SquaredPO: Displacement-Resistant Extensions of DPO with Nonconvex f-Divergences

Idan Pipano Shoham Sabach Kavosh Asadi
Mohammad Ghavamzadeh

TECHNION
Israel Institute
of Technology

Alignment: we start with an LLM π_{ref}, and wish to align it using a dataset of triplets (x, y_w, y_l).

How do we align?

- In RLHF, there are two steps:

How do we align?

- In RLHF, there are two steps:
  - we first learn a reward model r_ϕ

$$\min_\phi \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[-\log \sigma \left(r_\phi \left(x, y_w\right) - r_\phi \left(x, y_l\right)\right)\right]$$

How do we align?

- In RLHF, there are two steps:
  - we first learn a reward model r_ϕ

$$\min_\phi \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[-\log \sigma \left(r_\phi \left(x, y_w\right) - r_\phi \left(x, y_l\right)\right)\right]$$

  - and then train the LLM π_θ

$$\max_\theta \mathbb{E}_x \left[\mathbb{E}_{y\sim\pi_\theta(\cdot|x)} \left[r_\phi \left(x, y\right)\right] - \beta D_{\mathrm{KL}} \left[\pi_\theta \left(\cdot \mid x\right) \parallel \pi_{\mathrm{ref}} \left(\cdot \mid x\right)\right]\right]$$

How do we align?

- In RLHF, there are two steps:
  - we first learn a reward model r_ϕ

$$\min_\phi \mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}} \left[-\log \sigma \left(r_\phi \left(x, y_w\right) - r_\phi \left(x, y_l\right)\right)\right]$$

  - and then train the LLM π_θ

$$\max_\theta \mathbb{E}_x \left[\mathbb{E}_{y\sim\pi_\theta(\cdot|x)} \left[r_\phi \left(x, y\right)\right] - \beta D_{\mathrm{KL}} \left[\pi_\theta \left(\cdot \mid x\right) \parallel \pi_{\mathrm{ref}} \left(\cdot \mid x\right)\right]\right]$$

- DPO collapses the two-step RLHF process into a single loss
  - by substituting the optimal solution of the second RLHF loss into the first one.

f-DPO

- A recent generalization of DPO is f-DPO, obtained by replacing D_{KL} with the more general D_f.

f-DPO

- A recent generalization of DPO is f-DPO, obtained by replacing D_{KL} with the more general D_f.
- **Our first contribution**:

Theorem 1

f can be used in f-DPO if and only if $\lim_{t\to 0^+} f'(t) = -\infty$

f-DPO

- A recent generalization of DPO is f-DPO, obtained by replacing D_{KL} with the more general D_f.
- **Our first contribution**:

Theorem 1

f can be used in f-DPO if and only if $\lim_{t\to 0^+} f'(t) = -\infty$

- In particular, f does not have to be convex!

Likelihood Displacement

- Training an LLM π_θ with DPO causes both $\pi_\theta(y_l \mid x)$ and $\pi_\theta(y_w \mid x)$ (!) to decrease.

Likelihood Displacement

- Training an LLM π_θ with DPO causes both $\pi_\theta(y_l \mid x)$ and $\pi_\theta(y_w \mid x)$ (!) to decrease.
  - We call this phenomenon **likelihood displacement**.

Likelihood Displacement

- Training an LLM π_θ with DPO causes both $\pi_\theta(y_l \mid x)$ and $\pi_\theta(y_w \mid x)$ (!) to decrease.
  - We call this phenomenon **likelihood displacement**.
- Widely observed empirically, poorly understood theoretically.

Likelihood Displacement

- Training an LLM π_θ with DPO causes both $\pi_\theta(y_l \mid x)$ and $\pi_\theta(y_w \mid x)$ (!) to decrease.
  - We call this phenomenon **likelihood displacement**.
- Widely observed empirically, poorly understood theoretically.
- **Our second contribution:**

Theorem 2

A necessary condition on f to mitigate likelihood displacement in f-DPO is $\arg\min_t f(t) \geq 1$.

Likelihood Displacement

- Training an LLM π_θ with DPO causes both $\pi_\theta(y_l \mid x)$ and $\pi_\theta(y_w \mid x)$ (!) to decrease.
  - We call this phenomenon **likelihood displacement**.
- Widely observed empirically, poorly understood theoretically.
- **Our second contribution:**

Theorem 2

A necessary condition on f to mitigate likelihood displacement in f-DPO is $\arg\min_t f(t) \geq 1$.

- In particular, the f of DPO satisfies $\arg\min_t f(t) = e^{-1}$; hence, one can expect likelihood displacement in DPO.

Our method: SquaredPO

- Our theory imposes two desiderata on f for f-DPO.
  - $\lim_{t\to 0^+} f'(t) = -\infty$
  - $\arg\min_t f(t) \geq 1$

Our method: SquaredPO

- Our theory imposes two desiderata on f for f-DPO.
  - $\lim_{t\to 0^+} f'(t) = -\infty$
  - $\arg\min_t f(t) \geq 1$
- $f(t) = (\ln t)^2$ is an example of a function that satisfies them.

Our method: SquaredPO

- Our theory imposes two desiderata on f for f-DPO.
  - $\lim_{t\to 0^+} f'(t) = -\infty$
  - $\arg\min_t f(t) \geq 1$
- $f(t) = (\ln t)^2$ is an example of a function that satisfies them.
- We call the f-DPO loss with this f SquaredPO.

SquaredPO Mitigates Displacement

- Our method: SquaredPO = f-DPO with $f(t) = (\ln t)^2$

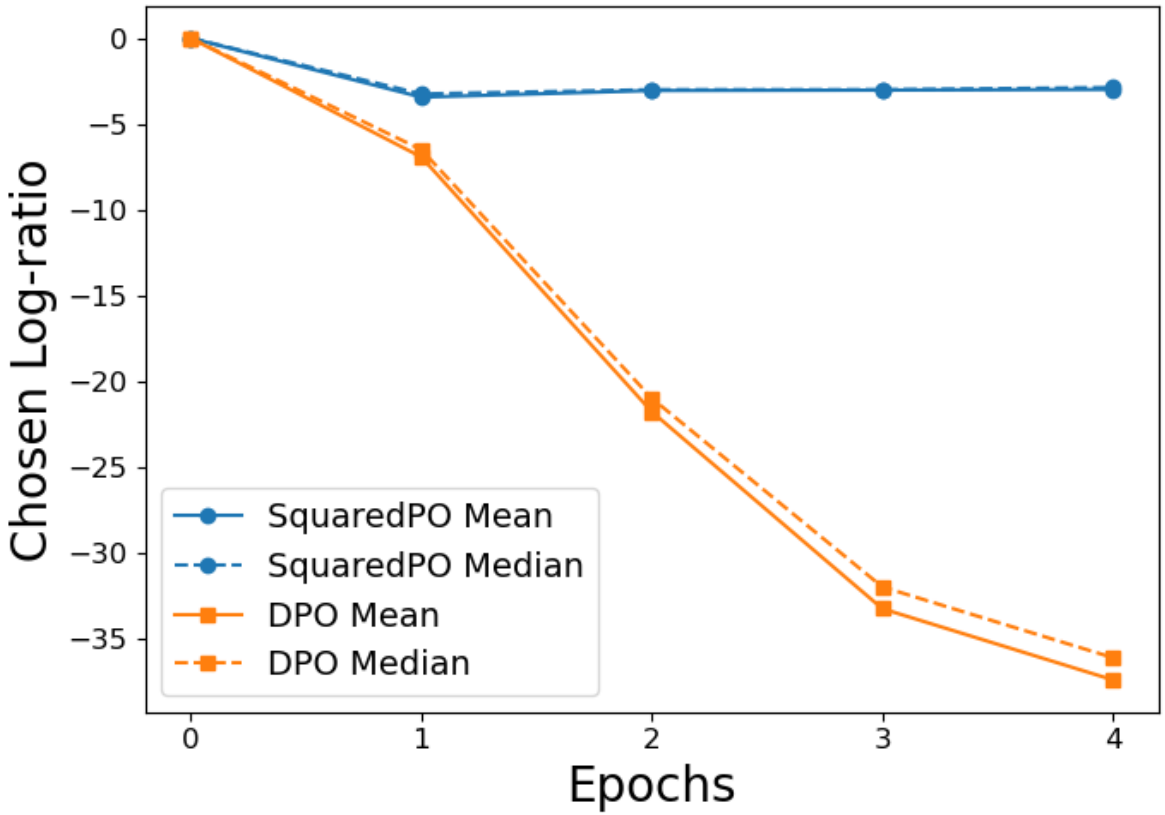

Figure: Chosen Log-ratio is $\log \frac{\pi_\theta(y_w|x)}{\pi_{\text{ref}}(y_w|x)}$.

- Our method: SquaredPO = f-DPO with $f(t) = (\ln t)^2$

Table: Win-rate on the validation set against π_{ref}.

Epochs	SquaredPO (%)	DPO (%)
1	50.8 ± 0.7	51.8 ± 1.0
2	50.6 ± 1.1	45.0 ± 1.1
4	51.0 ± 0.7	34.7 ± 1.3

SquaredPO's performance

- Our method: SquaredPO = f-DPO with $f(t) = (\ln t)^2$

Table: Win-rate on the validation set against π_{ref}.

Epochs	SquaredPO (%)	DPO (%)
1	50.8 ± 0.7	51.8 ± 1.0
2	50.6 ± 1.1	45.0 ± 1.1
4	51.0 ± 0.7	34.7 ± 1.3

- SquaredPO achieves performance comparable to DPO while being more robust to the number of training epochs.

TECHNION
Israel Institute
of Technology

Thank You for Listening!

The Paper 👆