Unavoidable Learning Constraints Alter the Foundations of Direct Preference Optimization

David Wipf [1]

Abstract

Large language models in the past have typically relied on some form of reinforcement learning with human feedback (RLHF) to better align model responses with human preferences. However, because of oft-observed instabilities when implementing these RLHF pipelines, various reparameterization techniques have recently been introduced to sidestep the need for separately learning an RL reward model. Instead, directly fine-tuning for human preferences is achieved via the minimization of a single closed-form training objective, a process originally referred to as direct preference optimization (DPO). Although effective in certain real-world settings, we detail how the foundational DPO reparameterization no longer holds 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 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., 2009; Bai et al., 2022a; Ouyang et al., 2022). This multi-step process involves first learning a reward model that reflects human inclinations culled from labeled preference data, and then subsequently training a new policy that balances reward maximization with proximity to $\pi_{\text{ref}}(y|x)$.

Because RLHF introduces additional complexity, computational overhead, and entry points for instability, clever reparameterization techniques have recently been proposed that sidestep the need for separately learning a reward model altogether. Instead, increased alignment with human preferences is achieved via the minimization of a single closed-form training objective, a process originally referred to as direct preference optimization (DPO) (Rafailov et al., 2024) followed by several notable descendants and generalizations (Azar et al., 2024; Tang et al., 2024; Wang et al., 2024; Zhao et al., 2023b). 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.

In particular, we prove that once inevitable model/learning constraints are introduced during training (explicitly or implicitly, e.g., early-stopping, weight decay, etc.), the core reparameterizations that underpin DPO models no longer strictly hold (Section 3). This then motivates alternative DPO derivations and supporting analyses that are not beholden to the impact of such constraints. We provide two such examples herein: (i) The re-derivation of DPO from a simple Gaussian estimation perspective independent of RLHF and attendant reparameterizations (Section 4); and (ii) The contextualization of DPO as an example of classical constrained optimization involving noise-adaptive, concave regularization (Section 5).

2. Background

We adopt $x \sim \mathcal{D}_x$ to denote an *input prompt* x drawn from some distribution $\mathcal{D}_x$. From here, conditioned on such prompts we may then generate *responses* y using, for example, a pre-trained reference language model/policy $\pi_{\text{ref}}(y|x)$. Moreover, given a pair of such responses $y_1 \neq y_2$, we adopt $y_1 \succ y_2$ to convey the notion that a human evaluator prefers y_1 over y_2. Given a population of such evaluations, we express the corresponding ground-truth human preference distribution as $p^*(y_1 \succ y_2|y_1, y_2, x)$. And finally, we define a set of human labeled tuples drawn from a training distribution $\mathcal{D}_{tr}$ as $\{y_w, y_l, x\} \sim \mathcal{D}_{tr}$, where $y_w \succ y_l$; subscripts here stand for 'win' and 'lose'.

[1]Amazon Web Services. Email: davidwipf@gmail.com.

Work presented at TF2M workshop at ICML 2024, Vienna, Austria. PMLR 235, 2024.

2.1. Reinforcement Learning with Human Feedback

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

$$p^*(y_1 \succ y_2|x) = \frac{\exp[r^*(y_1,x)]}{\exp[r^*(y_1,x)]+\exp[r^*(y_2,x)]} = \sigma\big[r^*(y_1,x)-r^*(y_2,x)\big], \quad (1)$$

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

$$\ell_{\text{BT}}(r_\phi) := \mathbb{E}_{\{y_w,y_l,x\}\sim\mathcal{D}_{\text{tr}}}\Big[-\log p_\phi(y_w \succ y_l|x)\Big] \quad (2)$$
$$= \mathbb{E}_{\{y_w,y_l,x\}\sim\mathcal{D}_{\text{tr}}}\Big[-\log\sigma\big[r_\phi(y_w,x)-r_\phi(y_l,x)\big]\Big].$$

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

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

$$\ell_{\text{RLHF}}(\pi_\theta,\pi_{\text{ref}},\hat{r}_\phi,\lambda) := \mathbb{E}_{y\sim\pi_\theta(y|x),x\sim\mathcal{D}_x}\Big[-\hat{r}_\phi(y,x)\Big] + \lambda\,\mathbb{E}_{x\sim\mathcal{D}_x}\Big[\mathbb{KL}\big[\pi_\theta(y|x)||\pi_{\text{ref}}(y|x)\big]\Big], \quad (3)$$

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

2.2. Direct Preference Optimization (DPO)

Consider now the reward-dependent RLHF loss ℓ_{RLHF} from (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 pivotal assumption in Section 3), the minimum of $\ell_{\text{RLHF}}(\pi_\theta,\pi_{\text{ref}},r,\lambda)$ w.r.t. π_θ can be directly computed as

$$\pi_r(y|x) := \arg\min_{\pi_\theta}\ell_{\text{RLHF}}(\pi_\theta,\pi_{\text{ref}},r,\lambda) = \frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\left[\frac{1}{\lambda}r(y,x)\right], \quad (4)$$

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

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

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 $\pi^{**}(y|x) := \arg\min_{\pi_\theta}\ell_{\text{RLHF}}(\pi_\theta,\pi_{\text{ref}},r^*,\lambda)$. 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

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

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.

3. Impact of Optimization Constraints

It has been previously shown that minimizing the DPO loss $\ell_{\text{DPO}}(\pi_\theta,\pi_{\text{ref}},\lambda)$ is effectively the same as minimizing the RLHF loss $\ell_{\text{RLHF}}(\pi_\theta,\pi_{\text{ref}},r^*,\lambda)$ with optimal reward model r^* (Rafailov et al., 2024). But there is a pivotal assumption underlying this association which previous analysis has not rigorously accounted for. Specifically, the key equality that facilitates the DPO reparameterization, namely (5), is predicated on the solution of an *uncononstrained* optimization problem from (4) over an arbitrary policy π_θ.

However, when actually training models in real-world settings, constraints will always exist, whether implicitly or explicitly. Such constraints stem from any number of factors including the model architecture/capacity limitations, 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 countermeasures like early stopping are imposed that effectively introduce a $\mathcal{S}_\pi$ and substantially alter 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 (as assumed by

DPO). 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. Consider now the following:

Proposition 3.1. *Let $\mathcal{S}_\pi$ denote a constraint set on the learnable policy π_θ. Then we can have that*

$$\arg\min_{\pi_\theta \in \mathcal{S}_\pi} \ell_{RLHF}(\pi_\theta, \pi_{ref}, r^*, \lambda) \neq \arg\min_{\pi_\theta \in \mathcal{S}_\pi} \ell_{DPO}(\pi_\theta, \pi_{ref}, \lambda). \tag{7}$$

As can be observed by the proof in Appendix C, 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 is generally quite distinct (Bishop, 2006); see also Figure 1 in Appendix A.

Consequently, once a constraint is introduced and the inequality from (7) activated, *we can no longer say that DPO provides an optimal implicit reward for the original RLHF problem*, i.e., the original connection is now ambiguous. And so the value of DPO in practice (and indeed it often does work well) cannot be unreservedly attributed to its motivational affiliation with an optimal RLHF solution, and instead, should be evaluated based on properties of $\min_{\pi_\theta \in \mathcal{S}_\pi} \ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda)$ itself. We take two steps in this direction as follows:

1. In Section 4 we rederive the DPO loss 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, 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 the reference policy performance.

4. Rederiving DPO from Scratch Via a Naive Gaussian Estimation Perspective

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

$$p^*(y_1 \succ y_2|x) = \mu\left[\frac{\pi^*(y_2|x)}{\pi^*(y_1|x)}\right], \tag{8}$$

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 (8) (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 functions $\xi_\theta(y_1, y_2, x) :=$

$$\mu\left[\frac{\pi_\theta(y_2|x)}{\pi_\theta(y_1|x)}\right], \quad \xi_{\text{ref}}(y_1, y_2, x) := \mu\left[\frac{\pi_{\text{ref}}(y_2|x)}{\pi_{\text{ref}}(y_1|x)}\right]. \tag{9}$$

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

$$= \mathcal{N}\left(\begin{bmatrix} \xi_\theta(y_1, y_2, x) \\ \xi_{\text{ref}}(y_1, y_2, x) \end{bmatrix} \middle| 0, \gamma(y_1, y_2, x) I\right), \tag{10}$$

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

$$\min_{\pi_\theta \in \mathcal{S}_\pi,\ \{\gamma(y_w, y_l, x) > 0\}} \Bigg\{ \mathbb{E}_{\{y_w, y_l, x\} \sim \mathcal{D}_{\text{tr}}} - \log \mathcal{N}\left(\begin{bmatrix} \xi_\theta(y_w, y_l, x) \\ \xi_{\text{ref}}(y_w, y_l, x) \end{bmatrix} \middle|\ 0, \gamma(y_w, y_l, x) I\right) \Bigg\}, \tag{11}$$

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 (11) 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 $\min_{\pi_\theta \in \mathcal{S}_\pi} \ell_{DPO}(\pi_\theta, \pi_{ref}, \lambda)$.*

5. The DPO Loss Induces Noise Adaptive Regularization

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

$$\ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda) \equiv \tag{12}$$
$$\mathbb{E}_{\{y_w, y_l, x\}\sim\mathcal{D}_{\text{tr}}}\left[\log\left(\left[\frac{\pi_{\text{ref}}(y_l|x)}{\pi_{\text{ref}}(y_w|x)}\right]^\lambda + \left[\frac{\pi_\theta(y_l|x)}{\pi_\theta(y_w|x)}\right]^\lambda\right)\right],$$

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

$$g(u_2) \leq g(u_1) + \frac{g'(u_1)}{f'(u_1)}\left[f(u_2) - f(u_1)\right], \tag{13}$$

where g' and f' denote the respective derivatives.

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

- $\log(\gamma + u)$ is a concave non-decreasing function of $u \in [0, \infty)$, which represents a well-known characteristic of sparsity-favoring penalty factors commonly used in robust estimation (Chartrand & Yin, 2008; Chen et al., 2017; Fan & Li, 2001; Rao et al., 2003).[1] 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 2 in Appendix B illustrates this phenomena by contrasting with two extremes producing the convex ℓ_1 norm and the non-convex ℓ_0 norm.

- Prior work (Candes et al., 2008; Wipf & Nagarajan, 2010) has investigated general optimization problems of the form
$$\min_{\{u_i\}\in\mathcal{S}_u} \sum_i \log(\gamma + |u_i|), \tag{14}$$
sometimes generalized to $\min_{\{u_i\}\in\mathcal{S}_u} \sum_i f(|u_i|, \gamma)$ over a concave, non-decreasing function f of $|u_i|$, where S_u is some constraint set.[2] Moreover, γ reflects a noise parameter or an analogous measure of uncertainty, with relative concavity dictated by γ as above. In these contexts, it has been argued that adjusting the curvature of the regularization factor based on noise levels can provide additional robustness to bad local minima and high noise regimes (Candes et al., 2008; Dai et al., 2018; Wipf & Zhang, 2014). The basic intuition here is that when noise is high, a more convex shape is preferable, while when the noise is low, a more concave alternative may be appropriate.

- Regarding DPO, it is natural to treat $\left[\frac{\pi_{\text{ref}}(y_l|x)}{\pi_{\text{ref}}(y_w|x)}\right]^\lambda$ as an analogous noise factor, given that whenever this ratio is large, it implies that our reference policy is poor. Hence, once we introduce a constraint $\mathcal{S}_\pi$ on π_θ (as will always occur in practice; see Section 3), 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 (14), involving a robust regularization factor with noise-adaptive curvature.

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.

We close by remarking that, although our focus was largely on the original DPO formulation (Rafailov et al., 2024), it nonetheless remains relevant to the foundations of follow-up work that relies on analogous DPO-like reparameterizations; recently published examples include (Azar et al., 2024; Wang et al., 2024). In these and other cases, constraints can obfuscate the degree to which model behavior can be directly traced back to a RLHF-based loss archetype.

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

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

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

dacher, N., Such, F. P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M. B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J. F. C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J. J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., and Zoph, B. Gpt-4 technical report, 2024.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A. Visualization of Constraint Impact

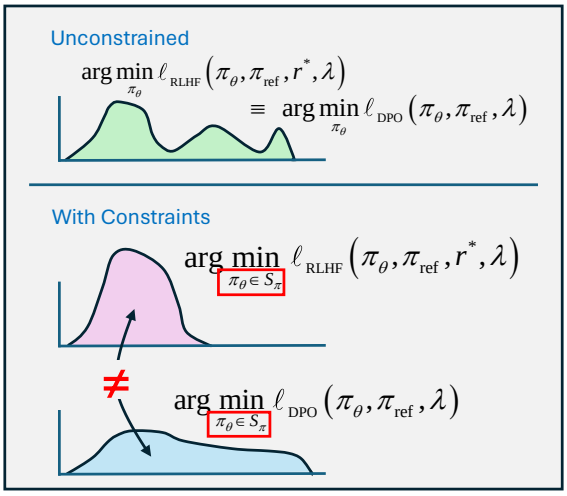

Figure 1: Visualization of learned policies with and without constraints as discussed in Section 3. On the top, no constraints are present and minimizing the respective DPO and RLHF losses leads to the same policy (green distribution over responses). In contrast, when the minimization is restricted to 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 optimal reward r^*, so the discrepancy is entirely a consequence of policy constraints, not sub-optimal reward usage.

B. Visualization of Different Penalty Factors

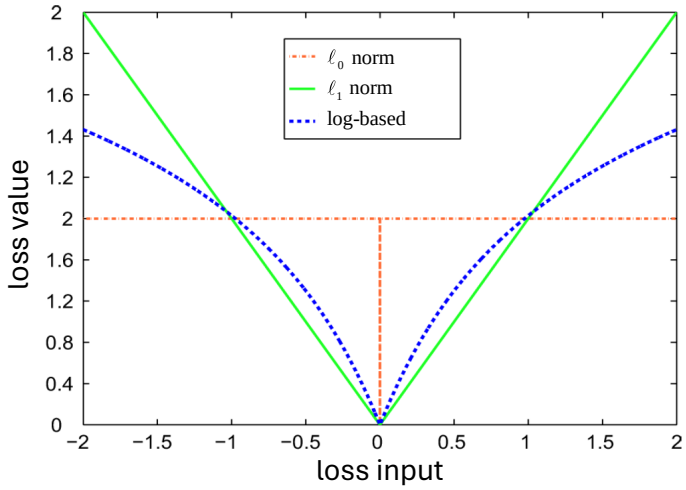

Figure 2: Visualization of different penalty factors associated with the DPO loss as discussed in Section 5. When $\gamma \to 0$, $\log(\gamma + |u|) \to \log|u| = \lim_{p\to 0} \frac{1}{p}[|u|^p - 1] \propto \mathbb{I}[u \neq 0]$ mimicking an ℓ_0 norm (red curve) w.r.t. relative concavity (if $u \geq 0$ 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 value, the relative concavity of $\log(\gamma + |u|)$ lies within these two extremes.

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;[3]
- 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

$$\begin{aligned} p^*(y_1 \succ y_2|x) &= \frac{\exp[r^*(y_1, x)]}{\exp[r^*(y_1, x)] + \exp[r^*(y_2, x)]} = \frac{\frac{\exp[r^*(y_1,x)]}{Z(x)}}{\frac{\exp[r^*(y_1,x)]}{Z(x)} + \frac{\exp[r^*(y_2,x)]}{Z(x)}} \\ &= \frac{\pi^*(y_1|x)}{\pi^*(y_1|x) + \pi^*(y_2|x)}, \end{aligned} \tag{15}$$

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

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

$$\begin{aligned} \ell_{\text{RLHF}}\left(\pi_\theta, \pi_{\text{ref}}, r^*, \lambda\right) &= \mathbb{E}_{y\sim\pi_\theta(y|x), x\sim\mathcal{D}_x}\Big[-r^*(y, x)\Big] + \lambda\, \mathbb{E}_{x\sim\mathcal{D}_x}\Big[\mathbb{KL}\big[\pi_\theta(y|x)||\pi_{\text{ref}}(y|x)\big]\Big] \\ &\equiv \mathbb{E}_{x\sim\mathcal{D}_x}\Big[\mathbb{KL}\big[\pi_\theta(y|x)||\pi^{**}(y|x)\big]\Big], \end{aligned} \tag{16}$$

where

$$\pi^{**}(y|x) := \frac{1}{Z(x)}\pi_{\text{ref}}(y|x)\exp\left[\frac{1}{\lambda}r^*(y, x)\right]. \tag{17}$$

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

$$\pi^{**}(y|x) = \frac{\exp\left[r^*(y, x)\right]}{\sum_y \exp\left[r^*(y, x)\right]}, \tag{18}$$

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

$$\frac{\pi^*(y_1|x)}{\pi^*(y_1|x) + \pi^*(y_2|x)} = p^*(y_1 \succ y_2|x) = \frac{\exp\left[r^*(y_1, x)\right]}{\exp\left[r^*(y_1, x)\right] + \exp\left[r^*(y_2, x)\right]}, \tag{19}$$

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

$$\ell_{\text{RLHF}}\left(\pi_\theta, \pi_{\text{ref}}, r^*, \lambda\right) = \mathbb{E}_{x\sim\mathcal{D}_x}\Big[\mathbb{KL}\big[\pi_\theta(y|x)||\pi^*(y|x)\big]\Big] \tag{20}$$

under the stated conditions.

[3] 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.

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

$$\begin{aligned}\ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda) &= \mathbb{E}_{\{y_w, y_l, x\}\sim\mathcal{D}_{\text{tr}}}\left[-\log\sigma\left(\lambda\log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \lambda\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)\right] \\ &= \mathbb{E}_{\{y_w, y_l, x\}\sim\mathcal{D}_{tr}}\left[\log\left(\frac{\pi_\theta(y_w|x)+\pi_\theta(y_l|x)}{\pi_\theta(y_w|x)}\right)\right]. \end{aligned} \tag{21}$$

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

$$\begin{aligned}\ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda) &= \mathbb{E}_{\{y_w, y_l, x\}\sim\mathcal{D}_{tr}}\left[\log\left(\frac{1}{\pi_\theta(y_w|x)}\right)\right] \\ &= \mathbb{E}_{x\sim\mathcal{D}_x}\left[p^*(z=1|y_1,y_2,x)\log\left(\frac{1}{\pi_\theta(y_1|x)}\right)\right. \\ &\qquad \left. + (p^*(z=0|y_1,y_2,x)\log\left(\frac{1}{\pi_\theta(y_2|x)}\right)\right] \\ &= \mathbb{E}_{x\sim\mathcal{D}_x}\left[\pi^*(y_1|x)\log\left(\frac{1}{\pi_\theta(y_1|x)}\right)\right. \\ &\qquad \left. + \pi^*(y_2|x)\log\left(\frac{1}{\pi_\theta(y_2|x)}\right)\right] \\ &= \mathbb{E}_{x\sim\mathcal{D}_x}\left[\pi^*(y_1|x)\log\left(\frac{\pi^*(y_1|x)}{\pi_\theta(y_1|x)}\right)\right. \\ &\qquad \left. + \pi^*(y_2|x)\log\left(\frac{\pi^*(y_2|x)}{\pi_\theta(y_2|x)}\right)\right] + C \\ &\equiv \mathbb{E}_{x\sim\mathcal{D}_x}\Big[\mathbb{KL}\big[\pi^*(y|x)||\pi_\theta(y|x)\big]\Big], \end{aligned} \tag{22}$$

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 (20) and (22) 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.,

$$\arg\min_{\pi_\theta} \ell_{\text{RLHF}}(\pi_\theta, \pi_{\text{ref}}, r^*, \lambda) = \arg\min_{\pi_\theta} \ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda). \tag{23}$$

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

D. Proof of Proposition 4.1

For an arbitrary real vector v we have that

$$\arg\min_{\gamma>0} -\log\mathcal{N}(v|0,\gamma I) \equiv \arg\min_{\gamma>0}\left[\frac{v^\top v}{\gamma} + \log|\gamma I|\right] = \frac{1}{2}v^\top v. \tag{24}$$

And therefore, we have

$$\min_{\gamma>0} -\log\mathcal{N}(v|0,\gamma I) \equiv \log(v^\top v) \tag{25}$$

excluding irrelevant constants. Returning to (11), if we first optimize over $\gamma(y_w, y_l, x)$ for each tuple, we obtain the loss factor

$$\log\left[\xi_{\text{ref}}(y_w,y_l,x)^2 + \xi_\theta(y_w,y_l,x)^2\right] = \log\left[\mu\left[\frac{\pi_\theta(y_l|x)}{\pi_\theta(y_w|x)}\right]^2 + \mu\left[\frac{\pi_{\text{ref}}(y_l|x)}{\pi_{\text{ref}}(y_w|x)}\right]^2\right]. \tag{26}$$

From here, by choosing $\mu(\cdot) = (\cdot)^{\frac{\lambda}{2}}$ we can modify (26) as

$$\begin{aligned} \log\left[\frac{\pi_\theta(y_l|x)^\lambda}{\pi_\theta(y_w|x)^\lambda} + \frac{\pi_{\text{ref}}(y_l|x)^\lambda}{\pi_{\text{ref}}(y_w|x)^\lambda}\right] &= \log\left[1 + \left(\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right)^\lambda \left(\frac{\pi_{\text{ref}}(y_w|x)}{\pi_\theta(y_w|x)}\right)^\lambda\right] + C \\ &\equiv -\log\sigma\left(\lambda \log\frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)} - \lambda\log\frac{\pi_\theta(y_l|x)}{\pi_{\text{ref}}(y_l|x)}\right), \end{aligned} \tag{27}$$

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 (27) over $\pi_\theta \in \mathcal{S}_\pi$ is equivalent to $\min_{\pi_\theta \in \mathcal{S}_\pi} \ell_{\text{DPO}}(\pi_\theta, \pi_{\text{ref}}, \lambda)$. ■