A note on DPO with noisy preferences & relationship to IPO

Eric Mitchell

November 25, 2023 (v1.1)

'OG' RLHF aims for reward maximization with a KL constraint to reference model π_{ref} (inputs x omitted):

$$\pi^* = \underset{\pi}{\text{argmax}}\, \mathbb{E}_{y\sim\pi}\left[r(y) - \beta \log \frac{\pi(y)}{\pi_{\text{ref}}(y)}\right] \tag{1}$$

DPO [3] derives a loss on the current policy π_θ (where our dataset says y_w is preferred to y_l, or $y_w \succ y_l$):

$$\mathcal{L}_{\text{DPO}}(\theta, y_w, y_l) = -\log\sigma\left(\beta\log\frac{\pi_\theta(y_w)}{\pi_{\text{ref}}(y_w)} - \beta\log\frac{\pi_\theta(y_l)}{\pi_{\text{ref}}(y_l)}\right), \tag{2}$$

i.e., the binary cross entropy with $\hat{p}_\theta(y_w \succ y_l) = \sigma\left(\beta\log\frac{\pi_\theta(y_w)}{\pi_{\text{ref}}(y_w)} - \beta\log\frac{\pi_\theta(y_l)}{\pi_{\text{ref}}(y_l)}\right)$ and target $p(y_w \succ y_l) = 1$.

What if preference labels are noisy? Say the labels have been flipped with some small probability $\epsilon \in (0, 0.5)$. We can use a *conservative* target distribution instead, $p(y_w \succ y_l) = 1 - \epsilon$, giving BCE loss:

$$\begin{aligned}
\mathcal{L}^{\epsilon}_{\text{DPO}}(\theta, y_w, y_l) &= -(1-\epsilon)\log\hat{p}_\theta(y_w \succ y_l) - \epsilon\log(1 - \hat{p}_\theta(y_w \succ y_l)) &(3)\\
&= (1-\epsilon)\mathcal{L}_{\text{DPO}}(\theta, y_w, y_l) + \epsilon\mathcal{L}_{\text{DPO}}(\theta, y_l, y_w) &(4)
\end{aligned}$$

The gradient of $\mathcal{L}^{\epsilon}_{\text{DPO}}(\theta, y_w, y_l)$ is simply the weighted sum of gradients $(1-\epsilon)\nabla_\theta\mathcal{L}(\theta, y_w, y_l) + \epsilon\nabla_\theta\mathcal{L}(\theta, y_l, y_w)$, which reduces to the simplified form (ignoring constants; see [3] for the gradient of the original DPO loss):

$$\begin{aligned}
\nabla_\theta\mathcal{L}^{\epsilon}_{\text{DPO}}(\theta, y_w, y_l) &= -\Big((1-\epsilon)(1-\hat{p}_\theta) - \epsilon\hat{p}_\theta\Big)\Big[\underbrace{\nabla_\theta\log\pi_\theta(y_w)}_{\text{upweight } y_w} - \underbrace{\nabla_\theta\log\pi_\theta(y_l)}_{\text{downweight } y_l}\Big] &(5)\\
&= \Big(\hat{p}_\theta - (1-\epsilon)\Big)\Big[\nabla_\theta\log\pi_\theta(y_w) - \nabla_\theta\log\pi_\theta(y_l)\Big]. &(6)
\end{aligned}$$

The gradient is zero when $\hat{p}_\theta(y_w \succ y_l) = (1-\epsilon)$, i.e., our (implicit) reward assigns the desired confidence level in this training example under the Bradley-Terry model [2]. For normal DPO, **the gradient is never zero!** Using the shorthand $h^{y_w,y_l}_{\pi_\theta} = \log\frac{\pi_\theta(y_w)}{\pi_{\text{ref}}(y_w)} - \log\frac{\pi_\theta(y_l)}{\pi_{\text{ref}}(y_l)}$, let's compare the conservative DPO (cDPO?) and IPO [1] loss gradient, where the IPO loss is given in Eq. 17 of [1] as $\mathcal{L}_{\text{IPO}}(\theta, y_w, y_l) = \left(h^{y_w,y_l}_{\pi} - \frac{1}{2\beta}\right)^2$:

$$\begin{aligned}
\nabla_\theta\mathcal{L}_{\text{IPO}}(\theta, y_w, y_l) &= \left(h^{y_w,y_l}_{\pi_\theta} - \frac{1}{2\beta}\right)\Big[\nabla_\theta\log\pi_\theta(y_w) - \nabla_\theta\log\pi_\theta(y_l)\Big] &(7)\\
\nabla_\theta\mathcal{L}^{\epsilon}_{\text{DPO}}(\theta, y_w, y_l) &= \Big(\sigma(\beta h^{y_w,y_l}_{\pi_\theta}) - (1-\epsilon)\Big)\Big[\nabla_\theta\log\pi_\theta(y_w) - \nabla_\theta\log\pi_\theta(y_l)\Big] &(8)
\end{aligned}$$

TL;DR: conservative DPO trains the model until a desired improvement in the *implicit probability assigned by the model to the observed preferences*[1] is met; IPO trains the model until a desired improvement in *implicit reward* is met. The ability for cDPO and IPO to optimize *only to a fixed delta from the reference model* and then stop (or even reverse!) likely makes these more stable than the original DPO loss after lots of training.

[1] Mohammad Gheshlaghi Azar, Mark Rowland, Bilal Piot, Daniel Guo, Daniele Calandriello, Michal Valko, and Rémi Munos. *A General Theoretical Paradigm to Understand Learning from Human Preferences.* 2023. arXiv: 2310.12036 [cs.AI].

[2] Ralph Allan Bradley and Milton E. Terry. "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons". In: *Biometrika* 39.3/4 (1952), pp. 324–345. DOI: https://doi.org/10.2307/2334029.

[3] 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". In: *Neural Information Processing Systems.* 2023.

[1] The Bradley-Terry model of human preferences [2] converts the β-scaled reward gap $h^{y_w,y_l}_{\pi_\theta}$ to a probability assigned by the model to the observed preference bit using the sigmoid of the scaled reward gap.