Common Learning Constraints Alter Interpretations of Direct Preference Optimization

Lemin Kong[1] **Xiangkun Hu**[2] **Tong He**[2] **David Wipf**[2]
[1]The Chinese University of Hong Kong [2]Amazon Web Services

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). Although effective in certain real-world settings, we detail how the foundational role of DPO reparameterizations (and equivalency to applying RLHF with an optimal reward) may be obfuscated once inevitable optimization constraints are introduced during model training. This then motivates alternative derivations and analysis of DPO that remain intact even in the presence of such constraints. As initial steps in this direction, we re-derive DPO from a simple Gaussian estimation perspective, with strong ties to compressive sensing and classical constrained optimization problems involving noise-adaptive, concave regularization.

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., 2020; 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). And as formal justification for applying such reparameterizations, results from Rafailov et al. (2024) indicate that the unconstrained DPO optimum solution can be viewed as minimizing the original RLHF objective defined w.r.t. an optimal reward function. While dramatically economizing model development, with recency comes the potential that the consequences of less obvious properties of DPO-based objectives may still be under-explored.

Specifically, we prove that once inevitable model/learning constraints are introduced during training (explicitly or implicitly, e.g., early-stopping, weight decay, etc.) to achieve good performance, the core reparameterizations that underpin DPO models no longer ensure equivalency with a closed-form RLHF instantiation involving an optimal reward (Section 3). Follow-up work relying on analogous reparameterizations (Azar et al., 2024; Wang et al., 2024) also exhibit the same phenomena. This then

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

motivates alternative DPO derivations and supporting analyses that are not beholden to the impact of such constraints. We provide two such examples herein:

1. The re-derivation of DPO from a simple Gaussian estimation perspective independent of RLHF and attendant reparameterizations (Section 4); and
2. The contextualization of DPO as an example of classical constrained optimization involving noise-adaptive, concave regularization (Section 5), as has been historically applied to tasks such as compressive sensing (Candès et al., 2006; Candès and Wakin, 2008; Donoho, 2006) and robust estimation (Chartrand and Yin, 2008; Chen et al., 2017; Fan and Li, 2001; Rao et al., 2003).

With regard to the latter in particular, although the constraint set and optimization variables differ, we show that the DPO loss is functionally equivalent to a specific, widely-applied compressive sensing objective originating from (Candes et al., 2008; Wipf and Nagarajan, 2010). And finally, we conclude by empirically exploring properties of the DPO loss through a novel repurposing of the above historical perspective. Overall, by elucidating subtle DPO characteristics, as well as forming new connections with established work from the past, we take strides towards elevating our understanding of mechanisms for human preference optimization.

2 PRELIMINARIES

This section presents the fundamentals of RLHF, DPO, and follow-up approaches relevant to our later analysis of the disruptive role of optimization constraints.

2.1 RLHF

Reward Function Estimation: When provided two candidate responses $y_1 \neq y_2$ sampled from an LLM-based policy using input prompt x, the Bradley-Terry (BT) model (Bradley and Terry, 1952) for human preferences specifies that

$$p^*(y_1 \succ y_2|x) = \sigma\big[r^*(y_1,x) - r^*(y_2,x)\big]. \tag{1}$$

In this expression, p^* denotes the ground-truth preference distribution as determined by human annotations, $y_1 \succ y_2$ indicates that y_1 is *preferred* relative to y_2, $r^*(y,x)$ is a latent reward model governing the distribution, and σ is the logistic function. Although $r^*(y,x)$ is unobservable and we cannot directly compute $p^*(y_1 \succ y_2|x)$, we can train an approximation $p_\phi(y_1 \succ y_2|x)$ defined by a parameterized surrogate reward $r_\phi(y,x)$. Specifically, based on (1) we can minimize the negative log-likelihood loss $\ell_{\rm BT}(r_\phi) :=$

$$\mathbb{E}_{\{y_w,y_l,x\}\sim\mathcal{D}_{\rm tr}}\Big[-\log\sigma\big[r_\phi(y_w,x) - r_\phi(y_l,x)\big]\Big], \tag{2}$$

over r_ϕ, where $\mathcal{D}_{tr}$ denotes a training distribution of response/prompt tuples $\{y_w, y_l, x\}$, and $y_w \succ y_l$ per ground-truth human preference annotation (subscripts here stand for 'win' and 'lose').

RL Fine-Tuning with Estimated Reward: The goal here is to improve upon a given $\pi_{\rm ref}(y|x)$ using a separate trainable model $\pi_\theta(y|x)$, the high-level desiderata being: (i) Maximize a 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 $\pi_{\rm ref}(y|x)$ to avoid overfitting merely to preference rewards. These objectives materialize through the minimization of

$$\begin{aligned}\ell_{\rm RLHF}\left(\pi_\theta, \pi_{\rm ref}, \hat{r}_\phi, \lambda\right) \;&:=\; \mathbb{E}_{y\sim\pi_\theta(y|x), x\sim\mathcal{D}_x}\Big[-\hat{r}_\phi(y,x)\Big] \\ &+\;\lambda\,\mathbb{E}_{x\sim\mathcal{D}_x}\Big[\mathbb{KL}\big[\pi_\theta(y|x)||\pi_{\rm ref}(y|x)\big]\Big],\end{aligned} \tag{3}$$

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

2.2 DPO

Consider now the reward-dependent RLHF loss $\ell_{\rm RLHF}$ from (3) 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 (we will return to this assumption in Section 3), the minimum of $\ell_{\rm RLHF}\left(\pi_\theta, \pi_{\rm ref}, r, \lambda\right)$ w.r.t. π_θ can be directly computed as

$$\begin{aligned}\pi_r(y|x) \;&:=\; \arg\min_{\pi_\theta} \ell_{\rm RLHF}\left(\pi_\theta, \pi_{\rm ref}, r, \lambda\right) \\ &=\; \frac{1}{Z(x)}\pi_{\rm ref}(y|x)\exp\left[\frac{1}{\lambda}r(y,x)\right],\end{aligned} \tag{4}$$

where $Z(x) := \sum_y \pi_{\rm ref}(y|x)\exp\left[\frac{1}{\lambda}r(y,x)\right]$ is the partition function ensuring that $\pi_r(y|x)$ forms a proper distribution (Peng et al., 2019; Peters and Schaal, 2007). From here, assuming $\pi_{\rm ref}(y|x) > 0$, we can rearrange (4) to equivalently establish that

$$r(y,x) = \lambda\log\frac{\pi_r(y|x)}{\pi_{\rm ref}(y|x)} + \lambda\log Z(x). \tag{5}$$

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 $\pi^{**}(y|x) := \arg\min_{\pi_\theta}\ell_{\rm RLHF}\left(\pi_\theta, \pi_{\rm ref}, r^*, \lambda\right)$. Hence instead of approximating $r^*(y,x)$ with $r_\phi(y,x)$ as in (1), we may equivalently approximate $\pi^{**}(y|x)$ with some $\pi_\theta(y|x)$

leading to the DPO loss

$$\ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda) \;:=\; \ell_{\text{BT}}\left(\lambda \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}\right)$$
$$= \; \mathbb{E}_{\{y_w,y_l,x\}\sim\mathcal{D}_{\text{tr}}}\left[-\log\sigma\left(\lambda\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} \right.\right.$$
$$\left.\left. - \;\lambda\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right], \quad (6)$$

noting that the partition function $Z(x)$ conveniently cancels out and can be excluded from further consideration. It is now possible to directly optimize (6) 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, ideally in a push towards $r^*(y,x)$.

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 (3) 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

$$r_{\text{IPO}}(y,x) := \mathbb{E}_{y'\sim\pi_{\text{ref}}(y|x)}\left[p^*(y \succ y'|x,y,y')\right], \quad (7)$$

over π_θ.[1] 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 (4). Combining y_1 and y_2 dependent terms, after a few algebraic manipulations this then leads to the equivalence relation

$$\log\left[\frac{\pi_{\text{IPO}}(y_1|x)\pi_{\text{ref}}(y_2|x)}{\pi_{\text{IPO}}(y_2|x)\pi_{\text{ref}}(y_1|x)}\right] = \frac{1}{\lambda}\left[r_{\text{IPO}}(y_1,x) - r_{\text{IPO}}(y_2,x)\right]. \quad (8)$$

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 $\ell_{\text{IPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda) \;:=$

$$\mathbb{E}_{\{y_1,y_2\}\sim\pi_{\text{ref}}(y|x), x\sim\mathcal{D}_x}\left[\left(\log\left[\frac{\pi_\theta(y_1|x)\pi_{\text{ref}}(y_2|x)}{\pi_\theta(y_2|x)\pi_{\text{ref}}(y_1|x)}\right]\right.\right.$$
$$\left.\left. - \;\frac{1}{\lambda}\left[r_{\text{IPO}}(y_1,x) - r_{\text{IPO}}(y_2,x)\right]\right)^2\right] \quad (9)$$
$$\equiv \; \mathbb{E}_{\{y_w,y_l,x\}\sim\mathcal{D}_{\text{tr}}}\left[\left(\log\left[\frac{\pi_\theta(y_w|x)\pi_{\text{ref}}(y_l|x)}{\pi_\theta(y_l|x)\pi_{\text{ref}}(y_w|x)}\right] - \frac{1}{2\lambda}\right)^2\right],$$

with the stated equivalence following under modest assumptions. Note that this closed-form consistency is a direct consequence of how r_{IPO} is defined in (7) and will not generally hold for *other* choices of the reward function. Regardless, it is straightforward to minimize $\ell_{\text{IPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda)$ via SGD as with DPO.

2.4 Additional DPO-like Methods

f-DPO (Wang et al., 2024) represents a natural generalization of DPO obtained by simply replacing the KL-divergence from (3) with a general f-divergence (Rubenstein et al., 2019). The f-DPO loss is then obtained using the same reparameterization scheme as DPO, adjusted for each choice of divergence; hence our constraint-based analysis below remains quite relevant to f-DPO. Other very recent approaches with varying degrees of similarity to DPO include Amini et al. (2024); Ethayarajh et al. (2024); Gorbatovski et al. (2024); Hong et al. (2024); Meng et al. (2024); Pal et al. (2024); Park et al. (2024); Tang et al. (2024); Zhao et al. (2023b). To the extent that these rely on analogous reparameterizations as with DPO, our analysis also remains loosely relevant, although we do not consider each on a case-by-case basis.

3 IMPACT OF OPTIMIZATION CONSTRAINTS

It follows from the design of DPO and its core reparameterization, that if $r^* = \arg\min_{r_\phi} \ell_{\text{BT}}(r_\phi)$, then

$$\arg\min_{\pi_\theta} \ell_{\text{RLHF}}(\pi_\theta, \pi_{\text{ref}}, r^*, \lambda) = \arg\min_{\pi_\theta} \ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda). \quad (10)$$

This establishes an explicit link between DPO and an idealized version of RLHF equipped with an optimal reward (approximate rewards are addressed in Appendix A). But there is a pivotal assumption underlying this association: the equality that facilitates the DPO reparameterization, namely (5), is predicated on the solution of an *unconstrained* optimization problem from (4) over an arbitrary policy π_θ.

However, when actually training models in real-world settings, constraints will *always* exist, whether implic-

[1] 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.

itly or explicitly. Such constraints stem from any number of factors including the model architecture/capacity limitations, weight decay, drop-out regularization, machine precision, and so on. Additionally, the DPO loss can have degenerate unconstrained minimizers that completely ignore π_{ref} on real-world datasets (Azar et al., 2024), and so counter-measures like early stopping are imposed that effectively introduce a constraint, which also substantially alters the estimated policy.

Therefore in reality we are never exactly minimizing the loss $\ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda)$ over any possible π_θ. Instead, we must consider properties of the *constrained* problem $\min_{\pi_\theta \in \mathcal{S}_\pi} \ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda)$, 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.

3.1 DPO Constraint-based Limitation

Consider now the following:

Proposition 3.1. *Let $\mathcal{S}_\pi$ denote a constraint set on the learnable policy π_θ, and assume $r^* = \arg\min_{r_\phi} \ell_{BT}(r_\phi)$. Then we can have that*

$$\arg\min_{\pi_\theta \in \mathcal{S}_\pi} \ell_{RLHF}\left(\pi_\theta, \pi_{ref}, r^*, \lambda\right) \neq \arg\min_{\pi_\theta \in \mathcal{S}_\pi} \ell_{DPO}(\pi_\theta, \pi_{ref}, \lambda). \tag{11}$$

As observed in the proof (see Appendix C for details), the difference between the two is akin to the difference between applying a constraint to a trainable policy in either the forward or backward KL divergence, which generally exhibit quite distinct mode covering or mode seeking behaviors respectively (Bishop, 2006). Figure 1 also presents a visualization of this phenomena as it pertains to the DPO-RLHF relationship, while empirical implications are deferred to Section 3.3. Overall, we have shown that once a constraint is introduced and the inequality from (11) activated, *we can no longer say that DPO equates to solving RLHF with an optimal reward*, i.e., this motivational connection is now more ambiguous as will be explored experimentally below. We include more fine-grained analysis in Appendix A, contrasting practical scenarios where an ideal RLHF reward may or not be present.

3.2 Extensions Beyond DPO

As other approaches such as IPO and f-DPO depend on analogous reparameterizations and potentially-tenuous connections with RLHF, it is natural to consider extending Proposition 3.1 beyond DPO. At least for IPO we have the following:

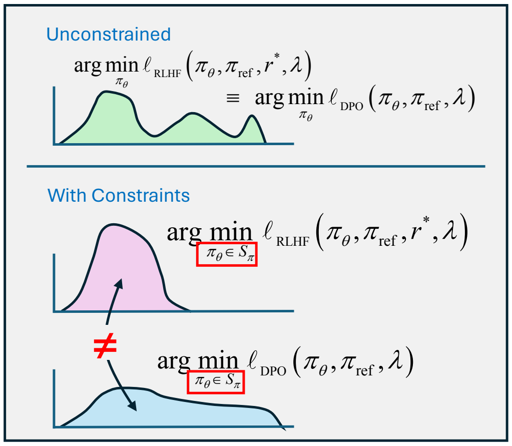
Unconstrained
$\arg\min_{\pi_\theta} \ell_{\text{RLHF}}(\pi_\theta, \pi_{\text{ref}}, r^*, \lambda)$
$\equiv \arg\min_{\pi_\theta} \ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda)$
With Constraints
$\arg\min_{\pi_\theta \in \mathcal{S}_\pi} \ell_{\text{RLHF}}(\pi_\theta, \pi_{\text{ref}}, r^*, \lambda)$
$\neq$
$\arg\min_{\pi_\theta \in \mathcal{S}_\pi} \ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda)$

Figure 1: *Visualization of learned policies with and without constraints.* On the top, no constraints are present and minimizing the respective DPO and RLHF losses leads to the same policy (green distribution over responses). In contrast, when the minimization is restricted to policies $\pi_\theta \in \mathcal{S}_\pi$, the RLHF solution (pink distribution) and DPO solution (blue distribution) are no longer the same. We emphasize that in all cases RLHF is instantiated with the same optimal reward r^*, so the discrepancy is entirely a consequence of policy constraints, not sub-optimal reward usage.

Proposition 3.2. *Let $\mathcal{S}_\pi$ denote a constraint set on the learnable policy π_θ. Then we can have that*

$$\arg\min_{\pi_\theta \in \mathcal{S}_\pi} \ell_{RLHF}\left(\pi_\theta, \pi_{ref}, r_{IPO}, \lambda\right) \neq \arg\min_{\pi_\theta \in \mathcal{S}_\pi} \ell_{IPO}(\pi_\theta, \pi_{ref}, \lambda). \tag{12}$$

We emphasize that although Proposition 3.2 is structured similarly to Proposition 3.1, the proof is considerably different, based on a construction that loosely resembles natural implications of an early-stopping constraint. From these results we also hypothesize that the same inherent, constraint-based limitations apply to f-DPO; however, we do not as of yet have a formal proof. Instead we rely on empirical corroborations as detailed next.

3.3 Empirical Study of Constraint Effects

To complement our analysis from above, we first empirically examine the practical implications of two types of commonly-used constraints: *weight decay* to control parameter magnitudes, and *early stopping* during training. In both cases we demonstrate that these constraints interfere with the equivalence between DPO, f-DPO, and IPO, and their respective RLHF counterparts when implemented with an optimal reward. Later we examine distinct mode covering (DPO) versus mode seeking (RLHF) behaviors that are induced by the imposition of constraints.

Following the design of Azar et al. (2024), we adopt a bandit setting with a discrete action space of three responses $\mathcal{Y} = \{y_a, y_b, y_c\}$. The dataset consists of labeled pairs $\{\{y_a, y_b\}, \{y_b, y_c\}, \{y_a, y_c\}\}$, representing a total order adhering to the BT model. To train our policy, we use a softmax-based parameterization $\pi_\theta(y_i)$, with $\theta \in \mathbb{R}^3$, and optimize the parameters using Adam across different preference losses. Experimental details are deferred to Appendix B.

Weight Decay: We train policies to separately minimize DPO, f-DPO, IPO, and their respective RLHF counterparts (6 models in total). For f-DPO we select the forward-KL as a contrast to the reverse-KL implicitly assumed by DPO (Wang et al., 2024). To implement weight decay, we apply the penalty term $\alpha\|\pi_\theta\|_2^2$ to all models, where $\alpha \geq 0$ is an adjustable hyperparameter. Figure 2 shows the distance (y-axis) between learned policies from DPO, f-DPO, IPO and their RLHF counterparts as α is varied (x-axis). Consistent with original model derivations and analyses in Azar et al. (2024); Rafailov et al. (2024); Wang et al. (2024), we observe negligible distances when $\alpha = 0$ given that the unconstrained DPO/IPO/f-DPO models are all designed to mimic RLHF with an optimal reward r^*. However, in accordance with our Propositions 3.1 and 3.2, as $\alpha > 0$ increases, these respective distances grow considerably, compromising the original relationships.

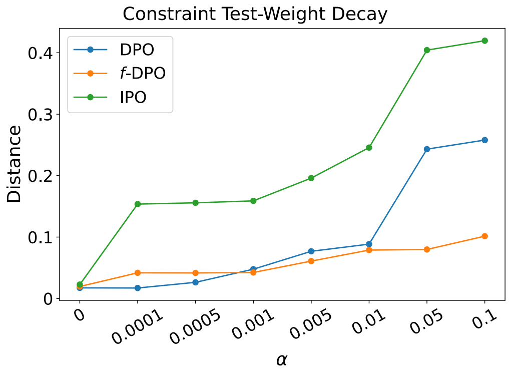

Figure 2: Effect of varying α on the distance between a policy generated by DPO, f-DPO, or IPO, and the corresponding RLHF-generated policy. A larger α is tantamount to a stronger constraint.

Early Stopping: Early stopping (during training) can also be viewed as a form of implicit constraint, whereby fewer epochs translate into harder constraints. It is also well-known that early stopping is essential to obtaining strong DPO performance in practice, and hence represents a particularly relevant form of constraint to test empirically. We hypothesize then that, as training epochs increase, and the implicit constraint loosens, the distance between the learned DPO policy and its RLHF counterpart will reduce; likewise for IPO and f-DPO. Figure 3 depicts these distances as a function of epoch, and consistent with expectations, as the number of epochs becomes sufficiently large all curves converge to zero indicating equivalence with RLHF. Of course in practice, since very few epochs are actually used, a substantial gap between each model and its RLHF counterpart will exist.

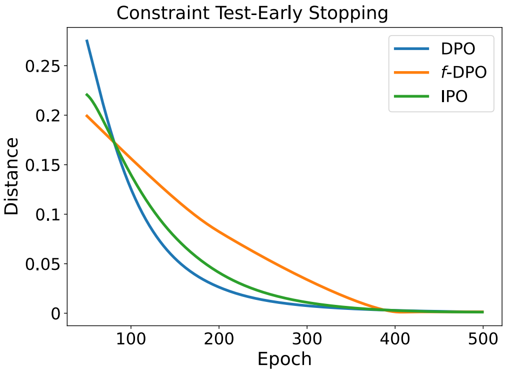

Figure 3: Effect of training epochs on the distance between a policy generated by DPO, f-DPO, or IPO, and the corresponding RLHF-generated policy. In this scenario, the largest implicit constraint occurs at zero epochs (i.e., with no training each model is constrained at its initialization point) and steadily relaxes, such that the observed distances converge to zero in all cases as predicted.

Mode Seeking vs Mode Covering: Beyond showing that DPO and RLHF policies diverge significantly with the introduction of constraints, we now explore more specific behavioral differences that stem from our prior theoretical contributions. In particular, as described in Section 3.1, under certain assumptions the RLHF loss reduces to a reverse KL divergence, which is mode seeking, while the DPO loss reduces to a forward KL divergence, which is mode-covering (sometimes called mean seeking). We hypothesize that this key behavioral distinction should be observable with a targeted empirical design as follows.

We again adopt a bandit setting with $\mathcal{Y} = \{y_a, y_b, y_c\}$, only now we stipulate a ground-truth preference distribution given by $p^*(y_a \succ y_b) = p^*(y_c \succ y_b) = 1$ and $p^*(y_a \succ y_c) = 1/2$. Per later discussion in Section 4, this distribution is directly induced via the optimal policy $\pi^*(y_a) = \pi^*(y_c) = 1/2$ and $\pi^*(y_b) = 0$, which can be viewed as having two modes (i.e., at y_a and y_c). We then assume the constrained policy $\pi_\theta = \theta \cdot [1, 0, 0]^\top + (1 - \theta) \cdot [1/3, 1/3, 1/3]^\top$, with trainable parameter $\theta \in [0, 1]$, and set π_{ref} to a uniform distribution. Starting with an initial $\theta = 0.5$, we optimize the constrained RLHF and DPO losses using projected Adam (Reddi et al., 2018). The resulting learned values of θ, denoted θ^{RLHF} and θ^{DPO}, closely conform with our theoretical prediction as follows:

- $\theta^{\text{RLHF}} = 1.00$, a *mode seeking* solution whereby all mass is confined to a single mode of π^*, and
- $\theta^{\text{DPO}} = 0.44$, a *mode covering* solution whereby significant mass is assigned to both modes of π^*.

These results highlight behavioral differences between the two frameworks: RLHF tends to produce a mode-seeking behavior while DPO more favors mode covering/mean seeking solutions.

3.4 Next Steps

The results of Section 3 as well as supporting analysis in Appendix A, establish that the value of DPO in practice (and indeed it often does work well) cannot be unreservedly attributed to its affiliation with an optimal RLHF solution (likewise for related models such as IPO and f-DPO, etc.). Because of this, we instead advocate that DPO be evaluated based on properties of $\min_{\pi_\theta \in \mathcal{S}_\pi} \ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda)$ itself that are not in contingent dependency on RLHF. We will take two steps in this direction as follows:

1. In Section 4 we rederive the DPO loss (6) from scratch based solely on a Gaussian estimation perspective that is *completely unrelated to RLHF-based reparameterizations*. Importantly, this derivation is orthogonal to whether or not constraints are included, and hence is not compromised when they inevitably are.
2. Of course what matters most are the properties of the underlying loss when deployed in practice, not necessarily the assumptions made in deriving the loss in the first place. To this end, based on the rederivation from above, Section 5 demonstrates how the constrained DPO loss can be interpreted as a well-studied instance of robust estimation using a noise-adaptive regularization factor, where the implicit noise is determined by π_{ref} performance.

4 REDERIVING DPO FROM SCRATCH VIA NAIVE GAUSSIAN ESTIMATION

Any preference probability given by the BT model in (1) can be equivalently re-expressed as

$$p^*(y_1 \succ y_2|x) \quad = \quad \mu\left[\frac{\pi^*(y_1|x)}{\pi^*(y_2|x)}\right], \qquad (13)$$

where $\pi^*(y|x)$ is a conditional probability of y given x 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 reasonable to consider other monotonically increasing choices to explore the additional generality of (13) (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 $\xi_\theta(y_1, y_2, x) :=$

$$\mu\left[\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)}\right], \quad \xi_{\text{ref}}(y_1, y_2, x) := \mu\left[\frac{\pi_{\text{ref}}(y_1|x)}{\pi_{\text{ref}}(y_2|x)}\right]. \quad (14)$$

Now suppose we assume the naive joint distribution given by $p\left(\begin{bmatrix} \xi_\theta(y_1, y_2, x) \\ \xi_{\text{ref}}(y_1, y_2, x) \end{bmatrix}\right)$

$$= \quad \mathcal{N}\left(\begin{bmatrix} \xi_\theta(y_1, y_2, x) \\ \xi_{\text{ref}}(y_1, y_2, x) \end{bmatrix} \middle| 0, \gamma(y_1, y_2, x) I\right), \qquad (15)$$

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

$$\min_{\pi_\theta \in \mathcal{S}_\pi,\ \{\gamma(y_w, y_l, x) > 0\}} \Bigg\{ \mathbb{E}_{\{y_w, y_l, x\} \sim \mathcal{D}_{\text{tr}}} \qquad (16)$$

$$- \log \mathcal{N}\left(\begin{bmatrix} \xi_\theta(y_w, y_l, x) \\ \xi_{\text{ref}}(y_w, y_l, x) \end{bmatrix} \middle| 0, \gamma(y_w, y_l, x) I\right) \Bigg\},$$

where I is a 2×2 identity matrix and $\mathcal{S}_\pi$ is any constraint set on π_θ as introduced in Section 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 per the following straightforward result:

Proposition 4.1. *Joinly minimizing (16) over $\pi_\theta \in \mathcal{S}_\pi$ and $\{\gamma(y_w, y_l, x) > 0\}$ with $\mu(\cdot) = (\cdot)^{\frac{\lambda}{2}}$ is equivalent to solving* $\min_{\pi_\theta \in \mathcal{S}_\pi} \ell_{DPO}(\pi_\theta, \pi_{ref}, \lambda)$.

We emphasize that the goal here is to show how DPO can be rederived from completely different assumptions, *not that we actually believe the necessary preference ratios involved actually adhere closely to a Gaussian distribution*. But the fact that we can produce DPO in this way reinforces the notion that the origin story itself (whether RLHF-based or Gaussian-based) may be less important than concrete, quantifiable properties of the underlying loss function that is produced in the end. We investigate such properties next.

5 THE DPO LOSS INDUCES NOISE ADAPTIVE REGULARIZATION

In this section we first elucidate previously-unexplored connections between the DPO loss and much older regularization techniques; later we empirically study implications of these connections.

5.1 Bridging Old and New

The results of the previous section provide an alternative lens with which to probe properties of the DPO loss. Of particular interest here, the proof of Proposition 4.1 involves re-expressing the DPO loss from (6) as

$$\ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda) \equiv \tag{17}$$
$$\mathbb{E}_{\{y_w, y_l, x\} \sim \mathcal{D}_{\text{tr}}} \left[\log \left(\left[\frac{\pi_{\text{ref}}(y_l|x)}{\pi_{\text{ref}}(y_w|x)} \right]^\lambda + \left[\frac{\pi_\theta(y_l|x)}{\pi_\theta(y_w|x)} \right]^\lambda \right) \right],$$

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 5.1. 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

$$g(u_2) \leq g(u_1) + \frac{g'(u_1)}{f'(u_1)} \left[f(u_2) - f(u_1) \right], \tag{18}$$

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:

- The function $\log(\gamma + u)$ is concave and non-decreasing function for $u \in [0, \infty)$, which represents a well-known characteristic of sparsity-favoring penalty factors commonly used for tasks such as compressive sensing Candès et al. (2006); Candès and Wakin (2008); Donoho (2006) and robust estimation (Chartrand and Yin, 2008; Chen et al., 2017; Fan and Li, 2001; Rao et al., 2003).[2] 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 5.1. Figure 4 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 and Nagarajan, 2010) has investigated general optimization problems of the form

$$\min_{\{u_i\} \in \mathcal{S}_u} \sum_i \log(\gamma + |u_i|), \tag{19}$$

  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 a constraint set.[3] 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 and Zhang, 2014), possibly by even allowing γ to vary across different u_i. The basic intuition here is that when noise is high, a more convex/conservative shape is preferable, while when the noise is low, a more concave alternative may be appropriate.

- Regarding DPO, it is natural to treat $\gamma \equiv \left[\frac{\pi_{\text{ref}}(y_l|x)}{\pi_{\text{ref}}(y_w|x)}\right]^\lambda$ as an analogous noise factor, given that

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

[3] In some applications the constraint set may be replaced by an additional regularization factor, and there is often an equivalency between the two. And more broadly, even earlier work applies analogous penalization to rank minimization problems (Fazel et al., 2003).

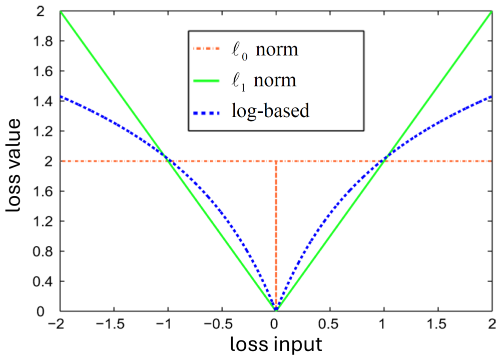

Figure 4: *Visualization of penalty effects associated with DPO loss.* When $\gamma \to 0$, $\log(\gamma + |u|) \to \log|u| = \lim_{p\to 0} \frac{1}{p}[|u|^p - 1] \propto \mathbb{I}[u \neq 0]$ mimicking an ℓ_0 norm (red curve) w.r.t. relative concavity (if $u \geq 0$ as with DPO, we can remove the absolute value, but we nonetheless include the general case here.). In contrast, $\lim_{\gamma\to\infty} \gamma \log(\gamma + |u|) = |u|$ 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, the relative concavity of $\log(\gamma + |u|)$ lies within these two extremes.

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), then solving $\min_{\pi_\theta \in \mathcal{S}_\pi} \ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda)$ can be viewed as a special case of (19), involving a robust regularization factor with noise-adaptive curvature. We explore this new interpretation next.

5.2 Empirical Consequences of (19) on DPO Regularization

To clearly distill the implications of (19), we consider a simplified bandit setting with response space $\mathcal{Y} = \{y_a, y_b\}$ and ground-truth preference distribution $p^*(y_a \succ y_b) = 1$. We parameterize $\pi_\theta(y_a) = \theta$ and $\pi_\theta(y_b) = 1 - \theta$, where $\theta \in [0, 1]$. Under these circumstances, and assuming $\lambda = 1/2$, we observe that (17) reduces to

$$\begin{aligned} \ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda) &= \log\left(\left[\frac{\pi_{\text{ref}}(y_b)}{\pi_{\text{ref}}(y_a)}\right]^{1/2} + \left[\frac{\pi_\theta(y_b)}{\pi_\theta(y_a)}\right]^{1/2}\right) \\ &= \log\left(\left[\frac{\pi_{\text{ref}}(y_b)}{\pi_{\text{ref}}(y_a)}\right]^{1/2} + \left[\frac{1-\theta}{\theta}\right]^{1/2}\right). \end{aligned} \tag{20}$$

We will now closely examine the behavior of this setup by varying the π_{ref} ratio under two types of constraints, namely, *weight decay* and *early stopping.*

Weight Decay: To incorporate a simple form of weight decay, we add the penalty factor θ^2 to (20). We

Table 1: Optimized values of θ for different reference policy ratios $\frac{\pi_{\text{ref}}(y_b)}{\pi_{\text{ref}}(y_a)}$.

π_{ref} ratio	0.01	0.05	0.1	0.5	1.0	5.0	10.0
Optimized θ	1.00	1.00	0.77	0.56	0.50	0.39	0.34

then vary the ratio $\frac{\pi_{\text{ref}}(y_b)}{\pi_{\text{ref}}(y_a)}$ from 0.01 to 10 to simulate different effective noise levels per Section 5.1 and optimize over θ in each case. We initialize θ so that $(1-\theta)/\theta$ equals the π_{ref} ratio and optimize via projected gradient descent. From the results summarized in Table 1 we observe that the noise-dependent regularization of DPO introduces a *strict thresholding effect*, whereby as long as the π_{ref} ratio is small enough (in this case ≤ 0.05), the optimal solution is $\theta = 1$, such that $\left[\frac{\pi_\theta(y_b)}{\pi_\theta(y_a)}\right]^{1/2} = \left[\frac{1-\theta}{\theta}\right]^{1/2}$ is exactly zero. Such exact thresholding behavior has been extensively studied in the literature (e.g., (Fan and Li, 2001)) and typically arises only with certain specialized loss functions. In the context of DPO, this thresholding with a weight decay constraint may or may not be desirable (depending on whether or not we wish to *selectively* override the influence of π_{ref}), but its emergence is nonetheless a notable observation unique to our work.

Early Stopping: We assess the impact of early stopping by measuring both the absolute and relative decreases of the DPO loss (20) after a single gradient descent step. These metrics provide insight into how quickly the optimization progresses across different π_{ref} ratios. From the results in Figure 5, we notice a remarkable difference, whereby early stopping disproportionately favors improving responses where π_{ref} is already good (i.e., a low ratio), at the expense of cases where π_{ref} is bad (i.e., a high ratio). We explore related behaviors using real data in Section 5.3 next.

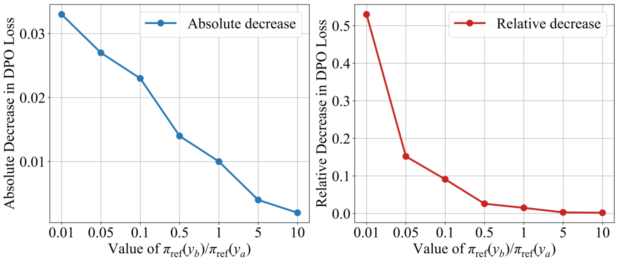

Figure 5: Decrease in DPO loss function (20) from a single step of gradient descent for different reference policy ratios $\frac{\pi_{\text{ref}}(y_b)}{\pi_{\text{ref}}(y_a)}$.

5.3 Real-World Data

The analysis from Section 5.1 predicts certain DPO regularization effects, which we have empirically corroborated using synthetic data in Section 5.2. For example, in cases where the implicit noise factor $\left[\frac{\pi_{\text{ref}}(y_l|x)}{\pi_{\text{ref}}(y_w|x)}\right]^\lambda$ is low, and the corresponding regularization is more concave, we should expect DPO training to reduce the corresponding learned policy ratio and loss function value more aggressively than when the opposite is true, the noise factor is high. We now turn to real-world data for further confirmation of this phenomena.

Setup: We train a Pythia 2.8B model (Biderman et al., 2023) using the Anthropic Helpfulness and Harmlessness preference benchmark (Bai et al., 2022a; Ganguli et al., 2022) as previously used in (Rafailov et al., 2024). As is customary, we first conduct supervised fine-tuning (SFT) by treating y_w as the training target; see Appendix B for details. Later we adopt this SFT model as π_{ref} for DPO training. We next compute the ratio $\frac{\pi_{\text{ref}}(y_l|x)}{\pi_{\text{ref}}(y_w|x)}$ for each sample in the dataset, splitting into two groups: samples where the ratio falls below the 0.5th percentile (*low-value* sample group) and those where the ratio exceeds the 99.5th percentile (*high-value* sample group) in terms of the overall distribution of ratio values. From here we compute

$$\frac{\pi_\theta(y_l|x)}{\pi_\theta(y_w|x)} \Big/ \frac{\pi_{\text{ref}}(y_l|x)}{\pi_{\text{ref}}(y_w|x)} \tag{21}$$

for all samples within each group, where π_θ is the model obtained from DPO after 3 epochs of training (note that at initialization this quantity should always equal 1 by design). We then statistically compare (21) computed using samples from the low-value and high-value groups. Our expectation here is that the more aggressive concave penalization associated with the low-value group will translate into greater suppression of (21) below 1. We apply two statistical techniques to evaluate such differences between groups.

Average Rank Sorting: Average rank sorting (Wilcoxon, 1945) ranks all data points together based on their relative sizes. For tied values, it assigns the average rank, ensuring a fair comparison between the two distributions. From Figure 6 we observe that the rank separation and smaller average rank for the low-value group confirm that it has been selectively reduced by DPO significantly more than the high-value group.

One-sided Mann-Whitney U Test: We also conduct a non-parametric Mann-Whitney U test (Mann and Whitney, 1947) designed to compare the distributions of two independent samples using another ranking-based statistic. This serves as an alternative to the t-test when the data are ordinal or otherwise not normally distributed. The large resulting U-statistic from Table 2 accompanied by an infinitesimally small p-value indicates a substantial difference between the low-value and high-value distributions.

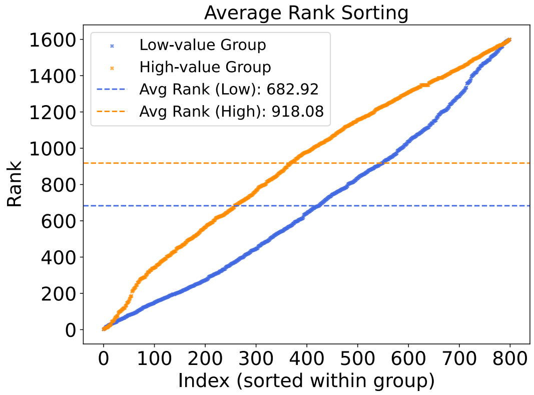

Figure 6: Average rank sorting of (21) computed over the low-value and high-value groups.

Table 2: Mann-Whitney U test results.

Statistic	Value
U-statistic	225937.0
p-value	1.22×10^{-24}

Interpretations: The complementary statistical tests from above confirm that when the implicit DPO noise factor is small, the percentage decrease in the ratio $\frac{\pi_\theta(y_l|x)}{\pi_\theta(y_w|x)}$ is larger compared to when the noise factor is large. Consistent with the notion of relative concavity from Section 5.1, when we initialize π_θ with π_{ref}, a small noise factor leads to a steeper initial gradient, which pushes the ratio $\frac{\pi_\theta(y_l|x)}{\pi_\theta(y_w|x)}$ more rapidly toward zero. Conversely, when the noise factor is large, the initial gradient is flatter, resulting in a slower decrease of the ratio. Of course one can argue whether or not this is necessarily a universally good outcome; however, our focus here is on quantifying and understanding actual DPO behavior, whatever that turns out to be.

6 CONCLUSION

We have argued that optimization constraints have the potential to interfere with the interpretation of DPO as implicitly minimizing the RLHF loss defined with an optimal reward function. As such constraints are unavoidable in practice, it therefore behooves us to consider alternative foundational entry points for quantifying DPO properties that withstand the introduction of constraints. We consider two such entry points herein, namely, a Gaussian estimation perspective and a complementary bridge to classical constrained optimization with noise-adaptive, concave regularization. The latter in particular allows us to elucidate previously-unexplored attributes of the DPO loss. Our analysis also motivates new alternatives to the DPO family that are initially derived with no direct affilation to RLHF; we reserve such considerations to future work.

References

Afra Amini, Tim Vieira, and Ryan Cotterell. Direct preference optimization with an offset. *arXiv preprint arXiv:2402.10571*, 2024.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR, 2024.

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

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

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR, 2023.

C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

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

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

Emanuel Candès, Justin Romberg, and Terence Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Information Theory*, 52(2): 489–509, Feb. 2006.

Emmanuel Candès and Michael Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

Emmanuel J Candes, Michael B Wakin, and Stephen P Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier analysis and applications*, 14:877–905, 2008.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3): 1–45, 2024.

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

Yichen Chen, Dongdong Ge, Mengdi Wang, Zizhuo Wang, Yinyu Ye, and Hao Yin. Strong NP-hardness for sparse optimization with concave penalty functions. In *International Confernece on Machine Learning*, 2017.

Bin Dai, Chen Zhu, Baining Guo, and David Wipf. Compressing neural networks using the variational information bottleneck. In *International Conference on Machine Learning*, pages 1135–1144. PMLR, 2018.

David L. Donoho. Compressed Sensing. *IEEE Trans. Information Theory*, 52(4), 2006.

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

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96 (456):1348–1360, 2001.

Maryam Fazel, Haitham Hindi, and Stephen Boyd. Log-det heuristic for matrix rank minimization with applications to hankel and euclidean distance matrices. In *Proceedings of the 2003 American Control Conference*, volume 3, pages 2156–2162. IEEE, 2003.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *arXiv preprint arXiv:2309.00770*, 2023.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv preprint arXiv:2209.07858*, 2022.

Alexey Gorbatovski, Boris Shaposhnikov, Alexey Malakhov, Nikita Surnachev, Yaroslav Aksenov, Ian Maksimov, Nikita Balagansky, and Daniil Gavrilov. Learn your reference model for real good alignment. *arXiv preprint arXiv:2404.09656*, 2024.

Jiwoo Hong, Noah Lee, and James Thorne. ORPO: Monolithic preference optimization without reference model. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11189. Association for Computational Linguistics, 2024.

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

Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, pages 50–60, 1947.

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

Nadim Nachar et al. The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. *Tutorials in quantitative Methods for Psychology*, 4(1):13–20, 2008.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

Arka Pal, Deep Karkhanis, Samuel Dooley, Manley Roberts, Siddartha Naidu, and Colin White. Smaug: Fixing failure modes of preference optimisation with dpo-positive. *arXiv preprint arXiv:2402.13228*, 2024.

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

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. *arXiv preprint arXiv:2403.19159*, 2024.

Xue Bin Peng, Aviral Kumar, Grace Zhang, and Sergey Levine. Advantage-weighted regression: Simple and scalable off-policy reinforcement learning. *arXiv preprint arXiv:1910.00177*, 2019.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. *arXiv preprint arXiv:2210.01241*, 2022.

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

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations*, 2018.

Paul Rubenstein, Olivier Bousquet, Josip Djolonga, Carlos Riquelme, and Ilya O Tolstikhin. Practical and consistent estimation of f-divergences. *Advances in Neural Information Processing Systems*, 32, 2019.

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

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021, 2020.

Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Rémi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Ávila Pires, and Bilal Piot. Generalized preference optimization: A unified approach to offline alignment. *arXiv preprint arXiv:2402.05749*, 2024.

Chaoqi Wang, Yibo Jiang, Chenghao Yang, Han Liu, and Yuxin Chen. Beyond reverse KL: Generalizing direct preference optimization with diverse divergence constraints. *International Conference on Learning Representations*, 2024.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

David Wipf and Srikantan Nagarajan. 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.

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

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, and Ji-Rong Wen. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023a.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. SLiC-HF: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023b.

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

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Not Applicable]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Yes]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A Additional Fine-Grained Analysis of DPO with Learning Constraints

Per the design of DPO and its adopted reparameterization, we know that if $r^* = \arg\min_{r_\phi} \ell_{\text{BT}}(r_\phi)$, then

$$\arg\min_{\pi_\theta} \ell_{\text{RLHF}}\left(\pi_\theta, \pi_{\text{ref}}, r^*, \lambda\right) = \arg\min_{\pi_\theta} \ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda). \tag{22}$$

As such, assuming r^* is reasonable, we may reliably utilize DPO to sidestep RL and still mimic an idealized version of RLHF that has been granted an optimal reward. However, once learning constraints are introduced (meaning we optimize over $\pi_\theta \in \mathcal{S}_\pi$ for DPO/RLHF, and possibly also a separate $r_\phi \in \mathcal{S}_r$ for RLHF), the interplay between DPO and RLHF becomes more nuanced. We address two distinct scenarios to highlight emergent ambiguities.

A.1 Ambiguity Involving Ideal Rewards

Consider an asymptotic regime whereby the minimum of $\ell_{\text{BT}}(r_\phi)$ perfectly reflects human preferences along some axis of interest. Additionally, suppose we introduce early stopping during training to prevent heavy drift away from π_{ref}. For example, this strategy may be computationally cheaper than using a larger value of λ and training until convergence (recall that a larger λ comports with a greater penalty on deviations from π_{ref} within the original RLHF framework; see (3)). However, as early stopping introduces an implicit constraint, we now enter the regime covered by Proposition 3.1 and equivalence between RLHF and DPO is no longer guaranteed. In fact their differences can be considerable as quantified in Section 3.3.

Constraints may also be invoked by necessity or other practical concerns as well. For example, model architectures, weight decay to stabilize training, etc., are frequently invoked with either RLHF or DPO. And again, by Proposition 3.1 these will generally introduce gaps between RLHF and DPO that may be consequential even when r^* itself is reasonable.

A.2 Ambiguity Selecting Among Approximate Rewards

We now address an alternative scenario whereby non-ideal, approximate rewards are in play via the incorporation of learning constraints.

On the Challenges of Choosing Constraints: To begin, we remark that *any* policy $\hat{\pi}_\theta$, *regardless of how it was originally obtained*, can be viewed as an optimal solution to the RLHF problem $\min_{\pi_\theta} \ell_{\text{RLHF}}\left(\pi_\theta, \pi_{\text{ref}}, \hat{r}_\phi, \lambda\right)$ with the right choice of reward $\hat{r}_\phi$. Specifically, we have that

$$\hat{\pi}_\theta \;=\; \arg\min_{\pi_\theta} \mathbb{E}_{y\sim\pi_\theta(y|x), x\sim\mathcal{D}_x}\Big[-\hat{r}_\phi(y,x)\Big] + \lambda\, \mathbb{E}_{x\sim\mathcal{D}_x}\Big[\mathbb{KL}\big[\pi_\theta(y|x) || \pi_{\text{ref}}(y|x)\big]\Big] \text{ s.t. } \hat{r}_\phi(y,x) = \lambda \log \frac{\hat{\pi}_\theta(y|x)}{\pi_{\text{ref}}(y|x)} \tag{23}$$

for any arbitrary $\hat{\pi}_\theta$. So without further specifications, minimizing an RLHF loss alone does not necessarily provide any means for preferring one $\hat{\pi}_\theta$ over another. Instead, it is the plausibility of the associated $\hat{r}_\phi$ that allows us to differentiate the associated policies or elevate one over another. For example, if $\hat{r}_\phi = r^* = \arg\min_{r_\phi} \ell_{\text{BT}}(r_\phi)$, then we have an optimal reward in terms of the sampled BT preference model, and the induced $\hat{\pi}_\theta$ will be desirable to the extent that we trust optimum of this model. But in cases where ℓ_{BT} is defined by limited labeled preference data, it could conceivably be the case that alternative rewards are preferable, such as instantiated by only *approximate* minimization of the BT preference model, possibly through early stopping during training. In practice, use of as few as 1 or 2 epochs of DPO training is commonplace, which is far from what would otherwise be necessary for convergence to an actual minimum of the BT preference model.

It is precisely here though that considerable ambiguity exists. If the goal is to determine our policy based on rewards that only partially reduce ℓ_{BT} by design, then what is the best way to introduce constraints to prevent $\hat{r}_\phi \to r^*$ during training? After all, there are countless ways to marginally reduce ℓ_{BT} without hitting a minimum. The original DPO framework is obviously one way, but a trivial recipe exists for extending, at least in principle, to broad model families with no direct relationship to RLHF.

A Trivial Hypothetical Recipe: Essentially any reasonable preference model should favor a policy that assigns higher probability to y_w than y_l for a given prompt x. For instance, consider a pair-wise preference loss of the general form given by

$$\ell_\eta(\pi_\theta) := \mathbb{E}_{\{y_w, y_l, x\}\sim\mathcal{D}_{\text{tr}}}\Big(\eta\big[\pi_\theta(y_w|x), \pi_\theta(y_l|x)\big]\Big), \tag{24}$$

where $\eta : [0,1]^2 \rightarrow \mathbb{R}$ is a monotonically increasing function of its first argument, and a monotonically decreasing function of its second argument. It is straightforward to show that the BT preference model is a special case of (24); likewise many other recently-proposed models such as SimPO (Meng et al., 2024). For our purposes here, the critical observation is that minimizing, even partially, any such (24) can in principle serve as a proxy for minimizing the BT preference model as follows:

Proposition A.1. *Suppose $\pi_{ref}(y|x) > 0$ and $\mathcal{D}_{tr}$ is such that for each prompt x, we have a single labeled response pair $\{y_w, y_l\}$.*[4] *Furthermore, let $\pi_\theta^{(0)}$ denote any policy used to initialize (24) whereby $\ell_\eta(\pi_\theta^{(0)}) > \arg\min_{\pi_\theta} \ell_\eta(\pi_\theta)$. Then there will always exist a loss value $\bar{\ell}_\eta$ such that*

$$\ell_{DPO}(\pi_\theta^{(0)}, \pi_{ref}, \lambda) \quad > \quad \ell_{DPO}(\bar{\pi}_\theta, \pi_{ref}, \lambda) \quad \forall\ \bar{\pi}_\theta \in \left\{\pi_\theta : \ell_\eta(\pi_\theta) < \bar{\ell}_\eta\right\}. \tag{25}$$

This result[5] allows us to apply the following recipe for "validating" an arbitrary preference model via the original DPO design criteria:

1. Train π_θ using (24) to obtain some $\bar{\pi}_\theta$ such that $\ell_{\text{DPO}}(\pi_\theta^{(0)}, \pi_{\text{ref}}, \lambda) > \ell_{\text{DPO}}(\bar{\pi}_\theta, \pi_{\text{ref}}, \lambda)$. By Proposition A.1 this should always be possible, at least if the stated dataset assumption is satisfied. This implies that we have at least approximately reduced the BT preference model reward implicitly used by DPO, which is the only criteria we have for assessing reward quality per the DPO framework.

2. Compute the implicit reward $\bar{r}_\phi(y, x) := \lambda \log \frac{\bar{\pi}_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$.

3. Conclude that optimality via (23) has been achieved with $\hat{\pi}_\theta = \bar{\pi}_\theta$ and $\hat{r}_\phi = \bar{r}_\phi$. Hence we have established a policy that is equivalent to obtaining the optimal RLHF policy associated with the given reward.

Per this recipe there is nothing that differentiates the original DPO model from any arbitrary model as specified here. And we cannot claim that the model which maximally reduces the BT preference loss used by DPO is preferable; if that were the case we should just optimize until we obtain r^*, which we have already conceded is undesirable under the current setting assumptions. Hence this challenges the notion that we can rely on (23) alone to justify DPO under the stated circumstances, as any such justification can apply equally well to an arbitrary model with no initial connection to RLHF.

Concluding Thoughts: We have argued that the ultimate justification for any preference optimization model is not that it minimizes (23), as any approach can do so for a specific reward. Instead, model justification rests on the legitimacy of the reward model used within (23). And then from here, once constraints are introduced and we no longer wish to optimally minimize the BT preference loss to obtain an optimal reward, it is not clear why one approach might be preferable to another for producing only marginally improved rewards (i.e., the size of the reward no longer establishes superiority of any given approach, as the maximal/optimal reward has already been deemed undesirable, which is why constraints were added in the first place). This is largely why we advocate for analyzing explicit properties of the DPO loss itself in Section 5, as these differentiating properties are not derived from (23) nor generally inherited by alternative approaches.

B Experiment Details

Code is available at `https://github.com/lmkong020/constraint-dpo-rlhf`.

B.1 Constraint Evaluation Tests from Section 3.3

Per our specifications from Section 3.3, we operate on a dataset of labeled response pairs drawn from $\{\{y_a, y_b\}, \{y_b, y_c\}, \{y_a, y_c\}\}$ following the original design from Azar et al. (2024). We assign ground-truth

[4] This is not uncommon in publicly-available human preference datasets (Bai et al., 2022a; Ganguli et al., 2022), where the empirical distribution of ground-truth preferences will be $p^*(y_w \succ y_l|x) = 1$ for all or most prompts $x \in \mathcal{D}_x$.

[5] The proof follows by simply noting that under the stated dataset assumptions, the optimal policy for any η will be such that $\pi_\theta(y_w|x) = 1$ and $\pi_\theta(y_l x) = 0$. But these values will also minimize the DPO loss provided that $\pi_{\text{ref}}(y|x) > 0$. Moreover, by assumption the initialized solution must have a higher loss. These conclusions then establish existence of a solution reducing both losses as desired; however, there could of course be others under more relaxed settings in practice.

preferences to these data as follows. First we define a latent policy π^* via the assignments

$$\pi^*(y_a) = 0.6, \quad \pi^*(y_b) = 0.3, \quad \pi^*(y_c) = 0.1. \tag{26}$$

We then generate preferences using a ground-truth preference distribution given by

$$p^*(y_1 \succ y_2) := \frac{\pi^*(y_1)}{\pi^*(y_1) + \pi^*(y_2)}, \tag{27}$$

noting that, w.l.o.g. any preference distribution expressible via the BT preference model can be represented in this way; see also (31) within the proof of Proposition 3.1 for the straightforward derivation in a general (non-bandit) setting. Equipped with (27) we can directly assign $\{y_w, y_l\}$ labels to each training sample pair, selected randomly from $\{\{y_a, y_b\}, \{y_b, y_c\}, \{y_a, y_c\}\}$. For each candidate model (i.e, DPO, f-DPO, IPO) we form a trainable policy as $\pi_\theta(y_i) = \text{softmax}[\theta_i]$ with $\theta \in \mathbb{R}^3$. Meanwhile, the reference policy is defined as

$$\pi_{\text{ref}}(y_a) = 0.4, \quad \pi_{\text{ref}}(y_b) = 0.4, \quad \pi_{\text{ref}}(y_c) = 0.2. \tag{28}$$

The experimental settings for investigating the mode-seeking and mode-covering behaviors of RLHF and DPO are detailed in Section 3.3. For the weight decay and early stopping experiments, we optimize model parameters using Adam with the respective preference losses. In contrast, for the experiments on mode-seeking and mode-covering behaviors, we use projected Adam to account for the constraint on model parameters. All experiments are conducted on a single A10 training for 1000 epochs (this was adequate for achieving convergence). Each RLHF counterpart was also trained accordingly to reach convergence. For the weight decay results in Figure 2, this process was repeated over a range of α values, in each case appending $\alpha\|\theta\|_2^2$ to the training loss. In contrast, for the early stopping results in Figure 3 only a single training run with $\alpha = 0$ is required.

B.2 DPO Regularization Tests from Section 5.2

For the experiment with a weight decay constraint, we add θ^2 to the loss function (20) and then optimize using the projected gradient descent for 500 iterations (this was adequate for achieving convergence). We repeat as the ratio $\frac{\pi_{\text{ref}}(y_b)}{\pi_{\text{ref}}(y_a)}$ is varied, with results displayed in Table 1.

Turning to the early stopping constraint, let θ_0 denote the initial value of θ and θ_1 the updated value after one step of gradient descent over the DPO loss function from (20). Figure 5 reports both the absolute and relative decreases in the DPO loss for different values of $\frac{\pi_{\text{ref}}(y_b)}{\pi_{\text{ref}}(y_a)}$, where

$$\begin{aligned} \text{Absolute Decrease} &:= \ell_{\text{DPO}}(\pi_{\theta_0}, \pi_{\text{ref}}, \lambda) - \ell_{\text{DPO}}(\pi_{\theta_1}, \pi_{\text{ref}}, \lambda) \quad \text{and} \\ \text{Relative Decrease} &:= \frac{\ell_{\text{DPO}}(\pi_{\theta_0}, \pi_{\text{ref}}, \lambda) - \ell_{\text{DPO}}(\pi_{\theta_1}, \pi_{\text{ref}}, \lambda)}{\ell_{\text{DPO}}(\pi_{\theta_0}, \pi_{\text{ref}}, \lambda)}. \end{aligned}$$

B.3 Tests with Anthropic HH Dataset from Section 5.3

For experiments involving the Anthropic HH benchmark[6], following Rafailov et al. (2024) we first conduct supervised fine-tuning for 2 epochs to obtain π_{ref}. Subsequently we train the DPO model for 3 epochs. All training was conducted using an 8×A100 40G GPU instance along with the Adam optimizer (Kingma and Ba, 2014), a learning rate of 10^{-6}, and batch size of 64. The statistical measures used to produce Figure 6 and Table 2 were computed as follows.

Average Rank Sorting: Beginning with two data groups, Group 1 and Group 2, we execute:

- Combine the data from both groups into a single dataset.
- Sort all values from the combined dataset in ascending order. Assign ranks starting from 1 for the smallest value. If two or more values are tied, assign them the average of the ranks they would have received.
- After assigning ranks, separate them back into their respective groups (e.g., Group 1 and Group 2). Then compute the average rank for Group 1 and Group 2 by summing the ranks within each group and dividing by the number of observations in that group.

[6] Released under a MIT License agreement available at `https://github.com/anthropics/hh-rlhf/blob/master/LICENSE`.

Mann-Whitney U Test: Operating with the same groups from before, we execute:

- Use the previous average rank sorting procedure to assign ranks to all values from both groups combined.

- The U statistic measures the number of times a value from one group precedes a value from the other group in rank order. This is computed group-wise as U_1 (for Group 1) and U_2 (for Group 2) using

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1, \; U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2, \tag{29}$$

  where R_1 and R_2 are the sums of ranks in Groups 1 and 2, and n_1 and n_2 are the sample sizes for each group. The smaller of U_1 and U_2 is then selected as the U statistic for the test. In our case, with $n_1 = n_2 = 800$, $R_1 = 546,337$, and $R_2 = 734,463$, we calculate $U_1 = 225,937$ and $U_2 = 414,063$, and select the smaller value, $U_1 = 225,937$, as the U statistic.

- Given our large sample size ($n_1 = n_2 = 800$), the distribution of the U statistic approximates a normal distribution. Therefore, we calculate the z-score and utilize the standard normal distribution to determine the p-value (Mann and Whitney, 1947; Nachar et al., 2008). First, we compute the mean (μ_U) and standard deviation (σ_U) of U:

$$\mu_U = \frac{n_1 n_2}{2}, \quad \sigma_U = \sqrt{\frac{n_1 n_2(n_1+n_2+1)}{12}}. \tag{30}$$

  Subsequently, we calculate the z-score as $z = (U - \mu_U)/\sigma_U$ and use the standard normal distribution to determine the corresponding p-value.

C Proof of Proposition 3.1

Our strategy here is to construct a situation whereby we can pinpoint emergent differences between RLHF and DPO losses in the presence of policy constraints. We note that while obviously simplified for transparency, the chosen formulation is nonetheless emblematic of behavior in broader regimes. To this end, we assume the following:

- For all $x \sim \mathcal{D}_x$, where $\mathcal{D}_x$ is an arbitrary prompt distribution, there exists two unique responses y_1 and y_2 with equal probability under π_{ref};

- Preference data $\{y_w, y_l, x\} \sim \mathcal{D}_{tr}$ are sampled according to $z \sim p^*(y_1 \succ y_2|x)$, $\{y_1, y_2\} \sim \pi_{\text{ref}}(y|x), x \sim \mathcal{D}_x$, where $z = \mathbb{I}[y_1 \succ y_2|y_1, y_2, x]$ is a binary indicator variable that determines y_w and y_l assignments;[7]

- 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 $x \in \mathcal{D}_x$.

With regard to the latter, we note that the preference distribution can be expressed as

$$\begin{aligned} p^*(y_1 \succ y_2|x) &= \frac{\exp[r^*(y_1,x)]}{\exp[r^*(y_1,x)] + \exp[r^*(y_2,x)]} = \frac{\frac{\exp[r^*(y_1,x)]}{Z(x)}}{\frac{\exp[r^*(y_1,x)]}{Z(x)} + \frac{\exp[r^*(y_2,x)]}{Z(x)}} \\ &= \frac{\pi^*(y_1|x)}{\pi^*(y_1|x) + \pi^*(y_2|x)}, \end{aligned} \tag{31}$$

where $\pi^*(y|x) := \frac{\exp[r^*(y_1,x)]}{Z(x)}$ and $Z(x) := \sum_y \exp[r^*(y,x)]$.

[7] We generally assume that $y_1 \neq y_2$; however, the $y_1 = y_2$ case can nonetheless be handled by simply assigning $p^*(y \succ y|x) = 1/2$, 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.

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

$$\begin{aligned} \ell_{\mathrm{RLHF}}\left(\pi_\theta, \pi_{\mathrm{ref}}, r^*, \lambda\right) &= \mathbb{E}_{y \sim \pi_\theta(y|x), x \sim \mathcal{D}_x}\Big[-r^*(y,x)\Big] + \lambda\, \mathbb{E}_{x \sim \mathcal{D}_x}\Big[\mathbb{KL}\big[\pi_\theta(y|x) || \pi_{\mathrm{ref}}(y|x)\big]\Big] \\ &\equiv \mathbb{E}_{x \sim \mathcal{D}_x}\Big[\mathbb{KL}\big[\pi_\theta(y|x) || \pi^{**}(y|x)\big]\Big], \end{aligned} \tag{32}$$

where

$$\pi^{**}(y|x) := \frac{1}{Z(x)} \pi_{\mathrm{ref}}(y|x) \exp\left[\frac{1}{\lambda} r^*(y,x)\right]. \tag{33}$$

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

$$\pi^{**}(y|x) = \frac{\exp\left[r^*(y,x)\right]}{\sum_y \exp\left[r^*(y,x)\right]}, \tag{34}$$

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

$$\frac{\pi^*(y_1|x)}{\pi^*(y_1|x) + \pi^*(y_2|x)} = p^*(y_1 \succ y_2|x) = \frac{\exp\left[r^*(y_1,x)\right]}{\exp\left[r^*(y_1,x)\right] + \exp\left[r^*(y_2,x)\right]}, \tag{35}$$

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

$$\ell_{\mathrm{RLHF}}\left(\pi_\theta, \pi_{\mathrm{ref}}, r^*, \lambda\right) = \mathbb{E}_{x \sim \mathcal{D}_x}\Big[\mathbb{KL}\big[\pi_\theta(y|x) || \pi^*(y|x)\big]\Big] \tag{36}$$

under the stated conditions.

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

$$\begin{aligned} \ell_{\mathrm{DPO}}(\pi_\theta, \pi_{\mathrm{ref}}, \lambda) &= \mathbb{E}_{\{y_w, y_l, x\} \sim \mathcal{D}_{\mathrm{tr}}}\left[-\log \sigma\left(\lambda \log \frac{\pi_\theta(y_w|x)}{\pi_{\mathrm{ref}}(y_w|x)} - \lambda \log \frac{\pi_\theta(y_l|x)}{\pi_{\mathrm{ref}}(y_l|x)}\right)\right] \\ &= \mathbb{E}_{\{y_w, y_l, x\} \sim \mathcal{D}_{tr}}\left[\log\left(\frac{\pi_\theta(y_w|x) + \pi_\theta(y_l|x)}{\pi_\theta(y_w|x)}\right)\right]. \end{aligned} \tag{37}$$

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

$$\begin{aligned} \ell_{\mathrm{DPO}}(\pi_\theta, \pi_{\mathrm{ref}}, \lambda) &= \mathbb{E}_{\{y_w, y_l, x\} \sim \mathcal{D}_{tr}}\left[\log\left(\frac{1}{\pi_\theta(y_w|x)}\right)\right] \\ &= \mathbb{E}_{x \sim \mathcal{D}_x}\Bigg[p^*(z=1|y_1,y_2,x) \log\left(\frac{1}{\pi_\theta(y_1|x)}\right) \\ &\qquad + (p^*(z=0|y_1,y_2,x) \log\left(\frac{1}{\pi_\theta(y_2|x)}\right)\Bigg] \\ &= \mathbb{E}_{x \sim \mathcal{D}_x}\Bigg[\pi^*(y_1|x) \log\left(\frac{1}{\pi_\theta(y_1|x)}\right) \\ &\qquad + \pi^*(y_2|x) \log\left(\frac{1}{\pi_\theta(y_2|x)}\right)\Bigg] \\ &= \mathbb{E}_{x \sim \mathcal{D}_x}\Bigg[\pi^*(y_1|x) \log\left(\frac{\pi^*(y_1|x)}{\pi_\theta(y_1|x)}\right) \\ &\qquad + \pi^*(y_2|x) \log\left(\frac{\pi^*(y_2|x)}{\pi_\theta(y_2|x)}\right)\Bigg] + C \\ &\equiv \mathbb{E}_{x \sim \mathcal{D}_x}\Big[\mathbb{KL}\big[\pi^*(y|x) || \pi_\theta(y|x)\big]\Big], \end{aligned} \tag{38}$$

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 (36) and (38) we observe that the only difference between the RLHF and DPO losses under the given conditions is whether a forward or backward KL divergence 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.,

$$\arg\min_{\pi_\theta} \ell_{\text{RLHF}}\left(\pi_\theta, \pi_{\text{ref}}, r^*, \lambda\right) \;=\; \arg\min_{\pi_\theta} \ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda). \tag{39}$$

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 seeking/following Bishop (2006). ■

D Proof of Proposition 3.2

Similar to the derivation in Appendix C, we assume the following:

- For all $x \sim \mathcal{D}_x$, where $\mathcal{D}_x$ is an arbitrary prompt distribution, there exists three unique responses y_1, y_2 and y_3 with equal probability $1/3$ under π_{ref};
- μ is the same as π_{ref};
- Preference data $\{y_w, y_l, x\} \sim \mathcal{D}_{tr}$ are sampled according to $z \sim p^*(y_1 \succ y_2|x)$, $\{y_1, y_2\} \sim \pi_{\text{ref}}(y|x), x \sim \mathcal{D}_x$, where $z = \mathbb{I}[y_1 \succ y_2|y_1, y_2, x]$ is a binary indicator variable that determines y_w and y_l assignments;
- 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 $x \in \mathcal{D}_x$.

We begin by considering a specific case of the constraint set:

$$\mathcal{S}_\pi = \left\{\pi_\theta \;:\; \frac{\pi_\theta(y_2|x)}{\pi_\theta(y_1|x)} \geq \frac{\pi^*(y_2|x)}{\pi^*(y_1|x)} + \varepsilon\right\}, \tag{40}$$

where π^* is the optimal unconstrained solution to both the original IPO and its RLHF counterpart; recall that these are equivalent (without constraints) by design when using the reward from (7). The constraint set from (40) includes policies π_θ where the ratio of the probability of generating y_2 to the probability of generating y_1 deviates by at least ε from the corresponding ratio in the optimal policy π^*. It represents a neighborhood around the optimal solution π^*, excluding π^* itself. Note that there is no constraint on $\pi_\theta(y_3|x)$.

RLHF loss processing: Following the derivation in Appendix C, the RLHF loss can be expressed as

$$\ell_{\text{RLHF}}\left(\pi_\theta, \pi_{\text{ref}}, r_{\text{IPO}}, \lambda\right) = \mathbb{E}_{x \sim \mathcal{D}_x}\Big[\mathbb{KL}\big[\pi_\theta(y|x)||\pi^{**}(y|x)\big]\Big], \tag{41}$$

where

$$\pi^{**}(y|x) \;:=\; \frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\left[\frac{1}{\lambda}r_{\text{IPO}}(y, x)\right] = \frac{\exp(p^*(y \succ \pi_{\text{ref}}|x))}{\sum_y \exp(p^*(y \succ \pi_{\text{ref}}|x))}.$$

Therefore, the optimal solution to the RLHF problem satisfies $\pi^* = \pi^{**}$. When considering the constraint case, the resulting RLHF problem can be written as

$$\begin{aligned}
&\min_{\pi_\theta}\; \pi_\theta(y_1|x)\log\left[\frac{\pi_\theta(y_1|x)}{\pi^*(y_1|x)}\right] + \pi_\theta(y_2|x)\log\left[\frac{\pi_\theta(y_2|x)}{\pi^*(y_2|x)}\right] + \pi_\theta(y_3|x)\log\left[\frac{\pi_\theta(y_3|x)}{\pi^*(y_3|x)}\right]\\
&\text{s.t. } \pi_\theta(y_2|x) \geq \left(\frac{\pi^*(y_2|x)}{\pi^*(y_1|x)} + \varepsilon\right)\pi_\theta(y_1|x),\\
&\quad\;\; \pi_\theta(y_1|x) + \pi_\theta(y_2|x) + \pi_\theta(y_3|x) = 1.
\end{aligned} \tag{42}$$

Consider the KKT conditions for problem (42) given by

$$
\begin{bmatrix} \log\left(\frac{\pi_\theta(y_1|x)}{\pi^*(y_1|x)}\right)+1 \\ \log\left(\frac{\pi_\theta(y_2|x)}{\pi^*(y_2|x)}\right)+1 \\ \log\left(\frac{\pi_\theta(y_3|x)}{\pi^*(y_3|x)}\right)+1 \end{bmatrix} + u \begin{bmatrix} \left(\frac{\pi^*(y_2|x)}{\pi^*(y_1|x)}+\varepsilon\right) \\ -1 \\ 0 \end{bmatrix} + v \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = 0, \tag{43}
$$

$$
u\left(\left(\frac{\pi^*(y_2|x)}{\pi^*(y_1|x)}+\varepsilon\right)\pi_\theta(y_1|x) - \pi_\theta(y_2|x)\right) = 0, \tag{44}
$$

$$
u \geq 0,\ \ v \in \mathbb{R}, \tag{45}
$$

$$
\left(\frac{\pi^*(y_2|x)}{\pi^*(y_1|x)}+\varepsilon\right)\pi_\theta(y_1|x) - \pi_\theta(y_2|x) \leq 0, \tag{46}
$$

$$
\pi_\theta(y_1|x) + \pi_\theta(y_2|x) + \pi_\theta(y_3|x) = 1. \tag{47}
$$

We may conclude that $u > 0$. Otherwise, if $u = 0$, by (43) and (47), the solution to the KKT conditions would be π^*, which violates the constraint from (40). Then we have $u > 0$ and the optimal solution to (42) satisfies

$$
\pi_\theta(y_2|x) = \left(\frac{\pi^*(y_2|x)}{\pi^*(y_1|x)}+\varepsilon\right)\pi_\theta(y_1|x). \tag{48}
$$

By substituting (48) into(47), we obtain the ratio of $\pi_\theta(y_1|x)$ to $\pi_\theta(y_3|x)$ as

$$
\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_3|x)} = \frac{\sqrt[\frac{\pi^*(y_2|x)}{\pi^*(y_1|x)}+\varepsilon+1]{\pi^*(y_1|x)\pi^*(y_2|x)^{\frac{\pi^*(y_2|x)}{\pi^*(y_1|x)}+\varepsilon}}}{\pi^*(y_3|x)}. \tag{49}
$$

IPO loss processing: When $\lambda = 1$ and $\pi_{\text{ref}}(y|x)$ is constant, the IPO loss is given by:

$$
\begin{aligned}
\ell_{\text{IPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda) &= \mathbb{E}_{x\sim\mathcal{D}_x}\left[\mathbb{E}_{y,y'\sim\pi_{\text{ref}}}\left[\left(\log\left(\frac{\pi_\theta(y|x)\pi_{\text{ref}}(y'|x)}{\pi_\theta(y'|x)\pi_{\text{ref}}(y|x)}\right) - \frac{p^*(y \succ \pi_{\text{ref}}|x) - p^*(y' \succ \pi_{\text{ref}}|x)}{\lambda}\right)^2\right]\right] \\
&= \mathbb{E}_{x\sim\mathcal{D}_x}\left[\mathbb{E}_{y,y'\sim\pi_{\text{ref}}}\left[\left(\log\left(\frac{\pi_\theta(y|x)}{\pi_\theta(y'|x)}\right) - (p^*(y \succ \pi_{\text{ref}}|x) - p^*(y' \succ \pi_{\text{ref}}|x))\right)^2\right]\right] \\
&= \mathbb{E}_{x\sim\mathcal{D}_x}\left[\frac{2}{9}\sum_{i<j}\left(\log\left(\frac{\pi_\theta(y_i|x)}{\pi_\theta(y_j|x)}\right) - (p^*(y_i \succ \mu|x) - p^*(y_j \succ \mu|x))\right)^2\right] + C,
\end{aligned} \tag{50}
$$

where C is a constant unrelated to π. Therefore, the constrained IPO problem is given as:

$$
\begin{aligned}
&\min_{\pi} \sum_{i<j}\left(\log\left(\frac{\pi_\theta(y_i|x)}{\pi_\theta(y_j|x)}\right) - (p^*(y_i \succ \mu|x) - p^*(y_j \succ \mu|x))\right)^2 \\
&\text{s.t. } \pi_\theta(y_2|x) \geq \left(\frac{\pi^*(y_2|x)}{\pi^*(y_1|x)}+\varepsilon\right)\pi_\theta(y_1|x), \\
&\qquad \pi_\theta(y_1|x) + \pi_\theta(y_2|x) + \pi_\theta(y_3|x) = 1.
\end{aligned} \tag{51}
$$

Similar to the analysis of RLHF loss processing, by examining the KKT conditions for problem (51), we observe that the optimal solution also satisfies (48). Next, we observe that

$$
\log\left(\frac{\pi_\theta(y_2|x)}{\pi_\theta(y_3|x)}\right) - \log\left(\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_3|x)}\right) = \log\left(\frac{\pi_\theta(y_2|x)}{\pi_\theta(y_1|x)}\right) = \log\left(\frac{\pi^*(y_2|x)}{\pi^*(y_1|x)}+\varepsilon\right).
$$

Let v_x denote $\log\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_3|x)}$, then solving (51) is equivalent to solving

$$
\min_{v_x} \left(v_x - (p^*(y_1 \succ \mu|x) - p^*(y_3 \succ \mu|x))\right)^2 + \left(v_x + \log\left(\frac{\pi^*(y_2|x)}{\pi^*(y_1|x)}+\varepsilon\right) - (p^*(y_2 \succ \mu|x) - p^*(y_3 \succ \mu|x))\right)^2 \tag{52}
$$

for every x. The optimal solution to (52) is

$$v_x^* = \frac{p^*(y_1 \succ \mu|x) + p^*(y_2 \succ \mu|x) - 2p^*(y_3 \succ \mu|x)}{2} - \frac{1}{2}\log\left(\frac{\pi^*(y_2|x)}{\pi^*(y_1|x)} + \varepsilon\right).$$

Thus, we obtain the ratio of $\pi(y_1|x)$ to $\pi(y_3|x)$ as

$$\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_3|x)} = \left(\frac{\pi^*(y_2|x)}{\pi^*(y_1|x)} + \varepsilon\right)^{-\frac{1}{2}} \exp\left(\frac{p^*(y_1 \succ \mu|x) + p^*(y_2 \succ \mu|x) - 2p^*(y_3 \succ \mu|x)}{2}\right). \tag{53}$$

Final Step: Comparing (49) and (53), we can see that

$$\arg\min_{\pi_\theta \in \mathcal{S}_\pi} \ell_{\mathrm{RLHF}}\left(\pi_\theta, \pi_{\mathrm{ref}}, r_{\mathrm{IPO}}, \lambda\right) \neq \arg\min_{\pi_\theta \in \mathcal{S}_\pi} \ell_{\mathrm{IPO}}(\pi_\theta, \pi_{\mathrm{ref}}, \lambda).$$

■

E Proof of Proposition 4.1

For an arbitrary real vector v we have that

$$\arg\min_{\gamma > 0} -\log \mathcal{N}(v|0, \gamma I) \equiv \arg\min_{\gamma > 0}\left[\frac{v^\top v}{\gamma} + \log|\gamma I|\right] = \frac{1}{2} v^\top v. \tag{54}$$

And therefore, we have

$$\min_{\gamma > 0} -\log \mathcal{N}(v|0, \gamma I) \equiv \log(v^\top v) \tag{55}$$

excluding irrelevant constants. Returning to (16), if we first optimize over $\gamma(y_w, y_l, x)$ for each tuple, we obtain the loss factor

$$\log\left[\xi_{\mathrm{ref}}(y_w, y_l, x)^2 + \xi_\theta(y_w, y_l, x)^2\right] = \log\left[\mu\left[\frac{\pi_\theta(y_l|x)}{\pi_\theta(y_w|x)}\right]^2 + \mu\left[\frac{\pi_{\mathrm{ref}}(y_l|x)}{\pi_{\mathrm{ref}}(y_w|x)}\right]^2\right]. \tag{56}$$

From here, by choosing $\mu(\cdot) = (\cdot)^{\frac{\lambda}{2}}$ we can modify (56) as

$$\begin{aligned} \log\left[\frac{\pi_\theta(y_l|x)^\lambda}{\pi_\theta(y_w|x)^\lambda} + \frac{\pi_{\mathrm{ref}}(y_l|x)^\lambda}{\pi_{\mathrm{ref}}(y_w|x)^\lambda}\right] &= \log\left[1 + \left(\frac{\pi_\theta(y_l|x)}{\pi_{\mathrm{ref}}(y_l|x)}\right)^\lambda \left(\frac{\pi_{\mathrm{ref}}(y_w|x)}{\pi_\theta(y_w|x)}\right)^\lambda\right] + C \\ &\equiv -\log\sigma\left(\lambda\log\frac{\pi_\theta(y_w|x)}{\pi_{\mathrm{ref}}(y_w|x)} - \lambda\log\frac{\pi_\theta(y_l|x)}{\pi_{\mathrm{ref}}(y_l|x)}\right), \end{aligned} \tag{57}$$

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 (6). From here it directly follows that minimizing (57) over $\pi_\theta \in \mathcal{S}_\pi$ is equivalent to $\min_{\pi_\theta \in \mathcal{S}_\pi} \ell_{\mathrm{DPO}}(\pi_\theta, \pi_{\mathrm{ref}}, \lambda)$. ■