# Distributional Preference Alignment of LLMs via Optimal Transport

Youssef Mroueh

UBC, Kantorovich Initiative

*Joint work with Igor Melnyk, Brian Belgodere, Mattia Rigotti, Apoorva Nitsure, Mikhail Yurochkin, Kristjan Greenewald, Jiri Navratil, Jarret Ross*

December 13, 2024

# Outline

# The Alignment Problem

- Find a policy that maximizes a reward $r$ while staying close to a reference policy $\pi_{\text{ref}}$:

$$\max_{\pi_\theta} \int r d\pi_\theta - \beta \text{KL}(\pi_\theta || \pi_{\text{ref}})$$

- The optimal policy is given by:

$$\pi_\theta(y|x) = \frac{\pi_{\text{ref}}(y|x) \exp(\frac{r(x,y)}{\beta})}{Z(x)}$$

# DPO as a Pointwise Preference Approach 1/2

**Direct Preference Optimization (DPO)**
[Rafailov et al.(2024)Rafailov, Sharma, Mitchell, Manning, Ermon, and Finn]

- The reward optimized by the LLM:

$$r_\theta(x, y) = \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)} + \beta \log(Z(x))$$

- $Z(x)$ is a normalization constant.
- Paired preference dataset: $(X, Y_+, Y_-) \sim \mu$
  - $Y_+$: positive (chosen) response
  - $Y_-$: negative (rejected) response

# DPO as a Pointwise Preference Approach 2/2

- Objective: Minimize the logarithmic sigmoid loss:

$$\min_{\theta \in \Theta} -\mathbb{E}_{(x,y_+,y_-)\sim\mu} \log(\sigma(\beta(r_\theta(x,y_+) - r_\theta(x,y_-))))$$

- Simplified form (normalization $Z(x)$ cancels out):

$$\min_{\theta} -\mathbb{E}_{(x,y_+,y_-)\sim\mu} \log\left(\sigma\left(\beta \log\left(\frac{\pi_\theta(y_+|x)}{\pi_{\text{ref}}(y_+|x)}\right) - \beta \log\left(\frac{\pi_\theta(y_-|x)}{\pi_{\text{ref}}(y_-|x)}\right)\right)\right)$$

- Interpretation as pointwise preference:

$$\boxed{\log\left(\frac{\pi_\theta(y_+|x)}{\pi_{\text{ref}}(y_+|x)}\right) \geq \log\left(\frac{\pi_\theta(y_-|x)}{\pi_{\text{ref}}(y_-|x)}\right), \quad \forall (x, y_+, y_-) \sim \mu.} \tag{1}$$

- Relaxation through logistic loss, suggests other algorithms (e.g., hinge loss in SLIC [Zhao et al.(2023)Zhao, Khalman, Joshi, Narayan, Saleh, and Liu]).

# Pointwise Preference

- Pointwise Preference positive versus negative:

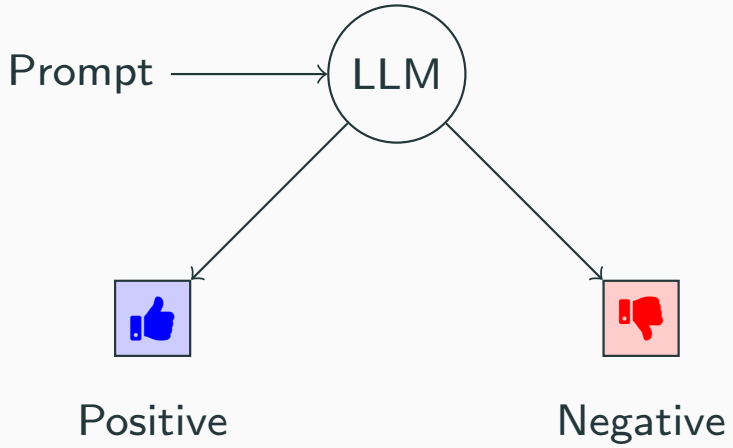
$$\boxed{\log\left(\frac{\pi_\theta(y_+|x)}{\pi_{\text{ref}}(y_+|x)}\right) \geq \log\left(\frac{\pi_\theta(y_-|x)}{\pi_{\text{ref}}(y_-|x)}\right), \quad \forall(x, y_+, y_-) \sim \mu.} \tag{2}$$

- Equivalently: pointwise Preference of margin (positive - negative) of Model versus Reference:

$$\boxed{\log\frac{\pi_\theta(y_+|x)}{\pi_\theta(y_-|x)} \geq \log\frac{\pi_{\text{ref}}(y_+|x)}{\pi_{\text{ref}}(y_-|x)}, \quad \forall(x, y_+, y_-) \sim \mu.} \tag{3}$$

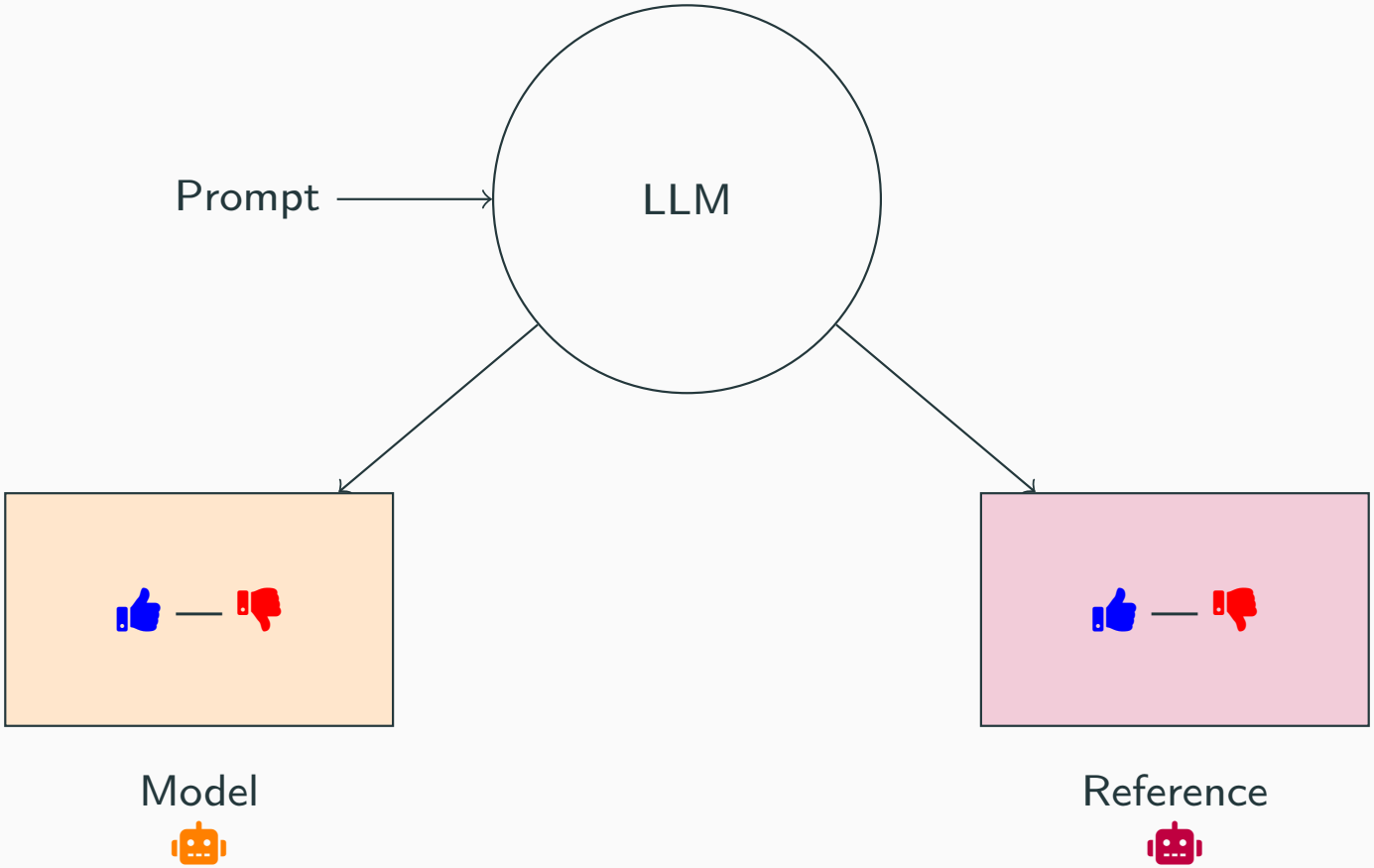
# Outline

# Pointwise Preference (Positive versus Negative)




$$\log\left(\frac{\pi_\theta(y_+|x)}{\pi_{\text{ref}}(y_+|x)}\right) \geq \log\left(\frac{\pi_\theta(y_-|x)}{\pi_{\text{ref}}(y_-|x)}\right)$$

# Pointwise Preference, Model Versus Reference based on Margins




$$\log \frac{\pi_\theta(y_+|x)}{\pi_\theta(y_-|x)} \geq \log \frac{\pi_{\text{ref}}(y_+|x)}{\pi_{\text{ref}}(y_-|x)}$$

# Partial Order on One Dimensional Distribution: First Order Stochastic Dominance

**Definition:**

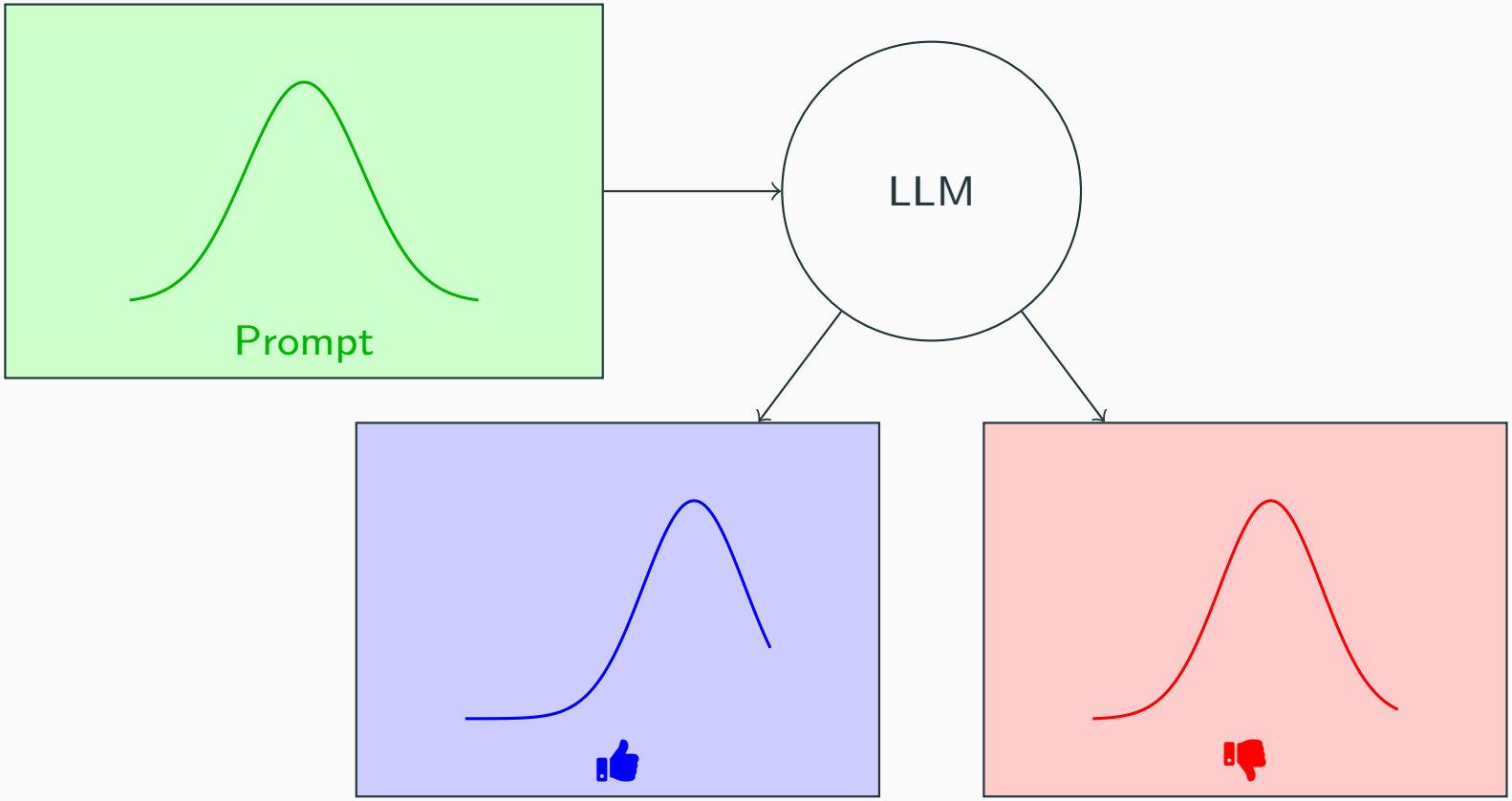
- For a real random variable $Z$, the left-continuous inverse of the Cumulative Distribution Function (CDF) $F_Z$ is denoted by $F_Z^{(-1)} : [0,1] \to \overline{\mathbb{R}}$.
- This inverse, also known as the quantile function $Q_Z(p)$, is defined as:
$$Q_Z(p) = F_Z^{(-1)}(p) = \inf\{\eta : F_Z(\eta) \geq p\} \text{ for } p \in [0,1].$$

**Definition (First Order Stochastic Dominance (FSD):)**

Given two random variables $Z_1$ and $Z_2$, $Z_1$ is said to dominate $Z_2$ in the first order if $Z_1$ has larger quantiles than $Z_2$ for all percentiles $p$:

$$Z_1 \underset{\text{FSD}}{\succcurlyeq} Z_2 \iff Q_{Z_1}(p) \geq Q_{Z_2}(p), \quad \forall p \in [0,1].$$

# Distributional Preference Unpaired (based on rewards of positive and negatives)

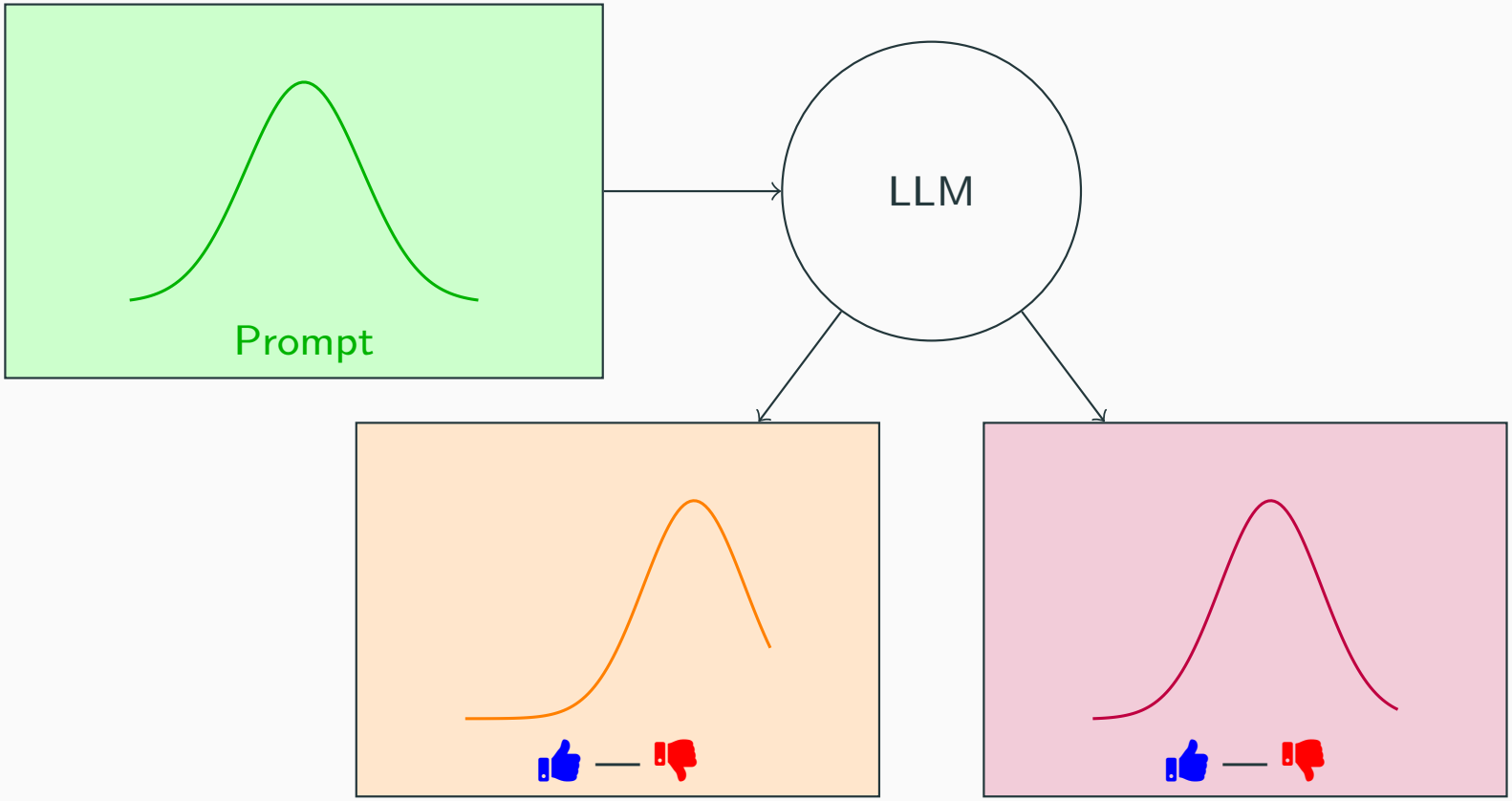



Positive Negative

$$\log \frac{\pi_\theta(Y_+|X_+)}{\pi_{\text{ref}}(Y_+|X_+)} \underset{\text{FSD}}{\succcurlyeq} \log \frac{\pi_\theta(Y_-|X_-)}{\pi_{\text{ref}}(Y_-|X_-)}.$$

does not need paired positive and negative answer for each prompt!

# Distributional Preference Paired
# (based on margins of rewards between Model and Reference)




$$\log \frac{\pi_\theta(Y_+|X)}{\pi_\theta(Y_-|X)} \succcurlyeq_{\text{FSD}} \log \frac{\pi_{\text{ref}}(Y_+|X)}{\pi_{\text{ref}}(Y_-|X)}$$

needs paired positive and negative answer for each prompt!

# Distributional Unpaired Preference

- No access to triplets of prompts and positive/negative responses $(x, y_+, y_-)$.
- Separate access to:
  - $\mu_+ \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$: Distribution of positive prompt/response pairs $(X_+, Y_+)$ to be highly rewarded.
  - $\mu_- \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$: Distribution of negative samples $(X_-, Y_-)$ to be associated with low reward.
- Define the distributional preference as follows:

**Definition (Distributional Preference in the Unpaired Setting)**

A policy $\pi$ prefers distributionally $\mu_+$ over $\mu_-$ with respect to a reference policy $\pi_{\text{ref}}$ if:

$$\log \frac{\pi_\theta(Y_+|X_+)}{\pi_{\text{ref}}(Y_+|X_+)} \underset{\text{FSD}}{\succcurlyeq} \log \frac{\pi_\theta(Y_-|X_-)}{\pi_{\text{ref}}(Y_-|X_-)}.$$

In other words, noting $r_u \circ \pi_\theta(x, y) = \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$, the distributional preference in the unpaired setting means that we have the following constraint:

$$(r_u \circ \pi_\theta)_\sharp \mu_+ \underset{\text{FSD}}{\succcurlyeq} (r_u \circ \pi_\theta)_\sharp \mu_-. \tag{4}$$

# Distributional Paired Preference

We define below more formally the paired distributional preference via stochastic dominance:

**Definition (Distributional Preference in the Paired Setting)**

We say that the policy $\pi_\theta$ distributionally dominates $\pi_{\text{ref}}$ in terms of log probability ratio of positive and negative responses if:

$$\log \frac{\pi_\theta(Y_+|X)}{\pi_\theta(Y_-|X)} \underset{\text{FSD}}{\succcurlyeq} \log \frac{\pi_{\text{ref}}(Y_+|X)}{\pi_{\text{ref}}(Y_-|X)}.$$

Noting $r_p \circ \pi_\theta(x, y_+, y_-) = \log \frac{\pi_\theta(y_+|x)}{\pi_\theta(y_-|x)}$ this can be written as follows:

$$(r_p \circ \pi_\theta)_\sharp \mu \underset{\text{FSD}}{\succcurlyeq} (r_p \circ \pi_{\text{ref}})_\sharp \mu. \tag{5}$$

# Outline

# Learning Policies with Distributional Constraints

- FSD unpaired Constraint:

  Find $\pi_\theta \in \mathcal{H}$ such that $(r_u \circ \pi_\theta)_\sharp \mu_+ \underset{\text{FSD}}{\succcurlyeq} (r_u \circ \pi_\theta)_\sharp \mu_-$ (FSD unpaired)

- FSD Paired Constraint:

  Find $\pi_\theta \in \mathcal{H}$ such that $(r_p \circ \pi_\theta)_\sharp \mu \underset{\text{FSD}}{\succcurlyeq} (r_p \circ \pi_{\text{ref}})_\sharp \mu$ (FSD paired)

- Both Problem can be abstracted out to:

  Find $\theta \in \Theta$ such that : $U_\theta \underset{\text{FSD}}{\succcurlyeq} V_\theta$

# A Convex Relaxation of FSD

$$\boxed{U_\theta \underset{\text{FSD}}{\succcurlyeq} V_\theta \iff Q_{U_\theta}(t) \geq Q_{V_\theta}(t), \forall t \in [0,1].}$$

We can relax this problem to minimizing the violation of the FSD order:

$$\boxed{\min_{\theta \in \Theta} \varepsilon(\theta) := \int_0^1 h(Q_{U_\theta}(t) - Q_{V_\theta}(t))dt,}$$

where $h$ penalizes the violation of dominance of $U_\theta$ on $V_\theta$

- **Indicator Loss**: $h(x) = \mathbb{1}_{x<0}$ .
- $\beta$-**squared Hinge Loss**: For a margin $\beta > 0$, $h(x) = (\beta - x)_+^2$.
- $\beta$-**logistic Loss**: $h(x) = \log(1 + \exp(-\beta x))$.
- $\beta$-**Least Squares**:
  - Although not a convex relaxation of the 0/1 loss, the least squares loss has been used in classification .
  - In the context of alignment, it was used in IPO [Azar et al.(2024)Azar, Guo, Piot, Munos, Rowland, Valko, and Calandriello].
  - $h(x) = (\beta - x)^2$.

# Quantile Violation as an Optimal Transport Problem

**Theorem (Theorem 2.9 and Proposition 2.17 in [Santambrogio(2015)])**

*Let $h : \mathbb{R} \to \mathbb{R}^+$ be a convex function we have for two real random variables $U, V$, with measures $\mu_U, \mu_V$:*

$$\int_0^1 h(Q_U(t) - Q_V(t))dt = \min_{\gamma \in \Pi(\mu_U, \mu_V)} \int h(u - v)d\gamma(u, v) = \mathrm{OT}_h(\mu_U, \mu_V)$$

*and $\gamma^* = (Q_U, Q_V)_\sharp \mathcal{L}_1([0, 1])$ is a minimizer (where $\mathcal{L}_1$ is the Lebesgue measure on $[0, 1]$ ). If furthermore $h$ is strictly convex $\gamma^*$ is the unique minimizer.*

$$\min_{\theta \in \Theta} \int_0^1 h(Q_{U_\theta}(t) - Q_{V_\theta}(t))dt = \min_{\theta \in \Theta} \mathrm{OT}_h(\mu_{U_\theta}, \mu_{V_\theta}) = \min_{\theta \in \Theta} \min_{\gamma \in \Pi(\mu_{U_\theta}, \mu_{V_\theta})} \int h(u - v)d\gamma(u, v)$$

Optimizing Relaxed FSD is an inner 1D Optimal Transport problem with smooth and convex cost

# Computational Algorithm via Sorting

- We consider empirical measures and solve for a fixed $\theta$. We simplify notation by omitting $\theta$.
- $\mathrm{OT}_h(\hat{\mu}_U, \hat{\mu}_V)$ is of interest, where:

$$\hat{\mu}_U = \frac{1}{n}\sum_{i=1}^{n} \delta_{u_i}, \quad \hat{\mu}_V = \frac{1}{n}\sum_{i=1}^{n} \delta_{v_i}.$$

- Due to the convexity of $h$ , $\mathrm{OT}_h(\hat{\mu}_{U_\theta}, \hat{\mu}_{V_\theta})$'s optimal coupling is given by the north-west corner solution [Peyré and Cuturi(2019)] (Chapter 3, Section 3.4.2), informally matching $i$-th smallest elements of $U$ and $V$.
- Formally, if we sort $u_i$ and $v_i$ into order statistics ($u^{(1)} \leq \cdots \leq u^{(n)}$, $v^{(1)} \leq \cdots \leq v^{(n)}$):

$$\mathrm{OT}_h(\hat{\mu}_U, \hat{\mu}_V) = \frac{1}{n}\sum_{i=1}^{n} h(u^{(i)} - v^{(i)}).$$

- Given empirical samples $\hat{\mu}_{U_\theta} = \frac{1}{n}\sum_{i=1}^{n} \delta_{u_\theta^i}$ and $\hat{\mu}_{V_\theta} = \frac{1}{n}\sum_{i=1}^{n} \delta_{v_\theta^i}$, with $u_\theta^{(i)}, v_\theta^{(i)}$ as order statistics:

**AOT**

$$\min_{\theta\in\Theta} \mathrm{OT}_h(\hat{\mu}_{U_\theta}, \hat{\mu}_{V_\theta}) = \min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} h(u_\theta^{(i)} - v_\theta^{(i)}) \quad \text{(AOT)}$$

# AOT: Alignment Via Optimal Transport

**Algorithm 1** AOT Unpaired

**Input:** $\pi_\theta$, $\pi_{\text{ref}}, \beta > 0, \varepsilon > 0$,
**Unpaired Preference Data:** "Chosen" $\hat{\mu}_+^n$ and "Rejected" $\hat{\mu}_-^n$.
**for** iter $\leftarrow 1, n_{\text{iter}}$ **do**
  **Get a Positive/Negative mini-batch**
  $\{(x_{i,+}, y_{i,+}) \sim \hat{\mu}_+^n, i = 1 \dots b\}$
  $\{(x_{i,-}, y_{i,-}) \sim \hat{\mu}_-^n, i = 1 \dots b\}$
  **Compute Rewards**
  $u_\theta^i = \log \frac{\pi_\theta(y_{i,+}|x_{i,+})}{\pi_{\text{ref}}(y_{i,+}|x_{i,+})}$
  $v_\theta^i = \log \frac{\pi_\theta(y_{i,-}|x_{i,-})}{\pi_{\text{ref}}(y_{i,-}|x_{i,-})}$
  **Sort Rewards**
  $(u^{(1)} \dots u^{(b)}) = \text{Sort}(u_\theta^i)$
  $(v^{(1)} \dots v^{(b)}) = \text{Sort}(v_\theta^i)$
  **Compute AOT logistic loss**
  $\ell_\theta = -\frac{1}{b} \sum_{i=1}^b \log \sigma(\beta(u_\theta^{(i)} - v_\theta^{(i)}))$
  **Update** $\theta$ **with** PagedAdamw32bit
**end for**
**Return** $\pi_\theta$

**Algorithm 2** AOT Paired

**Input:** $\pi_\theta$, $\pi_{\text{ref}}, \beta > 0$,
**Paired Preference Data:** $\hat{\mu}^n = \frac{1}{n} \sum_{i=1}^n \delta_{(x_i, y_{i,+}, y_{i,-})}$
**for** iter $\leftarrow 1, n_{\text{iter}}$ **do**
  **Get a Positive/Negative mini-batch**
  $\{(x_i, y_{i,+}, y_{i,-}) \sim \hat{\mu}^n, i = 1 \dots b\}$
  **Compute Margins for** $i = 1 \dots b$
  $u_\theta^i = \log \frac{\pi_\theta(y_{i,+}|x_i)}{\pi_\theta(y_{i,-}|x_i)}$
  $v_\theta^i = \log \frac{\pi_{\text{ref}}(y_{i,+}|x_i)}{\pi_{\text{ref}}(y_{i,-}|x_i)}$
  **Sort Margins**
  $(u^{(1)} \dots u^{(b)}) = \text{Sort}(u_\theta^i)$
  $(v^{(1)} \dots v^{(b)}) = \text{Sort}(v_\theta^i)$
  **Compute AOT logistic loss**
  $\ell_\theta = -\frac{1}{b} \sum_{i=1}^b \log \sigma(\beta(u_\theta^{(i)} - v_\theta^{(i)}))$
  $\theta$ **with** PagedAdamw32bit
**end for**
**Return** $\pi_\theta$

# Outline

# Parametric Rate for AOT

**Theorem (Informal)**

*Under appropriate Assumptions we have:*

$$\mathbb{E}\ \mathrm{OT}_h\left((r\circ\pi_{\hat{\theta}_n})_\sharp\mu_+,(r\circ\pi_{\hat{\theta}_n})_\sharp\mu_-\right)-\mathrm{OT}_h\left((r\circ\pi_{\theta^*})_\sharp\mu_+,(r\circ\pi_{\theta^*})_\sharp\mu_-\right)\lesssim n^{-\frac{1}{2}},$$

*where $\lesssim$ refers to inequality up to constants that depend only on constants in the assumptions.*

The model estimated from emprical samples with AOT generalizes on unseen samples with a rate $n^{-\frac{1}{2}}$

# Elements of the Proof 1/2

$$\min_{\pi_\theta \in \mathcal{H}} \int_0^1 h\left(Q_{(r\circ\pi_\theta)_\sharp \mu_+}(t) - Q_{(r\circ\pi_\theta)_\sharp \mu_-}(t)\right) dt = \min_{\pi_\theta \in \mathcal{H}} \mathrm{OT}_h\left((r\circ\pi_\theta)_\sharp \mu_+, (r\circ\pi_\theta)_\sharp \mu_-\right) \tag{uAOT$_h$}$$

Define the OT cost $c : [-M, M] \times [-M, M] \to [0, R]$ such that $c(z, z') = h(z - z')$, for $z, z \in [-M, M]$. Define the $c$-transform of a function $\varphi : [-M, M] \to \mathbb{R}$:

$$\varphi^c(z) = \inf_{z' \in [-M,M]} h(z - z') - \varphi(z).$$

In our setting, a function is called $c$-concave if there exists $\psi : [-M, M] \to \mathbb{R}$ such that $\varphi = \psi^c$. Define:

$$\mathcal{F}_c = \{\varphi : [-M, M] \to [-R, R], \varphi \text{ is } c\text{-concave, with } ||\varphi^c||_\infty \leq R\}$$

By duality (Theorem 5.10 in [Villani(2009)]) we have:

$$\mathrm{OT}_h\left((r\circ\pi_\theta)_\sharp \mu_+, (r\circ\pi_\theta)_\sharp \mu_-\right) = \sup_{\varphi \in \mathcal{F}_c} \int \varphi(r\circ\pi_\theta) d\mu_+ - \int \varphi^c(r\circ\pi_\theta) d\mu_-.$$

# Elements of the proof 2/2

Replacing the dual expression of $\mathrm{OT}_h$ in ($\mathrm{uAOT}_h$), we see that ($\mathrm{uAOT}_h$) can be cast as a min-max problem:

$$\min_{\pi_\theta \in \mathcal{H}} \sup_{\varphi \in \mathcal{F}_c} \int \varphi(r \circ \pi_\theta) d\mu_+ - \int \varphi^c(r \circ \pi_\theta) d\mu_-. \tag{6}$$

Given samples $\hat{\mu}_+^n = \frac{1}{n}\sum_{i=1}^n \delta_{(x_{i,+}, y_{i,+})}$ and $\hat{\mu}_-^n = \frac{1}{n}\sum_{i=1}^n \delta_{(x_{i,-}, y_{i,-})}$, the empirical problem is:

$$\min_{\pi_\theta \in \mathcal{H}} \sup_{\varphi \in \mathcal{F}_c} \int \varphi(r \circ \pi_\theta) d\hat{\mu}_+^n - \int \varphi^c(r \circ \pi_\theta) d\hat{\mu}_-^n. \tag{7}$$

Let $\pi_{\theta^*}$ be the population minimizer of ($\mathrm{uAOT}_h$) and $\pi_{\hat{\theta}_n}$ be the solution of the empirical problem (7). We have the following sample complexity bound for the violation of stochastic dominance in AOT unpaired:

$$\mathbb{E}\ \mathrm{OT}_h\left((r \circ \pi_{\hat{\theta}_n})_\sharp \mu_+, (r \circ \pi_{\hat{\theta}_n})_\sharp \mu_-\right) \leq \underbrace{\mathrm{OT}_h\left((r \circ \pi_{\theta^*})_\sharp \mu_+, (r \circ \pi_{\theta^*})_\sharp \mu_-\right)}_{\text{Optimal Almost FSD Violation}}$$

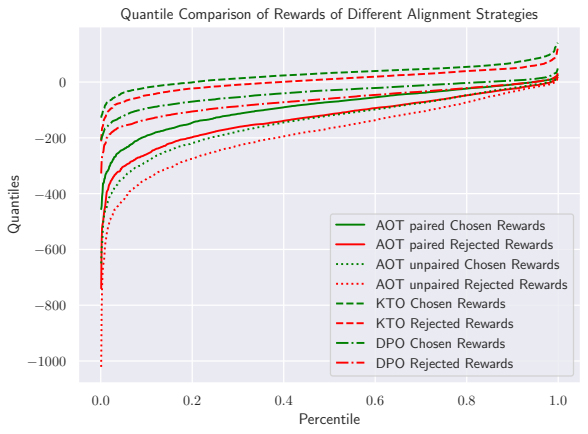
$$+ \underbrace{2\mathcal{R}_n(\mathcal{F}_c; (r \circ \pi_{\theta^*})_\sharp \mu_+) + 2\mathcal{R}_n(\mathcal{F}_c^c; (r \circ \pi_{\theta^*})_\sharp \mu_-)}_{\text{One dimensional OT sample complexity with optimal } \theta^*}$$

$$+ \underbrace{2\mathcal{R}_n(\mathcal{F}_c \circ r \circ \mathcal{H}; \mu_+) + 2\mathcal{R}_n(\mathcal{F}_c^c \circ r \circ \mathcal{H}; \mu_-)}_{\text{Complexity of learning in } \mathcal{H} \text{ via the 1D OT problem}},$$

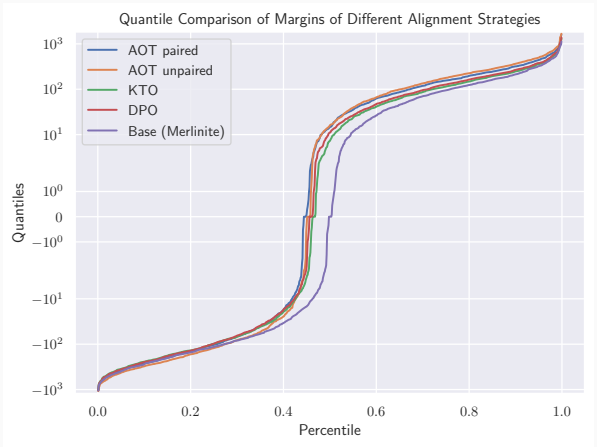
where $\mathcal{R}_n(\mathcal{F}; \nu) = \mathbb{E} \sup_{\varphi \in \mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n \sigma_i \varphi(Z_i)\right|$ is the Rademacher Complexity and for $i = 1 \ldots n$, $\sigma_i$ are independent Rademacher random variables and $Z_i \sim \nu$ iid.

# Outline

# Distributional Preference




(a) Stochastic Dominance of Reward of Chosen on Rejected: AOT achieves a larger margin between the quantile plots of chosen and rejected rewards.




(b) Stochastic Dominance of AOT's optimized policy margin (between Chosen on Rejected) on the margin of the reference policy.

**Figure 1:** AOT in the paired & unpaired settings enables first-order stochastic dominance of the chosen reward distribution on the rejected distribution (a). The margin between the quantiles of chosen and rejected rewards is larger than alternative strategies. In (b), we see that AOT's policy chosen to rejected log-likelihood ratio dominates that ratio for the base model and alternative strategies.

# Results on Merlinite 7B Alignment

| | AlpacaEval (GPT4) | ARC | Hellaswag | MMLU | Truthful | Winogrande | GSM8K |
|---|---|---|---|---|---|---|---|
| AOT paired | 29.9 | 82.5 | 66.1 | 62.9 | 50.8 | 74.4 | 53.1 |
| AOT unpaired | **31.3** | 82.5 | **66.2** | 62.8 | **51.1** | 74.4 | 51.8 |
| DPO | 27.4 | **82.8** | 65.8 | **63.1** | 50.6 | 74.3 | 52.0 |
| KTO | 24.9 | 82.7 | 65.4 | 63.0 | 48.7 | **74.9** | **53.9** |
| IPO | 27.7 | 82.4 | 65.1 | 63.0 | 46.5 | 74.0 | 52.3 |
| Merlinite-7B | 17.1 | 81.6 | 63.2 | 62.6 | 42.0 | 73.9 | 45.2 |

**Table 1:** `Merlinite-7B` trained on UltraFeedback Binarized. AOT results in the best performing LLM as compared to the alternative alignment algorithms on AlpacaEval, and is competitive across the other benchmarks that are evaluated in the zero shot regime.

# Ablations on AOT Losses and Batch size

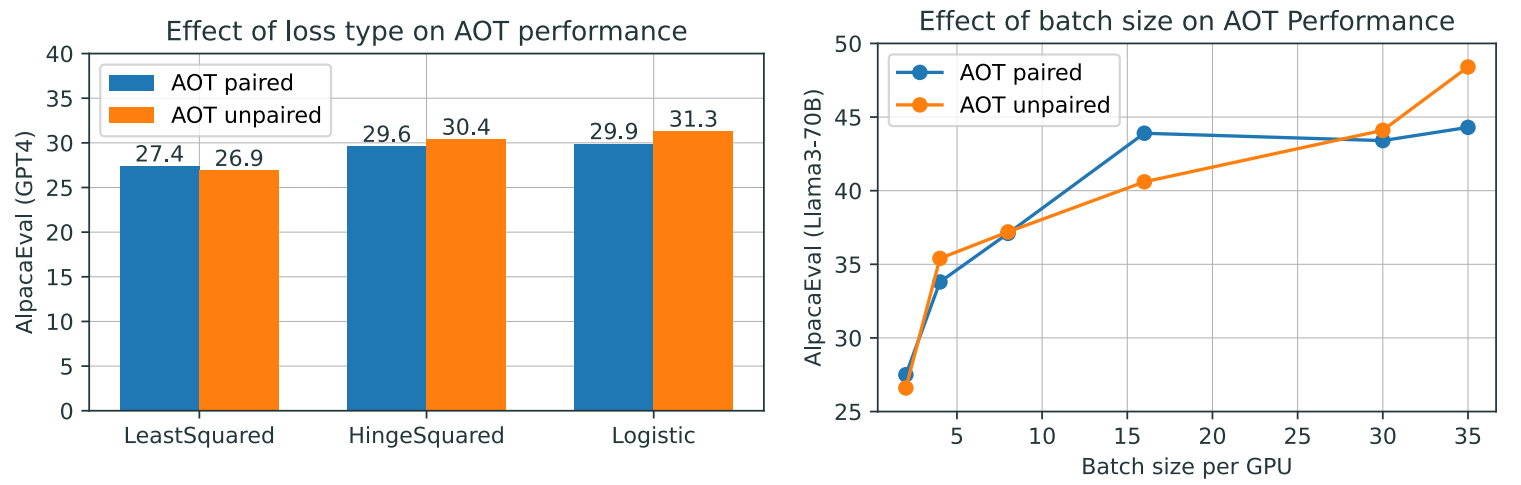



**Figure 2:** Impact of batch size and loss type on AOT performance. The batch size is the effective number of samples in the mini-batch per GPU. We found the logistic loss to be performing better than least squared or hinge squared losses (all using $\beta = 0.01$). As we increase batch size, we also observed improvement in AOT performance, which is expected as more samples per minibatch results in a better effect of stochastic dominance (conforming Corollary 5).

# Ablation Study on $\beta$

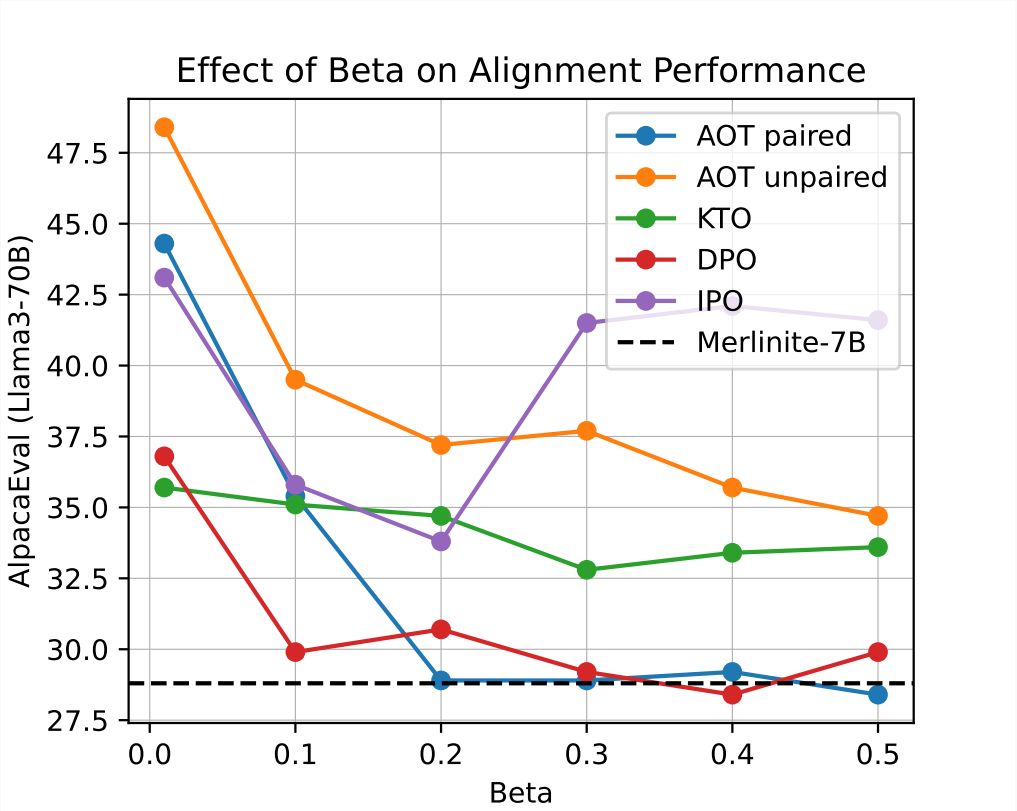



**Figure 3:** Impact of ($\beta$) parameter on performance of different alignment algorithms. $\beta$ controls the divergence of the policy model from the initial reference model (low beta - more divergence, high beta - less divergence). We see a general trend that with higher betas, LLMs alignment decreases the performance. Hence, for all experiments, we selected $\beta = 0.01$ as a default value.

Thank You !

*mroueh@us.ibm.com*

*Code in TRL:* HuggingFace Link

*Paper on arxiv:* `https://arxiv.org/abs/2406.05882`

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

G. Peyré and M. Cuturi.
**Computational optimal transport.**
*Foundations and Trends in Machine Learning*, 11(5-6):355–607, 2019.
ISSN 1935-8237.
doi: 10.1561/2200000073.

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

F. Santambrogio.
***Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*.**
Birkhäuser, Cham, 2015.
ISBN 9783319208275.
doi: 10.1007/978-3-319-20828-2.

C. Villani.

***Optimal Transport: old and new*, volume 338.**
Springer, 2009.

Y. Zhao, M. Khalman, R. Joshi, S. Narayan, M. Saleh, and P. J. Liu.
**Calibrating sequence likelihood improves conditional language generation.**
In *The Eleventh International Conference on Learning Representations*, 2023.
URL https://openreview.net/forum?id=0qSOodKmJaN.