DPO Unchained:

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

Wenxuan Zhou

Shujian Zhang

Brice Magdalou

John Lambert

Ehsan Amid

Richard Nock

Andrew Hard

Google DeepMind
Google Research

UNIVERSITÉ DE MONTPELLIER

cnrs

CORE AUTOMATION

What this talk is about

- **Normative properties** = guardrails of ML
- ⟶ criteria upon which are built ML's core components, starting with loss functions !
- ≠ knowledge carving, ~skewed by our understanding (biological structures in DNN architectures, etc).

R. Sutton, The Bitter Lesson, 2019

- Without such properties, ↘ ideas of output's quality
- Not "liked" because one usually thinks "straightjackets"

Summary of the talk

Beginning

- A neat+ technique for post-training LLMs dramatically simplifies SOTA via a normative *straightjacket* on human choice (**DPO**)

(DPO = Direct Preference Optimization)

Summary of the talk

Beginning

- A neat+ technique for post-training LLMs dramatically simplifies SOTA via a normative *straightjacket* on human choice (**DPO**)
- Followers+ *stretch it* beyond boundaries or *abandon it* (implicitly or explicitly) 😱

Google Research

Summary of the talk

Talk

End

- A neat+ technique for post-training LLMs dramatically simplifies SOTA via a normative *straightjacket* on human choice (**DPO**)
- Followers+ *stretch it* beyond boundaries or *abandon it* (implicitly or explicitly)

- Normative part wide enough to *disentangle* human choice and even free ML from classical constraints (e.g. *loss convexity*)

(and more, see paper)

Google Research

Google Research

$$p(y \succ y' | x)$$

Choice probabilities

"y is ***chosen*** over y' given x"

(*not* preferred)

policy

$$\pi_\theta : \mathcal{X} \to \Delta_{\mathcal{Y}}$$

Google Research

$$p(y \succ y' | x)$$

reward function

Modeling capability:

- ML / Architecture

policy

$$\pi_\theta : \mathcal{X} \to \Delta_\mathcal{Y}$$

Google Research

$$p(y \succ y' | x)$$

RLHF

Modeling capability:

- Normative
- ML / Architecture

reward function

DPO

Modeling capability:

- Normative

Modeling capability:

- ML / Architecture

policy

$$\pi_\theta : \mathcal{X} \to \Delta_\mathcal{Y}$$

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

loss function

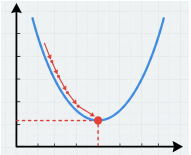
$$p(y \succ y'|x) = f(r(x,y), r(x,y'))$$

reward function

$$\pi_\theta : \mathcal{X} \to \Delta_{\mathcal{Y}}$$

Google Research

Savage



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

Properness, analytical

+Encourage predicting right value

$$p(y \succ y'|x) = f(r(x,y), r(x,y'))$$

Bradley-Terry-Luce

Utility = reward + Gumbel noise

+No abstention possible

$$\pi_\theta : \mathcal{X} \to \Delta_\mathcal{Y}$$

Google Research

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

$$\ell_i(\boldsymbol{q}) \doteq -\log q_i, \forall i$$ ⚠ separable

$$p(y \succ y'|x) = f(r(x,y), r(x,y'))$$

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

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

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

$$p(y \succ y'|x) = f(r(x,y), r(x,y'))$$

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

"$\ell \circ f$" → logistic loss: $I = \mathbb{E}_{(x,y,y')\sim\mathcal{D}}\left[\psi(\ldots)\right]$

$$\psi(-z) \doteq \log(1+\exp(-z))$$

⚠ ML's "final" loss ← **two** ≠ normative models (loss functions, human choice)

Google Research

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

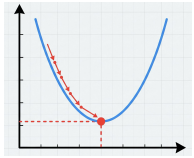
$$p(y \succ y'|x) = f(r(x,y), r(x,y'))$$

RLHF

$r_\omega(x,y)$

$\pi_\theta(y|x)$

Savage



$\pi_\theta : \mathcal{X} \to \Delta_\mathcal{Y}$

Google Research

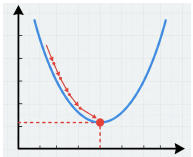
$$I = \mathbb{E}_{(x,y,y')\sim\mathcal{D}}\left[\ell(p(y \succ y'|x))\right]$$

$$p(y \succ y'|x) = f(r(x,y), r(x,y'))$$

DPO

$$r(x,y) = h(\pi_\theta(y|x))$$

Savage



$$\pi_\theta : \mathcal{X} \to \Delta_\mathcal{Y}$$

Google Research

DPO Unchained | Zhou et al. | ICML'26

Savage

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

$$p(y \succ y'|x) = f(r(x,y), r(x,y'))$$

BTL

DPO

$$r(x,y) = h(\pi_\theta(y|x))$$

Savage

Fully normative

$$\pi_\theta : \mathcal{X} \to \Delta_\mathcal{Y}$$

Google Research

DPO Unchained | Zhou et al. | ICML'26

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

$$p(y \succ y'|x) = f(r(x,y), r(x,y'))$$

$$r(x,y) = h(\pi_\theta(y|x))$$

$$\pi_\theta : \mathcal{X} \to \Delta_\mathcal{Y}$$

Google Research

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

$$p(y \succ y'|x) = f(r(x,y), r(x,y'))$$???

$$r(x,y) = h(\pi_\theta(y|x))$$

$$\pi_\theta : \mathcal{X} \to \Delta_\mathcal{Y}$$

Our first objective: broaden the *human choice* part so it matches the freedom of **Savage**'s model, while keeping compatibility with *ML*'s *pipeline*

Google Research

Transition from **BTL** to **KLST***

- A theory already exists to go “way” above BTL: Doignon-Falmagne (1974) and its simplification in **Krantz-Luce-Suppes-Tversky** (1989), **but** it relies on an axiom especially unfit for ML / RL, the solvability axiom, requiring $|\mathcal{Y}| = \infty$.

Krantz-Luce-Suppes-Tversky* (KLST*)

- We bypass the use of solvability by integrating a mechanism compatible with ML's sampling, **binary lotteries**. We end up with a model we call **KLST***.

Krantz-Luce-Suppes-Tversky* (KLST*)

- **Lotteries** and choice probabilities: $\alpha \in [0, 1]$

$\mathcal{Y}^{\alpha} \doteq \{(yy')_{\alpha} : y, y' \in \mathcal{Y}\}$ $\quad (y_1 y_2)_{\alpha}$ = lottery combining alternatives y_1 and y_2

$$\hookrightarrow p((y_1 y_2)_{\alpha} \succ L|x) \doteq \alpha \cdot p(y_1 \succ L|x) + (1 - \alpha) \cdot p(y_2 \succ L|x)$$

Krantz-Luce-Suppes-Tversky* (KLST*)

- **Lotteries** and choice probabilities: $\alpha \in [0, 1]$

$$\mathcal{Y}^{\alpha} \doteq \{(yy')_{\alpha} : y, y' \in \mathcal{Y}\}$$

$(y_1 y_2)_{\alpha}$ = lottery combining alternatives y_1 and y_2

$$\hookrightarrow p((y_1 y_2)_{\alpha} \succ L|x) \doteq \alpha \cdot p(y_1 \succ L|x) + (1 - \alpha) \cdot p(y_2 \succ L|x)$$

- **Preference** relationship:

$$L \rhd_x L' \Leftrightarrow p(L \succ L'|x) \geqslant 1/2$$

> ***Preference*** means choice probability ≥ 1/2

Krantz-Luce-Suppes-Tversky* (KLST*)

- **Lotteries** and choice probabilities: $\alpha \in [0, 1]$

$$\mathcal{Y}^\alpha \doteq \{(yy')_\alpha : y, y' \in \mathcal{Y}\} \qquad (y_1 y_2)_\alpha = \text{lottery combining alternatives } y_1 \text{ and } y_2$$

$$\hookrightarrow p((y_1 y_2)_\alpha \succ L|x) \doteq \alpha \cdot p(y_1 \succ L|x) + (1-\alpha) \cdot p(y_2 \succ L|x)$$

- **Preference** relationship:

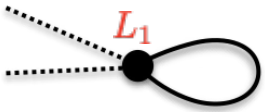
$$L \rhd_x L' \Leftrightarrow p(L \succ L'|x) \geqslant 1/2$$

- **Graph** $G_x^\alpha \doteq (\mathcal{Y}^\alpha, \mathcal{E}_x^\alpha)$; nodes = lotteries and **edges** = **zero abstention**

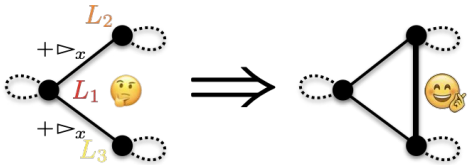
$$(L, L') \in \mathcal{E}_x^\alpha \Leftrightarrow p(L \succ L'|x) + p(L' \succ L|x) = 1$$

Local Choice Structure on Graph = 3 simple axioms

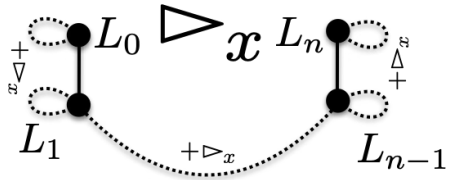
- Every lottery is bearable:



- Zero abstention **and** similar polarity for preference elicit zero abstention:



- Every preference is the closure of a chain of preferences + **zero abstention**:



KLST* Structure = 3 simple elements

- **Choice probabilities have a KLST* structure iff**
  - They are **expandable** (Lotteries)
  - The Lotteries graph has a **Local Choice Structure** for any $\alpha \in (0, 1)$
  - Choices satisfy **monotonicity** (includes a property on edges, not shown here): If
    - L_1 is preferred to L_2 **more than** L_4 is preferred to L_5 **and**
    - L_2 is preferred to L_3 **more than** L_5 is preferred to L_6 **, then**

    - L_1 is preferred to L_3 more than L_4 is preferred to L_6

Theorem: **Savage**-wide freedom for the human choice part

- Suppose choice probabilities have a **KLST*** structure. Then there exists
  - A strictly increasing 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 $p(y \succ y'|x) = F(u(x, y) - u(x, y'))$

Theorem: Savage-wide freedom for the human choice part

- Suppose choice probabilities have a **KLST*** structure. Then there exists
  - A strictly increasing 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 $p(y \succ y'|x) = F(u(x, y) - u(x, y'))$

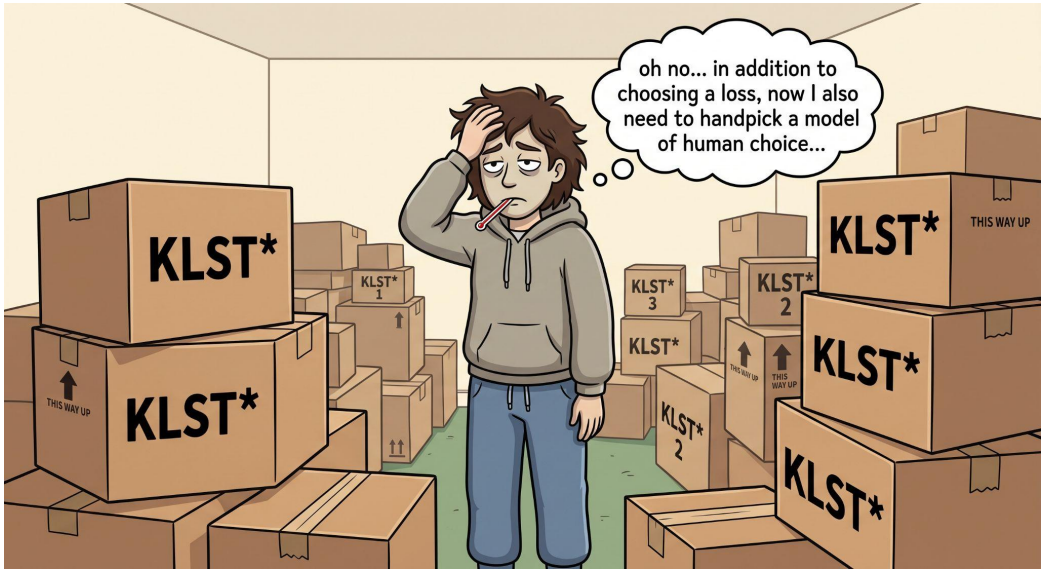
$$I = \mathbb{E}_{(x,y,y')\sim\mathcal{D}}\left[\ell(p(y \succ y'|x))\right]$$

$$p(y \succ y'|x) = f(r(x, y), r(x, y'))$$

"**Savage**-wide" because Savage classically enforce 2nd-order regularity

Hold on !

Did we just end here ?



Reminder

ℓ and f yield ML's "final" loss

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

$$p(y \succ y'|x) = f(r(x,y), r(x,y'))$$

"$\ell \circ f$" → logistic loss: $I = \mathbb{E}_{(x,y,y')\sim\mathcal{D}}\left[\psi(\ldots)\right]$

$$\psi(-z) \doteq \log(1+\exp(-z))$$

⚠ ML's "final" loss ← **two** ≠ normative models (loss functions, human choice)

Theorem

- For any strictly increasing $F : \mathbb{R} \to [0, 1]$ and any strictly increasing $\psi : \mathbb{R} \to \mathbb{R}$,

  there exists a *strictly proper loss* (ℓ_0, ℓ_1) such that $\psi = \ell_0 \circ F$

Partial losses, C=2 classes

Theorem

- For any strictly increasing $F : \mathbb{R} \to [0, 1]$ and any strictly increasing $\psi : \mathbb{R} \to \mathbb{R}$,

  there exists a *strictly proper loss* (ℓ_0, ℓ_1) such that $\psi = \ell_0 \circ F$

(+Proof is constructive)

Theorem

- For **any** strictly increasing $F : \mathbb{R} \to [0, 1]$

and any strictly increasing $\psi : \mathbb{R} \to \mathbb{R}$,

  there exists a *strictly proper loss* (ℓ_0, ℓ_1) such that $\psi = \ell_0 \circ F$

Theorem

- For any strictly increasing $F : \mathbb{R} \to [0, 1]$
  and **any** strictly increasing $\psi : \mathbb{R} \to \mathbb{R}$; ← ML's "final" loss

  there exists a *strictly proper loss* (ℓ_0, ℓ_1) such that $\psi = \ell_0 \circ F$

Theorem

- For any strictly increasing $F : \mathbb{R} \to [0, 1]$ and any strictly increasing $\psi : \mathbb{R} \to \mathbb{R}$,

  Savage

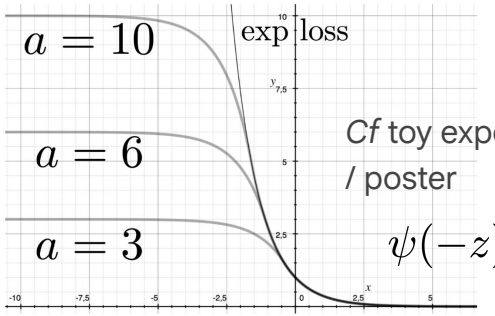
  there exists a *strictly proper loss* (ℓ_0, ℓ_1) such that $\psi = \ell_0 \circ F$

Theorem

- For **any** strictly increasing $F : \mathbb{R} \to [0, 1]$ and **any** strictly increasing $\psi : \mathbb{R} \to \mathbb{R}$,

  there exists a *strictly proper loss* (ℓ_0, ℓ_1) such that $\psi = \ell_0 \circ F$

Any ML's strictly monotonic loss ψ can embed **any KLST*** model



Cf toy experiment in paper / poster

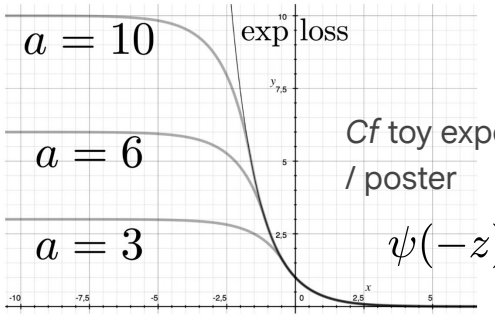
Google Research

Theorem

- For any strictly increasing $F : \mathbb{R} \to [0, 1]$ and **any** strictly increasing $\psi : \mathbb{R} \to \mathbb{R}$,

  there exists a *strictly proper loss* (ℓ_0, ℓ_1) such that $\psi = \ell_0 \circ F$

Convexity of ML's ψ not mandatory



Cf toy experiment in paper / poster

Google Research

But not everything is

Theorem

- In any of the **Savage** parts, if one wants a *separable* loss, then **one** choice only, the log-loss: $\ell_i(\boldsymbol{q}) \doteq -\log q_i, \forall i$

Theorem

- In any of the **Savage** parts, if one wants a *separable* loss, then **one** choice only, the log-loss: $\ell_i(\boldsymbol{q}) \doteq -\log q_i, \forall i$
- Even if
  - we authorize a ≠ partial loss for each coord.

Theorem

- In any of the **Savage** parts, if one wants a *separable* loss, then **one** choice only, the log-loss: $\ell_i(\boldsymbol{q}) \doteq -\log q_i, \forall i$
- Even if
  - we authorize a ≠ partial loss for each coord.
  - +we alleviate differentiability (partial losses)

Theorem

- In any of the **Savage** parts, if one wants a *separable* loss, then **one** choice only, the log-loss: $\ell_i(\boldsymbol{q}) \doteq -\log q_i, \forall i$
- Even if
  - we authorize a ≠ partial loss for each coord.
  - +we alleviate differentiability (partial losses)
  - +we alleviate continuity (partial losses)

Theorem

- In any of the **Savage** parts, if one wants a *separable* loss, then **one** choice only, the log-loss: $\ell_i(\boldsymbol{q}) \doteq -\log q_i, \forall i$
- Even if
  - we authorize a ≠ partial loss for each coord.
  - +we alleviate differentiability (partial losses)
  - +we alleviate continuity (partial losses)
  - +we require properness for 1 target $\boldsymbol{p}$ only with support on 3+ elements of $\mathcal{Y}$

(Substantially extends a series of results with assumptions: McCarthy, 1956; Savage, 1971; Bernardo, 1979; Dawid, 2007)

Conclusion

- A Normative straightjacket can become ML's freedom
- In DPO's pipeline, **KLST*** 's modeling capability matches **Savage**'s (⟶ do not expect more capability by extending **KLST***)
- BUT **Savage** hides "in plain sight" a key analytical assumption… replace it and more can happen

A. Huang et al., Correcting the Mythos of KL-Regularization: Direct Alignment without Overoptimization via Chi-Squared Preference Optimization, 2025

Thank You

Wenxuan Zhou	Google DeepMind
Shujian Zhang	Google DeepMind
Brice Magdalou	CEE-M, Univ. Montpellier, CNRS, INRAé, Institut Agro
John Lambert	Google DeepMind
Ehsan Amid	Core Automation
Richard Nock	Google Research
Andrew Hard	Google DeepMind

Google Research