Mallows meets DPO (Direct Preference Optimization)

In this note, we explain how Mallows ranking models are integrated into DPO [RS23], in the context of RLHF.

(I) Mallows models.

Given n items indexed by $\{1, 2, \cdots, n\}$, $\sigma: \{1, \cdots, n\} \to \{1, \cdots, n\}$ specifies the ranking of these items.

E.g. $\sigma(i) = j$ means that item i has rank j. Smaller the rank is, better the item is.

$$i \succ i' \iff i \text{ wins over } i'$$
$$\iff \sigma(i) < \sigma(i').$$

$$\mathbb{P}(\sigma) \propto \phi^{d(\sigma, \sigma_0)}$$

ϕ: dispersion parameter; σ_0: central ranking.

where $d(\cdot, \cdot)$ is a right invariant metrics. Refer to [T18] for further details.

Examples of "d":

(a) $d(\sigma, \sigma_0) = \sum_i (\sigma(i) - \sigma_0(i))^2$ — Mallows L^2 or Mallows θ model

Such d is called Spearman's rank correlation.

(b) $d(\sigma, \sigma_0) = \text{inv}(\sigma_0 \sigma_0^{-1})$ — Mallows ϕ model

number of inversions

Such d is called Kendall's tau.

(c) $d(\sigma, \sigma_0)$ = minimum transportations taking σ to σ_0 — Ewens model.

Such d is called Cayley distance.

• • • Many other choices of "d", see [D88].

Example (c) is trivial, as $\mathbb{P}(i > i') = \frac{1}{2}$ independent of i, i'.

We focus on Example (a) and (b), where we recover the DPO [RS23] from Example (a), and design a noval DPO from Example (b).

Example (a): The Mallows θ model is also Bradley-Terry with

$$\mathbb{P}(i > i') = \mathbb{P}(\sigma(i) < \sigma(i'))$$

$$= \frac{1}{1 + e^{-2(\sigma_0(i') - \sigma_0(i))\log\phi}}$$

$$:= g(-r(x,i') + r(x,i)),$$

where $g(x) = \frac{1}{1+e^{-x}}$ is the link function

$r(x,i) = +2\sigma_0(i|x)\log\phi$ is the reward

By [RS23], RL gives $r(x,i) = \beta \log \frac{\pi(i|x)}{\pi_{ref}(i|x)} + \beta \log 2$

$$\mathbb{P}(i > i' | x) = g(r(x,i) - r(x,i'))$$

This recovers the DPO in [RS23].

Example ⑥. It is known that [M57]:

$$\mathbb{P}(i > i') = \mathbb{P}(\sigma(i) < \sigma(i'))$$

$$= \begin{cases} \dfrac{\sigma_0(i') - \sigma_0(i) + 1}{1 - e^{(\sigma_0(i') - \sigma_0(i) + 1)\log\phi}} - \dfrac{\sigma_0(i') - \sigma_0(i)}{1 - e^{(\sigma_0(i') - \sigma_0(i))\log\phi}} & \text{if } \sigma_0(i') > \sigma_0(i) \\ 1 - \quad " & \text{if not} \end{cases}$$

$$=: g(-r(x,i') + r(x,i)),$$

where $g(x) = \begin{cases} \frac{x+1}{1-\phi^{x+1}} - \frac{x}{1-\phi^x} & x>0 \\ 1- \quad " & x<0 \end{cases}$ is the link function

$r(x,i) = -\sigma_0(i|x)$ is the reward.

Same as before, replacing $r(x,i)$ with $\beta \log \frac{\pi(i|x)}{\pi_{ref}(i|x)} + \beta \log Z$

$$\mathbb{P}(i \succeq i' | x) = g(r(x,i) - r(x,i'))$$

we derive a novel DPO.

Subtlety: In the link function, there is still the dispersion parameter "ϕ".

One can estimate $\hat{\phi}$ as in the reward model.

This is not hard, see [MB10, Section 3.2].

Ref:

[RS23]. Rafailov et al, Direct Preference Opt ... Neurips '23.

[T18]. Tang, Mallows ranking model ... ICML '18.

[D88]. Diaconis, Group representation in Prob & Statistics IMS lecture note '88.

[M57]. Mallows, Non-null ranking models, Biometrika '57.

[MB10]. Meila & Bao, An exponential model for infinite rankings JMLR '10.