DPO Unchained: Your Training Algorithm is Secretly Disentangled in Human Choice Theory (and its Loss' Convexity is Dispensable)

Google Research · Google DeepMind · Université de Montpellier · cnrs · CORE AUTOMATION

Wenxuan Zhou · Shujian Zhang · Brice Magdalou · John Lambert · Ehsan Amid · Richard Nock · Andrew Hard

Summary

- Normative properties are guardrails, *often hidden*, safeguarding training (e.g. losses)
- Direct Preference Optimization (DPO) simplified Reinforcement Learning from Human Feedback (RLHF) by exploiting a model of Human Choice with ~ 0 leeway
- Followers either stayed close to DPO or abandoned normative safeguards
- We show how go "far" away while keeping normative guardrails, + surprises as bonus

states / context / prompts: $x \in \mathcal{X}$; actions / alternatives / answers: $y \in \mathcal{Y}$; **policy**: $\pi : \mathcal{X} \to \Delta_{\mathcal{Y}}$ (simplex)

The data we have access to

$\sim \mathcal{D}$

From the data, can learn choice probabilities → loss

$$I = \mathbb{E}_{(x,y,y')\sim\mathcal{D}}\left[\ell^{①}(p(y \succ y'|x))\right]$$

Model choice probabilities using rewards

$$p(y \succ y'|x) = {}^{②}f(r(x,y), r(x,y'))$$

RLHF | DPO

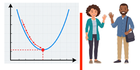
Learn rewards vs **policy** | Model rewards using **policy**

$$r_\omega(x,y) \overset{③}{\updownarrow} \pi_\theta(y|x) \qquad r(x,y) = {}^{③}h(\pi_\theta(y|x))$$

The output we seek: a **policy**

DPO's Normative components 101

- ①,②,③ carve normative components from two theories: loss function & human choice



① ③ Savage

- Loss $L(\boldsymbol{p}, \boldsymbol{q}) : \Delta_n \times \Delta_n \to \mathbb{R}$
- **Analytic form**: $L(\boldsymbol{p}, \boldsymbol{q}) \doteq \boldsymbol{p}^\top \boldsymbol{\ell}(\boldsymbol{q}) = \mathbb{E}_{i\sim \mathrm{CAT}(\boldsymbol{p})}[\ell_i(\boldsymbol{q})]$
  $\boldsymbol{p}$-proper iff $L(\boldsymbol{p}, \boldsymbol{q}) \geqslant L(\boldsymbol{p}, \boldsymbol{p}), \forall \boldsymbol{q} \in \Delta_n$, proper iff $\boldsymbol{p}$-proper for all $\boldsymbol{p}$
- Used as is in ① ; in ③ folds properness in
  $\pi = \arg\max_{\pi'} \mathbb{E}_{\pi'}[r] - R(\pi'\|\pi_{\mathrm{ref}})$
  where $R(\pi'\|\pi_{\mathrm{ref}})$ is the regret of a proper loss
- **Infinite** #choices for ① ③ ○○○○○○○○ Vs ○ **One** choice for ②

DPO: $\ell_i(\boldsymbol{q}) \doteq -\log q_i, \forall i$ (log-loss)

② Bradley-Terry-Luce

- Choice probability $p(y \succ y'|x)$
- **Analytic form**: $u(x,y) = r(x,y) + \varepsilon, \quad \varepsilon \sim \mathrm{Gumbel}$ (u: utility)
- No abstention:
  $p(y \succ y'|x) + p(y' \succ y|x) = 1$
- gives
  $p(y \succ y'|x) \doteq p(u(x,y) > u(x,y')) = \sigma(r(y,x) - r(y',x))$

sigmoid: $\sigma(z) \doteq 1/(1+\exp(-z))$

Final loss
$$I(\theta) = \mathbb{E}_{(x,y,y')\sim\mathcal{D}}\left[\psi\left(-\left(\log\frac{\pi_\theta(y|x)}{\pi_{\mathrm{ref}}(y|x)} - \log\frac{\pi_\theta(y'|x)}{\pi_{\mathrm{ref}}(y'|x)}\right)\right)\right]$$
logistic loss: $\psi(-z) \doteq \log(1+\exp(-z))$ (reference (pretrained) policy)

CAUTION: The final loss "connects" **2 ≠ normative components**

$\psi = \ell_0 \circ \sigma$ (Savage ∘ BTL)

Our initial objective: broaden human choice part so it matches the ML freedom of Savage's properness
Our contribution: objective achieved + a few ML surprises along the way

Getting above **BTL** and on par with **Savage**: **KLST***

- The theory exists to get above BTL: Doignon-Falmagne (1974) simplified by Krantz-Luce-Suppes-Tversky (1989) **but** relies on an axiom, solvability, imposing $|\mathcal{Y}| = \infty$
- We bypass the use of solvability via a mechanism compatible with sampling: (binary) **lotteries** (lottery)

$\mathcal{Y}^\alpha \doteq \{(yy')_\alpha : y, y' \in \mathcal{Y}\}$ → lottery $(y_1 y_2)_\alpha$ → choice proba: $p((y_1y_2)_\alpha \succ L|x) \doteq \alpha \cdot p(y_1 \succ L|x) + (1-\alpha)\cdot p(y_2 \succ L|x)$

- Our model, **KLTS*** relies on three building blocks: *expandability* (lotteries), *local choice structure* and *monotonicity*
- *Local Choice Structure* (LCS): for any $\alpha \in (0,1)$, graph $G_x^\alpha \doteq (\mathcal{Y}^\alpha, \mathcal{E}_x^\alpha)$

Zero-Abstention (ZA) encoded in edges: $(L, L') \in \mathcal{E}_x^\alpha \Leftrightarrow p(L \succ L'|x) + p(L' \succ L|x) = 1$

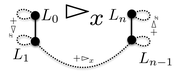
Preference (P) "$\rhd_x$" = choice proba ≥ 1/2: $L \rhd_x L' \Leftrightarrow p(L \succ L'|x) \geqslant 1/2$

- LCS imposes three invariants in graph

Bearability: $(L, L) \in \mathcal{E}_x^\alpha, \forall L \in \mathcal{Y}^\alpha$

ZA + P → ZA: for any wedge $\{L_2, L_1, L_3\}$ of G_x^α
$(L_1 \rhd_x L_2) \wedge (L_1 \rhd_x L_3)$ or $(L_2 \rhd_x L_1) \wedge (L_3 \rhd_x L_1)$ $\Rightarrow (L_2, L_3) \in \mathcal{E}_x^\alpha$

P → ZA + P: whenever $L \rhd_x L'$, there exists a path in G_x^α $(L = L_0, L_1), (L_1, L_2), ..., (L_{n-1}, L_n = L')$ such that $L_i \rhd_x L_{i+1}, \forall i \in [n-1]$



- *Monotonicity*

$\exists \alpha \in (0,1) : \forall L_1, L_2, L_3, L_4, L_5, L_6 \in \mathcal{Y}^\alpha$ such that $\{L_1, L_2, L_3\}$ triangle, $\{L_4, L_5, L_6\}$ wedge, $p(L_1 \succ L_2|x) \geqslant p(L_4 \succ L_5|x)$ and $p(L_2 \succ L_3|x) \geqslant p(L_5 \succ L_6|x)$, then

$p(L_1 \succ L_3|x) \geqslant p(L_4 \succ L_6|x)$ and $(L_4, L_6) \in \mathcal{E}_x^\alpha$

- Fundamental: implicitly creates utilities / rewards
- Analogy:

$z_1 - z_2 \geqslant z_4 - z_5$
$z_2 - z_3 \geqslant z_5 - z_6$
$\Rightarrow \quad z_1 - z_3 \geqslant z_4 - z_6$

Theorems+

Extension of **BTL** to match **Savage**

- Assume choice probabilities have a **KLTS*** structure. Then there exists
  - A strictly ↗ function $F : \mathbb{R} \to [0,1]$ such that $F(-z) + F(z) \leqslant 1, \forall z$
  - A function $u(x,y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$

  such that $p(y \succ y'|x) = F(u(x,y) - u(x,y'))$

Working the ML match with **Savage**

- For ***any*** strictly ↗ function $\psi : \mathbb{R} \to \mathbb{R}$, and ***any*** strictly ↗ function $F : \mathbb{R} \to [0,1]$, there exists a strictly proper loss (ℓ_0, ℓ_1) such that (can trivially switch)

$$\psi = \ell_0 \circ F$$

- Consequences: to be feasible, loss design ψ ***only*** needs to be strictly monotonic (*no convexity required*) and works with ***any*** **KLST*** model

Technical Lemma that brings the connection

- For ***any*** strictly ↗ function $\ell : [0,1] \to \overline{\mathbb{R}}$, any $K \in \mathbb{R}, a \in [0,1]$, the following loss (ℓ_0, ℓ_1) is strictly proper:

$$\left(\ell(p), K - \left(\int_a^p \frac{\ell(t)}{t^2}\mathrm{d}t + \frac{1-p}{p}\cdot\ell(p)\right)\right)$$

No differentiability assumption of the partial losses, not even continuity

Warning: nomenclature follows Reid & Williamson' 2011

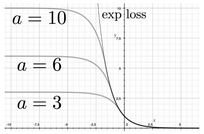
A few additional results, summarized

- DPO's separable loss (log) is the *only one* that fits; extensions of DPO can be safeguarded (margins, correction for length, etc…)

Toy Experiment

- Non-convex loss from exp loss balancing strong convexity vs Lipschitzness

$$\psi_a(-z) \doteq \begin{cases} a - (a^2/4)\cdot\exp(z) & \text{if} \quad z < \log(2/a) \\ \exp(-z) & \text{otherwise} \end{cases}$$



- We use Gemma2_2b_it, rater is gemini_2.5_flash_lite, test on Alpaca Eval v2
- Average over two runs, wins computed vs using exp loss

a	win% (us)
10	44.60%
6	**54.50%**
3	**53.00%**