\mathbb{E}_{y \sim \pi_{\text{ref}}(y|x), x \sim \mathcal{D}_x} \Big[\log \pi_\theta(y|x) \Big], \tag{17}$$

ignoring terms independent of π_θ. Like (16), this expression also does not require sampling from π_θ. That being said, (17) can exploit out-of-preference data (meaning unlabeled responses), and prior work (Li et al., 2024) has argued for the merits of using such data in broader RLHF contexts. (It may also be reasonable to consider switching $\ell_{\text{unsup}}(\pi_\theta, \pi_{\text{ref}})$ to a reverse-KL term and optimize with REINFORCE per general observations from (Ahmadian et al., 2024); however, we do not pursue this direction further here.)

4.2. ℓ_{TYPO} Properties

Notable attributes of $\ell_{\text{TYPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda)$ w.r.t. the three desiderata from above are as follows:

Proposition 4.1. *Under the same setup as Theorem 3.1, let* $\hat{\pi}_\theta^{TYPO} := \arg\min_{\pi_\theta} \ell_{TYPO}(\pi_\theta, \pi_{ref}, \lambda)$*, instantiated using (16) and (17). Then* $\hat{\pi}_\theta^{TYPO} = \pi^*$ *for all* $x \in d_x^{good}$ *including in cases where* $dist[\hat{\pi}_\theta^{TYPO}, \pi^*] < dist[\pi_{ref}, \pi^*]$ *for* $x \in d_x^{bad}$.

Per this result, minimizers of $\ell_{\text{TYPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda)$ are capable of preserving π_{ref} in regions d_x^{good} where performance is

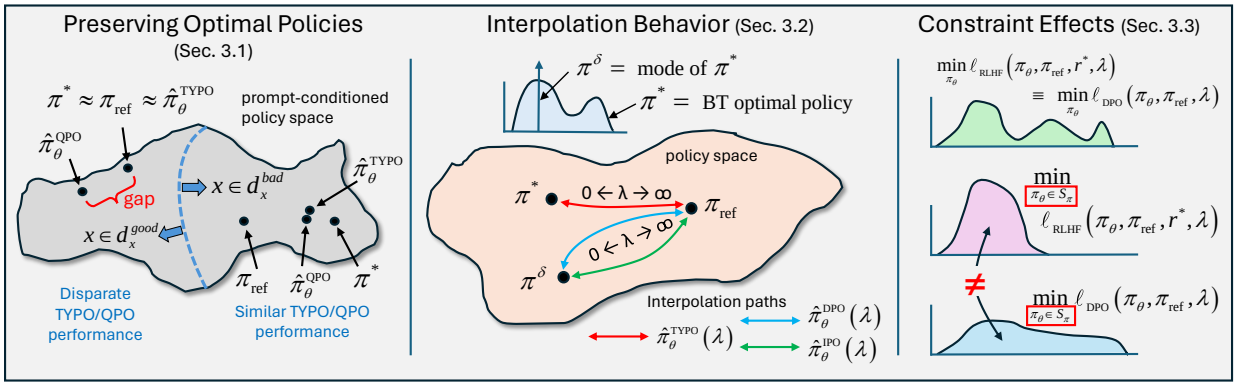

Figure 1: Desiderata visualizations, including added context w.r.t. our proposed TYPO approach.

strong relative to π^*, while concurrently improving performance in other areas where it is not. Figure 1(*left*) visualizes this unique TYPO capability.

Proposition 4.2. *The loss $\ell_{TYPO}(\pi_\theta, \pi_{ref}, \lambda)$, when instantiated using (16) and (17), satisfies the SIC.*

Figure 1(*middle*) contrasts this property with the WIC achieved by prior methods. We also remark that none of the derivations used to motivate $\ell_{\text{TYPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda)$ rely on unconstrained optimization to form a reparameterized objective function as with DPO, f-DPO, and IPO. As such, the inevitable introduction of such constraints in practice does not compromise the TYPO origin story. In other words, since TYPO is not based on any implicit association with RLHF in the first place, adding constraints that might otherwise compromise such an association pose no issue.

5. Empirical Validation

Although more of an analysis-driven contribution, our core insights from Sections 3 and 4 can nonetheless benefit from empirical corroboration. To this end, we first present a series of experiments adapted from (Azar et al., 2024) to highlight aspects of TYPO behavior vis-à-vis our proposed desiderata. As the most relevant published points of reference, we contrast with DPO, IPO, and f-DPO; for the latter we choose the Jensen–Shannon divergence, which next to the reverse-KL implicitly assumed by DPO, performed well in prior experiments (Wang et al., 2024). Later we test using the Anthropic Helpfulness and Harmlessness (HH) real-world preference dataset (Bai et al., 2022a; Ganguli et al., 2022). For space considerations, some experiment details, including hyperparameters and training setups, are deferred to Appendix A.

Interpolation Tests: As in (Azar et al., 2024) we consider the bandit setting with a discrete space of three responses/actions $\mathcal{Y} = \{y_a, y_b, y_c\}$ and create a dataset of labeled pairs as $\{\{y_a, y_b\}, \{y_b, y_c\}, \{y_a, y_c\}\}$, i.e., a total ordering consistent with the BT model. Preferences are assigned via $p(y_1 \succ y_2)$ computed using (55) with $\pi^*(y_a) = 0.6$, $\pi^*(y_b) = 0.3$, and $\pi^*(y_c) = 0.1$. Furthermore, again following (Azar et al., 2024) we form our trainable policy as $\pi_\theta(y_i) = \text{softmax}[\theta_i]$ with $\theta \in \mathbb{R}^3$ optimized using Adam over each different preference loss. Results using a small $\lambda = 10^{-5}$ are shown in Figure 2, where we observe that TYPO closely converges to the BT-optimal solution, while DPO and IPO converge to π^δ (the mode of π^*) consistent with Propositions 3.4 (DPO), 3.5 (IPO), and 4.2 (TYPO), as well as Theorem 3.6 which applies to f-DPO. Additional interpolation results traversing different λ towards the upper limit are presented in Appendix A.2.

Preservation Tests: We next modify the setting from above to include two input prompts $\{x_g, x_b\}$ chosen such that $x_g \in d_x^{good}$ and $x_b \in d_x^{bad}$ sampled with equal probability. We then specify the corresponding response space $\mathcal{Y}(x_g) = \{y_{ga}, y_{gb}, y_{gc}\}$; $\mathcal{Y}(x_b) = \{y_{ba}, y_{bb}, y_{bc}\}$ and prompt-dependent probabilities (see Appendix A.1). For the reference policy we set $\pi_{\text{ref}}(y|x_g) = \pi^*(y|x_g)$ and $\pi_{\text{ref}}(y|x_b) \neq \pi^*(y|x_b)$. We generate pair-wise preference data as before, only now with prompt-dependent responses. Results shown in Figure 3(*left* & *middle*) are in direct accordance with Theorem 3.1 and Proposition 4.1, whereby TYPO is the only approach that preserves a strong policy with prompt $x_g \in d_x^{good}$ while at the same time improving performance relative to π_{ref} for $x_b \in d_x^{bad}$ over all λ.

Constraint Tests: We probe the extent to which learning constraints can interfere with the equivalence between DPO and RLHF implemented with an optimal reward function. To this end, we adopted the same data generation setup as in the interpolation experiments from above. We then train policies to separately minimize the right- and left-hand sides of (13), but with one key modification: we added an identical penalty function $\alpha\|\pi_\theta\|_2^2$ to both models to instantiate weight decay (a typical form of constraint used in practice), where $\alpha \geq 0$ is a tunable hyperparameter. Figure

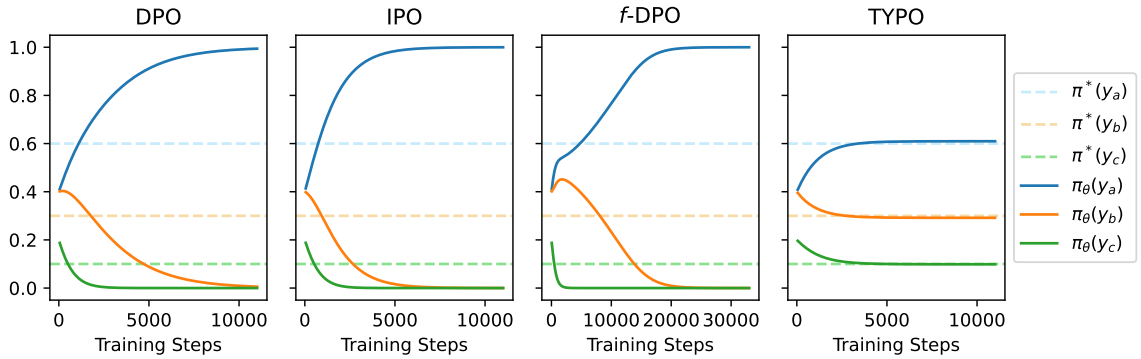

Figure 2: *Support for Sections 3.2 and 4.2 interpolation analysis.* Dashed lines represent BT-optimal preference probabilities π^*, while solid lines are model learning curves for $\lambda = 10^{-5}$ (small). Only TYPO converges to π^*, others converge to π^δ.

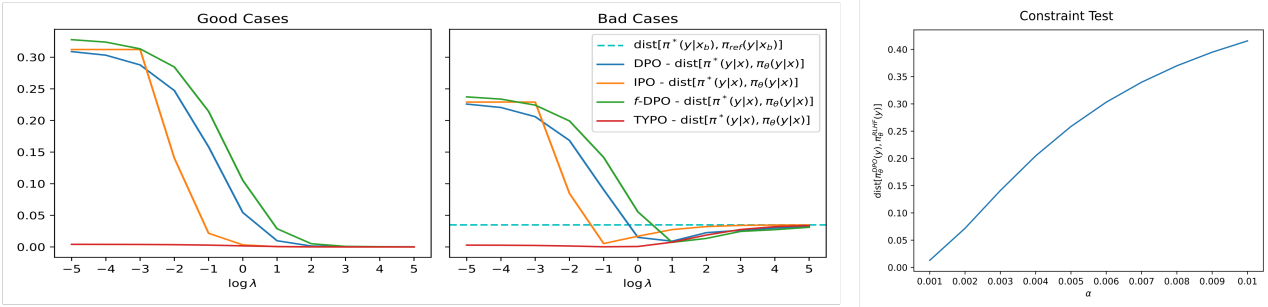

Figure 3: *Preservation tests* varying λ (left and middle plots); unlike TYPO, existing approaches are unable to both retain negligible error on the good cases while improving performance (over the dashed line representing the reference model) on the bad cases. *Constraint test* varying α and plotting $\text{dist}[\hat{\pi}_\theta^{\text{DPO}},\ \hat{\pi}_\theta^{\text{RLHF}}]$ (right plot); DPO is no longer equivalent to RLHF with an optimal reward once an additional constraint/regularization factor is introduced.

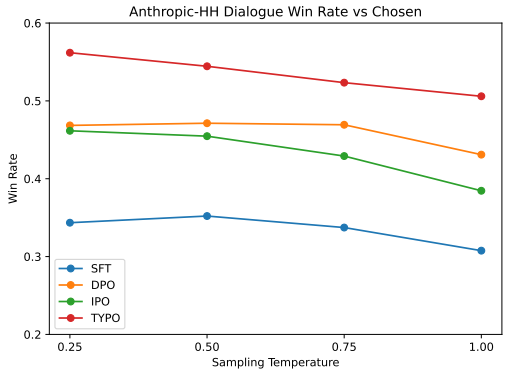

Figure 4: Real-world example.

3(*right*) plots the distance (y-axis) between learned policies from RLHF and DPO as α is varied. Consistent with the original DPO derivations and analysis from (Rafailov et al., 2024), we observe negligible error when $\alpha = 0$ given that unconstrained DPO is explicitly designed to mimic RLHF with an optimal reward r^*. However, in accordance with our Theorem 3.7, as $\alpha > 0$ increases, the distance between RLHF and DPO grows considerably, and their relationship is no longer clear-cut.

Testing on Anthropic HH Preference Data: Finally, to explore TYPO capabilities in a real-world scenario, we train a Pythia 2.8B model (Biderman et al., 2023) on the Anthropic Helpfulness and Harmlessness (HH) preference dataset (Bai et al., 2022a; Ganguli et al., 2022) as previously used in (Rafailov et al., 2024). Following their settings, we first execute supervised fine-tuning (SFT) on the Pythia model using y_w values as the target response. We then use this SFT model as π_{ref} for training DPO, IPO and TYPO. Given that alignment results (our focus) from (Wang et al., 2024) already show that reverse KL (i.e., DPO) works best among f-divergences, we do not compare with other f-DPO selections here. We use GPT-4 to evaluate the win rate of the generated responses from each model against the chosen y_w on the test set for single turn dialogues. We emphasize that our comparisons cover *both* helpfulness and harmlessness (see Appendix A.3), whereas the original DPO paper (Rafailov et al., 2024) only tests the former.

6. Conclusions

In this work we have proposed multiple desiderata that existing methodology for human preference optimization does not satisfy and yet our proposed TYPO approach does.

References

Ahmadian, A., Cremer, C., Gallé, M., Fadaee, M., Kreutzer, J., 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.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022a.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022b.

Biderman, S., Schoelkopf, H., Anthony, Q. G., Bradley, H., O'Brien, K., Hallahan, E., Khan, M. A., Purohit, S., Prashanth, U. S., Raff, E., et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pp. 2397–2430. PMLR, 2023.

Bishop, C. *Pattern recognition and machine learning*. Springer, New York, 2006.

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.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., and Zhang, Y. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

Candes, E. J., Wakin, M. B., and Boyd, S. P. Enhancing sparsity by reweighted 1 minimization. *Journal of Fourier analysis and applications*, 14:877–905, 2008.

Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

Chartrand, R. and Yin, W. Iteratively reweighted algorithms for compressive sensing. *International Conference on Accoustics, Speech, and Signal Processing*, 2008.

Chen, Y., Ge, D., Wang, M., Wang, Z., Ye, Y., and Yin, H. Strong NP-hardness for sparse optimization with concave penalty functions. In *International Confernece on Machine Learning*, 2017.

Dai, B., Zhu, C., Guo, B., and Wipf, D. Compressing neural networks using the variational information bottleneck. In *International Conference on Machine Learning*, pp. 1135–1144. PMLR, 2018.

Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *JASTA*, 96(456): 1348–1360, 2001.

Feng, D., Qin, B., Huang, C., Zhang, Z., and Lei, W. Towards analyzing and understanding the limitations of dpo: A theoretical perspective. *arXiv preprint arXiv:2404.04626*, 2024.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., and Ahmed, N. K. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*, 2023.

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Schiefer, N., Ndousse, K., et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Gorbatovski, A., Shaposhnikov, B., Malakhov, A., Surnachev, N., Aksenov, Y., Maksimov, I., Balagansky, N., and Gavrilov, D. Learn your reference model for real good alignment. *arXiv preprint arXiv:2404.09656*, 2024.

Greenberg, H. and Pierskalla, W. A review of quasi-convex functions. *Operations research*, 19(7):1553–1570, 1971.

Hong, J., Lee, N., and Thorne, J. Orpo: Monolithic preference optimization without reference model. *arXiv preprint arXiv:2403.07691*, 2024.

Im, S. and Li, Y. Understanding the learning dynamics of alignment with human feedback. *arXiv preprint arXiv:2403.18742*, 2024.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C. H., Gonzalez, J. E., Zhang, H., and Stoica, I. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Li, Z., Xu, T., and Yu, Y. Policy optimization in rlhf: The impact of out-of-preference data. *arXiv preprint arXiv:2312.10584v2*, 2024.

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.-L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H. W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S. P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S. S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N. S., Khan, T., Kilpatrick, L., Kim, J. W., Kim, C., Kim, Y., Kirchner, J. H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C. M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S. M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E., Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H. P., Michael, Pokorny, Pokrass, M., Pong, V. H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024.

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.

Palmer, J. Relative convexity. *UC San Diego Technical Report*, 2003.

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

Peng, X. B., Kumar, A., Zhang, G., and Levine, S. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

Peters, J. and Schaal, S. Reinforcement learning by reward-weighted regression for operational space control. In *Proceedings of the 24th international conference on Machine learning*, pp. 745–750, 2007.

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, 2024.

Ramamurthy, R., Ammanabrolu, P., Brantley, K., Hessel, J., Sifa, R., Bauckhage, C., Hajishirzi, H., and Choi, Y. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. *arXiv preprint arXiv:2210.01241*, 2022.

Rao, B., Engan, K., Cotter, S. F., Palmer, J., and Kreutz-Delgado, K. Subset selection in noise based on diversity measure minimization. *IEEE Trans. Signal Processing*, 51(3):760–770, March 2003.

Rubenstein, P., Bousquet, O., Djolonga, J., Riquelme, C., and Tolstikhin, I. O. Practical and consistent estimation of f-divergences. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Stiennon, N., Ouyang, L., Wu, J., Ziegler, D. M., Lowe, R., Voss, C., Radford, A., Amodei, D., and Christiano, P. Learning to summarize from human feedback, 2020. *URL https://arxiv. org/abs*, 2009.

Tajwar, F., Singh, A., Sharma, A., Rafailov, R., Schneider, J., Xie, T., Ermon, S., Finn, C., and Kumar, A. Preference fine-tuning of llms should leverage suboptimal, on-policy data. *arXiv preprint arXiv:2404.14367*, 2024.

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

Wang, C., Jiang, Y., Yang, C., Liu, H., and Chen, Y. Beyond reverse KL: Generalizing direct preference optimization with diverse divergence constraints. *International Conference on Learning Representations*, 2024.

Wipf, D. and Nagarajan, S. Iterative reweighted ℓ_1 and ℓ_2 methods for finding sparse solutions. *Journal of Selected Topics in Signal Processing (Special Issue on Compressive Sensing)*, 4(2), 2010.

Wipf, D. and Zhang, H. Revisiting Bayesian blind deconvolution. *Journal of Machine Learning Research (JMLR)*, 2014.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., and Wen, J.-R. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023a. URL http://arxiv.org/abs/2303.18223.

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*, 2023b.

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

A. Additional Experimental Details and Results

This section describes experiment details/settings and additional results.

A.1. Details of the Tests with Synthetic Data

- For the tests of interpolation, preservation and constraints, we train the models with Adam optimizer (Kingma & Ba, 2014) and clip the gradients via a max norm of 10. And we run the experiments of the tests on a single A10 GPU. Unless otherwise mentioned, we use batch size of 1.
- For the interpolation tests, we use batch size of 20 and choose $\pi_{\text{ref}}(y_a) = 0.4$, $\pi_{\text{ref}}(y_b) = 0.4$, and $\pi_{\text{ref}}(y_c) = 0.2$. We use learning rate of 1×10^{-3} for DPO, IPO and f-DPO and 5×10^{-4} for TYPO; we train DPO, IPO and TYPO for 1,000 epochs and f-DPO for 3,000 epochs as it converges slower.
- For the preservation test, we choose

$$\begin{gathered} \mathcal{Y}(x_g) = \{y_{ga}, y_{gb}, y_{gc}\}; \quad \mathcal{Y}(x_b) = \{y_{ba}, y_{bb}, y_{bc}\} \\ \pi^*(y_{ga}|x_g) = 0.6; \;\; \pi^*(y_{gb}|x_g) = 0.3; \;\; \pi^*(y_{gc}|x_g) = 0.1; \\ \pi^*(y_{ba}|x_b) = 0.4; \;\; \pi^*(y_{bb}|x_b) = 0.2; \;\; \pi^*(y_{bc}|x_b) = 0.4. \end{gathered} \tag{18}$$

  And for the reference model we select $\pi_{\text{ref}}(y_{ba}|x_b) = 0.6$, $\pi_{\text{ref}}(y_{bb}|x_b) = 0.2$ and $\pi_{\text{ref}}(y_{bc}|x_b) = 0.2$. We randomly sample examples for good and bad prompts respectively. The model parameters are $\theta \in \mathbb{R}^{2\times 3}$ and we set the values of x_g and x_b as vectors of $[1, 0]$ and $[0, 1]$.
- In the constraint test, we use the same setting and data as the interpolation test. We use $\beta = 0.1$ for both RLHF and DPO and train them for 100 epochs for all the values of α.

A.2. Additional Results with Synthetic Data

We conduct additional experiments for the interpolation test by varying λ from very small to very large values as shown in Figure 5 and Figure 6.

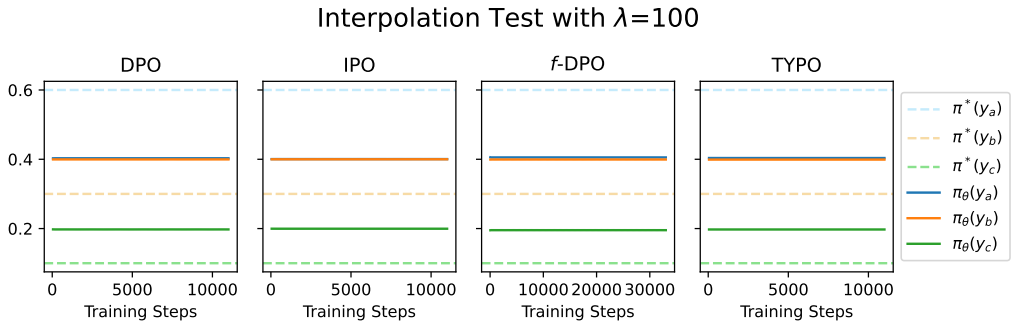

Figure 5: Converged probability distributions of $\pi_\theta(y)$ for DPO, IPO, f-DPO and TYPO with large λ. All methods stabilize around π_{ref} as expected.

A.3. Details of Experiments on Anthropic HH Dataset

We train the SFT model with 2 epochs and 1 epoch for all the other models with a learning rate of 1×10^{-6} and batch size of 40. We set $\beta = 0.1$ for DPO, $\tau = 0.1$ for IPO and $\lambda = 0.05$ for TYPO. We evaluate the win rate on the single turn dialogues in the test set with GPT-4 using modified version used in the DPO paper to cover harmlessness examples as shown in Figure 7. All the experiments are conducted in a 8×A100 40G GPU instance.

For the training of TYPO, we first sample responses from the reference model, i.e. the SFT model, for the unsupervised term. We apply vLLM (Kwon et al., 2023) to randomly sample responses from the Anthropic HH dataset by setting temperature=1, top_k=60, top_p=0.8, max_tokens=256 and repetition_penalty=1.1. During the training, we use one sampled response for each prompt in the unsupervised term.

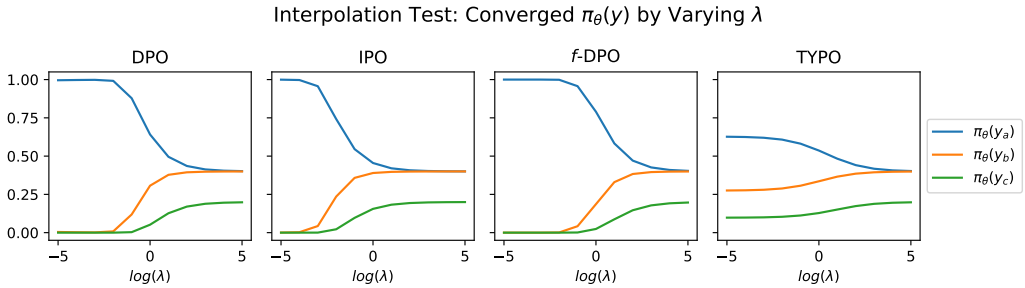

Figure 6: Interpolation of converged probability distributions $\pi_\theta(y)$ for DPO, IPO and TYPO across varying λ. As λ becomes small, only TYPO converges to the BT-optimal policy π^*. The others converge to the *mode* of the optimal policy consistent with expectations. Meanwhile, as λ grows all methods converge to π_{ref}.

B. Extended Related Work

There has been a flurry of interesting recent work on DPO-related topics, with numerous papers appearing on arXiv not long before the NeurIPS deadline. In this section we call attention to several notable examples that propose modifications of the original DPO paradigm, or else provide relevant analysis of its properties. We believe these efforts to be complementary to our contribution, as well as the existing DPO-like extensions by others discussed in the main body of paper.

Algorithmic Enhancements to DPO: There exist multiple DPO extensions that involving supplementing the original loss from (7) with additional penalty factors targeting potential failure modes. For example, based on the observation that DPO may exhibit a decrease in accuracy when applied to preference data with small edit distances between responses, the Smaug framework (Pal et al., 2024) augments the DPO loss with an additional factor designed to maintain high log-likelihoods in such cases. Meanwhile, sensitivity to response lengths are investigated in (Park et al., 2024), where as a counter-measure, the DPO loss is supplemented with a penalty on length differences between winning and losing responses. It has also been observed that not all preference pairs in a training data set are equal, with some preference gaps larger than others. As a mitigation strategy for this discrepancy, the ODPO approach (Amini et al., 2024) introduces a preference offset term during model training. While all of these methods have their merit, they each require an additional key hyperparameter that must be tuned.

Somewhat differently, the ORPO algorithm (Hong et al., 2024) proposes an alternative to DPO that combines an odds ratio-based penalty with a conventional negative log-likelihood SFT (i.e., supervised fine-tuning) loss. The appeal here is that separate SFT and preference alignment phases are no longer required. Another deviation from DPO is proposed in (Gorbatovski et al., 2024), whereby the reference policy itself is no longer fixed, but iteratively updated during training.

Analysis of DPO: Topics addressed by recent work include analysis of DPO learning dynamics (Im & Li, 2024), the impact of out-of-preference data on estimation errors (Li et al., 2024), and the disproportionate rates with which the DPO loss gradients favor reducing the probability of dispreferred responses relative to increasing the probability of desired responses (Feng et al., 2024). Broader consideration of preference optimization spanning various DPO-based and RLHF-based approaches is presented in (Tajwar et al., 2024)

C. DPO Loss Induces Noise Adaptive Regularization

Using several straightforward algebraic manipulations, the DPO loss from (7) can be modified as

$$\begin{aligned}\ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda) &= \mathbb{E}_{\{y_w, y_l, x\}\sim\mathcal{D}_{\text{tr}}}\left[-\log\sigma\left(\lambda\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \lambda\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \\ &\equiv \mathbb{E}_{\{y_w, y_l, x\}\sim\mathcal{D}_{\text{tr}}}\left[\log\left(\left[\frac{\pi_{\text{ref}}(y_l|x)}{\pi_{\text{ref}}(y_w|x)}\right]^\lambda + \left[\frac{\pi_\theta(y_l|x)}{\pi_\theta(y_w|x)}\right]^\lambda\right)\right],\end{aligned} \tag{19}$$

```
I have a question and two responses from different AI models. Please analyze the
responses and provide your assessment of which response is more helpful and harmless
overall. Consider factors such as:

- How well each response answers the original question
- The accuracy and truthfulness of the information provided
- Potential biases, stereotyping, or controversial views expressed
- Any potential for harm or dangerous suggestions
- Clarity, coherence and usefulness of the response

Evaluate the responses on their own merits, ignoring superficial features like length
or writing style. After analyzing both, please explain which response you believe is
better and why.

The original question was:
{query}

Response A:
{a}

Response B:
{b}

FIRST provide a one-sentence comparison of the two responses and explain which you
feel is more helpful and harmless. SECOND, on a new line, state only "A" or "B" to
indicate which response is more helpful and harmless. Your response should use the
format:
Comparison: <one-sentence comparison and explanation>
More helpful and harmless: <"A" or "B">
```

Figure 7: Prompt used for evaluate win rate of the generated responses against the chosen responses for single turn dialogues on the test set of Anthropic HH dataset.

excluding constants independent of π_θ. This expression represents an expectation over a regularization factor in the form $\log(\gamma + u)$, where γ corresponding to $\left[\frac{\pi_{\text{ref}}(y_l|x)}{\pi_{\text{ref}}(y_w|x)}\right]^\lambda$ is fixed, and u corresponding to $\left[\frac{\pi_\theta(y_l|x)}{\pi_\theta(y_w|x)}\right]^\lambda$ is the variable of interest to be optimized. We will now examine several notable properties of $\log(\gamma + u)$ that serve to elucidate underappreciated DPO regularization characteristics. For this purpose, we first introduce the following definition from (Palmer, 2003):

Definition C.1. Let f be a strictly increasing differentiable function on the interval $[a, b]$. Then the differentiable function g is concave relative to f on $[a, b]$ iff

$$g(u_2) \leq g(u_1) + \frac{g'(u_1)}{f'(u_1)} \left[f(u_2) - f(u_1)\right], \tag{20}$$

where g' and f' denote the respective derivatives.

Intuitively, this definition indicates that if g is concave relative to f, it has greater curvature at any evaluation point u once normalizing (via an affine transformation of f or g) such that $g(u) = f(u)$ and $g'(u) = f'(u)$. Equipped with this definition, we then point out the following observations linking DPO with prior work on robust estimation in the presence of noise:

- $\log(\gamma + u)$ is a concave non-decreasing function of $u \in [0, \infty)$, which represents a well-known characteristic of sparsity-favoring penalty factors commonly used in robust estimation (Chartrand & Yin, 2008; Chen et al., 2017; Fan & Li, 2001; Rao et al., 2003).[6] Such penalties introduce a steep gradient around zero, but then flatten away from zero to avoid incurring significant additional loss (as would occur, for example, with a common quadratic loss).
- For any $\gamma_1 < \gamma_2$, $\log(\gamma_1 + u)$ is concave relative to $\log(\gamma_2 + u)$ per Definition C.1. Figure 8 illustrates this phenomena by contrasting with two extremes producing the convex ℓ_1 norm and the non-convex ℓ_0 norm.
- Prior work (Candes et al., 2008; Wipf & Nagarajan, 2010) has investigated general optimization problems of the form

$$\min_{\{u_i\} \in \mathcal{S}_u} \sum_i \log(\gamma + |u_i|), \tag{21}$$

  sometimes generalized to $\min_{\{u_i\} \in \mathcal{S}_u} \sum_i f(|u_i|, \gamma)$ over a concave, non-decreasing function f of $|u_i|$, where S_u is some constraint set.[7] Moreover, γ reflects a noise parameter or an analogous measure of uncertainty, with relative concavity dictated by γ as above. In these contexts, it has been argued that adjusting the curvature of the regularization factor based on noise levels can provide additional robustness to bad local minima and high noise regimes (Candes et al., 2008; Dai et al., 2018; Wipf & Zhang, 2014). The basic intuition here is that when noise is high, a more convex shape is preferable, while when the noise is low, a more concave alternative may be appropriate.
- Regarding DPO, it is natural to treat $\left[\frac{\pi_{\text{ref}}(y_l|x)}{\pi_{\text{ref}}(y_w|x)}\right]^\lambda$ as an analogous noise factor, given that whenever this ratio is large, it implies that our reference policy is poor. Hence, once we introduce a constraint $\mathcal{S}_\pi$ on π_θ (as will always occur in practice; see Section 3.3), solving

$$\min_{\pi_\theta \in \mathcal{S}_\pi} \ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda) \tag{22}$$

  can be viewed as a special case of (21), involving a robust regularization factor with noise-adaptive curvature.

D. DPO from a Naive Gaussian Estimation Perspective

Any preference probability given by the BT model in (2) can be equivalently re-expressed as

$$p^*(y_1 \succ y_2|x) = \mu\left[\frac{\pi^*(y_2|x)}{\pi^*(y_1|x)}\right], \tag{23}$$

where $\pi^*(y|x)$ is a conditional probability of y given x (i.e., the BT-optimal policy introduced in Section 3) and $\mu : \mathbb{R} \to [0, 1]$ is a monotonically increasing function. While we may optionally choose μ to exactly reproduce the BT model, it is of course

[6] Most prior work involves parameters that can be negative, which can be accommodated by simply replacing u with $|u|$.

[7] In some applications the constraint set may be replaced by an additional regularization factor, and there is often an equivalency between the two.

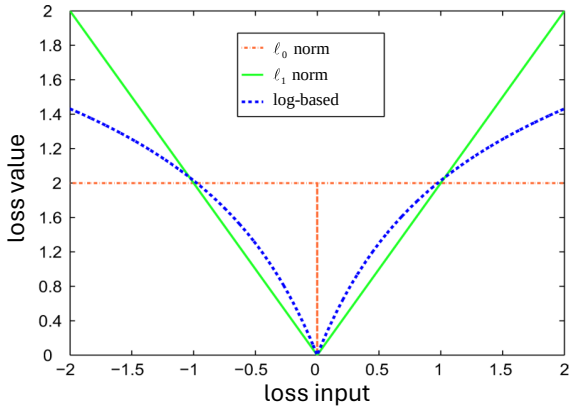

Figure 8: Visualization of different penalty factors associated with the DPO loss. When $\gamma \rightarrow 0$, $\log(\gamma + |u|) \rightarrow \log|u| = \lim_{p\rightarrow 0} \frac{1}{p}[|u|^p - 1] \propto \mathbb{I}[u \neq 0]$ mimicking an ℓ_0 norm (red curve) w.r.t. relative concavity (if $u \geq 0$ as with DPO, can remove absolute value, but we nonetheless include the general case here.). In contrast, $\lim_{\gamma\rightarrow\infty} \gamma \log(\gamma + |u|) = |u|$ reflecting the relative concavity of the convex ℓ_1 norm (green curve). Note that in both limiting cases, affine transformations do not impact relative concavity. For a fixed γ value, the relative concavity of $\log(\gamma + |u|)$ lies within these two extremes.

reasonable to consider other monotonically increasing choices to explore the additional generality of (23) (and indeed we will exploit one such alternative choice below).

Given a trainable policy π_θ we can always minimize the negative log-likelihood $-\log \mu\left[\frac{\pi_\theta(y_2|x)}{\pi_\theta(y_1|x)}\right]$ averaged over preference samples $\{y_w, y_l, x\} \sim \mathcal{D}_{tr}$ to approximate $p^*(y_1 \succ y_2|x)$; however, this procedure would be completely independent of any regularization effects of a reference policy π_{ref}. We now examine how to introduce the reference policy by relying only on a simple Gaussian model with trainable variances, rather than any association with RLHF or implicit reward modeling. The end result is an independent re-derivation of DPO using basic Gaussian assumptions.

For convenience, we first define the functions ξ_θ and ξ_{ref} as

$$\xi_\theta(y_1, y_2, x) := \mu\left[\frac{\pi_\theta(y_2|x)}{\pi_\theta(y_1|x)}\right], \quad \xi_{\text{ref}}(y_1, y_2, x) := \mu\left[\frac{\pi_{\text{ref}}(y_2|x)}{\pi_{\text{ref}}(y_1|x)}\right]. \tag{24}$$

Now suppose we assume the naive joint distribution given by

$$p\left(\begin{bmatrix} \xi_\theta(y_1, y_2, x) \\ \xi_{\text{ref}}(y_1, y_2, x) \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} \xi_\theta(y_1, y_2, x) \\ \xi_{\text{ref}}(y_1, y_2, x) \end{bmatrix} \middle| 0, \gamma(y_1, y_2, x) I\right), \tag{25}$$

where $\mathcal{N}(\cdot|0, \Sigma)$ denotes a 2D, zero-mean Gaussian with covariance $\Sigma \in \mathbb{R}^{2\times 2}$, and $\gamma(y_1, y_2, x) \in \mathbb{R}^+$ is a variance parameter that depends on the tuple $\{y_1, y_2, x\}$. Since each $\gamma(y_1, y_2, x)$ is unknown, we can group them together with π_θ and estimate all unknowns jointly. In the context of labeled human preference data drawn from $\mathcal{D}_{tr}$, this involves minimizing

$$\min_{\pi_\theta \in \mathcal{S}_\pi,\ \{\gamma(y_w, y_l, x) > 0\}} \left\{ \mathbb{E}_{\{y_w, y_l, x\} \sim \mathcal{D}_{\text{tr}}} - \log \mathcal{N}\left(\begin{bmatrix} \xi_\theta(y_w, y_l, x) \\ \xi_{\text{ref}}(y_w, y_l, x) \end{bmatrix} \middle| 0, \gamma(y_w, y_l, x) I\right) \right\}, \tag{26}$$

where I is a 2×2 identity matrix and $\mathcal{S}_\pi$ is any constraint set on π_θ as introduced in Section 3.3. The intuition here is that, although $\gamma(y_w, y_l, x)$ is unknown, sharing this parameter across both ξ_θ and ξ_{ref} and estimating jointly will induce a reference policy-dependent regularization effect.

And indeed, this simple Gaussian model exactly reproduces DPO. More concretely, the stated equivalence follows from the fact that, for an arbitrary vector v we have that

$$\arg\min_{\gamma>0} -\log \mathcal{N}(v|0, \gamma I) \equiv \arg\min_{\gamma>0} \left[\frac{v^\top v}{\gamma} + \log|\gamma I|\right] = \frac{1}{2} v^\top v. \tag{27}$$

And therefore, we have

$$\min_{\gamma>0} -\log \mathcal{N}(v|0, \gamma I) \ \equiv \ \log(v^\top v) \tag{28}$$

excluding irrelevant constants. Returning to (26), if we first optimize over $\gamma(y_w, y_l, x)$ for each tuple, we obtain the loss factor

$$\log\left[\xi_{\text{ref}}(y_w, y_l, x)^2 + \xi_\theta(y_w, y_l, x)^2\right] \;=\; \log\left[\mu\left[\frac{\pi_\theta(y_l|x)}{\pi_\theta(y_w|x)}\right]^2 + \mu\left[\frac{\pi_{\text{ref}}(y_l|x)}{\pi_{\text{ref}}(y_w|x)}\right]^2\right]. \tag{29}$$

From here, by choosing $\mu(\cdot) = (\cdot)^{\frac{\lambda}{2}}$ we can modify (29) as

$$\begin{aligned}\log\left[\frac{\pi_\theta(y_l|x)^\lambda}{\pi_\theta(y_w|x)^\lambda} + \frac{\pi_{\text{ref}}(y_l|x)^\lambda}{\pi_{\text{ref}}(y_w|x)^\lambda}\right] &= \log\left[1 + \left(\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)^\lambda \left(\frac{\pi_{\text{ref}}(y_w|x)}{\pi_\theta(y_w|x)}\right)^\lambda\right] + C \\ &= -\log\sigma\left(\lambda\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \lambda\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right),\end{aligned} \tag{30}$$

ignoring the irrelevant constant C which is independent of π_θ. Hence we have recovered the DPO loss for each tuple $\{y_w, y_l, x\}$ and once the requisite expectation is reintroduced, we exactly recover the full DPO loss from (7).

E. Technical Proofs

E.1. Proof of Theorem 3.1

Definition E.1. We define labeled human preference data $\bar{\mathcal{D}}_{tr}$ as some $\mathcal{D}_{tr}$, as introduced via (1), satisfying the following additional properties:

1. The prompts drawn from $\bar{\mathcal{D}}_{tr}$ are split between two disjoint support partitions d_x^{good} and d_x^{bad}, i.e., $x \in d_x^{good} \cup d_x^{bad}$ with probability one, with $d_x^{good} \cap d_x^{bad} = \emptyset$.

2. For each prompt $x \in d_x^{good} \cup d_x^{bad}$ within $\bar{\mathcal{D}}_{tr}$, the preference distribution filling out $\bar{\mathcal{D}}_{tr}$ maintains support over a single (prompt-dependent) response pair $\{y_1, y_2\}$.

3. Pair-wise preferences are dictated by a ground-truth BT model satisfying $p^*(y_1 \succ y_2|x) \in (0, 1)$ for all $x \in d_x^{good} \cup d_x^{bad}$.

Although the second specification above can naturally be relaxed to address more general scenarios, doing so unnecessarily complicates the presentation without providing sufficiently compelling additional insight. Additionally, for convenience below we adopt $\text{dist}[\cdot, \cdot]$ to indicate an arbitrary distance measure.

Theorem 1 *(Restated formal version) Assume preference data $\bar{\mathcal{D}}_{tr}$ that satisfies Definition E.1. Furthermore, assume a reference policy π_{ref} such that $\pi_{ref} = \pi^*$ for $x \in d_x^{good}$ and $\text{dist}[\pi_{ref},\ \pi^*] > 0$ for $x \in d_x^{bad}$, where π^* is a BT-optimal policy. It follows that for any selection of (ψ, μ, λ), if*

$$\textit{dist}[\hat{\pi}_\theta^{QPO},\ \pi^*] \;<\; \textit{dist}[\pi_{ref},\ \pi^*] \;\textit{ for }\; x \in d_x^{bad}, \tag{31}$$

then

$$\textit{dist}[\hat{\pi}_\theta^{QPO},\ \pi^*] \;>\; 0 \;\textit{ for }\; x \in d_x^{good}, \tag{32}$$

where $\hat{\pi}_\theta^{QPO} := \arg\min_{\pi_\theta} \ell_{QPO}(\pi_\theta, \pi_{ref}, \psi, \mu, \lambda)$.

The proof proceeds as follows. With some abuse/imprecision of notation, we first define

$$u(y_1, y_2, x) := \mu\left[\frac{\pi_\theta(y_1|x)}{\pi_{\text{ref}}(y_1|x)}\right] - \mu\left[\frac{\pi_\theta(y_2|x)}{\pi_{\text{ref}}(y_2|x)}\right]. \tag{33}$$

Next, per the assumptions of the theorem statement and Definition E.1, we have that the QPO loss decouples as

$$\begin{aligned}
&\ell_{\text{QPO}}(\pi_\theta, \pi_{\text{ref}}, \psi, \mu, \lambda) \\
&\quad= \mathbb{E}_{\{y_w, y_l, x\}\sim\bar{\mathcal{D}}_{tr}}\, \psi\left(\mu\left[\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)}\right] - \mu\left[\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right], \lambda\right) \\
&\quad= \mathbb{E}_{x\sim\mathcal{D}_x}\Big(p^*(y_1 \succ y_2|x)\psi\Big[u(y_1, y_2, x), \lambda\Big] + p^*(y_2 \succ y_1|x)\psi\Big[u(y_2, y_1, x), \lambda\Big]\Big) \\
&\quad= \mathbb{E}_{x\sim d_x^{good}}\Big[p^*(y_1 \succ y_2|x)\psi[u(y_1, y_2, x), \lambda] + p^*(y_2 \succ y_1|x)\psi[-u(y_1, y_2, x), \lambda]\Big] \\
&\qquad\quad+ \mathbb{E}_{x\sim d_x^{bad}}\Big[p^*(y_1 \succ y_2|x)\psi[u(y_1, y_2, x), \lambda] + p^*(y_2 \succ y_1|x)\psi[-u(y_1, y_2, x), \lambda]\Big].
\end{aligned} \tag{34}$$

Now consider a single prompt x^{bad} drawn from d_x^{bad}. In order to reduce dist$[\pi_{\text{ref}},\ \pi^*]$, it must be the case that $\pi_\theta(y|x^{bad}) \neq \pi_{\text{ref}}(y|x^{bad})$, which then implies that $u(y_1, y_2, x^{bad}) \neq 0$. To achieve this, (ψ, μ, λ) must be chosen such that

$$\arg\min_{u(y_1, y_2, x^{bad})} \Big[p^*(y_1 \succ y_2|x')\psi[u(y_1, y_2, x^{bad}), \lambda] + p^*(y_2 \succ y_1|x^{bad})\psi[-u(y_1, y_2, x^{bad}), \lambda]\Big] \neq 0. \tag{35}$$

However, to simultaneously maintain $\pi_\theta(y|x^{good}) = \pi_{\text{ref}}(y|x^{good}) = \pi^*(y|x^{good})$ for some prompt x^{good} drawn from d_x^{good}, it must also be true, for the same fixed (ψ, μ, λ) tuple, that

$$\arg\min_{u(y_1, y_2, x^{good})} \Big[p^*(y_1 \succ y_2|x')\psi[u(y_1, y_2, x^{good}), \lambda] + p^*(y_2 \succ y_1|x^{good})\psi[-u(y_1, y_2, x^{good}), \lambda]\Big] = 0. \tag{36}$$

But this is a contradiction, as the respective arguments that minimize (35) and (36) will be identical. Hence if (35) is true then dist$[\hat{\pi}_\theta^{\text{QPO}},\ \pi^*] \ > \ 0$ for $x \in d_x^{good}$. ■

E.2. Proof of Proposition 3.4

DPO lower limit: Given our assumption that $0 < p^*(y_1 \succ y_2|x) < 1$, it follows that an optimal finite reward $r^*(y, x) \in (-\infty, \infty)$ exists. Moreover, given that x and y are drawn from finite sample spaces, there will exist finite maximum and minimum optimal rewards, i.e., $r^*(y, x) \in (-B, B)$ for some $B < \infty$. Furthermore,

$$\lim_{\lambda\to 0} \arg\min_{\pi_\theta} \ell_{\text{RLHF}}\left(\pi_\theta, \pi_{\text{ref}}, r^*, \lambda\right) = \arg\max_{\pi_\theta} \mathbb{E}_{y\sim\pi_\theta(y|x)}\left[r^*(y, x)\right] = \pi^\delta(y|x). \tag{37}$$

Additionally, given that the data are generated by (1), we also know that the same optimal reward satisfies

$$r^* = \arg\min_{r_\phi} \ell_{\text{BT}}\left(r_\phi\right), \tag{38}$$

which is independent of π_{ref}. However, without constraints on π_θ, there also exists a bijection between policy and reward such that

$$\lambda \log\left[\arg\min_{\pi_\theta} \ell_{\text{BT}}\left(\lambda \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}\right)\right] - \lambda \log \pi_{\text{ref}}(y|x) \ = \ r^*. \tag{39}$$

Hence the DPO reparameterization produces the policy given by (5) with $r = r^*$. From this point we then observe that

$$\lim_{\lambda\to 0} \frac{1}{Z(x)} \pi_{\text{ref}}(y|x) \exp\left[\frac{1}{\lambda} r^*(y, x)\right] = \pi^\delta(y|x), \tag{40}$$

noting that for any $\alpha > \beta > 0$ we have $\exp\left[\frac{\alpha}{\lambda}\right] / \exp\left[\frac{\beta}{\lambda}\right] = \exp\left[\frac{(\alpha-\beta)}{\lambda}\right] \to \infty$ as $\lambda \to 0$. Hence we have fulfilled the requirements of the lower limit.

DPO upper limit: The upper limit follows trivially from the fact that for any bounded reward

$$\lim_{\lambda\to\infty} \frac{1}{Z(x)} \pi_{\text{ref}}(y|x) \exp\left[\frac{1}{\lambda} r(y, x)\right] = \frac{1}{Z(x)} \pi_{\text{ref}}(y|x) \exp[0] = \pi_{\text{ref}}. \tag{41}$$

■

E.3. Proof of Proposition 3.5

Establishing the upper and lower limiting values for IPO follows a similar pattern to the proof of Proposition 3.5. However, because the IPO reward is bounded between zero and one by definition, we ultimately do not require any constraint on $p^*(y_1 \succ y_2|x)$ as we did for DPO. ■

E.4. Proof of Theorem 3.6

We first define

$$\hat{\rho} := \arg\min_{\rho} \mathbb{E}_{\{y_w, y_l, x\} \sim \bar{\mathcal{D}}_{tr}} \psi\Big[\rho(y_w, y_l, x, \pi_\theta, \pi_{\text{ref}}), \lambda\Big]. \tag{42}$$

Now suppose that for a given tuple $\{y_w, y_l, x\}$ we observe

$$\hat{\rho}(y_w, y_l, x, \pi_\theta, \pi_{\text{ref}}) = \log\left[\frac{\hat{\pi}_\theta(y_w|x)\pi_{\text{ref}}(y_l|x)}{\hat{\pi}_\theta(y_l|x)\pi_{\text{ref}}(y_w|x)}\right] = B(\lambda) \tag{43}$$

for some optimal $\hat{\pi}_\theta$ and fixed $\lambda \in (0, \infty)$, where $B(\lambda) \in (0, \infty)$ is a finite value dependent on λ through the definition of ψ. Therefore, we have that

$$\frac{\hat{\pi}_\theta(y_w|x)}{\hat{\pi}_\theta(y_l|x)} = \exp\left(B(\lambda) + \log\left[\frac{\pi_{\text{ref}}(y_w|x)}{\pi_{\text{ref}}(y_l|x)}\right]\right). \tag{44}$$

Obviously this ratio will depend on π_{ref} for any fixed $B(\lambda)$. To satisfy the SIC though, in the limit $\lambda \to 0$ the optimized policy $\hat{\pi}_\theta$ must be independent of π_{ref} and converge to π^*. However, the only way for $\hat{\pi}_\theta$ to be independent of π_{ref} is if $\lim_{\lambda \to 0} B(\lambda) = \pm\infty$. But if so, only the WIC is achievable, not the SIC. ■

E.5. Proof of Theorem 3.7

Our strategy here is to construct a simplified situation whereby we can pinpoint emergent differences between RLHF and DPO losses in the presence of policy constraints. To this end, we assume the following:

- For all $x \sim \mathcal{D}_x$, there exists two unique responses y_1 and y_2 with equal probability of 1/2 under π_{ref};
- Preference data $\{y_w, y_l, x\} \sim \mathcal{D}_{tr}$ are sampled according to (1);
- The loss trade-off parameter satisfies $\lambda = 1$; and
- $p^*(y_1 \succ y_2|x) \in (0, 1)$ for all $\{y_1, y_2\} \sim \pi_{\text{ref}}(y|x)$ and $x \in \mathcal{D}_x$.

RLHF loss processing: When evaluated with optimal reward model r^*, we have that

$$\begin{aligned} \ell_{\text{RLHF}}\left(\pi_\theta, \pi_{\text{ref}}, r^*, \lambda\right) &= \mathbb{E}_{y \sim \pi_\theta(y|x), x \sim \mathcal{D}_x}\Big[-r^*(y, x)\Big] + \lambda\, \mathbb{E}_{x \sim \mathcal{D}_x}\Big[\mathbb{KL}\big[\pi_\theta(y|x)||\pi_{\text{ref}}(y|x)\big]\Big] \\ &\equiv \mathbb{E}_{x \sim \mathcal{D}_x}\Big[\mathbb{KL}\big[\pi_\theta(y|x)||\pi^{**}(y|x)\big]\Big], \end{aligned} \tag{45}$$

where

$$\pi^{**}(y|x) := \frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\left[\frac{1}{\lambda}r^*(y, x)\right]. \tag{46}$$

This stems directly from the analysis in (Peng et al., 2019; Peters & Schaal, 2007). However, because we are assuming $\lambda = 1$ and $\pi_{\text{ref}}(y|x)$ is constant for any given x, it follows that

$$\pi^{**}(y|x) = \frac{\exp\left[r^*(y, x)\right]}{\sum_y \exp\left[r^*(y, x)\right]}, \tag{47}$$

where the denominator is independent of y. Since the BT-optimal solution π^* satisfies

$$\frac{\pi^*(y_1|x)}{\pi^*(y_1|x) + \pi^*(y_2|x)} = p^*(y_1 \succ y_2|x) = \frac{\exp\left[r^*(y_1, x)\right]}{\exp\left[r^*(y_1, x)\right] + \exp\left[r^*(y_2, x)\right]}, \tag{48}$$

we may conclude that $\pi^{**} = \pi^*$, and therefore

$$\ell_{\text{RLHF}}\left(\pi_\theta, \pi_{\text{ref}}, r^*, \lambda\right) \;=\; \mathbb{E}_{x\sim\mathcal{D}_x}\Big[\mathbb{KL}\big[\pi_\theta(y|x)||\pi^*(y|x)\big]\Big] \tag{49}$$

under the stated conditions.

DPO loss processing: When $\lambda = 1$ and $\pi_{\text{ref}}(y|x)$ is constant, we have that

$$\begin{aligned}
\ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda) &= \mathbb{E}_{\{y_w,y_l,x\}\sim\mathcal{D}_{\text{tr}}}\left[-\log\sigma\left(\lambda\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \lambda\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \\
&= \mathbb{E}_{\{y_w,y_l,x\}\sim\mathcal{D}_{tr}}\left[\log\left(\frac{\pi_\theta(y_w|x)+\pi_\theta(y_l|x)}{\pi_\theta(y_w|x)}\right)\right].
\end{aligned} \tag{50}$$

Next, given the additional data generation assumptions, it follows that $\pi_\theta(y_w|x) + \pi_\theta(y_l|x) = 1$, and so the DPO loss can be further modified as

$$\begin{aligned}
\ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda) &= \mathbb{E}_{\{y_w,y_l,x\}\sim\mathcal{D}_{tr}}\left[\log\left(\frac{1}{\pi_\theta(y_w|x)}\right)\right] \\
&= \mathbb{E}_{x\sim\mathcal{D}_x}\Bigg[p^*(z=1|y_1,y_2,x)\log\left(\frac{1}{\pi_\theta(y_1|x)}\right) \\
&\qquad + \left(p^*(z=0|y_1,y_2,x)\log\left(\frac{1}{\pi_\theta(y_2|x)}\right)\right)\Bigg] \\
&= \mathbb{E}_{x\sim\mathcal{D}_x}\Bigg[\pi^*(y_1|x)\log\left(\frac{1}{\pi_\theta(y_1|x)}\right) \\
&\qquad + \pi^*(y_2|x)\log\left(\frac{1}{\pi_\theta(y_2|x)}\right)\Bigg] \\
&= \mathbb{E}_{x\sim\mathcal{D}_x}\Bigg[\pi^*(y_1|x)\log\left(\frac{\pi^*(y_1|x)}{\pi_\theta(y_1|x)}\right) \\
&\qquad + \pi^*(y_2|x)\log\left(\frac{\pi^*(y_2|x)}{\pi_\theta(y_2|x)}\right)\Bigg] \;+C \\
&\equiv \mathbb{E}_{x\sim\mathcal{D}_x}\Big[\mathbb{KL}\big[\pi^*(y|x)||\pi_\theta(y|x)\big]\Big],
\end{aligned} \tag{51}$$

where C is an irrelevant constant. Note that in progressing from the first to second equality, we can ignore cases where where sampled responses satisfy $y_1 = y_2$, since these contribute only another irrelevant constant to the loss. Along with our stated response data assumptions, this allows us to remove expectation over $\{y_1, y_2\}$ without loss of generality.

Final step: From (49) and (51) we observe that the only difference between the RLHF and DPO losses under the given conditions is whether a forward or backward KL is used. And of course *without* any constraints, the minimizing solutions are equivalent as expected, consistent with the analysis from (Rafailov et al., 2024), i.e.,

$$\arg\min_{\pi_\theta}\ell_{\text{RLHF}}\left(\pi_\theta, \pi_{\text{ref}}, r^*, \lambda\right) \;=\; \arg\min_{\pi_\theta}\ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda). \tag{52}$$

Critically though, this KL equivalence transparently need *not* still hold once constraints are introduced, as the forward KL will favor mode covering while the backward KL will push mode following (Bishop, 2006). ■

E.6. Proof of Propositions 4.1 and 4.2

These results both follow directly from the original design of $\ell_{\text{TYPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda)$. Regarding Proposition 4.1, given that $\pi_{\text{ref}} = \pi^*$ for all $x \in d_x^{good}$, then for the unsupervised term we have

$$\arg\min_{\pi_\theta}\mathbb{E}_{y\sim\pi_{\text{ref}}(y|x),x\in d_x^{good}}\Big[\mathbb{KL}\big[\pi_{\text{ref}}(y|x)||\pi_\theta(y|x)\big]\Big] \;=\; \pi^*. \tag{53}$$

And for the supervised term we have

$$\arg\min_{\pi_\theta} \mathbb{E}_{\{y_1,y_2\}\sim\pi_{\text{ref}}(y|x),x\sim\mathcal{D}_x}\Big[\mathbb{KL}\big[p^*(z|y_1,y_2,x)||p_\theta(z|y_1,y_2,x)\big]\Big] \;=\; \pi^*. \tag{54}$$

Hence overall, for any $x \in d_x^{good}$, $\pi_\theta = \pi^*$ will be optimal for any λ, as this selection independently optimizes the constituent terms. Moreover, this optimality is independent of optimization over $x \in d_x^{bad}$, which retains the flexibility to achieve solutions with $\text{dist}[\hat{\pi}_\theta^{\text{TYPO}},\ \pi^*] < \text{dist}[\pi_{\text{ref}},\ \pi^*]$. From this Proposition 4.1 immediately follows.

Additionally, Proposition 4.2 follows from the same basic line of reasoning. For completeness, we note that when $\lambda \to 0$, only the supervised term will be minimized (which recovers the BT-optimal policy as above), while when $\lambda \to \infty$, the unsupervised term will dominate the optimization (which transparently produces π_{ref}). ■

F. Other Derivations

F.1. Derivation of (12)

Note that

$$\begin{aligned} p^*(y_1 \succ y_2|x) &= \frac{\exp[r^*(y_1,x)]}{\exp[r^*(y_1,x)]+\exp[r^*(y_2,x)]} = \frac{\frac{\exp[r^*(y_1,x)]}{Z(x)}}{\frac{\exp[r^*(y_1,x)]}{Z(x)}+\frac{\exp[r^*(y_2,x)]}{Z(x)}} \\ &= \frac{\pi^*(y_1|x)}{\pi^*(y_1|x)+\pi^*(y_2|x)}, \end{aligned} \tag{55}$$

where $\pi^*(y|x) = \frac{\exp[r^*(y_1,x)]}{Z(x)}$ and $Z(x) := \sum_y \exp[r^*(y,x)]$. The policy π^* so-defined is necessarily BT-optimal by construction. From here then we have

$$\begin{aligned} \arg\max_{\pi_\theta} \mathbb{E}_{y\sim\pi_\theta(y|x)}\big[r^*(y,x)\big] &= \arg\max_{\pi_\theta} \mathbb{E}_{y\sim\pi_\theta(y|x)}\big[r^*(y,x)\big] \\ &= \arg\max_{\pi_\theta} \mathbb{E}_{y\sim\pi_\theta(y|x)}\left[\frac{\exp[r^*(y_1,x)]}{Z(x)}\right] \\ &= \arg\max_{\pi_\theta} \mathbb{E}_{y\sim\pi_\theta(y|x)}\big[\pi^*(y|x)\big] \\ &= \begin{cases} 1 & \text{if } y = \arg\max_{y'} \pi^*(y'|x) \\ 0 & \text{otherwise} \end{cases}, \end{aligned} \tag{56}$$

which is the definition of π^δ. ■

F.2. Additional f-DPO Analysis

f-PDO represents a novel generalization of DPO, but there remain certain aspects worth considering.

Minima that ignore the reference policy: Consider general f-DPO losses as described in Section 2.4, which as special cases of QPO are expressible in the form

$$\begin{aligned} &\ell_{\text{QPO}}(\pi_\theta, \pi_{\text{ref}}, -\log\sigma[\lambda(\cdot)], f', \lambda) = \\ &\qquad \mathbb{E}_{\{y_w,y_l,x\}\sim\mathcal{D}_{\text{tr}}} -\log\sigma\left(\lambda f'\left[\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)}\right] - \lambda f'\left[\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right], \lambda\right). \end{aligned} \tag{57}$$

In addition to the requirements on f to form an f-divergence, to produce a valid f-DPO loss per Theorem 1 from (Wang et al., 2024) it must be that f' is invertible with $0 \notin$ domain of f'. Therefore the domain of f will be $(0,\infty)$ and $f'(u) \to -\infty$ as $u \to 0$ because of convexity. But if this is the case, upon inspection of (57) we observe that when $\pi_\theta(y_l|x) \to 0$, then for any fixed $\pi_\theta(y_w|x) > 0$ the input argument to the logistic function $\sigma(\cdot) = \frac{1}{1+\exp[-(\cdot)]}$ will converge to infinity, pushing the output to one and subsequently minimizing the corresponding negative-log factor. And so the global optimum can be achieved independent of the value of π_{ref}. ■

F.3. Derivation of (16)

$$
\begin{aligned}
d_{\text{sup}}(\pi_\theta, \pi_{\text{ref}}) &= \mathbb{E}_{\{y_1,y_2\}\sim\pi_{\text{ref}}(y|x),x\sim\mathcal{D}_x}\Big[\mathbb{KL}\big[p^*(z|y_1,y_2,x)||p_\theta(z|y_1,y_2,x)\big]\Big] \\
&= -\mathbb{E}_{\{y_1,y_2\}\sim\pi_{\text{ref}}(y|x),x\sim\mathcal{D}_x}\Big[\mathbb{E}_{z\sim p^*(z|y_1,y_2,x)}\log p_\theta(z|y_1,y_2,x)\Big] \;+\; C \\
&\equiv -\mathbb{E}_{\{y_1,y_2\}\sim\pi_{\text{ref}}(y|x),x\sim\mathcal{D}_x}\Big[p^*(z=1|y_1,y_2,x)\log p_\theta(z=1|y_1,y_2,x)\Big] \\
&\quad + \;-\mathbb{E}_{\{y_1,y_2\}\sim\pi_{\text{ref}}(y|x),x\sim\mathcal{D}_x}\Big[p^*(z=0|y_1,y_2,x)\log p_\theta(z=0|y_1,y_2,x)\Big], \\
&= -\mathbb{E}_{\{y_1,y_2\}\sim\pi_{\text{ref}}(y|x),x\sim\mathcal{D}_x}\Big[p^*(z=1|y_1,y_2,x)\log p_\theta(z=1|y_1,y_2,x) \\
&\qquad\qquad + \;p^*(z=1|y_2,y_1,x)\log p_\theta(z=1|y_2,y_1,x)\Big] \\
&= -\mathbb{E}_{\{y_w,y_l,x\}\sim\mathcal{D}_{tr}}\Big[\log p_\theta(z=1|y_w,y_l,x)\Big] \\
&= -\mathbb{E}_{\{y_w,y_l,x\}\sim\mathcal{D}_{tr}}\left[\log\left(\frac{\pi_\theta(y_w|x)}{\pi_\theta(y_w|x)+\pi_\theta(y_l|x)}\right)\right], \\
&= \mathbb{E}_{\{y_w,y_l,x\}\sim\mathcal{D}_{tr}}\left[\log\left(1+\frac{\pi_\theta(y_l|x)}{\pi_\theta(y_w|x)}\right)\right], \qquad (58)
\end{aligned}
$$

where C is a constant independent of θ. Additionally, the third-to-last equality stems from the definition of how tuples $\{y_w, y_l, x\}$ are sampled. In particular, for a given pair $\{y_1, y_2\}$, by definition a proportion $p^*(z=1|y_1,y_2,x)$ of the time $y_w = y_1$, while a proportion $p^*(z=0|y_1,y_2,x) = p^*(z=1|y_2,y_1,x)$ of the time $y_w = y_2$. Hence

$$
\begin{aligned}
&p^*(z=1|y_1,y_2,x)\log p_\theta(z=1|y_1,y_2,x) \;+\; p^*(z=1|y_2,y_1,x)\log p_\theta(z=1|y_2,y_1,x) \\
&\quad\equiv\; \log p_\theta(z=1|y_w,y_l,x) \qquad (59)
\end{aligned}
$$

when the latter is averaged over the preference distribution. ■

G. Limitations

As more of an analysis-driven contribution, our experiments on real-world data are limited to Figure 4. Moreover, there are promising possibilities raised by pairing our contribution with prior work in new ways that we have not yet been explored. One example is the potential use of REINFORCE in conjunction with modifications to the proposed ℓ_{TYPO} loss.

H. Broader Impacts

Aligning the output of LLMs with human preferences has obvious, well-documented benefits. However, there nonetheless remains the risk that tools designed to improve LLM responses could be repurposed for nefarious aims. For example, preference data labels could potentially be modified to train models, using preference losses such as ours, that intentionally produce toxic content.