Understanding the Gaps between Two-stage and Direct Preference-based Policy Learning

Ruizhe Shi

University of Washington

January 30, 2026

A quick overview of preference-based policy learning

Reward-based policy learning

- **State set** (prompt): $\mathcal{X}$.
- **Trajectory set** (response): $\mathcal{Y}$.
- **Reference policy** (base model): $\pi_{\text{ref}} : \mathcal{X} \to \Delta(\mathcal{Y})$.
- **Reward oracle**: $r^\star : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

A quick overview of preference-based policy learning

Reward-based policy learning

- **State set** (prompt): $\mathcal{X}$.
- **Trajectory set** (response): $\mathcal{Y}$.
- **Reference policy** (base model): $\pi_{\text{ref}} : \mathcal{X} \to \Delta(\mathcal{Y})$.
- **Reward oracle**: $r^\star : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

Maximize the objective function below:

$$V_{\pi_\theta}^{r^\star} := \mathbb{E}_{x\sim\rho, y\sim\pi_\theta(\cdot|x)} \left[\underbrace{r^\star(x,y)}_{\text{maximize reward}} - \beta \underbrace{\mathrm{KL}\left(\pi_\theta(\cdot|x) \| \pi_{\text{ref}}(\cdot|x)\right)}_{\text{deviation penalty}} \right].$$

A quick overview of preference-based policy learning

Reward-based policy learning

- **State set** (prompt): $\mathcal{X}$.
- **Trajectory set** (response): $\mathcal{Y}$.
- **Reference policy** (base model): $\pi_{\mathsf{ref}} : \mathcal{X} \to \Delta(\mathcal{Y})$.
- **Reward oracle**: $r^\star : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

Maximize the objective function below:

$$V_{\pi_\theta}^{r^\star} := \mathbb{E}_{x\sim\rho, y\sim\pi_\theta(\cdot|x)} \left[\underbrace{r^\star(x,y)}_{\text{maximize reward}} - \beta \underbrace{\mathsf{KL}\left(\pi_\theta(\cdot|x) \| \pi_{\mathsf{ref}}(\cdot|x)\right)}_{\text{deviation penalty}} \right].$$

Closed-form solution

Set **optimal policy** $\pi^\star := \operatorname{argmax}_\pi V_\pi^{r^\star}$, then $\pi^\star(y|x) \propto \pi_{\mathsf{ref}}(y|x)\exp(r^\star(x,y)/\beta)$.

Preference-based policy learning

~~**Reward oracle**: $r^\star : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.~~

Preference annotation: Given x, y_0, y_1, Bradley-Terry model determines a preference signal b:

$$b = \begin{cases} 0 & \text{w.p. } \sigma(r^\star(x, y_0) - r^\star(x, y_1)) \ \textit{(y_0 is preferred)}\,, \\ 1 & \text{w.p. } \sigma(r^\star(x, y_1) - r^\star(x, y_0)) \ \textit{(y_1 is preferred)}\,. \end{cases}$$

Preference-based policy learning

~~**Reward oracle**: $r^\star : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.~~

Preference annotation: Given x, y_0, y_1, Bradley-Terry model determines a preference signal b:

$$b = \begin{cases} 0 & \text{w.p. } \sigma(r^\star(x, y_0) - r^\star(x, y_1)) \ (y_0 \textit{ is preferred}) \ , \\ 1 & \text{w.p. } \sigma(r^\star(x, y_1) - r^\star(x, y_0)) \ (y_1 \textit{ is preferred}) \ . \end{cases}$$

Empirically, we are given a human preference dataset $\mathcal{D} = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^n$, where $y_w^{(i)}$ is preferred to $y_l^{(i)}$ given $x^{(i)}$ following BT model.

Preference-based policy learning

~~Reward oracle: $r^\star : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.~~

Preference annotation: Given x, y_0, y_1, Bradley-Terry model determines a preference signal b:

$$b = \begin{cases} 0 & \text{w.p. } \sigma(r^\star(x, y_0) - r^\star(x, y_1)) \ (y_0 \textit{ is preferred}) , \\ 1 & \text{w.p. } \sigma(r^\star(x, y_1) - r^\star(x, y_0)) \ (y_1 \textit{ is preferred}) . \end{cases}$$

Empirically, we are given a human preference dataset $\mathcal{D} = \{x^{(i)}, y_w^{(i)}, y_l^{(i)}\}_{i=1}^n$, where $y_w^{(i)}$ is preferred to $y_l^{(i)}$ given $x^{(i)}$ following BT model.Still maximize the objective function below:

$$V_{\pi_\theta}^{r^\star} := \mathop{\mathbb{E}}_{x \sim \rho, y \sim \pi_\theta(\cdot|x)} \left[r^\star(x, y) - \beta \mathrm{KL}\left(\pi_\theta(\cdot|x) \| \pi_{\text{ref}}(\cdot|x) \right) \right] .$$

Two-stage approach: Reinforcement learning from human feedback (RLHF)

Step 1: Train a reward model r_{RLHF} by optimizing a cross-entropy loss:

$$\mathcal{L}_{\mathrm{RM}}(\phi) = -\frac{1}{n}\sum_{i=1}^{n} \log \sigma\left(r_\phi(x^{(i)}, y_w^{(i)}) - r_\phi(x^{(i)}, y_l^{(i)})\right) \quad \textit{(note } \mathbb{P}(y_w^{(i)} \succ y_l^{(i)}) = \sigma(r^\star(x^{(i)}, y_w^{(i)}) - r^\star(x^{(i)}, y_l^{(i)}))\textit{)}$$

Step 2: Train a policy model π_{RLHF} by RL:

$$\mathcal{J}_{\mathrm{RL}}(\theta) = V_{\pi_\theta}^{r_{\mathrm{RLHF}}} = \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_\theta(\cdot|x)}\left[r_{\mathrm{RLHF}}(x, y) - \beta \mathrm{KL}\left(\pi_\theta(\cdot|x) \| \pi_{\mathrm{ref}}(\cdot|x)\right)\right] .$$

Two-stage approach: Reinforcement learning from human feedback (RLHF)

Step 1: Train a reward model r_{RLHF} by optimizing a cross-entropy loss:

$$\mathcal{L}_{\mathrm{RM}}(\phi) = -\frac{1}{n}\sum_{i=1}^{n} \log\sigma\left(r_\phi(x^{(i)}, y_w^{(i)}) - r_\phi(x^{(i)}, y_l^{(i)})\right) \quad \textit{(note } \mathbb{P}(y_w^{(i)} \succ y_l^{(i)}) = \sigma(r^\star(x^{(i)}, y_w^{(i)}) - r^\star(x^{(i)}, y_l^{(i)}))\textit{)}$$

Step 2: Train a policy model π_{RLHF} by RL:

$$\mathcal{J}_{\mathrm{RL}}(\theta) = V_{\pi_\theta}^{r_{\mathrm{RLHF}}} = \mathop{\mathbb{E}}_{x\sim\rho, y\sim\pi_\theta(\cdot|x)}\left[r_{\mathrm{RLHF}}(x,y) - \beta \mathrm{KL}\left(\pi_\theta(\cdot|x)\|\pi_{\mathrm{ref}}(\cdot|x)\right)\right] .$$

Direct approach: Direct preference optimization (DPO)

Train a policy model π_{DPO} by optimizaing a cross-entropy loss:

$$\mathcal{L}_{\mathrm{DPO}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log\sigma\left(\beta\log\frac{\pi_\theta(y_w^{(i)}|x)}{\pi_{\mathrm{ref}}(y_w^{(i)}|x)} - \beta\log\frac{\pi_\theta(y_l^{(i)}|x)}{\pi_{\mathrm{ref}}(y_l^{(i)}|x)}\right) .$$

Direct approach: Direct preference optimization (DPO)

Train a policy model π_{DPO} by optimizing a cross-entropy loss:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log \sigma \left(\beta \log \frac{\pi_\theta(y_w^{(i)}|x)}{\pi_{\text{ref}}(y_w^{(i)}|x)} - \beta \log \frac{\pi_\theta(y_l^{(i)}|x)}{\pi_{\text{ref}}(y_l^{(i)}|x)} \right)$$

$$= -\frac{1}{n}\sum_{i=1}^{n} \log \sigma(\hat{r}_\theta(x^{(i)}, y_w^{(i)}) - \hat{r}_\theta(x^{(i)}, y_l^{(i)}))$$

Key idea:

- Set $\hat{r}_\theta(x, y) := \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$ as a surrogate reward model.

Direct approach: Direct preference optimization (DPO)

Train a policy model π_{DPO} by optimizing a cross-entropy loss:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log\sigma\left(\beta\log\frac{\pi_\theta(y_w^{(i)}|x)}{\pi_{\text{ref}}(y_w^{(i)}|x)} - \beta\log\frac{\pi_\theta(y_l^{(i)}|x)}{\pi_{\text{ref}}(y_l^{(i)}|x)}\right)$$
$$= -\frac{1}{n}\sum_{i=1}^{n} \log\sigma(\hat{r}_\theta(x^{(i)}, y_w^{(i)}) - \hat{r}_\theta(x^{(i)}, y_l^{(i)}))$$

Key idea:

- Set $\hat{r}_\theta(x, y) := \beta\log\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$ as a surrogate reward model.
- $\hat{r}_{\text{DPO}}(x, y) := \beta\log\frac{\pi_{\text{DPO}}(y|x)}{\pi_{\text{ref}}(y|x)}$ is learned in the same way as r_{RLHF}.

Direct approach: Direct preference optimization (DPO)

Train a policy model π_{DPO} by optimizing a cross-entropy loss:

$$\mathcal{L}_{\text{DPO}}(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log\sigma\left(\beta\log\frac{\pi_\theta(y_w^{(i)}|x)}{\pi_{\text{ref}}(y_w^{(i)}|x)} - \beta\log\frac{\pi_\theta(y_l^{(i)}|x)}{\pi_{\text{ref}}(y_l^{(i)}|x)}\right)$$
$$= -\frac{1}{n}\sum_{i=1}^{n} \log\sigma(\hat{r}_\theta(x^{(i)}, y_w^{(i)}) - \hat{r}_\theta(x^{(i)}, y_l^{(i)}))$$

Key idea:

- Set $\hat{r}_\theta(x,y) := \beta\log\frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$ as a surrogate reward model.
- $\hat{r}_{\text{DPO}}(x,y) := \beta\log\frac{\pi_{\text{DPO}}(y|x)}{\pi_{\text{ref}}(y|x)}$ is learned in the same way as r_{RLHF}.
- We have $\pi_{\text{DPO}}(y|x) \propto \pi_{\text{ref}}(y|x)\exp(\hat{r}_{\text{DPO}}(y|x)/\beta) \implies \pi_{\text{DPO}} = \operatorname{argmax}_\pi V_\pi^{\hat{r}_{\text{DPO}}}$.

Direct approach: Direct preference optimization (DPO)

Train a policy model π_{DPO} by optimizing a cross-entropy loss:

$$\begin{aligned}
\mathcal{L}_{\text{DPO}}(\theta) &= -\frac{1}{n}\sum_{i=1}^{n} \log\sigma\left(\beta \log\frac{\pi_\theta(y_w^{(i)}|x)}{\pi_{\text{ref}}(y_w^{(i)}|x)} - \beta\log\frac{\pi_\theta(y_l^{(i)}|x)}{\pi_{\text{ref}}(y_l^{(i)}|x)}\right) \\
&= -\frac{1}{n}\sum_{i=1}^{n} \log\sigma(\hat{r}_\theta(x^{(i)}, y_w^{(i)}) - \hat{r}_\theta(x^{(i)}, y_l^{(i)}))
\end{aligned}$$

Key idea:

- Set $\hat{r}_\theta(x, y) := \beta \log \frac{\pi_\theta(y|x)}{\pi_{\text{ref}}(y|x)}$ as a surrogate reward model.
- $\hat{r}_{\text{DPO}}(x, y) := \beta \log \frac{\pi_{\text{DPO}}(y|x)}{\pi_{\text{ref}}(y|x)}$ is learned in the same way as r_{RLHF}.
- We have $\pi_{\text{DPO}}(y|x) \propto \pi_{\text{ref}}(y|x)\exp(\hat{r}_{\text{DPO}}(y|x)/\beta) \implies \pi_{\text{DPO}} = \operatorname{argmax}_\pi V_\pi^{\hat{r}_{\text{DPO}}}$.

DPO can also be online, covered later.

Central Question

Which approach is better?

Starting point: Infinite data and computation

- With no restriction on representation power, solutions of RLHF and DPO are **equivalent**:

Starting point: Infinite data and computation

- With no restriction on representation power, solutions of RLHF and DPO are **equivalent**:

$$\textbf{Trivially,}\ \text{RLHF}\begin{cases} r_{\text{RLHF}} = r^\star\ , \\ \pi_{\text{RLHF}} = \underset{\pi}{\text{argmax}}\, V^{\pi}_{r^\star} = \pi^\star\ ; \end{cases} \ \text{DPO}\begin{cases} \hat{r}_{\text{DPO}} = r^\star\ , \\ \pi_{\text{DPO}} = \underset{\pi}{\text{argmax}}\, V^{\pi}_{r^\star} = \pi^\star\ . \end{cases}$$

Starting point: Infinite data and computation

- With no restriction on representation power, solutions of RLHF and DPO are **equivalent**:

$$\textbf{Trivially, } \text{RLHF} \begin{cases} r_{\text{RLHF}} = r^\star , \\ \pi_{\text{RLHF}} = \underset{\pi}{\operatorname{argmax}} V_{r^\star}^{\pi} = \pi^\star ; \end{cases} \quad \text{DPO} \begin{cases} \hat{r}_{\text{DPO}} = r^\star , \\ \pi_{\text{DPO}} = \underset{\pi}{\operatorname{argmax}} V_{r^\star}^{\pi} = \pi^\star . \end{cases}$$

What's the benefit of Online DPO (data are iteratively generated by π_θ and then annotated with preference) compared with Offline DPO?

Starting point: Infinite data and computation

- With no restriction on representation power, solutions of RLHF and DPO are **equivalent**:

$$\textbf{Trivially, } \text{RLHF}\begin{cases} r_{\text{RLHF}} = r^\star\,, \\ \pi_{\text{RLHF}} = \operatorname{argmax}_{\pi} V_{r^\star}^{\pi} = \pi^\star\,; \end{cases} \text{DPO}\begin{cases} \hat{r}_{\text{DPO}} = r^\star\,, \\ \pi_{\text{DPO}} = \operatorname{argmax}_{\pi} V_{r^\star}^{\pi} = \pi^\star\,. \end{cases}$$

What's the benefit of Online DPO (data are iteratively generated by π_θ and then annotated with preference) compared with Offline DPO?

  - can enhance the convergence rate of gradient descent in tabular setting. [Theorem 1-4, Shi et al. 2024]
  - has the same gradient as RL up to a second-order deviation. [Theorem 4.1, Feng et al. 2025; Theorem 2]

- To reveal a separation in their solutions, we need to look into the model class:

Model class

- **Reward model class**: $\mathcal{F} = \{r_\phi : \phi \in \mathbb{R}^{d_R}\}$, $d_R \in \mathbb{Z}_+$ is the parameter size;
- **Policy model class**: $\Pi = \{\pi_\theta : \theta \in \mathbb{R}^{d_P}\}$, $d_P \in \mathbb{Z}_+$ is the parameter size.

- To reveal a separation in their solutions, we need to look into the model class:

Model class

- **Reward model class**: $\mathcal{F} = \{r_\phi : \phi \in \mathbb{R}^{d_R}\}$, $d_R \in \mathbb{Z}_+$ is the parameter size;
- **Policy model class**: $\Pi = \{\pi_\theta : \theta \in \mathbb{R}^{d_P}\}$, $d_P \in \mathbb{Z}_+$ is the parameter size.

$$\text{A unification of RLHF and DPO: } \begin{cases} \pi_{\text{RLHF}} = \operatorname*{argmax}_{\pi_\theta \in \Pi} V^{\pi_\theta}_{r_{\text{RLHF}}} \\ \pi_{\text{DPO}} = \operatorname*{argmax}_{\pi_\theta \in \Pi} V^{\pi_\theta}_{\hat{r}_{\text{DPO}}} \end{cases}$$

We are comparing their reward model qualities, and what's the difference?

- To reveal a separation in their solutions, we need to look into the model class:

Model class

- **Reward model class**: $\mathcal{F} = \{r_\phi : \phi \in \mathbb{R}^{d_R}\}$, $d_R \in \mathbb{Z}_+$ is the parameter size;
- **Policy model class**: $\Pi = \{\pi_\theta : \theta \in \mathbb{R}^{d_P}\}$, $d_P \in \mathbb{Z}_+$ is the parameter size.

$$\text{A unification of RLHF and DPO: } \begin{cases} \pi_{\text{RLHF}} = \underset{\pi_\theta \in \Pi}{\operatorname{argmax}}\ V^{\pi_\theta}_{r_{\text{RLHF}}} \leftarrow \textbf{lie in } \mathcal{F} \\ \pi_{\text{DPO}} = \underset{\pi_\theta \in \Pi}{\operatorname{argmax}}\ V^{\pi_\theta}_{\hat{r}_{\text{DPO}}} \leftarrow \textbf{mapped from } \Pi \end{cases}$$

A simple observation:
The reward models r_{RLHF} and $\hat{r}_{\text{DPO}}$ are from different model classes.

Applications of the simple observation

A unification of RLHF and DPO: $\begin{cases} \pi_{\text{RLHF}} = \operatorname*{argmax}_{\pi_\theta \in \Pi} V^{\pi_\theta}_{r_{\text{RLHF}}} \leftarrow \textbf{lie in } \mathcal{F} \\ \pi_{\text{DPO}} = \operatorname*{argmax}_{\pi_\theta \in \Pi} V^{\pi_\theta}_{\hat{r}_{\text{DPO}}} \leftarrow \textbf{mapped from } \Pi \end{cases}$

- $r^\star \notin \mathcal{F}, \pi^\star \in \Pi$, *i.e.*, the reward model is mis-specified$\rightarrow V^{\pi_{\text{RLHF}}}_{r^\star} \leqslant V^{\pi_{\text{DPO}}}_{r^\star}$
  - $\exists$ a bandit environment, s.t. $V^{\pi_{\text{RLHF}}}_{r^\star} < V^{\pi_{\text{DPO}}}_{r^\star}$. [Prop. 5]

Applications of the simple observation

A unification of RLHF and DPO: $\begin{cases} \pi_{\text{RLHF}} = \operatorname{argmax}_{\pi_\theta \in \Pi} V^{\pi_\theta}_{r_{\text{RLHF}}} \leftarrow \textbf{lie in } \mathcal{F} \\ \pi_{\text{DPO}} = \operatorname{argmax}_{\pi_\theta \in \Pi} V^{\pi_\theta}_{\hat{r}_{\text{DPO}}} \leftarrow \textbf{mapped from } \Pi \end{cases}$

- $r^\star \notin \mathcal{F}, \pi^\star \in \Pi$, *i.e.*, the reward model is mis-specified$\rightarrow V^{\pi_{\text{RLHF}}}_{r^\star} \leqslant V^{\pi_{\text{DPO}}}_{r^\star}$
  - $\exists$ a bandit environment, s.t. $V^{\pi_{\text{RLHF}}}_{r^\star} < V^{\pi_{\text{DPO}}}_{r^\star}$. [Prop. 5]
- $r^\star \in \mathcal{F}, \pi^\star \notin \Pi$, *i.e.*, the policy model is mis-specified$\rightarrow V^{\pi_{\text{RLHF}}}_{r^\star} \geqslant V^{\pi_{\text{DPO}}}_{r^\star}$
  - $\exists$ a bandit environment, s.t. $V^{\pi_{\text{RLHF}}}_{r^\star} > V^{\pi_{\text{DPO}}}_{r^\star}$. [Prop. 3]
  - Online DPO cannot close the gap. [Prop. 4]

Applications of the simple observation

A unification of RLHF and DPO:
$$\begin{cases} \pi_{\text{RLHF}} = \operatorname{argmax}_{\pi_\theta \in \Pi} V^{\pi_\theta}_{r_{\text{RLHF}}} \leftarrow \textbf{lie in } \mathcal{F} \\ \pi_{\text{DPO}} = \operatorname{argmax}_{\pi_\theta \in \Pi} V^{\pi_\theta}_{\hat{r}_{\text{DPO}}} \leftarrow \textbf{mapped from } \Pi \end{cases}$$

- $r^\star \notin \mathcal{F}, \pi^\star \in \Pi$, *i.e.*, the reward model is mis-specified$\rightarrow V^{\pi_{\text{RLHF}}}_{r^\star} \leqslant V^{\pi_{\text{DPO}}}_{r^\star}$
  - $\exists$ a bandit environment, s.t. $V^{\pi_{\text{RLHF}}}_{r^\star} < V^{\pi_{\text{DPO}}}_{r^\star}$. [Prop. 5]
- $r^\star \in \mathcal{F}, \pi^\star \notin \Pi$, *i.e.*, the policy model is mis-specified$\rightarrow V^{\pi_{\text{RLHF}}}_{r^\star} \geqslant V^{\pi_{\text{DPO}}}_{r^\star}$
  - $\exists$ a bandit environment, s.t. $V^{\pi_{\text{RLHF}}}_{r^\star} > V^{\pi_{\text{DPO}}}_{r^\star}$. [Prop. 3]
  - Online DPO cannot close the gap. [Prop. 4]
- $r^\star \notin \mathcal{F}, \pi^\star \notin \Pi, \mathcal{F} \cong \Pi \rightarrow V^{\pi_{\text{RLHF}}}_{r^\star} = V^{\pi_{\text{DPO}}}_{r^\star}$
  - $\exists$ a bandit environment, s.t. Online DPO outperforms RLHF. [Prop. 7]
  - Online data (carrying the information of the current policy) benefits reward learning.

Following Question

In practice, the access to finite data induces estimation error.
Then which approach is better?

Following Question

In practice, the access to finite data induces estimation error.
Then which approach is better?

Two-stage approach: RLHF

- **Reward learning. (restricted by finite data)**
- Policy optimization. (no information bottleneck)

Direct approach: DPO

- **Surrogate reward learning. (restricted by finite data)**
- Policy transformation: $\pi_{\text{DPO}}(y|x) \propto \pi_{\text{ref}}(y|x) \exp(\hat{r}_{\text{DPO}}(x, y)/\beta)$. (directly optimal)

Optimal solution under token-level parameterization

- **Optimal reward**:

$$r^\star(x, y) \ .$$

- **Optimal policy**:

Optimal solution under token-level parameterization

- **Optimal reward**:

$$r^\star(x, y) \ .$$

- **Optimal policy**:

$$\pi^\star(y_t|x, y_{0\dots t-1}) \propto \pi_{\text{ref}}(y_t|x, y_{0\dots t-1}) \exp\left(\frac{q^\star(y_t|x, y_{0\dots t-1})}{\beta}\right) \ ,$$

where $q^\star$ is the **soft Q function**:

$$q^\star(y_t|x, y_{0\dots t-1}) = \begin{cases} \beta \log \sum_{s \in \mathcal{V}} \pi_{\text{ref}}(s|x, y_{0\dots t}) \exp(q^\star(s|x, y_{0\dots t})/\beta) & y_t \text{ is not terminal toke} \\ r^\star(x, y_{0\dots t}) & y_t \text{ is terminal token.} \end{cases}$$

Optimal solution under token-level parameterization

- **Optimal reward**:

$$r^\star(x, y) \ .$$

- **Optimal policy**:

$$\pi^\star(y_t|x, y_{0\ldots t-1}) \propto \pi_{\text{ref}}(y_t|x, y_{0\ldots t-1}) \exp\left(\frac{q^\star(y_t|x, y_{0\ldots t-1})}{\beta}\right) \ ,$$

where $q^\star$ is the **soft Q function**:

$$q^\star(y_t|x, y_{0\ldots t-1}) = \begin{cases} \beta \log \sum_{s\in\mathcal{V}} \pi_{\text{ref}}(s|x, y_{0\ldots t}) \exp(q^\star(s|x, y_{0\ldots t})/\beta) & y_t \text{ is not terminal toke} \\ r^\star(x, y_{0\ldots t}) & y_t \text{ is terminal token.} \end{cases}$$

- **Observation**: $q^\star$ is harder to estimate than $r^\star$ (the intrinsic structure of the reward function, *e.g.* linearity and sparsity, is distorted).

Target (prompt omitted): $\exists q^\star$, s.t.

$$\pi^\star(y_t|x, y_{0\dots t-1}) \propto \pi_{\text{ref}}(y_t|x, y_{0\dots t-1}) \exp\left(\frac{q^\star(y_t|x, y_{0\dots t-1})}{\beta}\right),$$

and

$$q^\star(y_t|x, y_{0\dots t-1}) = \begin{cases} \beta \log \sum_{s\in\mathcal{V}} \pi_{\text{ref}}(s|x, y_{0\dots t}) \exp(q^\star(s|x, y_{0\dots t})/\beta) & y_t \text{ is not terminal token;} \\ r^\star(x, y_{0\dots t}) & y_t \text{ is terminal token.} \end{cases}$$

Proof. Set the $q^\star$ function as

$$q^\star(y_0) = \beta \log Z + \beta \log \frac{\pi^\star(y_0)}{\pi_{\text{ref}}(y_0)}, \quad q^\star(y_t|y_{0\dots t-1}) = q^\star(y_{t-1}|y_{0\dots t-2}) + \beta \log \frac{\pi^\star(y_t|y_{0\dots t-1})}{\pi_{\text{ref}}(y_t|y_{0\dots t-1})}.$$

Target (prompt omitted): $\exists q^\star$, s.t.

$$\pi^\star(y_t|x, y_{0\ldots t-1}) \propto \pi_{\text{ref}}(y_t|x, y_{0\ldots t-1}) \exp\left(\frac{q^\star(y_t|x, y_{0\ldots t-1})}{\beta}\right),$$

and

$$q^\star(y_t|x, y_{0\ldots t-1}) = \begin{cases} \beta \log \sum_{s\in\mathcal{V}} \pi_{\text{ref}}(s|x, y_{0\ldots t}) \exp(q^\star(s|x, y_{0\ldots t})/\beta) & y_t \text{ is not terminal token;} \\ r^\star(x, y_{0\ldots t}) & y_t \text{ is terminal token.} \end{cases}$$

Proof. Set the $q^\star$ function as

$$q^\star(y_0) = \beta \log Z + \beta \log \frac{\pi^\star(y_0)}{\pi_{\text{ref}}(y_0)}, \quad q^\star(y_t|y_{0\ldots t-1}) = q^\star(y_{t-1}|y_{0\ldots t-2}) + \beta \log \frac{\pi^\star(y_t|y_{0\ldots t-1})}{\pi_{\text{ref}}(y_t|y_{0\ldots t-1})}.$$

Then we have $\pi_{\text{ref}}(y_t|y_{0\ldots t-1}) \exp\left(\frac{q^\star(y_t|y_{0\ldots t-1}) - q^\star(y_{t-1}|y_{0\ldots t-2})}{\beta}\right) = \pi^\star(y_t|y_{0\ldots t-1})$

Target (prompt omitted): $\exists q^\star$, s.t.

$$\pi^\star(y_t|x, y_{0\dots t-1}) \propto \pi_{\text{ref}}(y_t|x, y_{0\dots t-1}) \exp\left(\frac{q^\star(y_t|x, y_{0\dots t-1})}{\beta}\right) ,$$

and

$$q^\star(y_t|x, y_{0\dots t-1}) = \begin{cases} \beta \log \sum_{s\in\mathcal{V}} \pi_{\text{ref}}(s|x, y_{0\dots t}) \exp(q^\star(s|x, y_{0\dots t})/\beta) & y_t \text{ is not terminal token;} \\ r^\star(x, y_{0\dots t}) & y_t \text{ is terminal token.} \end{cases}$$

Proof. Set the $q^\star$ function as

$$q^\star(y_0) = \beta \log Z + \beta \log \frac{\pi^\star(y_0)}{\pi_{\text{ref}}(y_0)} , \; q^\star(y_t|y_{0\dots t-1}) = q^\star(y_{t-1}|y_{0\dots t-2}) + \beta \log \frac{\pi^\star(y_t|y_{0\dots t-1})}{\pi_{\text{ref}}(y_t|y_{0\dots t-1})} .$$

Then we have $\pi_{\text{ref}}(y_t|y_{0\dots t-1}) \exp\left(\frac{q^\star(y_t|y_{0\dots t-1}) - q^\star(y_{t-1}|y_{0\dots t-2})}{\beta}\right) = \pi^\star(y_t|y_{0\dots t-1})$, and thus $\sum_s \pi_{\text{ref}}(s|y_{0\dots t-1}) \exp\left(\frac{q^\star(y_t|y_{0\dots t-1}) - q^\star(y_{t-1}|y_{0\dots t-2})}{\beta}\right) = 1$ (just sum up)

Target (prompt omitted): $\exists q^\star$, s.t.

$$\pi^\star(y_t|x, y_{0\ldots t-1}) \propto \pi_{\text{ref}}(y_t|x, y_{0\ldots t-1}) \exp\left(\frac{q^\star(y_t|x, y_{0\ldots t-1})}{\beta}\right),$$

and

$$q^\star(y_t|x, y_{0\ldots t-1}) = \begin{cases} \beta \log \sum_{s\in\mathcal{V}} \pi_{\text{ref}}(s|x, y_{0\ldots t}) \exp(q^\star(s|x, y_{0\ldots t})/\beta) & y_t \text{ is not terminal token;} \\ r^\star(x, y_{0\ldots t}) & y_t \text{ is terminal token.} \end{cases}$$

Proof. Set the $q^\star$ function as

$$q^\star(y_0) = \beta \log Z + \beta \log \frac{\pi^\star(y_0)}{\pi_{\text{ref}}(y_0)}, \quad q^\star(y_t|y_{0\ldots t-1}) = q^\star(y_{t-1}|y_{0\ldots t-2}) + \beta \log \frac{\pi^\star(y_t|y_{0\ldots t-1})}{\pi_{\text{ref}}(y_t|y_{0\ldots t-1})}.$$

Then we have $\pi_{\text{ref}}(y_t|y_{0\ldots t-1}) \exp\left(\frac{q^\star(y_t|y_{0\ldots t-1}) - q^\star(y_{t-1}|y_{0\ldots t-2})}{\beta}\right) = \pi^\star(y_t|y_{0\ldots t-1})$, and thus $\sum_s \pi_{\text{ref}}(s|y_{0\ldots t-1}) \exp\left(\frac{q^\star(y_t|y_{0\ldots t-1}) - q^\star(y_{t-1}|y_{0\ldots t-2})}{\beta}\right) = 1$ (just sum up), which yields:

$$q^\star(y_{t-1}|y_{0\ldots t-2}) = \beta \log \sum_s \pi_{\text{ref}}(s|y_{0\ldots t-1}) \exp(q^\star(s|y_{0\ldots t-1})/\beta). \text{ (non-terminal token)}$$

Target (prompt omitted): $\exists q^\star$, s.t.

$$\pi^\star(y_t|x, y_{0\ldots t-1}) \propto \pi_{\text{ref}}(y_t|x, y_{0\ldots t-1}) \exp\left(\frac{q^\star(y_t|x, y_{0\ldots t-1})}{\beta}\right),$$

and

$$q^\star(y_t|x, y_{0\ldots t-1}) = \begin{cases} \beta \log \sum_{s\in\mathcal{V}} \pi_{\text{ref}}(s|x, y_{0\ldots t}) \exp(q^\star(s|x, y_{0\ldots t})/\beta) & y_t \text{ is not terminal token;} \\ r^\star(x, y_{0\ldots t}) & y_t \text{ is terminal token.} \end{cases}$$

Proof. (con'd) Recall the $q^\star$ function as

$$q^\star(y_0) = \beta \log Z + \beta \log \frac{\pi^\star(y_0)}{\pi_{\text{ref}}(y_0)}, \quad q^\star(y_t|y_{0\ldots t-1}) = q^\star(y_{t-1}|y_{0\ldots t-2}) + \beta \log \frac{\pi^\star(y_t|y_{0\ldots t-1})}{\pi_{\text{ref}}(y_t|y_{0\ldots t-1})}.$$

Target (prompt omitted): $\exists q^\star$, s.t.

$$\pi^\star(y_t|x, y_{0\ldots t-1}) \propto \pi_{\text{ref}}(y_t|x, y_{0\ldots t-1}) \exp\left(\frac{q^\star(y_t|x, y_{0\ldots t-1})}{\beta}\right) ,$$

and

$$q^\star(y_t|x, y_{0\ldots t-1}) = \begin{cases} \beta \log \sum_{s\in\mathcal{V}} \pi_{\text{ref}}(s|x, y_{0\ldots t}) \exp(q^\star(s|x, y_{0\ldots t})/\beta) & y_t \text{ is not terminal token;} \\ r^\star(x, y_{0\ldots t}) & y_t \text{ is terminal token.} \end{cases}$$

Proof. (con'd) Recall the $q^\star$ function as

$$q^\star(y_0) = \beta \log Z + \beta \log \frac{\pi^\star(y_0)}{\pi_{\text{ref}}(y_0)} , \quad q^\star(y_t|y_{0\ldots t-1}) = q^\star(y_{t-1}|y_{0\ldots t-2}) + \beta \log \frac{\pi^\star(y_t|y_{0\ldots t-1})}{\pi_{\text{ref}}(y_t|y_{0\ldots t-1})} .$$

And for a y with y_N as the terminal token, note that $\pi^\star(y) = \frac{1}{Z}\pi_{\text{ref}}(y)\exp(r^\star(y)/\beta)$

Target (prompt omitted): $\exists q^\star$, s.t.

$$\pi^\star(y_t|x, y_{0\ldots t-1}) \propto \pi_{\text{ref}}(y_t|x, y_{0\ldots t-1}) \exp\left(\frac{q^\star(y_t|x, y_{0\ldots t-1})}{\beta}\right) ,$$

and

$$q^\star(y_t|x, y_{0\ldots t-1}) = \begin{cases} \beta \log \sum_{s\in\mathcal{V}} \pi_{\text{ref}}(s|x, y_{0\ldots t}) \exp(q^\star(s|x, y_{0\ldots t})/\beta) & y_t \text{ is not terminal token;} \\ r^\star(x, y_{0\ldots t}) & y_t \text{ is terminal token.} \end{cases}$$

Proof. (con'd) Recall the $q^\star$ function as

$$q^\star(y_0) = \beta \log Z + \beta \log \frac{\pi^\star(y_0)}{\pi_{\text{ref}}(y_0)} , \quad q^\star(y_t|y_{0\ldots t-1}) = q^\star(y_{t-1}|y_{0\ldots t-2}) + \beta \log \frac{\pi^\star(y_t|y_{0\ldots t-1})}{\pi_{\text{ref}}(y_t|y_{0\ldots t-1})} .$$

And for a y with y_N as the terminal token, note that $\pi^\star(y) = \frac{1}{Z}\pi_{\text{ref}}(y)\exp(r^\star(y)/\beta)$, we have:

$$r^\star(y) = \beta \log Z + \beta \log \frac{\pi^\star(y)}{\pi_{\text{ref}}(y)} = \beta \log Z + \beta \log \frac{\pi^\star(y_0)}{\pi_{\text{ref}}(y_0)} + \sum_{t=1}^{N} \beta \log \frac{\pi^\star(y_t|y_{0\ldots t-1})}{\pi_{\text{ref}}(y_t|y_{0\ldots t-1})}$$

$$= q^\star(y_0) + \sum_{t=1}^{N} q^\star(y_t|y_{0\ldots t-1}) - q^\star(y_{t-1}|y_{0\ldots t-2}) = q^\star(y_N|y_{0\ldots N-1}) . \text{ (terminal token)} \quad \square$$

A representative token-level parameterization (prompt omitted)

Reward Model: The common reward model shares the same architecture with LM but replaces the last layer with a linear head (here $\theta_t, \psi(y_{0\dots t} \in \mathbb{R}^d)$:

$$r_\theta(y) = \beta \sum_{t=0}^{N} \theta_t^\top \psi(y_{0\dots t}) \; .$$

Policy Model: One needs to go through the softmax results of all tokens and multiply them:

$$\pi_\theta(y) = \prod_{t=0}^{N} \pi_\theta(y_t | y_{0\dots t-1}) = \prod_{t=0}^{N} \frac{\pi_{\text{ref}}(y_t|y_{0\dots t-1}) \exp(\theta_t^\top \psi(y_{0\dots t}))}{\sum_s \pi_{\text{ref}}(s|y_{0\dots t-1}) \exp(\theta_t^\top \psi(y_{0\dots t-1}, s))} \; .$$

- $\theta_r^\star$: the optimal solution for reward learning;
- $\theta_p^\star$: the optimal solution for policy learning.

Difference in solution structure

Dual-token Sparse Prediction (DTSP)

The policy is required to sequentially output two tokens y, ω, and the ground-truth reward is:

$$r^\star(y, \omega) = \beta \mathbf{r}_{\text{sparse}}^\top \psi(y) + \beta {\mathbf{e}_1}^\top \psi(y, \omega) \ ,$$

where $\psi(y), \psi(y, \omega) \in \mathbb{R}^d$, $\mathbf{r}_{\text{sparse}}, \mathbf{r}_{\text{dense}} \in \mathbb{R}^d$, $\|\mathbf{r}_{\text{sparse}}\|_0 = k$, $k \ll d$.

Difference in solution structure

Dual-token Sparse Prediction (DTSP)

The policy is required to sequentially output two tokens y, ω, and the ground-truth reward is:

$$r^\star(y,\omega) = \beta \mathbf{r}_{\text{sparse}}^\top \psi(y) + \beta {\mathbf{e}_1}^\top \psi(y,\omega) \ ,$$

where $\psi(y), \psi(y,\omega) \in \mathbb{R}^d$, $\mathbf{r}_{\text{sparse}}, \mathbf{r}_{\text{dense}} \in \mathbb{R}^d$, $\|\mathbf{r}_{\text{sparse}}\|_0 = k$, $k \ll d$.

For the second token, $\theta_r^\star$ and $\theta_p^\star$ share the same optimal solution:

$$(\theta_{r,1}^\star)^\top \psi(y,\omega) = {\mathbf{e}_1}^\top \psi(y,\omega) + C_1 \ , \ (\theta_{p,1}^\star)^\top \psi(y,\omega) = {\mathbf{e}_1}^\top \psi(y,\omega) + C_2 \ .$$

Difference in solution structure

Dual-token Sparse Prediction (DTSP)

The policy is required to sequentially output two tokens y, ω, and the ground-truth reward is:

$$r^\star(y,\omega) = \beta \mathbf{r}_{\text{sparse}}^\top \psi(y) + \beta {\mathbf{e}_1}^\top \psi(y,\omega) \ ,$$

where $\psi(y), \psi(y,\omega) \in \mathbb{R}^d$, $\mathbf{r}_{\text{sparse}}, \mathbf{r}_{\text{dense}} \in \mathbb{R}^d$, $\|\mathbf{r}_{\text{sparse}}\|_0 = k$, $k \ll d$.

For the second token, $\theta_r^\star$ and $\theta_p^\star$ share the same optimal solution:

$$(\theta_{r,1}^\star)^\top \psi(y,\omega) = {\mathbf{e}_1}^\top \psi(y,\omega) + C_1 \ , \ (\theta_{p,1}^\star)^\top \psi(y,\omega) = {\mathbf{e}_1}^\top \psi(y,\omega) + C_2 \ .$$

While for the first token y, there is a distinction:

$$(\theta_{r,0}^\star)^\top \psi(y) = \mathbf{r}_{\text{sparse}}^\top \psi(y) + C_3 \ , \ (\theta_{p,0}^\star)^\top \psi(y) = \mathbf{r}_{\text{sparse}}^\top \psi(y) + \underbrace{\log \mathop{\mathbb{E}}_{w \sim \pi_{\text{ref}}(\cdot|y)} \exp(\psi(y,\omega)_1)}_{\text{log partition function}} + C_4 \ ,$$

Difference in solution structure

$$(\theta^\star_{r,0})^\top \psi(y) = \mathbf{r}_{\text{sparse}}^\top \psi(y) + C_3 \ , \ (\theta^\star_{p,0})^\top \psi(y) = \mathbf{r}_{\text{sparse}}^\top \psi(y) + \underbrace{\log \mathop{\mathbb{E}}_{w \sim \pi_{\text{ref}}(\cdot|y)} \exp(\psi(y,\omega)_1)}_{\text{log partition function}} + C_4 \ ,$$

The log partition function can be

- non-linear function of $\psi(y) \to$ DPO is prone to model mis-specification, and thus requires a large parameter size;

Difference in solution structure

$$(\theta^\star_{r,0})^\top \psi(y) = \mathbf{r}_{\text{sparse}}^\top \psi(y) + C_3\ ,\ (\theta^\star_{p,0})^\top \psi(y) = \mathbf{r}_{\text{sparse}}^\top \psi(y) + \underbrace{\log \mathop{\mathbb{E}}_{w\sim\pi_{\text{ref}}(\cdot|y)} \exp(\psi(y,\omega)_1)}_{\text{log partition function}} + C_4\ ,$$

The log partition function can be

- non-linear function of $\psi(y)$ → DPO is prone to model mis-specification, and thus requires a large parameter size;
- **dense linear function of $\psi(y)$ → DPO can not efficiently leverage sparsity.**

Difference in solution structure

$$(\theta^\star_{r,0})^\top \psi(y) = \mathbf{r}_{\text{sparse}}^\top \psi(y) + C_3 \ , \ (\theta^\star_{p,0})^\top \psi(y) = \mathbf{r}_{\text{sparse}}^\top \psi(y) + \underbrace{\log \mathop{\mathbb{E}}_{w \sim \pi_{\text{ref}}(\cdot|y)} \exp(\psi(y,\omega)_1)}_{\text{log partition function}} + C_4 \ ,$$

The log partition function can be

- non-linear function of $\psi(y) \to$ DPO is prone to model mis-specification, and thus requires a large parameter size;
- **dense linear function of $\psi(y) \to$ DPO can not efficiently leverage sparsity.**

Is there a sample complexity separation between reward learning and DPO?

Task construction

- Recall

$$(\theta^{\star}_{r,0})^{\top}\psi(y) = \mathbf{r}_{\text{sparse}}^{\top}\psi(y) + C_3 \ ,$$
$$(\theta^{\star}_{p,0})^{\top}\psi(y) = \mathbf{r}_{\text{sparse}}^{\top}\psi(y) + \log \mathop{\mathbb{E}}_{w\sim\pi_{\text{ref}}(\cdot|y)} \exp(\psi(y,\omega)_1) + C_4 \ ;$$

Task construction

- Recall

$$(\theta^\star_{r,0})^\top \psi(y) = \mathbf{r}_{\text{sparse}}^\top \psi(y) + C_3 \ ,$$
$$(\theta^\star_{p,0})^\top \psi(y) = \mathbf{r}_{\text{sparse}}^\top \psi(y) + \log \mathop{\mathbb{E}}_{w\sim\pi_{\text{ref}}(\cdot|y)} \exp(\psi(y,\omega)_1) + C_4 \ ;$$

- Set $\psi(y,\omega) = \psi(\omega) + \left(\mathbf{r}_{\text{dense}}^\top \psi(y)\right) \mathbf{e}_1$, and $\pi_{\text{ref}}(\cdot|y)$ as uniform;

Task construction

- Recall

$$(\theta^\star_{r,0})^\top \psi(y) = \mathbf{r}_{\text{sparse}}^\top \psi(y) + C_3 \ ,$$
$$(\theta^\star_{p,0})^\top \psi(y) = \mathbf{r}_{\text{sparse}}^\top \psi(y) + \log \mathop{\mathbb{E}}_{w\sim\pi_{\text{ref}}(\cdot|y)} \exp(\psi(y,\omega)_1) + C_4 \ ;$$

- Set $\psi(y,\omega) = \psi(\omega) + \left(\mathbf{r}_{\text{dense}}^\top \psi(y)\right) \mathbf{e}_1$, and $\pi_{\text{ref}}(\cdot|y)$ as uniform;
- Then we have

$$(\theta^\star_{r,0})^\top \psi(y) = \mathbf{r}_{\text{sparse}}^\top \psi(y) + C_3 \ ,$$
$$(\theta^\star_{p,0})^\top \psi(y) = (\underbrace{\mathbf{r}_{\text{sparse}} + \mathbf{r}_{\text{dense}}}_{\text{sparsity is distorted}})^\top \psi(y) + C_4 \ .$$

Technical tools: estimation for single token

Definition (Reward quality measure: Data-induced semi-norm)

Under single-token setting, the empirical error of an estimate $\hat{\theta}$ is defined as $\|\hat{\theta} - \theta^\star\|^2_{\Sigma_\mathcal{D}} := \frac{1}{n}\sum_{i=1}^n \left[(r_{\hat{\theta}}(y_w^{(i)}) - r_{\hat{\theta}}(y_l^{(i)})) - (r^\star(y_w^{(i)}) - r^\star(y_l^{(i)}))\right]^2$, where $\Sigma_\mathcal{D} := \frac{1}{n}\sum_{i=1}^n (\psi(y_w^{(i)}) - \psi(y_l^{(i)}))(\psi(y_w^{(i)}) - \psi(y_l^{(i)}))^\top$ is the **Gram matrix**.

Technical tools: estimation for single token

Definition (Reward quality measure: Data-induced semi-norm)

Under single-token setting, the empirical error of an estimate $\hat{\theta}$ is defined as
$\|\hat{\theta} - \theta^\star\|^2_{\Sigma_\mathcal{D}} := \frac{1}{n}\sum_{i=1}^n \left[(r_{\hat{\theta}}(y_w^{(i)}) - r_{\hat{\theta}}(y_l^{(i)})) - (r^\star(y_w^{(i)}) - r^\star(y_l^{(i)}))\right]^2$, where
$\Sigma_\mathcal{D} := \frac{1}{n}\sum_{i=1}^n (\psi(y_w^{(i)}) - \psi(y_l^{(i)}))(\psi(y_w^{(i)}) - \psi(y_l^{(i)}))^\top$ is the **Gram matrix**.

Lemma (Lower bound, Theorem 1.a, Shah et al. 2015)

For a sample size $n \geqslant c_1 tr(\Sigma_\mathcal{D}^\dagger)$, **any estimate** $\hat{\theta}$ *based on n samples* **has a lower bound** *as:*

$$\sup_{\theta^\star \in \Theta_B} \mathbb{E}\left[\|\hat{\theta} - \theta^\star\|^2_{\Sigma_\mathcal{D}}\right] = \Omega\left(\frac{d}{n}\right).$$

Technical tools: estimation for single token

Definition (Reward quality measure: Data-induced semi-norm)

Under single-token setting, the empirical error of an estimate $\hat{\theta}$ is defined as
$\|\hat{\theta} - \theta^\star\|^2_{\Sigma_\mathcal{D}} := \frac{1}{n}\sum_{i=1}^n \left[(r_{\hat{\theta}}(y_w^{(i)}) - r_{\hat{\theta}}(y_l^{(i)})) - (r^\star(y_w^{(i)}) - r^\star(y_l^{(i)}))\right]^2$, where
$\Sigma_\mathcal{D} := \frac{1}{n}\sum_{i=1}^n (\psi(y_w^{(i)}) - \psi(y_l^{(i)}))(\psi(y_w^{(i)}) - \psi(y_l^{(i)}))^\top$ is the **Gram matrix**.

Lemma (Lower bound, Theorem 1.a, Shah et al. 2015)

For a sample size $n \geqslant c_1 tr(\Sigma_\mathcal{D}^\dagger)$, **any estimate** *$\hat{\theta}$ based on n samples* **has a lower bound** *as:*

$$\sup_{\theta^\star \in \Theta_B} \mathbb{E}\left[\|\hat{\theta} - \theta^\star\|^2_{\Sigma_\mathcal{D}}\right] = \Omega\left(\frac{d}{n}\right).$$

Lemma (Upper bound, Lemma 3.1, Zhu et al. 2023)

$$\|\hat{\theta}_{\text{MLE}} - \theta^\star\|^2_{\Sigma_\mathcal{D}} = \mathcal{O}\left(\frac{d + \log(1/\delta)}{n}\right), \ w.p.\ 1-\delta .$$

Technical tools: sparse recovery

Lemma (Theorem 3.3, Yao et al. 2025)

Consider $\|\theta^\star\|_0 = k$, $k \ll d$, *the* ℓ_1**-regularized estimate** $\hat{\theta}_{\ell_1}$*:*

$$\hat{\theta}_{\ell_1} \in \underset{\theta \in \Theta_B}{\operatorname{argmin}} \mathcal{L}_{\mathrm{MLE}}(\theta) + \gamma \|\theta\|_1 .$$

with an appropriate $\gamma = \Theta\left(\sqrt{\frac{\log(d)+\log(1/\delta)}{n}}\right)$ **has an upper bound** *as:*

$$\|\hat{\theta}_{\ell_1} - \theta^\star\|_{\Sigma_{\mathcal{D}}}^2 = \mathcal{O}\left(\sqrt{\frac{k\log(d) + k\log(1/\delta)}{n}}\right) , \ w.p. \ 1-\delta .$$

Theorem (Separation of RLHF and DPO in sample complexity)

Under token-level linear parameterization and mild assumptions, there exists an environment for DTSP task, s.t. by estimating from a preference dataset $\mathcal{D}$ with n samples under $\theta_1 = e_1$ constraint, the estimation error of the reward model $\hat{\theta}_r$ **can be reduced to** $\tilde{\mathcal{O}}(\sqrt{k \log d/n})$ *using a (computationally efficient) ℓ_1-regularized estimator, i.e., w.p. $1-\delta$,*

$$\frac{1}{n}\sum_{i=1}^{n}\left[(r^\star(y_w^{(i)}) - r^\star(y_l^{(i)})) - (r_{\hat{\theta}_r}(y_w^{(i)}) - r_{\hat{\theta}_r}(y_l^{(i)}))\right]^2 = \mathcal{O}\left(\sqrt{\frac{k\log(d) + k\log(1/\delta)}{n}}\right),$$

while the estimation error of the DPO model $\hat{\theta}_p$ is **lower bounded by** $\Omega(d/n)$*:*

$$\frac{1}{n}\sum_{i=1}^{n}\left[(r^\star(y_w^{(i)}) - r^\star(y_l^{(i)})) - (r_{\hat{\theta}_p}(y_w^{(i)}) - r_{\hat{\theta}_p}(y_l^{(i)}))\right]^2 = \Omega\left(\frac{d}{n}\right).$$

Experimental Verification

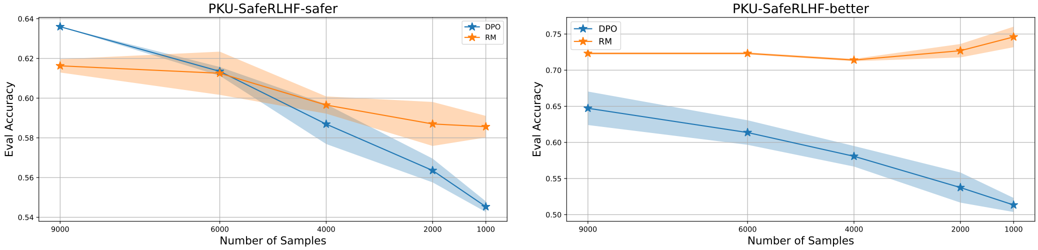

Figure: **Experimental Results on Statistical Efficiency.** We experiment on two preference types, and pure reward learning is shown to be more data-efficient than surrogate reward learning.