

Part I

Appendix

Table of Contents

A Broader impact statement	26
B Additional Discussion of Related Work	27
C Influence Function to Measure the Impact of the data contamination for DPO	27
D Computation of Hölder-DPO Objective	28
E Proofs	28
E.1 Proof for Theorem 2 (The Case of DPO)	28
E.2 Proof for Theorem 3 (The Case of DPO)	29
E.3 Proof for Theorem 4	30
E.4 Proof for Corollary 1	35
F Contamination ratio estimation and outlier detection	35
F.1 Model extension approach to both parameter estimation and contamination rate estimation	35
F.2 Contamination ratio estimation	36
F.3 Estimator of ξ^*	37
F.4 Choice of ϕ	38
G IF Analysis for the DPO variants (summarized in Theorem 3)	40
G.1 rDPO do not satisfy the redescending property	40
G.2 cDPO do not satisfy the redescending property	42
G.3 IPO do not satisfy the redescending property	43
G.4 Dr. DPO do not satisfy the redescending property	44
H Additional Experimental details	47
H.1 Dataset and model details	47
H.2 Training and hyperparameter details	47
H.3 Anthropic HH dataset valuation	48
H.4 Additional hyperparameter experiments	49

A Broader impact statement

Robust model alignment and reliable human feedback valuation are critical for the safe deployment of large language models (LLMs). While our work may have various societal implications, we do not identify any specific risks that require immediate attention.

Conceptually, our definition of “robustness” aligns with a utilitarian perspective: we assume that the majority opinion in the training dataset represents the true target distribution, and that minority opinions deviating from this majority are noise to be filtered out. This assumption is often reasonable in applications like toxicity removal or instruction following, where the normative standard tends to be implicitly shared by the majority.

However, we acknowledge that this framework does not universally apply. In contexts such as opinion formation or deliberative dialogue, disregarding minority voices is inappropriate and potentially harmful. In such cases, distributionally robust methods (e.g., [85, 21]) that account for underrepresented groups are more suitable. Fundamentally, these challenges relate to the broader problem of aggregating heterogeneous preferences into a single model—a problem well-known in

social choice theory. According to the impossibility theorem [5], no aggregation rule (or "social welfare functional") can simultaneously satisfy all desirable rationality axioms. As such, there is no universally optimal solution, and algorithmic choices must be guided by task-specific priorities (see, e.g., [2]).

B Additional Discussion of Related Work

Theoretical Role of Influence Functions. While influence functions (IFs) have been widely used for dataset valuation and model interpretation [41, 61], their use in deriving *robustness guarantees* remains less explored. Classical works in robust statistics [49, 44] provide foundational tools for analyzing model behavior under infinitesimal contamination. Our work draws a conceptual bridge between these classical formulations and the practical challenges in LLM alignment, utilizing IFs to derive sufficient conditions for redescending robustness and contamination detection.

Limitations of Prior Robust DPOs. Recent works propose various robust DPO formulations, including DRO-based [105], noise-aware [23], and filtering-enhanced methods [66]. However, many of these approaches either break the connection to reward learning or rely on strong assumptions or additional supervision. Our Hölder-DPO maintains a clean theoretical formulation with a robustness guarantee, while being computationally efficient and easy to implement.

Empirical Noise Levels and Filtering Limitations. Empirical studies report that preference datasets often contain 20–40% noise [35, 62], with performance degrading sharply under modest increases in noise. Common mitigation strategies, such as regularization [35] and teacher-based filtering [35, 17], suffer from limited generalization and inefficiency against symmetric noise. Our approach requires neither external LLMs nor manual heuristics, offering a lightweight alternative grounded in divergence-based theory.

Broader Applicability. Our framework naturally extends to settings like:

- **Group-specific Alignment:** Supporting heterogeneous user preferences [85, 19].
- **Personalized Objectives:** Aligning LLMs to individual user intents [81].
- **DPO with Divergence Constraints:** Leveraging f -divergences to balance alignment and diversity [100].

C Influence Function to Measure the Impact of the data contamination for DPO

Given the application of DPO and under the first-order optimality conditions, the model alignment results for $p_{\mathcal{D}}(s)$ and $p_{\mathcal{D}}^{(\epsilon)}(\tilde{s})$ are expressed as:

$$\theta^* = \operatorname{argmin}_{\theta} \mathbb{E}_{p_{\mathcal{D}}(s)} [-\log \sigma(g_{\theta}(s))] \quad \text{and} \quad \theta^*(\epsilon) = \operatorname{argmin}_{\theta} \mathbb{E}_{p_{\mathcal{D}}^{(\epsilon)}(\tilde{s})} [-\log \sigma(g_{\theta}(\tilde{s}))]. \quad (13)$$

From our definition of contamination data in Section 2.2, its influence on model alignment can be quantified by measuring the deviation in the optimized parameters, given by $\|\theta^*(\epsilon) - \theta^*\|$. To evaluate this quantity, we apply a Taylor expansion with respect to ϵ around $\epsilon = 0$ for $\theta^*(\epsilon)$, yielding:

$$\theta^*(\epsilon) = \theta^* + \epsilon \cdot \left. \frac{\partial \theta^*(\epsilon)}{\partial \epsilon} \right|_{\epsilon=0} + \mathcal{O}(\epsilon^2) \Rightarrow \theta^*(\epsilon) - \theta^* = \epsilon \cdot \left. \frac{\partial \theta^*(\epsilon)}{\partial \epsilon} \right|_{\epsilon=0} + \mathcal{O}(\epsilon^2).$$

From this result, we can approximately evaluate the influence of s_{flip} as follows:

$$\text{IF}(z, \theta, p_{\mathcal{D}}) := \left. \frac{\partial \theta^*(\epsilon)}{\partial \epsilon} \right|_{\epsilon=0}, \quad (14)$$

which is called as the *influence function* (IF) in the robust statistic context [44]. In the next section, we provide a detailed discussion on the robustness of DPO to contamination data through an analysis of this IF.

D Computation of Hölder-DPO Objective

Here, we describe the computation of the Hölder-DPO objective introduced in Eq. (8). Given the dataset $\mathcal{D} = \{s^{(i)}\}_{i=1}^N$ with $s^{(i)} \sim p_{\mathcal{D}}$, our optimization objective is:

$$\operatorname{argmin}_{\theta} S_{\gamma}(p_{\mathcal{D}} \|\sigma(g_{\theta})) \Rightarrow \operatorname{argmin}_{\theta} \phi \left(\frac{\mathbb{E}_{p_{\mathcal{D}}}[\sigma(g_{\theta}(s))^{\gamma}]}{\int \sigma(g_{\theta}(s))^{1+\gamma} ds} \right) \cdot \left(\int \sigma(g_{\theta}(s))^{1+\gamma} ds \right).$$

The numerator $\mathbb{E}_{p_{\mathcal{D}}}[\sigma(g_{\theta}(s))^{\gamma}]$ is directly estimated from the empirical distribution. However, the integral in the denominator is generally intractable. To address this, we approximate it using the empirical measure $d\hat{g}(s) := \frac{1}{N} \sum_{i=1}^N \delta(s - s^{(i)})$, yielding:

$$\int \sigma(g_{\theta}(s))^{1+\gamma} d\hat{g}(s) = \frac{1}{N} \sum_{i=1}^N \sigma(g_{\theta}(s^{(i)}))^{1+\gamma},$$

where δ is the Dirac's delta function.

Thus, our estimator can be expressed as:

$$\hat{S}_{\gamma}(p_{\mathcal{D}} \|\sigma(g_{\theta})) = \phi \left(\frac{\frac{1}{N} \sum_{i=1}^N \sigma(g_{\theta}(s^{(i)}))^{\gamma}}{\frac{1}{N} \sum_{i=1}^N \sigma(g_{\theta}(s^{(i)}))^{1+\gamma}} \right) \cdot \left(\frac{1}{N} \sum_{i=1}^N \sigma(g_{\theta}(s^{(i)}))^{1+\gamma} \right).$$

For the special case where $\phi(h) = \gamma - (1+\gamma)h$, corresponding to the scaled density power divergence, the objective simplifies to:

$$\hat{S}_{\text{DP}}(p_{\mathcal{D}} \|\sigma(g_{\theta})) = -\frac{(1+\gamma)}{N} \sum_{i=1}^N \sigma(g_{\theta}(s^{(i)}))^{\gamma} + \frac{\gamma}{N} \sum_{i=1}^N \sigma(g_{\theta}(s^{(i)}))^{1+\gamma}.$$

An alternative approach for constructing a more precise estimator is to employ importance sampling. However, in the context of LLM fine-tuning, selecting an appropriate proposal distribution is nontrivial. While this issue lies beyond the scope of the present work, it constitutes an important direction for future research. Importantly, our empirical results confirm that this objective still retains the key robustness properties of DP divergence: it remains resistant to label-flipped dataset and enables accurate estimation of the contamination ratio in practice (see Section 5).

E Proofs

E.1 Proof for Theorem 2 (The Case of DPO)

Theorem 5 (IF for DPO). *Suppose θ^* denotes the optimal parameters learned from clean dataset $p_{\mathcal{D}}$, and $\theta^*(\epsilon)$ denotes those learned from ϵ -contaminated dataset $p_{\mathcal{D}}^{(\epsilon)}$. Assume that the Hessian $\nabla_{\theta}^2 \mathcal{L}(s, \pi_{\theta})|_{\theta=\theta^*}$ is positive definite. Then, the s_{flip} -dependent component of the IF for DPO is given by:*

$$\text{IF}_{\text{DPO}}(x, \theta, p_{\mathcal{D}}) \propto \mathbb{E}_{p(s_{\text{flip}})}[\nabla_{\theta} \mathcal{L}(s_{\text{flip}}, \pi_{\theta^*})], \quad (15)$$

where $\nabla_{\theta} \mathcal{L}(s_{\text{flip}}, \pi_{\theta^*})$ is in Eq. (3).

Proof. The gradient of Eq. (2) under $p_{\mathcal{D}}^{(\epsilon)}$ is given by

$$\nabla_{\theta} \tilde{\mathcal{L}}(\tilde{s}, \pi_{\theta}) = -\beta \mathbb{E}_{p_{\mathcal{D}}^{(\epsilon)}} \left[\sigma(-g_{\theta}(\tilde{s})) \left(\nabla_{\theta} \log \pi_{\theta}(\tilde{y}_{\text{win}} | \tilde{x}) - \nabla_{\theta} \log \pi_{\theta}(\tilde{y}_{\text{lose}} | \tilde{x}) \right) \right],$$

where $\tilde{s} = \{\tilde{x}, \tilde{y}_{\text{win}}, \tilde{y}_{\text{lose}}\} \sim p_{\mathcal{D}}^{(\epsilon)}$.

From the definition of $\theta^*(\epsilon)$, we have $0 = \nabla_{\theta} \tilde{\mathcal{L}}(\tilde{s}, \pi_{\theta})|_{\theta=\theta^*(\epsilon)}$. By taking the derivation of this term w.r.t. ϵ , we obtain

$$\begin{aligned}
0 &= \frac{\partial}{\partial \epsilon} \nabla_{\theta} \tilde{\mathcal{L}}(\tilde{s}, \pi_{\theta}) \Big|_{\theta=\theta^*(\epsilon)} \\
&= -\beta \frac{\partial}{\partial \epsilon} \mathbb{E}_{p_{\tilde{\mathcal{D}}}^{(\epsilon)}} \left[\underbrace{\sigma(-g_{\theta^*(\epsilon)}(\tilde{s})) \left(\nabla_{\theta} \log \pi_{\theta^*(\epsilon)}(\tilde{y}_{\text{win}} | \tilde{x}) - \nabla_{\theta} \log \pi_{\theta^*(\epsilon)}(\tilde{y}_{\text{lose}} | \tilde{x}) \right)}_{=: F_{\theta^*(\epsilon)}(\tilde{s})} \right] \\
&= -\beta \left\{ \int \left\{ \frac{\partial}{\partial \epsilon} p_{\tilde{\mathcal{D}}}^{(\epsilon)}(\tilde{s}) \right\} F_{\theta^*(\epsilon)}(\tilde{s}) d\tilde{s} + \mathbb{E}_{p_{\tilde{\mathcal{D}}}^{(\epsilon)}} \left[\frac{\partial}{\partial \epsilon} F_{\theta^*(\epsilon)}(\tilde{s}) \right] \right\} \\
&= -\beta \left\{ \int \left\{ \frac{\partial}{\partial \epsilon} p_{\tilde{\mathcal{D}}}^{(\epsilon)}(\tilde{s}) \right\} F_{\theta^*(\epsilon)}(\tilde{s}) d\tilde{s} + \mathbb{E}_{p_{\tilde{\mathcal{D}}}^{(\epsilon)}} \left[\frac{\partial \theta^*(\epsilon)}{\partial \epsilon} \frac{\partial F_{\theta^*(\epsilon)}(\tilde{s})}{\partial \theta^*(\epsilon)} \right] \right\} \\
&= -\beta \left\{ \int \left\{ \frac{\partial}{\partial \epsilon} p_{\tilde{\mathcal{D}}}^{(\epsilon)}(\tilde{s}) \right\} F_{\theta^*(\epsilon)}(\tilde{s}) d\tilde{s} + \mathbb{E}_{p_{\tilde{\mathcal{D}}}^{(\epsilon)}} \left[\frac{\partial \theta^*(\epsilon)}{\partial \epsilon} H_{\theta^*(\epsilon)}(\tilde{s}) \right] \right\}, \tag{16}
\end{aligned}$$

where $H_{\theta^*(\epsilon)}(\tilde{s}) := \frac{\partial F_{\theta^*(\epsilon)}(\tilde{s})}{\partial \theta^*(\epsilon)}$.

From the definition of $p_{\tilde{\mathcal{D}}}^{(\epsilon)}(\tilde{s})$, we obtain

$$\int \left\{ \frac{\partial}{\partial \epsilon} p_{\tilde{\mathcal{D}}}^{(\epsilon)}(\tilde{s}) \right\} F_{\theta^*(\epsilon)}(\tilde{s}) d\tilde{s} = \mathbb{E}_{p(s_{\text{flip}})} [F_{\theta^*(\epsilon)}(s_{\text{flip}})] - \mathbb{E}_{p_{\mathcal{D}}} [F_{\theta^*(\epsilon)}(s)],$$

since $\frac{\partial}{\partial \epsilon} p_{\tilde{\mathcal{D}}}^{(\epsilon)}(\tilde{s}) = p(s_{\text{flip}}) - p_{\mathcal{D}}(s)$, where $F_{\theta^*(\epsilon)}(s_{\text{flip}}) := \sigma(-g_{\theta^*(\epsilon)}(s_{\text{flip}}))(\nabla_{\theta} \log \pi_{\theta^*(\epsilon)}(y_{\text{win}}^{\text{flip}} | z) - \nabla_{\theta} \log \pi_{\theta^*(\epsilon)}(y_{\text{lose}}^{\text{flip}} | z))$. By taking $\epsilon \rightarrow 0$, we have

$$\left(\int \left\{ \frac{\partial}{\partial \epsilon} p_{\tilde{\mathcal{D}}}^{(\epsilon)}(\tilde{s}) \right\} F_{\theta^*(\epsilon)}(\tilde{s}) d\tilde{s} \right) \Big|_{\epsilon=0} = \mathbb{E}_{p(s_{\text{flip}})} [F_{\theta^*}(s_{\text{flip}})],$$

since $\theta^{(*)}(\epsilon) \rightarrow \theta^{(*)}$ and thus $\mathbb{E}_{p_{\mathcal{D}}} [F_{\theta^*}(s)] = \nabla_{\theta} \mathcal{L}(\pi_{\theta}; \pi_{\text{ref}})|_{\theta=\theta^*} = 0$ from the first-order optimal condition in Eq. (13).

Furthermore, we also obtain

$$\mathbb{E}_{p_{\tilde{\mathcal{D}}}^{(\epsilon)}} \left[\frac{\partial \theta^*(\epsilon)}{\partial \epsilon} H_{\theta^*(\epsilon)}(\tilde{s}) \right] \Big|_{\epsilon=0} = \mathbb{E}_{p_{\mathcal{D}}} \left[\frac{\partial \theta^*(\epsilon)}{\partial \epsilon} H_{\theta^*}(s) \right],$$

where $H_{\theta^*}(s) := \frac{\partial F_{\theta^*}(s)}{\partial \theta^*}$.

Then, Eq. (16) under $\epsilon \rightarrow 0$ can be rewritten as

$$0 = \left(\frac{\partial}{\partial \epsilon} \nabla_{\theta} \tilde{\mathcal{L}}(\pi_{\theta}; \pi_{\text{ref}}) \Big|_{\theta=\theta^*(\epsilon)} \right) \Big|_{\epsilon=0} = -\beta \left\{ \mathbb{E}_{p(s_{\text{flip}})} [F_{\theta^*}(s_{\text{flip}})] + \mathbb{E}_{p_{\mathcal{D}}} \left[\frac{\partial \theta^*(\epsilon)}{\partial \epsilon} \Big|_{\epsilon=0} H_{\theta^*}(s) \right] \right\}.$$

By solving the above equality w.r.t. $\frac{\partial \theta^*(\epsilon)}{\partial \epsilon}$, we obtain

$$\frac{\partial \theta^*(\epsilon)}{\partial \epsilon} \Big|_{\epsilon=0} = - \left(\mathbb{E}_{p_{\mathcal{D}}} [H_{\theta^*}(s)] \right)^{-1} \mathbb{E}_{p(s_{\text{flip}})} [F_{\theta^*}(s_{\text{flip}})].$$

This completes the proof. \square

E.2 Proof for Theorem 3 (The Case of DPO)

Corollary 2 (DPO is not robust). Suppose that the policy gradient $\nabla_{\theta} \log \pi_{\theta}(y | x)$ is bounded by C and satisfies L -Lipchitz in θ , where $0 < C < \infty$ and $0 < L < \infty$. Then, under Theorem 2, the IF of DPO does not satisfy the rescending property in Definition 2, i.e., $\lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \|\text{IF}_{\text{DPO}}(x, \theta, p_{\mathcal{D}})\| \neq 0$.

Proof. From the positive definite assumption on the Hessian in Theorem 5 and the L -Lipschitz assumption on the gradient, it follows that $\mathbb{E}_{p_{\mathcal{D}}}[H_{\theta^*}(s)]$ is a positive definite matrix. Let $L' = \lambda_{\min}(\mathbb{E}_{p_{\mathcal{D}}}[H_{\theta^*}(s)]) > 0$ be its minimum eigenvalue. Then the norm of its inverse is bounded: $\|(\mathbb{E}_{p_{\mathcal{D}}}[H_{\theta^*}(s)])^{-1}\| = 1/L'$. Furthermore, from the assumption that $\|\nabla_{\theta} \log \pi_{\theta}(y \mid x)\| \leq C$ ($0 < C < \infty$), we have $\|\nabla_{\theta} \log \pi_{\theta}(y_{\text{win}} \mid x) - \nabla_{\theta} \log \pi_{\theta}(y_{\text{lose}} \mid x)\| \leq 2C$ from the triangle inequality. Taking the limit and applying Jensen's inequality and the bounded convergence theorem, we have:

$$\begin{aligned} & \lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \|\text{IF}_{\text{DPO}}(x, \theta, p_{\mathcal{D}})\| \\ & \leq \lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \left\| \left(\mathbb{E}_{p_{\mathcal{D}}}[H_{\theta^*}(s)] \right)^{-1} \right\| \cdot \lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \left\| \mathbb{E}_{p(s_{\text{flip}})}[F_{\theta^*}(s_{\text{flip}})] \right\| \\ & \leq (1/L') \cdot \lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \mathbb{E}_{p(s_{\text{flip}})}[\sigma(-g_{\theta^*}(s_{\text{flip}}))] \cdot 2C \\ & = (1/L') \cdot 1 \cdot 2C = 2C/L'. \end{aligned}$$

The fact $0 < 2C/L' < \infty$ according to $0 < C < \infty$ and $0 < L' < \infty$ completes the proof. \square

E.3 Proof for Theorem 4

Before showing Theorem 4, we introduce the following proposition regarding the gradient of $S_{\gamma}(p_{\mathcal{D}} \parallel \sigma(g_{\theta}))$.

Proposition 2. *Suppose that $\nabla_{\theta} g_{\theta}(s)$ is bounded. Then, under $p_{\mathcal{D}}$, the gradient of $S_{\gamma}(p_{\mathcal{D}} \parallel \sigma(g_{\theta}))$ w.r.t. θ is obtained as*

$$\begin{aligned} \nabla_{\theta} S_{\gamma}(p_{\mathcal{D}} \parallel \sigma(g_{\theta})) &= \gamma \phi'(h_{\theta}) \mathbb{E}_{p_{\mathcal{D}}}[\sigma(g_{\theta}(s))^{\gamma} (1 - \sigma(g_{\theta}(s))) \nabla_{\theta} g_{\theta}(s)] \\ &+ (1 + \gamma) \cdot \left(\phi(h_{\theta}) - h_{\theta} \cdot \phi'(h_{\theta}) \right) \cdot \left(\int \sigma(g_{\theta}(s))^{1+\gamma} (1 - \sigma(g_{\theta}(s))) \nabla_{\theta} g_{\theta}(s) ds \right), \end{aligned}$$

where $\phi'(h_{\theta}) := \nabla_{h_{\theta}} \phi(h_{\theta})$ and $h_{\theta} := \frac{\mathbb{E}_{p_{\mathcal{D}}}[\sigma(g_{\theta}(s))^{\gamma}]}{\int \sigma(g_{\theta}(s))^{1+\gamma} ds}$.

Proof. We start by introducing the following shorthand notation:

$$A(\theta) := \mathbb{E}_{p_{\mathcal{D}}}[\sigma(g_{\theta}(s))^{\gamma}], \quad B(\theta) := \int \sigma(g_{\theta}(s))^{1+\gamma} ds,$$

and $h_{\theta} := \frac{A(\theta)}{B(\theta)}$. Then, $S_{\gamma}(p_{\mathcal{D}} \parallel \sigma(g_{\theta}))$ can be expressed as $S_{\gamma}(p_{\mathcal{D}} \parallel \sigma(g_{\theta})) = \phi(h_{\theta}) \cdot B(\theta)$. Taking the derivative with respect to θ using the product rule, we have

$$\nabla_{\theta} S_{\gamma}(p_{\mathcal{D}} \parallel \sigma(g_{\theta})) = \nabla_{\theta} (\phi(h_{\theta}) \cdot B(\theta)) = \phi'(h_{\theta}) \cdot \nabla_{\theta} h_{\theta} \cdot B(\theta) + \phi(h_{\theta}) \cdot \nabla_{\theta} B(\theta). \quad (17)$$

We first evaluate $\nabla_{\theta} h_{\theta}$, which can be calculated as

$$\nabla_{\theta} h_{\theta} = \frac{B(\theta) \nabla_{\theta} A(\theta) - A(\theta) \nabla_{\theta} B(\theta)}{B(\theta)^2}.$$

By substituting this into Eq. (17), we obtain

$$\begin{aligned} \nabla_{\theta} S_{\gamma}(p_{\mathcal{D}} \parallel \sigma(g_{\theta})) &= \phi'(h_{\theta}) \cdot \frac{B(\theta) \nabla_{\theta} A(\theta) - A(\theta) \nabla_{\theta} B(\theta)}{B(\theta)} + \phi(h_{\theta}) \cdot \nabla_{\theta} B(\theta) \\ &= \phi'(h_{\theta}) \cdot \left(\nabla_{\theta} A(\theta) - h_{\theta} \nabla_{\theta} B(\theta) \right) + \phi(h_{\theta}) \cdot \nabla_{\theta} B(\theta) \\ &= \phi'(h_{\theta}) \cdot \nabla_{\theta} A(\theta) + \left(\phi(h_{\theta}) - h_{\theta} \cdot \phi'(h_{\theta}) \right) \cdot \nabla_{\theta} B(\theta). \end{aligned}$$

We next evaluate $\nabla_\theta A(\theta) = \nabla_\theta \mathbb{E}_{p_D} [\sigma(g_\theta(s))^\gamma]$. Since p_D does not depend on θ , from the chain rule, we can see $\nabla_\theta \mathbb{E}_{p_D} [\sigma(g_\theta(s))^\gamma] = \mathbb{E}_{p_D} [\nabla_\theta \sigma(g_\theta(s))^\gamma]$, where

$$\begin{aligned}\nabla_\theta \sigma(g_\theta(s))^\gamma &= \gamma \sigma(g_\theta(s))^{\gamma-1} \nabla_\theta \sigma(g_\theta(s)) \\ &= \gamma \sigma(g_\theta(s))^{\gamma-1} \cdot \left(\sigma(g_\theta(s))(1 - \sigma(g_\theta(s))) \nabla_\theta g_\theta(s) \right) \\ &= \gamma \sigma(g_\theta(s))^\gamma (1 - \sigma(g_\theta(s))) \nabla_\theta g_\theta(s).\end{aligned}$$

As for $\nabla_\theta B(\theta)$, we obtain

$$\begin{aligned}\nabla_\theta B(\theta) &= \nabla_\theta \int \sigma(g_\theta(s))^{1+\gamma} ds = \int \nabla_\theta \sigma(g_\theta(s))^{1+\gamma} ds \\ &= (1 + \gamma) \int \sigma(g_\theta(s))^{1+\gamma} (1 - \sigma(g_\theta(s))) \nabla_\theta g_\theta(s) ds,\end{aligned}$$

where the second equality comes from the dominated convergence theorem due to the fact that the derivative of $\sigma(g_\theta)$ is bounded and continuous under the bounded $\nabla_\theta g(s)$.

By substituting the results of $\nabla_\theta A(\theta)$ and $\nabla_\theta B(\theta)$ into Eq. (17), we have

$$\begin{aligned}\nabla_\theta S_\gamma(p_D \| \sigma(g_\theta)) &= \gamma \phi'(h_\theta) \mathbb{E}_{p_D} [\sigma(g_\theta(s))^\gamma (1 - \sigma(g_\theta(s))) \nabla_\theta g_\theta(s)] \\ &\quad + (1 + \gamma) \cdot \left(\phi(h_\theta) - h_\theta \cdot \phi'(h_\theta) \right) \cdot \left(\int \sigma(g_\theta(s))^{1+\gamma} (1 - \sigma(g_\theta(s))) \nabla_\theta g_\theta(s) ds \right).\end{aligned}$$

This completes the proof. \square

Furthermore, we show the following lemma to precisely derive the IF for Hölder-DPO.

Lemma 1. Suppose that $\theta^* = \operatorname{argmin}_\theta S_\gamma(p_D \| \sigma(g_\theta))$. Let $0 < \gamma < \infty$, and let $0 < \sigma(g_\theta(s))$. Assume that $\phi(h)$ satisfies $\phi'(h) \neq 0$ for $h > 0$ and $\phi(h) \neq c \cdot h$ for any constant c . Then, the first-order optimal condition of Hölder-DPO, $0 = \nabla_\theta S_\gamma(p_D \| \sigma(g_\theta))|_{\theta=\theta^*}$, holds if and only if $\mathbb{E}_{p_D} [F_{\theta^*}^{(\gamma)}(s)] = 0$ and $\int F_{\theta^*}^{(1+\gamma)}(s) ds = 0$, where $F_{\theta^*}^{(\gamma)}(s) := \sigma(g_{\theta^*}(s))^\gamma (1 - \sigma(g_{\theta^*}(s))) \nabla_\theta g_{\theta^*}(s)$.

Proof. From Proposition 2, the first-order optimal condition is given by:

$$\begin{aligned}0 &= \nabla_\theta S_\gamma(p_D \| \sigma(g_\theta)) \Big|_{\theta=\theta^*} \\ &= \underbrace{\gamma \phi'(h_\theta)}_{C_1} \underbrace{\mathbb{E}_{p_D} [F_{\theta^*}^{(\gamma)}(s)]}_X + \underbrace{(1 + \gamma) (\phi(h_\theta) - \phi'(h_\theta) h_\theta)}_{C_2} \underbrace{\left(\int F_{\theta^*}^{(1+\gamma)}(s) ds \right)}_Y.\end{aligned}$$

This is a linear combination $C_1 X + C_2 Y = 0$. The “if and only if” statement holds if we can show that both coefficients C_1 and C_2 are non-zero.

First, we show $h_{\theta^*} > 0$. By definition, $h_{\theta^*} = A(\theta^*)/B(\theta^*)$, where $A(\theta^*) = \mathbb{E}_{p_D} [\sigma(g_{\theta^*}(s))^\gamma]$ and $B(\theta^*) = \int \sigma(g_{\theta^*}(s))^{1+\gamma} ds$. From the premise $0 < \sigma(g_\theta(s))$ and $\gamma > 0$, we have $\sigma(g_{\theta^*}(s))^\gamma > 0$ and $\sigma(g_{\theta^*}(s))^{1+\gamma} > 0$. Therefore, $A(\theta^*) > 0$ and $B(\theta^*) > 0$, which implies $h_{\theta^*} > 0$. Given $\gamma > 0$, $h_{\theta^*} > 0$, and our assumption $\phi'(h) \neq 0$ for $h > 0$, the first coefficient $C_1 = \gamma \phi'(h_{\theta^*}) \neq 0$.

The second coefficient C_2 is zero if and only if $\phi(h_{\theta^*}) - h_{\theta^*} \phi'(h_{\theta^*}) = 0$. This condition holds if $\phi(h)$ is a homogeneous function of degree 1, i.e., $\phi(h) = c \cdot h$. By our assumption $\phi(h) \neq c \cdot h$, this implies $\phi(h_{\theta^*}) - h_{\theta^*} \phi'(h_{\theta^*}) \neq 0$, and thus $C_2 \neq 0$. Since both coefficients C_1 and C_2 are non-zero, the optimality condition $C_1 X + C_2 Y = 0$ holds if and only if $X = \mathbb{E}_{p_D} [F_{\theta^*}^{(\gamma)}(s)] = 0$ and $Y = \int F_{\theta^*}^{(1+\gamma)}(s) ds = 0$. This completes the proof. \square

We remark that the conditions of $\phi(h)$, $\phi'(h) \neq 0$ for $h > 0$ and $\phi(h) \neq c \cdot h$ for any constant c , introduced in Lemma 1, are satisfied by the DP divergence and the PS score (which is closely related to the γ -divergence). The following lemma formally verifies this. This fact indicates that constructing ϕ so as to satisfy the above condition is one of keys to guarantee robustness.

Lemma 2. Let $\gamma > 0$. The $\phi(h)$ functions for the DP-divergence and the PS-divergence (Remark 1) satisfy the assumptions required in Lemma 1, namely $\phi'(h) \neq 0$ and $\phi(h) - h\phi'(h) \neq 0$ for all $h > 0$.

Proof. **For DP-divergence:** Let $\phi(h) = \gamma - (1 + \gamma)h$.

- (i) The derivative is $\phi'(h) = -(1 + \gamma)$. Since $\gamma > 0$, $\phi'(h)$ is a non-zero constant, thus $\phi'(h) \neq 0$ for all $h > 0$.
- (ii) We check the second coefficient term from Lemma 1:

$$\begin{aligned}\phi(h) - h\phi'(h) &= (\gamma - (1 + \gamma)h) - h(-(1 + \gamma)) \\ &= \gamma - (1 + \gamma)h + (1 + \gamma)h \\ &= \gamma\end{aligned}$$

Since $\gamma > 0$, we have $\phi(h) - h\phi'(h) \neq 0$.

For PS score: Let $\phi(h) = -h^{1+\gamma}$.

- (i) The derivative is $\phi'(h) = -(1 + \gamma)h^\gamma$. Since $\gamma > 0$ and we evaluate for $h > 0$, $h^\gamma > 0$. Thus, $\phi'(h)$ is strictly negative, and $\phi'(h) \neq 0$.
- (ii) We check the second coefficient term:

$$\begin{aligned}\phi(h) - h\phi'(h) &= (-h^{1+\gamma}) - h(-(1 + \gamma)h^\gamma) \\ &= -h^{1+\gamma} + (1 + \gamma)h^{1+\gamma} \\ &= (-1 + 1 + \gamma)h^{1+\gamma} \\ &= \gamma h^{1+\gamma}\end{aligned}$$

Since $\gamma > 0$ and $h > 0$, we have $\gamma h^{1+\gamma} > 0$, which implies $\phi(h) - h\phi'(h) \neq 0$.

In both cases, the assumptions hold. This completes the proof. \square

Now we show the full proof of Theorem 4.

Theorem 4 (IF for Hölder-DPO). Suppose that $\theta^*(\epsilon) = \operatorname{argmin}_\theta S_\gamma(\tilde{p}_D^{(\epsilon)} \parallel \sigma(g_\theta))$ and $\theta^* = \operatorname{argmin}_\theta S_\gamma(p_D \parallel \sigma(g_\theta))$. Let $\phi(h)$ be twice-differentiable, and let $0 < \sigma(g_\theta(s))$ and $0 < \gamma < \infty$. Assume that $\phi(h)$ satisfies $\phi'(h) \neq 0$ for $h > 0$ and $\phi(h) \neq c \cdot h$ for any constant c ⁶ and the Hessian $\nabla_\theta^2 \mathcal{L}(s, \pi_\theta)|_{\theta=\theta^*}$ is positive definite. Then, the IF of the Hölder-DPO, excluding terms independent of s_{flip} , is given by:

$$\text{IF}_{\text{H-DPO}}(s_{flip}, \theta, p_D) \propto \mathbb{E}_{p(s_{flip})}[F_{\theta^*}^{(\gamma)}(s_{flip})], \quad (10)$$

where $F_{\theta^*}^{(\gamma)}(s_{flip}) := \sigma(g_{\theta^*}(s_{flip}))^\gamma \nabla_\theta \mathcal{L}(s_{flip}, \pi_{\theta^*})$.

Proof. From Proposition 2, the gradient of $S_\gamma(\tilde{p}_D^{(\epsilon)} \parallel \sigma(g_\theta))$ w.r.t. θ is

$$\nabla_\theta S_\gamma(\tilde{p}_D^{(\epsilon)} \parallel \sigma(g_\theta)) = \gamma \phi'(\tilde{h}_\theta) \mathbb{E}_{\tilde{p}_D^{(\epsilon)}}[F_\theta^{(\gamma)}(\tilde{s})] + (1 + \gamma) \cdot \left(\phi(\tilde{h}_\theta) - \tilde{h}_\theta \cdot \phi'(\tilde{h}_\theta) \right) \cdot \left(\int F_\theta^{(1+\gamma)}(\tilde{s}) d\tilde{s} \right),$$

where $F_\theta^{(\gamma)}(\tilde{s}) := \sigma^\gamma(g_\theta(\tilde{s})) (1 - \sigma(g_\theta(\tilde{s}))) \nabla_\theta g_\theta(\tilde{s}) = \sigma^\gamma \nabla_\theta \mathcal{L}(\tilde{s}, \pi_\theta)$, $F_\theta^{(1+\gamma)}(\tilde{s}) := \sigma(g_\theta(\tilde{s})) \cdot F_\theta^{(\gamma)}(\tilde{s})$, $\tilde{h}_\theta := \mathbb{E}_{\tilde{p}_D^{(\epsilon)}}[\sigma(g_\theta(\tilde{s}))^\gamma] / (\int \sigma(g_\theta(\tilde{s}))^{1+\gamma} d\tilde{s})$, and $\tilde{s} = \{\tilde{x}, \tilde{y}_{\text{win}}, \tilde{y}_{\text{lose}}\} \sim \tilde{p}_D^{(\epsilon)}$.

From the definition of $\theta^*(\epsilon)$, we have

$$\begin{aligned}0 &= \nabla_\theta S_\gamma(\tilde{p}_D^{(\epsilon)} \parallel \sigma(g_\theta)) \Big|_{\theta=\theta^*(\epsilon)} \\ &= \gamma \phi'(\tilde{h}_{\theta^*(\epsilon)}) \mathbb{E}_{\tilde{p}_D^{(\epsilon)}}[F_{\theta^*(\epsilon)}^{(\gamma)}(\tilde{s})] + (1 + \gamma) \left(\phi(\tilde{h}_{\theta^*(\epsilon)}) - \phi'(\tilde{h}_{\theta^*(\epsilon)}) \tilde{h}_{\theta^*(\epsilon)} \right) \left(\int F_{\theta^*(\epsilon)}^{(1+\gamma)}(\tilde{s}) d\tilde{s} \right),\end{aligned}$$

⁶These assumptions are satisfied by the DP divergence and the PS score, as formally shown in Lemma 2.

where $F_{\theta^*(\epsilon)}^{(\gamma)}(\tilde{s}) := \sigma(g_{\theta^*(\epsilon)}(\tilde{s}))^\gamma (1 - \sigma(g_{\theta^*(\epsilon)}(\tilde{s}))) \{\nabla_{\theta} g_{\theta}(\tilde{s})|_{\theta=\theta^*(\epsilon)}\}$ and $F_{\theta^*(\epsilon)}^{(1+\gamma)}(\tilde{s}) := \sigma(g_{\theta^*(\epsilon)}(\tilde{s})) \cdot F_{\theta^*(\epsilon)}^{(\gamma)}(\tilde{s})$.

By taking the derivation of this term w.r.t. ϵ in the above, we obtain

$$0 = \gamma \left\{ \underbrace{\frac{\partial}{\partial \epsilon} \phi'(\tilde{h}_{\theta^*(\epsilon)}) \mathbb{E}_{\tilde{p}_{\mathcal{D}}^{(\epsilon)}}[F_{\theta^*(\epsilon)}^{(\gamma)}(\tilde{s})]}_{\text{(I)}} \right. \\ \left. + (1 + \gamma) \underbrace{\left\{ \frac{\partial}{\partial \epsilon} \left(\phi(\tilde{h}_{\theta^*(\epsilon)}) - \phi'(\tilde{h}_{\theta^*(\epsilon)}) \tilde{h}_{\theta^*(\epsilon)} \right) \left(\int F_{\theta^*(\epsilon)}^{(1+\gamma)}(\tilde{s}) d\tilde{s} \right) \right\}}_{\text{(II)}} \right\}. \quad (18)$$

For the term (I), we have

$$\begin{aligned} & \frac{\partial}{\partial \epsilon} \phi'(h_{\theta^*(\epsilon)}) \mathbb{E}_{\tilde{p}_{\mathcal{D}}^{(\epsilon)}}[F_{\theta^*(\epsilon)}^{(\gamma)}(\tilde{s})] \\ &= \phi''(h_{\theta^*(\epsilon)}) \cdot \frac{\partial h_{\theta^*(\epsilon)}}{\partial \epsilon} \mathbb{E}_{\tilde{p}_{\mathcal{D}}^{(\epsilon)}}[F_{\theta^*(\epsilon)}^{(\gamma)}(\tilde{s})] \\ & \quad + \phi'(h_{\theta^*(\epsilon)}) \left\{ \int \left\{ \frac{\partial}{\partial \epsilon} p_{\mathcal{D}}^{(\epsilon)}(\tilde{s}) \right\} F_{\theta^*(\epsilon)}^{(\gamma)}(\tilde{s}) d\tilde{s} + \mathbb{E}_{\tilde{p}_{\mathcal{D}}^{(\epsilon)}} \left[\frac{\partial}{\partial \epsilon} F_{\theta^*(\epsilon)}^{(\gamma)}(\tilde{s}) \right] \right\} \\ &= \phi''(h_{\theta^*(\epsilon)}) \cdot \frac{\partial h_{\theta^*(\epsilon)}}{\partial \epsilon} \cdot \mathbb{E}_{\tilde{p}_{\mathcal{D}}^{(\epsilon)}}[F_{\theta^*(\epsilon)}^{(\gamma)}(\tilde{s})] \\ & \quad + \phi'(h_{\theta^*(\epsilon)}) \left\{ \mathbb{E}_{p(s_{\text{flip}})}[F_{\theta^*(\epsilon)}^{(\gamma)}(s_{\text{flip}})] - \mathbb{E}_{p_{\mathcal{D}}}[F_{\theta^*(\epsilon)}^{(\gamma)}(s)] + \mathbb{E}_{\tilde{p}_{\mathcal{D}}^{(\epsilon)}} \left[\frac{\partial \theta^*(\epsilon)}{\partial \epsilon} \cdot \frac{\partial F_{\theta^*(\epsilon)}^{(\gamma)}(\tilde{s})}{\partial \theta^*(\epsilon)} \right] \right\} \\ &= \phi''(h_{\theta^*(\epsilon)}) \cdot \frac{\partial \theta^*(\epsilon)}{\partial \epsilon} \cdot \frac{\partial h_{\theta^*(\epsilon)}}{\partial \theta^*(\epsilon)} \cdot \mathbb{E}_{\tilde{p}_{\mathcal{D}}^{(\epsilon)}}[F_{\theta^*(\epsilon)}^{(\gamma)}(\tilde{s})] \\ & \quad + \phi'(h_{\theta^*(\epsilon)}) \left\{ \mathbb{E}_{p(s_{\text{flip}})}[F_{\theta^*(\epsilon)}^{(\gamma)}(s_{\text{flip}})] - \mathbb{E}_{p_{\mathcal{D}}}[F_{\theta^*(\epsilon)}^{(\gamma)}(s)] + \mathbb{E}_{\tilde{p}_{\mathcal{D}}^{(\epsilon)}} \left[\frac{\partial \theta^*(\epsilon)}{\partial \epsilon} \cdot H_{\theta^*(\epsilon)}^{(\gamma)}(\tilde{s}) \right] \right\} \\ &= \frac{\partial \theta^*(\epsilon)}{\partial \epsilon} \left\{ \phi''(h_{\theta^*(\epsilon)}) \cdot \frac{\partial h_{\theta^*(\epsilon)}}{\partial \theta^*(\epsilon)} \cdot \mathbb{E}_{\tilde{p}_{\mathcal{D}}^{(\epsilon)}}[F_{\theta^*(\epsilon)}^{(\gamma)}(\tilde{s})] + \phi'(h_{\theta^*(\epsilon)}) \cdot \mathbb{E}_{\tilde{p}_{\mathcal{D}}^{(\epsilon)}}[H_{\theta^*(\epsilon)}^{(\gamma)}(\tilde{s})] \right\} \\ & \quad + \phi'(h_{\theta^*(\epsilon)}) \left\{ \mathbb{E}_{p(s_{\text{flip}})}[F_{\theta^*(\epsilon)}^{(\gamma)}(s_{\text{flip}})] - \mathbb{E}_{p_{\mathcal{D}}}[F_{\theta^*(\epsilon)}^{(\gamma)}(s)] \right\}, \end{aligned}$$

where $H_{\theta^*(\epsilon)}^{(\gamma)}(\tilde{s}) := \partial F_{\theta^*(\epsilon)}^{(\gamma)} / \partial \theta^*(\epsilon)$. By taking $\epsilon \rightarrow 0$, we have

$$\begin{aligned} & \left. \frac{\partial}{\partial \epsilon} \phi'(h_{\theta^*(\epsilon)}) \mathbb{E}_{\tilde{p}_{\mathcal{D}}^{(\epsilon)}}[F_{\theta^*(\epsilon)}^{(\gamma)}(\tilde{s})] \right|_{\epsilon=0} \\ &= \left. \frac{\partial \theta^*(\epsilon)}{\partial \epsilon} \right|_{\epsilon=0} \cdot \left\{ \phi''(h_{\theta^*}) \cdot \frac{\partial h_{\theta^*}}{\partial \theta^*} \cdot \mathbb{E}_{p_{\mathcal{D}}}[F_{\theta^*}^{(\gamma)}(s)] + \phi'(h_{\theta^*}) \cdot \mathbb{E}_{p_{\mathcal{D}}}[H_{\theta^*}^{(\gamma)}(s)] \right\} \\ & \quad + \phi'(h_{\theta^*}) \left\{ \mathbb{E}_{p(s_{\text{flip}})}[F_{\theta^*}^{(\gamma)}(s_{\text{flip}})] - \mathbb{E}_{p_{\mathcal{D}}}[F_{\theta^*}^{(\gamma)}(s)] \right\}. \quad (19) \end{aligned}$$

For the term (II), we obtain

$$\begin{aligned}
& \frac{\partial}{\partial \epsilon} (\phi(h_{\theta^*(\epsilon)}) - \phi'(h_{\theta^*(\epsilon)})h_{\theta^*(\epsilon)}) \left(\int F_{\theta^*(\epsilon)}^{(1+\gamma)}(\tilde{s}) d\tilde{s} \right) \\
&= \left\{ \frac{\partial}{\partial \epsilon} (\phi(h_{\theta^*(\epsilon)}) - \phi'(h_{\theta^*(\epsilon)})h_{\theta^*(\epsilon)}) \right\} \left(\int F_{\theta^*(\epsilon)}^{(1+\gamma)}(\tilde{s}) d\tilde{s} \right) \\
&\quad + (\phi(h_{\theta^*(\epsilon)}) - \phi'(h_{\theta^*(\epsilon)})h_{\theta^*(\epsilon)}) \left(\int \frac{\partial}{\partial \epsilon} F_{\theta^*(\epsilon)}^{(1+\gamma)}(\tilde{s}) d\tilde{s} \right) \\
&= \left(-\phi''(h_{\theta^*(\epsilon)}) \cdot \frac{\partial h_{\theta^*(\epsilon)}}{\partial \epsilon} \cdot h_{\theta^*(\epsilon)} \right) \left(\int F_{\theta^*(\epsilon)}^{(1+\gamma)}(\tilde{s}) d\tilde{s} \right) \\
&\quad + (\phi(h_{\theta^*(\epsilon)}) - \phi'(h_{\theta^*(\epsilon)})h_{\theta^*(\epsilon)}) \left(\int \frac{\partial \theta^*(\epsilon)}{\partial \epsilon} H_{\theta^*(\epsilon)}^{(1+\gamma)}(\tilde{s}) d\tilde{s} \right) \\
&= \frac{\partial \theta^*(\epsilon)}{\partial \epsilon} \left\{ \left(-\phi''(h_{\theta^*(\epsilon)}) \cdot \frac{\partial h_{\theta^*(\epsilon)}}{\partial \theta^*(\epsilon)} \cdot h_{\theta^*(\epsilon)} \right) \left(\int F_{\theta^*(\epsilon)}^{(1+\gamma)}(\tilde{s}) d\tilde{s} \right) \right. \\
&\quad \left. + (\phi(h_{\theta^*(\epsilon)}) - \phi'(h_{\theta^*(\epsilon)})h_{\theta^*(\epsilon)}) \left(\int H_{\theta^*(\epsilon)}^{(1+\gamma)}(\tilde{s}) d\tilde{s} \right) \right\},
\end{aligned}$$

where we use the dominated convergence theorem to exchange the integral and derivative in the second equality and $H_{\theta^*(\epsilon)}^{(1+\gamma)}(\tilde{s}) := \partial F_{\theta^*(\epsilon)}^{(1+\gamma)} / \partial \theta^*(\epsilon)$. By taking $\epsilon \rightarrow 0$, we obtain

$$\begin{aligned}
& \frac{\partial}{\partial \epsilon} (\phi(h_{\theta^*(\epsilon)}) - \phi'(h_{\theta^*(\epsilon)})h_{\theta^*(\epsilon)}) \left(\int F_{\theta^*(\epsilon)}^{(1+\gamma)}(\tilde{s}) d\tilde{s} \right) \Big|_{\epsilon=0} \\
&= \frac{\partial \theta^*(\epsilon)}{\partial \epsilon} \Big|_{\epsilon=0} \cdot \left\{ \left(-\phi''(h_{\theta^*}) \cdot \frac{\partial h_{\theta^*}}{\partial \theta^*} \cdot h_{\theta^*} \right) \left(\int F_{\theta^*}^{(1+\gamma)}(s) ds \right) \right. \\
&\quad \left. + (\phi(h_{\theta^*}) - \phi'(h_{\theta^*})h_{\theta^*}) \left(\int H_{\theta^*}^{(1+\gamma)}(s) ds \right) \right\}. \tag{20}
\end{aligned}$$

Substituting Eqs. (19) and (20) into Eq. (18) gives us:

$$\begin{aligned}
& -\gamma \phi'(h_{\theta^*}) \left\{ \mathbb{E}_{p(s_{\text{flip}})}[F_{\theta^*}^{(\gamma)}(s_{\text{flip}})] - \mathbb{E}_{p_{\mathcal{D}}}[F_{\theta^*}^{(\gamma)}(s)] \right\} \\
&= \frac{\partial \theta^*(\epsilon)}{\partial \epsilon} \Big|_{\epsilon=0} \cdot \left\{ \gamma \phi''(h_{\theta^*}) \cdot \frac{\partial h_{\theta^*}}{\partial \theta^*} \cdot \mathbb{E}_{p_{\mathcal{D}}}[F_{\theta^*}^{(\gamma)}(s)] + \gamma \phi'(h_{\theta^*}) \cdot \mathbb{E}_{p_{\mathcal{D}}}[H_{\theta^*}^{(\gamma)}(s)] \right. \\
&\quad + (1+\gamma) \left(-\phi''(h_{\theta^*}) \cdot \frac{\partial h_{\theta^*}}{\partial \theta^*} \cdot h_{\theta^*} \right) \left(\int F_{\theta^*}^{(1+\gamma)}(s) ds \right) \\
&\quad \left. + (1+\gamma) (\phi(h_{\theta^*}) - \phi'(h_{\theta^*})h_{\theta^*}) \left(\int H_{\theta^*}^{(1+\gamma)}(s) ds \right) \right\}.
\end{aligned}$$

From Lemma 1, we further obtain

$$-\gamma \phi'(h_{\theta^*}) \mathbb{E}_{p(s_{\text{flip}})}[F_{\theta^*}^{(\gamma)}(s_{\text{flip}})] = \frac{\partial \theta^*(\epsilon)}{\partial \epsilon} \Big|_{\epsilon=0} \cdot C_{\theta^*}^{(\gamma)}(s),$$

where $C_{\theta^*}^{(\gamma)}(s)$ is the constant term w.r.t. the contaminated input s_{flip} defined as

$$C_{\theta^*}^{(\gamma)}(s) := \gamma \phi'(h_{\theta^*}) \cdot \mathbb{E}_{p_{\mathcal{D}}}[H_{\theta^*}^{(\gamma)}(s)] + (1+\gamma) (\phi(h_{\theta^*}) - \phi'(h_{\theta^*})h_{\theta^*}) \left(\int H_{\theta^*}^{(1+\gamma)}(s) ds \right).$$

We finally obtain

$$\frac{\partial \theta^*(\epsilon)}{\partial \epsilon} \Big|_{\epsilon=0} = - \left(C_{\theta^*}^{(\gamma)}(s) \right)^{-1} \cdot \gamma \phi'(h_{\theta^*}) \mathbb{E}_{p(s_{\text{flip}})}[F_{\theta^*}^{(\gamma)}(s_{\text{flip}})]. \tag{21}$$

From the assumption, $(C_{\theta^*}^{(\gamma)}(s))^{-1}$ exists because $H_{\theta^*}^{(\gamma)}(s)$ is positive definite from the fact that the gradient of Hölder-DPO is $\sigma(g_{\theta^*}(s_{\text{flip}}))^\gamma \nabla_\theta \mathcal{L}(s_{\text{flip}}, \pi_{\theta^*})$, where $\nabla_\theta \mathcal{L}(s_{\text{flip}}, \pi_{\theta^*})$ is the gradient of DPO loss whose Hessian is assumed as positive definite. This completes the proof. \square

E.4 Proof for Corollary 1

Corollary 1 (Hölder-DPO is robust). *Suppose that the policy gradient $\nabla_\theta \log \pi_\theta(y | x)$ is bounded by C and satisfies L -Lipchitz in θ , where $0 < C < \infty$ and $0 < L < \infty$. Then, under Theorem 4, the IF of Hölder-DPO satisfies the redescending property in Definition 2.*

Proof. From Theorem 4, the IF for Hölder-DPO is given by:

$$\text{IF}_{\text{H-DPO}} = - \left(C_{\theta^*}^{(\gamma)}(s) \right)^{-1} \cdot \gamma \phi'(h_{\theta^*}) \mathbb{E}_{p(s_{\text{flip}})} [F_{\theta^*}^{(\gamma)}(s_{\text{flip}})].$$

First, we analyze the denominator $C_{\theta^*}^{(\gamma)}(s)$. Following Theorem 5, we can see that the Hessian-related term $C_{\theta^*}^{(\gamma)}(s)$ is positive definite. Let $L' = \lambda_{\min}(C_{\theta^*}^{(\gamma)}(s)) > 0$ be its minimum eigenvalue. Then the norm of its inverse is bounded: $\|(C_{\theta^*}^{(\gamma)}(s))^{-1}\| \leq 1/L'$.

Next, we analyze the numerator, $\mathbb{E}_{p(s_{\text{flip}})} [F_{\theta^*}^{(\gamma)}(s_{\text{flip}})]$, where $F_{\theta^*}^{(\gamma)}(s_{\text{flip}}) := \sigma(g_{\theta^*}(s_{\text{flip}}))^\gamma (1 - \sigma(g_{\theta^*}(s_{\text{flip}}))) \nabla_\theta g_{\theta^*}(s_{\text{flip}})$. From the assumption that $\|\nabla_\theta \log \pi_\theta\| \leq C$, the term $\|\nabla_\theta g_{\theta^*}\|$ is also bounded by $C' = 2\beta C$.

We now take the limit required by Definition 2:

$$\begin{aligned} & \lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \|\text{IF}_{\text{H-DPO}}\| \\ & \leq \lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \left\| \left(C_{\theta^*}^{(\gamma)}(s) \right)^{-1} \cdot \gamma \phi'(h_{\theta^*}) \right\| \cdot \lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \left\| \mathbb{E}_{p(s_{\text{flip}})} [F_{\theta^*}^{(\gamma)}(s_{\text{flip}})] \right\| \\ & \leq K \cdot \lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \mathbb{E}_{p(s_{\text{flip}})} \left[\underbrace{\sigma(g_{\theta^*}(s_{\text{flip}}))^\gamma}_{\rightarrow 0} \cdot \underbrace{(1 - \sigma(g_{\theta^*}(s_{\text{flip}})))}_{\rightarrow 1} \cdot \underbrace{\|\nabla_\theta g_{\theta^*}\|}_{\leq C'} \right], \end{aligned}$$

where $K = (1/L') \cdot |\gamma \phi'(h_{\theta^*})|$ is a finite non-zero constant.

By the bounded convergence theorem, we can move the limit inside the expectation:

$$\begin{aligned} \lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \|\text{IF}_{\text{H-DPO}}\| & \leq K \cdot \mathbb{E}_{p(s_{\text{flip}})} \left[\lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \left(\sigma(g_{\theta^*})^\gamma \cdot (1 - \sigma(g_{\theta^*})) \cdot \|\nabla_\theta g_{\theta^*}\| \right) \right] \\ & \leq K \cdot \mathbb{E}_{p(s_{\text{flip}})} [0 \cdot 1 \cdot C'] \\ & = 0. \end{aligned}$$

The IF of Hölder-DPO converges to 0, satisfying the redescending property. This completes the proof. \square

F Contamination ratio estimation and outlier detection

Estimating the contamination ratio ϵ and identifying the contamination data based on this are crucial for achieving appropriate model alignment. In this section, we show that our Hölder-DPO can be extended to incorporate Enlarged models, enabling these objectives to be realized.

F.1 Model extension approach to both parameter estimation and contamination rate estimation

Here, we reorganize the framework using *model extension* proposed by Kanamori and Fujisawa [54] for simultaneously estimating model parameters and contamination rates in the context of the DPO.

According to Eq. (6), the essence of DPO-based approaches, including our Hölder-DPO, lies in minimizing a divergence D between $p_{\mathcal{D}}(s)$ and $\sigma(g_\theta(s))$ with respect to θ , i.e., estimating $p_{\mathcal{D}}(s)$ through the model parameters θ . When $p_{\mathcal{D}}(s)$ is contaminated as defined in Definition 1, this optimization problem can be reformulated as:

$$\theta^*(\epsilon) = \operatorname{argmin}_{\theta} D[\tilde{p}_{\mathcal{D}}^{(\epsilon)} \| \sigma(g_\theta(s))] = \operatorname{argmin}_{\theta} D[(1 - \epsilon)p_{\mathcal{D}}(s) + \epsilon \delta_{s_{\text{flip}}} \| \sigma(g_\theta(s))].$$

If a divergence that automatically mitigates the impact of δ_{flip} is chosen, the optimization reduces to:

$$\theta^*(\epsilon) = \underset{\theta}{\operatorname{argmin}} D[\tilde{p}_{\mathcal{D}}^{(\epsilon)} \|\sigma(g_{\theta}(s))\|] \approx \underset{\theta}{\operatorname{argmin}} D[(1 - \epsilon)p_{\mathcal{D}}(s) \|\sigma(g_{\theta}(s))\|],$$

indicating that $\sigma(g_{\theta}(s))$ estimates $(1 - \epsilon)p_{\mathcal{D}}(s)$, rather than the original target $p_{\mathcal{D}}(s)$.

To address this gap, we consider the following extended model: $m_{\eta}(s) = \xi\sigma(g_{\theta}(s))$, where $\eta = (\xi, \theta)$. In this case, the DPO-based model alignment under the robust divergence can be formulated as:

$$\theta^*(\epsilon) = \underset{\theta}{\operatorname{argmin}} \min_{\xi} D[\tilde{p}_{\mathcal{D}}^{(\epsilon)} \|m_{\eta}(s)\|] \approx \underset{\theta}{\operatorname{argmin}} \min_{\xi} D[(1 - \epsilon)p_{\mathcal{D}}(s) \|\xi\sigma(g_{\theta}(s))\|].$$

With this formulation, ξ is expected to play the role of determining the ratio of clean data $1 - \epsilon$ and $\sigma(g_{\theta}(s))$ is expected to serve as an estimator of the clean data distribution $p_{\mathcal{D}}(s)$.

F.2 Contamination ratio estimation

F.2.1 Model extension to estimate θ and ϵ :

Recall from Eq. (6) that DPO-based methods, including our Hölder-DPO, estimate the preference data distribution $p_{\mathcal{D}}$ by minimizing $D[p_{\mathcal{D}} \|\sigma(g_{\theta})\|]$, where D denotes a *generic* divergence measuring the discrepancy between $p_{\mathcal{D}}$ and the model output $\sigma(g_{\theta})$. When the data is contaminated as $p_{\mathcal{D}}^{(\epsilon)}(\tilde{s}) = (1 - \epsilon)p_{\mathcal{D}}(s) + \epsilon p(s_{\text{flip}})$, the objective becomes minimizing $D[p_{\mathcal{D}}^{(\epsilon)} \|\sigma(g_{\theta})\|]$. If D approximately nullifies the contribution of the contamination term $p(s_{\text{flip}})$, i.e., $D[p(s_{\text{flip}}) \|\sigma(g_{\theta}(s_{\text{flip}}))\|] \approx 0$ ⁷, then minimizing $D[p_{\mathcal{D}}^{(\epsilon)} \|\sigma(g_{\theta})\|]$ aligns $\sigma(g_{\theta})$ with $(1 - \epsilon)p_{\mathcal{D}}$. This results in a *mismatch in scale* relative to $p_{\mathcal{D}}$ (see Appendix F.1 for details). To correct for this mismatch, we *extend the model* to $m_{\eta} = \xi \cdot \sigma(g_{\theta})$ with $\eta = (\xi, \theta)$, introducing a scaling parameter $\xi > 0$ to explicitly account for the *clean-data proportion* $(1 - \epsilon)$. The revised objective $\underset{\eta}{\operatorname{argmin}} D[p_{\mathcal{D}}^{(\epsilon)} \|m_{\eta}\|]$ enables $\sigma(g_{\theta})$ to serve as an estimator of $p_{\mathcal{D}}(s)$, while ξ absorbs the $(1 - \epsilon)$ scaling.

F.2.2 Extended model for estimating the contamination ratio

In this section, we discuss the case when we conduct our Hölder-DPO under the extended model m_{η} . According to the discussion in Section 4.1, the optimization problem of Hölder-DPO with m_{η} can be formulated as

$$\underset{\theta}{\operatorname{argmin}} \min_{\xi} D_H[\tilde{p}_{\mathcal{D}}^{(\epsilon)} \|m_{\eta}(\tilde{s})\|] = \underset{\theta}{\operatorname{argmin}} \min_{\xi} S_{\gamma}(\tilde{p}_{\mathcal{D}}^{(\epsilon)} \|m_{\eta}(\tilde{s})\|). \quad (22)$$

Recalling Definition 3 and the discrete probabilistic nature of σ , we can see that

$$\begin{aligned} S_{\gamma}(\tilde{p}_{\mathcal{D}}^{(\epsilon)} \|m_{\eta}(\tilde{s})\|) &= \phi\left(\frac{\mathbb{E}_{\tilde{p}_{\mathcal{D}}^{(\epsilon)}}[m_{\eta}^{\gamma}(\tilde{s})]}{\int m_{\eta}^{1+\gamma}(\tilde{s}) d\tilde{s}}\right) \left(\int m_{\eta}^{1+\gamma}(\tilde{s}) d\tilde{s}\right) \\ &= \phi\left(\frac{\mathbb{E}_{\tilde{p}_{\mathcal{D}}^{(\epsilon)}}[\sigma(g_{\theta}(\tilde{s}))^{\gamma}]}{\xi \cdot \int \sigma(g_{\theta}(\tilde{s}))^{1+\gamma} d\tilde{s}}\right) \left(\xi^{1+\gamma} \cdot \int \sigma(g_{\theta}(\tilde{s}))^{1+\gamma} d\tilde{s}\right) \\ &\geq -\left(\frac{\mathbb{E}_{\tilde{p}_{\mathcal{D}}^{(\epsilon)}}[\sigma(g_{\theta}(\tilde{s}))^{\gamma}]}{\xi \cdot \int \sigma(g_{\theta}(\tilde{s}))^{1+\gamma} d\tilde{s}}\right)^{1+\gamma} \left(\xi^{1+\gamma} \cdot \int \sigma(g_{\theta}(\tilde{s}))^{1+\gamma} d\tilde{s}\right) \\ &= -\frac{\mathbb{E}_{\tilde{p}_{\mathcal{D}}^{(\epsilon)}}[\sigma(g_{\theta}(\tilde{s}))^{\gamma}]}{\left(\int \sigma(g_{\theta}(\tilde{s}))^{1+\gamma} d\tilde{s}\right)^{\frac{\gamma}{1+\gamma}}} \\ &\quad \underbrace{\hspace{10em}}_{=: S_{\text{PS}}(\tilde{p}_{\mathcal{D}}^{(\epsilon)} \|\sigma(g_{\theta}(\tilde{s}))\|)} \\ &= S_{\text{PS}}(\tilde{p}_{\mathcal{D}}^{(\epsilon)} \|\sigma(g_{\theta}(\tilde{s}))\|) = -\exp\left\{-\gamma(1 + \gamma) \cdot S_{\log}(\tilde{p}_{\mathcal{D}}^{(\epsilon)} \|\sigma(g_{\theta}(\tilde{s}))\|)\right\}, \quad (23) \end{aligned}$$

⁷DP- and γ -divergences satisfy this property under Assumption 1 (see Appendix F.4).

where the third inequality comes from the fact that $\phi(h) \geq -h^{1+\gamma}$ for all $h \geq 0$, the term $S_{\text{PS}}(\tilde{p}_{\mathcal{D}}^{(\epsilon)} \parallel \sigma(g_{\theta}(\tilde{s})))$ in the forth line is called as the pseudo-spherical (PS) score, and $S_{\log}(\tilde{p}_{\mathcal{D}}^{(\epsilon)} \parallel \sigma(g_{\theta}(\tilde{s})))$ is the γ -score associated with γ -divergence defined as

$$S_{\log}(\tilde{p}_{\mathcal{D}}^{(\epsilon)} \parallel \sigma(g_{\theta}(\tilde{s}))) := -\frac{1}{\gamma} \log \left(\mathbb{E}_{\tilde{p}_{\mathcal{D}}^{(\epsilon)}} [\sigma(g_{\theta}(\tilde{s}))^{\gamma}] \right) + \frac{1}{1+\gamma} \log \int \sigma(g_{\theta}(\tilde{s}))^{1+\gamma} d\tilde{s}. \quad (24)$$

From the Eq. (23), we can see that the lower bound of $S_{\gamma}(\tilde{p}_{\mathcal{D}}^{(\epsilon)} \parallel m_{\eta}(\tilde{s}))$ is independent of ξ . Therefore, the optimal solution of $\xi^* = \min_{\xi} S_{\gamma}(\tilde{p}_{\mathcal{D}}^{(\epsilon)} \parallel m_{\eta}(\tilde{s}))$ is obtained as the equality condition of Eq. (23), that is,

$$\xi^* = \frac{\mathbb{E}_{\tilde{p}_{\mathcal{D}}^{(\epsilon)}} [\sigma(g_{\theta}(\tilde{s}))^{\gamma}]}{\int \sigma(g_{\theta}(\tilde{s}))^{1+\gamma} d\tilde{s}} \quad (\forall \theta), \quad (25)$$

where we used the fact that $\phi(1) = -1$.

At the end of this section, we verify that ξ^* in Eq. (25) serves as an estimate of the contamination rate. When robust model alignment using Hölder-DPO effectively reduce the effects of contamination and the value of $\sigma_{\theta}(g_{\theta}(\tilde{s}))$ closely approximates the target distribution $p_{\mathcal{D}}(s)$, i.e., $\sigma_{\theta}(g_{\theta}(\tilde{s})) \approx p_{\mathcal{D}}(s)$, we have:

$$\begin{aligned} \xi^* &\approx \frac{\mathbb{E}_{\tilde{p}_{\mathcal{D}}^{(\epsilon)}} [p_{\mathcal{D}}(s)^{\gamma}]}{\mathbb{E}_{p_{\mathcal{D}}} [p_{\mathcal{D}}(s)^{\gamma}]} = \frac{(1-\epsilon)\mathbb{E}_{p_{\mathcal{D}}} [p_{\mathcal{D}}(s)^{\gamma}] + \epsilon\mathbb{E}_{p(s_{\text{flip}})} [p_{\mathcal{D}}(s_{\text{flip}})^{\gamma}]}{\mathbb{E}_{p_{\mathcal{D}}} [p_{\mathcal{D}}(s)^{\gamma}]} \\ &= (1-\epsilon) + \frac{\epsilon}{\mathbb{E}_{p_{\mathcal{D}}} [p_{\mathcal{D}}(s)^{\gamma}]} \cdot \mathbb{E}_{p(s_{\text{flip}})} [p_{\mathcal{D}}(s_{\text{flip}})^{\gamma}]. \end{aligned}$$

When the contamination distribution $\delta_{s_{\text{flip}}}$ is located at the tail of the target distribution $p_{\mathcal{D}}(s)$, meaning that the probability of a sample drawn from $\delta_{s_{\text{flip}}}$ under $p_{\mathcal{D}}(s)$ is sufficiently small, we can see $\mathbb{E}_{\delta_z(s)} [p_{\mathcal{D}}(s)^{\gamma}] \approx 0$. Then, we have $\xi^* \approx (1-\epsilon)$ and thus the contamination rate can be estimated by $\min\{0, 1 - \xi^*\}$.

F.3 Estimator of ξ^*

Since the exact form of $p_{\mathcal{D}}^{(\epsilon)}$ is unknown and computing the integral with respect to $\sigma(g_{\theta})$ in Eq. (22) is intractable, following the same strategy in our objective, we estimate ξ empirically as:

$$\hat{\xi} = \frac{\frac{1}{N} \sum_{i=1}^N \sigma(g_{\theta}(\tilde{s}^{(i)}))^{\gamma}}{\frac{1}{N} \sum_{i=1}^N \sigma(g_{\theta}(\tilde{s}^{(i)}))^{1+\gamma}} = \frac{\sum_{i=1}^N \sigma(g_{\theta}(\tilde{s}^{(i)}))^{\gamma}}{\sum_{i=1}^N \sigma(g_{\theta}(\tilde{s}^{(i)}))^{1+\gamma}}, \quad (26)$$

where the final expression is derived via empirical approximation.

While ξ^* can be estimated via Eq. (26), the resulting estimator $\hat{\xi}^*$ is not necessarily guaranteed to lie within the valid range $[0, 1]$. In fact, the following lemma demonstrates that the estimator $\hat{\xi}^*$ is not operate properly within the valid interval $[0, 1]$.

Lemma 3. Suppose that $0 < \sigma(g_{\theta}(\tilde{s}^{(i)})) \leq 1$ for all $i = 1, \dots, N$. Let $0 < \gamma < \infty$. Then, the estimator $\hat{\xi}^*$ defined in Eq. (26) satisfies $\hat{\xi}^* > 1$.

Proof. If $\hat{\xi}^* \leq 1$, we have

$$\sum_{i=1}^N \sigma(g_{\theta}(\tilde{s}^{(i)}))^{\gamma} \leq \sum_{i=1}^N \sigma(g_{\theta}(\tilde{s}^{(i)}))^{1+\gamma}.$$

However, since $0 < \sigma(g_{\theta}(\tilde{s}^{(i)})) \leq 1$ for all i , we have

$$\sum_{i=1}^N \sigma(g_{\theta}(\tilde{s}^{(i)}))^{\gamma} \geq \sum_{i=1}^N \sigma(g_{\theta}(\tilde{s}^{(i)}))^{1+\gamma} \Rightarrow \hat{\xi} \geq 1,$$

which is contradicted by the above condition. This completes the proof. \square

One possible approach to mitigate this issue is to introduce a scaling parameter v and define $\bar{\xi} := v\xi$ as the scaling term in the extended model m_η , where $\eta = (\xi, \theta)$. Note that v is not optimized; rather, it serves solely as a hyperparameter for scaling the estimated rate ξ^* . In this case, by following the same discussion on Section F.2.2, we obtain, for any v and θ ,

$$\hat{\xi}^* = v^{-1} \cdot \frac{\sum_{i=1}^N \sigma(g_\theta(\tilde{s}^{(i)}))^\gamma}{\sum_{i=1}^N \sigma(g_\theta(\tilde{s}^{(i)}))^{1+\gamma}}.$$

Since the above expression holds for any v , one can select v to ensure that $\hat{\xi} \in [0, 1]$. However, manually tuning v introduces arbitrariness and risks biasing the contamination estimate. Given that the original $\hat{\xi}$ is derived from model likelihoods $\sigma(g_\theta(\tilde{s}^{(i)}))$, it is desirable for the scaling factor to be informed by model-based quantities. The following lemma shows that scaling by the mean model likelihood yields a principled correction to $\hat{\xi}$.

Lemma 4. *Suppose that $0 < \sigma(g_\theta(\tilde{s}^{(i)})) \leq 1$ for all $i = 1, \dots, N$. Let $0 < \gamma < \infty$. Then, we have $0 < v \cdot \hat{\xi} \leq 1$ when we set $v^{-1} = \frac{1}{N} \sum_{i=1}^N \sigma(g_\theta(\tilde{s}^{(i)}))$.*

Proof. Because of $0 < \sigma(g_\theta(\tilde{s}^{(i)})) \leq 1$ for all i , we have $v \cdot \hat{\xi} > 0$. We can reorganize $\{\sigma(g_\theta(\tilde{s}^{(i)}))\}_{i=1}^N$ and $\{\sigma(g_\theta(\tilde{s}^{(i)}))^\gamma\}_{i=1}^N$ so as to be the similarly-ordered-sequences (both are increasing functions of $\sigma(g_\theta(\tilde{s}^{(i)}))$). Then, for two similarly ordered, non-negative sequences $\{\sigma(g_\theta(\tilde{s}^{(i)}))\}_{i=1}^N$ and $\{\sigma(g_\theta(\tilde{s}^{(i)}))^\gamma\}_{i=1}^N$, we obtain

$$\begin{aligned} \frac{1}{N} \sum_{i=1}^N \sigma(g_\theta(\tilde{s}^{(i)}))^{1+\gamma} &= \frac{1}{N} \sum_{i=1}^N \sigma(g_\theta(\tilde{s}^{(i)})) \cdot \sigma(g_\theta(\tilde{s}^{(i)}))^\gamma \\ &\geq \left(\frac{1}{N} \sum_{i=1}^N \sigma(g_\theta(\tilde{s}^{(i)})) \right) \left(\frac{1}{N} \sum_{i=1}^N \sigma(g_\theta(\tilde{s}^{(i)}))^\gamma \right), \end{aligned}$$

where the final inequality comes from Chebyshev's sum inequality. Dividing both sides by $\frac{1}{N} \sum_{i=1}^N \sigma(g_\theta(\tilde{s}^{(i)}))^{1+\gamma}$ completes the proof. \square

From this lemma, we adopt the following estimator for the clean data proportion:

$$\hat{\xi}^* = \left(\frac{1}{N} \sum_{i=1}^N \sigma(g_\theta(\tilde{s}^{(i)})) \right) \cdot \frac{\sum_{i=1}^N \sigma(g_\theta(\tilde{s}^{(i)}))^\gamma}{\sum_{i=1}^N \sigma(g_\theta(\tilde{s}^{(i)}))^{1+\gamma}} = \frac{\frac{1}{N} \sum_{i=1}^N \bar{\sigma}(g_\theta(\tilde{s}^{(i)}))^\gamma}{\sum_{i=1}^N \bar{\sigma}(g_\theta(\tilde{s}^{(i)}))^{1+\gamma}}, \quad (27)$$

where the normalized likelihood is defined as

$$\bar{\sigma}(g_\theta(\tilde{s}^{(i)})) := \frac{\sigma(g_\theta(\tilde{s}^{(i)}))}{\sum_{i=1}^N \sigma(g_\theta(\tilde{s}^{(i)}))}.$$

Recalling that $\hat{\xi}$ estimates the *clean* data ratio, Eq. (27) behaves as desired. Consider a simple case where the model $\sigma(g_\theta)$ perfectly distinguishes between clean and contaminated data, i.e., $\sigma(g_\theta(s_{\text{flip}})) = 1$ for clean samples and $\sigma(g_\theta(s_{\text{flip}})) = 0$ for flipped (noisy) ones. Suppose that M out of N total samples are contaminated. Then:

$$\frac{\sum_{i=1}^N \sigma(g_\theta(\tilde{s}^{(i)}))^\gamma}{\sum_{i=1}^N \sigma(g_\theta(\tilde{s}^{(i)}))^{1+\gamma}} = \frac{N-M}{N-M} = 1, \quad \frac{1}{N} \sum_{i=1}^N \sigma(g_\theta(\tilde{s}^{(i)})) = \frac{N-M}{N}.$$

Multiplying these two terms yields $\hat{\xi}^* = \frac{N-M}{N}$, which exactly recovers the true clean-data proportion.

F.4 Choice of ϕ

To implement Hölder-DPO in practice, one must specify a concrete choice of the function ϕ . For robust LM alignment, a natural choice is the DP divergence with $\phi(h) = \gamma - (1+\gamma)h$, or alternatively, the PS score (equivalently, the γ -score) with $\phi(h) = -h^{1+\gamma}$. The following lemma shows that, when the goal is to achieve *both robustness and contamination ratio estimation* simultaneously, the DP divergence is the preferable choice among these two options.

Lemma 5. Under Assumption 1 and Definition 1. Then, under the extended model m_η , the objective $S_\gamma(p_{\mathcal{D}}^{(\epsilon)} \| m_\eta(\tilde{s}))$ can be approximated around $\theta = \theta^*$ as

$$S_\gamma(p_{\mathcal{D}}^{(\epsilon)} \| m_\eta(\tilde{s})) \approx \begin{cases} (1 - \epsilon) S_{\text{PS}}(p_{\mathcal{D}}(s) \| \sigma(g_\theta(s))) & \text{if } \phi(h) = -h^{1+\gamma}, \\ S_{\text{DP}}((1 - \epsilon)p_{\mathcal{D}} \| m_\eta(s)) & \text{if } \phi(h) = \gamma - (1 + \gamma)h, \end{cases}$$

where S_{DP} and S_{PS} denote the DP divergence and the PS score, respectively, as defined in Appendix F.4.

Under Lemma 5, the DPO objective based on the PS score reduces to $(1 - \epsilon) S_{\text{PS}}(p_{\mathcal{D}} \| \sigma(g_\theta))$ in a neighborhood of θ^* . Consequently, provided that the optimized parameter θ lies within this neighborhood, $\min_\eta S_\gamma(p_{\mathcal{D}}^{(\epsilon)} \| m_\eta)$ recovers $\min_\theta S_{\text{PS}}(p_{\mathcal{D}} \| \sigma(g_\theta))$, independently of the contamination ratio ϵ . Because the scale parameter ξ disappears from this reduced objective, the contamination proportion cannot be identified in the PS-score variant. By contrast, with the DP-divergence-based DPO, the reduced objective $S_{\text{DP}}((1 - \epsilon)p_{\mathcal{D}} \| m_\eta)$ retains ξ , enabling the optimization to *jointly recover* both the target parameter θ^* and the clean-data proportion $1 - \epsilon$. In summary, once a robust solution near θ^* is obtained, the optimized model $\sigma(g_\theta)$ closely approximates $p_{\mathcal{D}}$, while the scaling parameter ξ serves as an accurate estimator of $1 - \epsilon$. For further details, see Appendix F.4. Whether this property holds in practice depends on the ability to estimate θ robustly under $p_{\mathcal{D}}^{(\epsilon)}$; hence, the theoretical robustness guarantee of Hölder-DPO presented in Section 4.1 plays a crucial role.

When applying Hölder-DPO, it is necessary to select an appropriate function ϕ . For the purpose of performing robust DPO, one may consider using the DP divergence with $\phi(h) = \gamma - (1 + \gamma)h$, or the γ -score obtained by setting $\phi(h) = -h^{1+\gamma}$. However, if the goal is to simultaneously *achieve robustness and estimate the contamination ratio*, the following discussion shows that using the DP divergence is preferable.

When we set $\phi(h) = -h^{1+\gamma}$, the Hölder-DPO objective function can be decomposed as:

$$\begin{aligned} S_\gamma(\tilde{p}_{\mathcal{D}}^{(\epsilon)} \| m_\eta(\tilde{s})) &= - \frac{\mathbb{E}_{p_{\mathcal{D}}^{(\epsilon)}}[m_\eta(\tilde{s})^\gamma]}{\underbrace{(\mathbb{E}_{m_\eta}[m_\eta(\tilde{s})^\gamma])^{\frac{\gamma}{1+\gamma}}}_{=: S_{\text{PS}}(\tilde{p}_{\mathcal{D}}^{(\epsilon)} \| m_\eta(\tilde{s}))}} \\ &= -(1 - \epsilon) \frac{\mathbb{E}_{p_{\mathcal{D}}}[m_\eta(s)^\gamma]}{(\mathbb{E}_{m_\eta}[m_\eta(s)^\gamma])^{\frac{\gamma}{1+\gamma}}} - \epsilon \frac{\mathbb{E}_{p(s_{\text{flip}})}[m_\eta(s_{\text{flip}})^\gamma]}{(\mathbb{E}_{m_\eta}[m_\eta(s_{\text{flip}})^\gamma])^{\frac{\gamma}{1+\gamma}}} \\ &= -(1 - \epsilon) \frac{\mathbb{E}_{p_{\mathcal{D}}}[\sigma(g_\theta(s))^\gamma]}{(\mathbb{E}_{\sigma(g_\theta)}[\sigma(g_\theta(s))^\gamma])^{\frac{\gamma}{1+\gamma}}} - \epsilon \frac{\mathbb{E}_{p(s_{\text{flip}})}[\sigma(g_\theta(s_{\text{flip}}))^\gamma]}{(\mathbb{E}_{\sigma(g_\theta)}[\sigma(g_\theta(s_{\text{flip}}))^\gamma])^{\frac{\gamma}{1+\gamma}}} \\ &= (1 - \epsilon) S_{\text{PS}}(p_{\mathcal{D}}(s) \| \sigma(g_\theta(s))) - \epsilon \frac{\mathbb{E}_{p(s_{\text{flip}})}[\sigma(g_\theta(s_{\text{flip}}))^\gamma]}{(\mathbb{E}_{\sigma(g_\theta)}[\sigma(g_\theta(s_{\text{flip}}))^\gamma])^{\frac{\gamma}{1+\gamma}}}. \end{aligned}$$

Under Assumption 1, the optimal solution of $\arg\min_\theta S_\gamma(\tilde{p}_{\mathcal{D}}^{(\epsilon)} \| m_\eta(\tilde{s}))$ will be close to that of $\arg\min_\theta S_{\text{PS}}(p_{\mathcal{D}}(s) \| \sigma(g_\theta(s)))$. This implies that Hölder-DPO with $\phi(h) = -h^{1+\gamma}$ is *robust to heavy contamination*, since it does not require the contamination ratio ϵ to be small.

However, this objective function ignores the parameter ξ , which was introduced in the extended model to estimate the contamination ratio. This is because

$$\begin{aligned} S_{\text{PS}}(\tilde{p}_{\mathcal{D}}^{(\epsilon)} \| m_\eta(\tilde{s})) &= - \frac{\mathbb{E}_{p_{\mathcal{D}}^{(\epsilon)}}[m_\eta(\tilde{s})^\gamma]}{(\mathbb{E}_{m_\eta}[m_\eta(\tilde{s})^\gamma])^{\frac{\gamma}{1+\gamma}}} \\ &= - \frac{\mathbb{E}_{p_{\mathcal{D}}^{(\epsilon)}}[\sigma(g_\theta(\tilde{s}))^\gamma]}{(\int \sigma(g_\theta(\tilde{s}))^{1+\gamma} d\tilde{s})^{\frac{\gamma}{1+\gamma}}} \\ &= S_{\text{PS}}(\tilde{p}_{\mathcal{D}}^{(\epsilon)} \| \sigma(g_\theta(\tilde{s}))), \end{aligned}$$

which implies that the parameter ξ in Eq. (25) does not influence the optimization, and therefore cannot serve as an estimator of $(1 - \epsilon)$. In fact, even when using the enlarged model m_η , we have,

for all $\xi > 0$,

$$\operatorname{argmin}_{\eta=\{\xi, \theta\}} S_{\text{PS}}(\tilde{p}_{\mathcal{D}}^{(\epsilon)} \| m_{\eta}(\tilde{s})) = \operatorname{argmin}_{\theta} S_{\text{PS}}(\tilde{p}_{\mathcal{D}}^{(\epsilon)} \| \sigma(g_{\theta}(\tilde{s}))),$$

which confirms that ξ has no effect on the solution.

On the other hand, the enlarged model becomes effective when we use the DP divergence. When we set $\phi(h) = \gamma - (1 + \gamma)h$, the Hölder-DPO objective becomes:

$$\begin{aligned} S_{\gamma}(\tilde{p}_{\mathcal{D}}^{(\epsilon)} \| m_{\eta}(\tilde{s})) &= \underbrace{\gamma \mathbb{E}_{m_{\eta}}[m_{\eta}(\tilde{s})^{\gamma}] - (1 + \gamma) \mathbb{E}_{p_{\mathcal{D}}^{(\epsilon)}}[m_{\eta}(\tilde{s})^{\gamma}]}_{=: S_{\text{DP}}(\tilde{p}_{\mathcal{D}}^{(\epsilon)} \| m_{\eta}(\tilde{s}))} \\ &= \gamma \mathbb{E}_{m_{\eta}}[m_{\eta}(s)^{\gamma}] - (1 - \epsilon)(1 + \gamma) \mathbb{E}_{p_{\mathcal{D}}}[m_{\eta}(s)^{\gamma}] - \epsilon(1 + \gamma) \mathbb{E}_{p(s_{\text{flip}})}[m_{\eta}(s_{\text{flip}})^{\gamma}] \\ &= S_{\text{DP}}((1 - \epsilon)p_{\mathcal{D}} \| m_{\eta}(s)) - \epsilon(1 + \gamma) \xi \mathbb{E}_{p(s_{\text{flip}})}[\sigma(g_{\theta}(\tilde{s}))^{\gamma}]. \end{aligned}$$

If $\mathbb{E}_{p(s_{\text{flip}})}[\sigma(g_{\theta}(s_{\text{flip}}))^{\gamma}] \approx 0$ around $\theta = \theta^*$, then the optimal solution of $\operatorname{argmin}_{\eta} S_{\gamma}(\tilde{p}_{\mathcal{D}}^{(\epsilon)} \| m_{\eta}(\tilde{s}))$ is close to that of $\operatorname{argmin}_{\eta} S_{\text{DP}}((1 - \epsilon)p_{\mathcal{D}} \| m_{\eta}(s))$. Recalling that the DP divergence is strictly proper over the set of non-negative functions [53], this implies that minimizing the DP divergence with the extended model allows for estimation of both the target parameter θ^* and the clean-data ratio $1 - \epsilon$.

G IF Analysis for the DPO variants (summarized in Theorem 3)

G.1 rDPO do not satisfy the redescending property

The objective of rDPO [23] is as follows:

$$\tilde{\mathcal{L}}_{\text{rDPO}}(\pi_{\theta}; \pi_{\text{ref}}) := \frac{(1 - c) \mathbb{E}_{p_{\mathcal{D}}^{(\epsilon)}}[-\log \sigma(g_{\theta}(\tilde{s}))] - c \mathbb{E}_{p_{\mathcal{D}}^{(\epsilon)}}[-\log \sigma(-g_{\theta}(\tilde{s}))]}{1 - 2c}, \quad (28)$$

where $0 \leq c < 1/2$.

We first show the IF for the rDPO.

Theorem 6. Suppose θ^* denotes the optimal parameters learned from the clean dataset $p_{\mathcal{D}}$, and $\theta^*(\epsilon)$ denotes those learned from the ϵ -contaminated dataset $p_{\mathcal{D}}^{(\epsilon)}$. Let the Hessian $H_{\theta^*}^{(\text{rDPO})}(s) := \nabla_{\theta}^2 \mathcal{L}_{\text{rDPO}}(s, \pi_{\theta})|_{\theta=\theta^*}$ is positive definite. Then, the IF for the rDPO is given by:

$$\text{IF}_{\text{rDPO}}(x, \theta, p_{\mathcal{D}}) = - \left(\mathbb{E}_{p_{\mathcal{D}}} \left[H_{\theta^*}^{(\text{rDPO})}(s) \right] \right)^{-1} \mathbb{E}_{p(s_{\text{flip}})} [F_{\theta^*}^{(\text{rDPO})}(s_{\text{flip}})], \quad (29)$$

where $F_{\theta^*}^{(\text{rDPO})}(s_{\text{flip}}) := \xi_{\theta^*}(s_{\text{flip}}) \left(\nabla_{\theta} \log \pi_{\theta^*}(y_{\text{win}}^{\text{flip}} | x) - \nabla_{\theta} \log \pi_{\theta^*}(y_{\text{lose}}^{\text{flip}} | x) \right)$ and $\xi_{\theta^*}(s_{\text{flip}}) := \frac{1-c}{1-2c} \sigma(-g_{\theta^*}(s_{\text{flip}})) + \frac{c}{1-2c} \sigma(g_{\theta^*}(s_{\text{flip}}))$.

Proof. The gradient of Eq. (28) under $p_{\mathcal{D}}^{(\epsilon)}$ is given by

$$\nabla_{\theta} \tilde{\mathcal{L}}_{\text{rDPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\beta \mathbb{E}_{p_{\mathcal{D}}^{(\epsilon)}} \left[\xi_{\theta^*(\epsilon)}(\tilde{s}) \left(\nabla_{\theta} \log \pi_{\theta}(\tilde{y}_{\text{win}} | \tilde{x}) - \nabla_{\theta} \log \pi_{\theta}(\tilde{y}_{\text{lose}} | \tilde{x}) \right) \right],$$

where

$$\xi_{\theta^*(\epsilon)}(\tilde{s}) := \frac{1-c}{1-2c} \sigma(-g_{\theta^*(\epsilon)}(\tilde{s})) + \frac{c}{1-2c} \sigma(g_{\theta^*(\epsilon)}(\tilde{s})).$$

From the definition of $\theta^*(\epsilon)$, we have $0 = \nabla_{\theta} \tilde{\mathcal{L}}_{\text{rDPO}}(\pi_{\theta}; \pi_{\text{ref}})|_{\theta=\theta^*(\epsilon)}$. By taking the derivation of this term w.r.t. ϵ , we obtain

$$\begin{aligned} 0 &= \frac{\partial}{\partial \epsilon} \nabla_{\theta} \tilde{\mathcal{L}}_{\text{rDPO}}(\pi_{\theta}; \pi_{\text{ref}}) \Big|_{\theta=\theta^*(\epsilon)} \\ &= -\beta \frac{\partial}{\partial \epsilon} \mathbb{E}_{p_{\mathcal{D}}^{(\epsilon)}} \left[\underbrace{\xi_{\theta^*(\epsilon)}(\tilde{s}) \left(\nabla_{\theta} \log \pi_{\theta}(\tilde{y}_{\text{win}} | \tilde{x}) - \nabla_{\theta} \log \pi_{\theta}(\tilde{y}_{\text{lose}} | \tilde{x}) \right)}_{=: F_{\theta^*(\epsilon)}^{(\text{rDPO})}(\tilde{s})} \right] \\ &= -\beta \left\{ \int \left\{ \frac{\partial}{\partial \epsilon} p_{\mathcal{D}}^{(\epsilon)}(\tilde{s}) \right\} F_{\theta^*(\epsilon)}^{(\text{rDPO})}(\tilde{s}) d\tilde{s} + \mathbb{E}_{p_{\mathcal{D}}^{(\epsilon)}} \left[\frac{\partial \theta^*(\epsilon)}{\partial \epsilon} H_{\theta^*(\epsilon)}^{(\text{rDPO})}(\tilde{s}) \right] \right\}, \end{aligned} \quad (30)$$

where $H_{\theta^*(\epsilon)}^{(\text{rDPO})}(\tilde{s}) := \frac{\partial F_{\theta^*(\epsilon)}^{(\text{rDPO})}(\tilde{s})}{\partial \theta^*(\epsilon)}$.

From Definition 1, we obtain

$$\int \left\{ \frac{\partial}{\partial \epsilon} p_{\mathcal{D}}^{(\epsilon)}(\tilde{s}) \right\} F_{\theta^*(\epsilon)}^{(\text{rDPO})}(\tilde{s}) d\tilde{s} = \mathbb{E}_{p(s_{\text{flip}})}[F_{\theta^*(\epsilon)}^{(\text{rDPO})}(s_{\text{flip}})] - \mathbb{E}_{p_{\mathcal{D}}}[F_{\theta^*(\epsilon)}^{(\text{rDPO})}(s)],$$

where $F_{\theta^*(\epsilon)}^{(\text{rDPO})}(s_{\text{flip}}) := \xi_{\theta^*(\epsilon)}(s_{\text{flip}})(\nabla_{\theta} \log \pi_{\theta^*(\epsilon)}(y_{\text{win}}^{\text{flip}} | x) - \nabla_{\theta} \log \pi_{\theta^*(\epsilon)}(y_{\text{lose}}^{\text{flip}} | x))$. By taking $\epsilon \rightarrow 0$, we have

$$\left(\int \left\{ \frac{\partial}{\partial \epsilon} p_{\mathcal{D}}^{(\epsilon)}(\tilde{s}) \right\} F_{\theta^*(\epsilon)}^{(\text{rDPO})}(\tilde{s}) d\tilde{s} \right) \Big|_{\epsilon=0} = \mathbb{E}_{p(s_{\text{flip}})}[F_{\theta^*}^{(\text{rDPO})}(s_{\text{flip}})],$$

since $\theta^{(*)}(\epsilon) \rightarrow \theta^{(*)}$ and thus $\mathbb{E}_{p_{\mathcal{D}}}[F_{\theta^*}^{(\text{rDPO})}(s)] = \nabla_{\theta} \mathcal{L}_{\text{rDPO}}(\pi_{\theta}; \pi_{\text{ref}})|_{\theta=\theta^*} = 0$ from the first-order optimal condition.

Furthermore, we also obtain

$$\mathbb{E}_{p_{\mathcal{D}}^{(\epsilon)}} \left[\frac{\partial \theta^*(\epsilon)}{\partial \epsilon} H_{\theta^*(\epsilon)}^{(\text{rDPO})}(\tilde{s}) \right] \Big|_{\epsilon=0} = \mathbb{E}_{p_{\mathcal{D}}} \left[\frac{\partial \theta^*(\epsilon)}{\partial \epsilon} H_{\theta^*}^{(\text{rDPO})}(s) \right],$$

where $H_{\theta^*}^{(\text{rDPO})}(s) := \frac{\partial F_{\theta^*}^{(\text{rDPO})}(s)}{\partial \theta^*}$.

Then, Eq. (30) under $\epsilon \rightarrow 0$ can be rewritten as

$$0 = \left(\frac{\partial}{\partial \epsilon} \nabla_{\theta} \tilde{\mathcal{L}}_{\text{rDPO}}(\pi_{\theta}; \pi_{\text{ref}}) \Big|_{\theta=\theta^*(\epsilon)} \right) \Big|_{\epsilon=0} = -\beta \left\{ \mathbb{E}_{p(s_{\text{flip}})}[F_{\theta^*}^{(\text{rDPO})}(s_{\text{flip}})] + \mathbb{E}_{p_{\mathcal{D}}} \left[\frac{\partial \theta^*(\epsilon)}{\partial \epsilon} H_{\theta^*}^{(\text{rDPO})}(s) \right] \right\}.$$

By solving the above equality w.r.t. $\frac{\partial \theta^*(\epsilon)}{\partial \epsilon}$, we obtain

$$\frac{\partial \theta^*(\epsilon)}{\partial \epsilon} \Big|_{\epsilon=0} = - \left(\mathbb{E}_{p_{\mathcal{D}}} [H_{\theta^*}^{(\text{rDPO})}(s)] \right)^{-1} \mathbb{E}_{p(s_{\text{flip}})}[F_{\theta^*}^{(\text{rDPO})}(s_{\text{flip}})].$$

This completes the proof. \square

Corollary 3 (rDPO is not robust). *Suppose that the policy gradient $\nabla_{\theta} \log \pi_{\theta}(y | x)$ is bounded by C and satisfies L -Lipchitz in θ , where $0 < C < \infty$ and $0 < L < \infty$. Let $0 \leq c < 1/2$. Then, under Theorem 6, the IF of rDPO do not satisfy the robustness condition in Definition 2, i.e., $\lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \|\text{IF}_{\text{rDPO}}(x, \theta, p_{\mathcal{D}})\| \neq 0$.*

Proof. From Theorem 6, the IF for rDPO is given by

$$\text{IF}_{\text{rDPO}} = - \left(\mathbb{E}_{p_{\mathcal{D}}} [H_{\theta^*}^{(\text{rDPO})}(s)] \right)^{-1} \mathbb{E}_{p(s_{\text{flip}})}[F_{\theta^*}^{(\text{rDPO})}(s_{\text{flip}})].$$

From the positive definite assumption on the Hessian $H_{\theta^*}^{(\text{rDPO})}(s)$, it follows that its expectation $\mathbb{E}_{p_{\mathcal{D}}}[H_{\theta^*}^{(\text{rDPO})}(s)]$ is also a positive definite matrix. Let $L' = \lambda_{\min}(\mathbb{E}_{p_{\mathcal{D}}}[H_{\theta^*}^{(\text{rDPO})}(s)]) > 0$ be its minimum eigenvalue. Then the norm of its inverse is bounded: $\|(\mathbb{E}_{p_{\mathcal{D}}}[H_{\theta^*}^{(\text{rDPO})}(s)])^{-1}\| \leq 1/L'$.

Furthermore, from the assumption that $\|\nabla_\theta \log \pi_\theta(y \mid x)\| \leq C$, we have $\|\nabla_\theta \log \pi_\theta(y_{\text{win}}^{\text{flip}} \mid x) - \nabla_\theta \log \pi_\theta(y_{\text{lose}}^{\text{flip}} \mid x)\| \leq 2C$.

Taking the limit required by Definition 2, we have:

$$\begin{aligned}
& \lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \|\text{IF}_{\text{rDPO}}\| \\
& \leq \lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \left\| \left(\mathbb{E}_{p_{\mathcal{D}}} \left[H_{\theta^*}^{(\text{rDPO})}(s) \right] \right)^{-1} \right\| \cdot \lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \left\| \mathbb{E}_{p(s_{\text{flip}})} [F_{\theta^*}^{(\text{rDPO})}(s_{\text{flip}})] \right\| \\
& \leq (1/L') \cdot \lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \mathbb{E}_{p(s_{\text{flip}})} \left[\underbrace{\xi_{\theta^*}(s_{\text{flip}})}_{\text{IF Weight}} \cdot \underbrace{\|\nabla_\theta \log \pi_\theta(y_{\text{win}}^{\text{flip}} \mid x) - \nabla_\theta \log \pi_\theta(y_{\text{lose}}^{\text{flip}} \mid x)\|}_{\leq 2C} \right] \\
& \leq (1/L') \cdot \lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \mathbb{E}_{p(s_{\text{flip}})} [\xi_{\theta^*}(s_{\text{flip}})] \cdot 2C,
\end{aligned}$$

where

$$\xi_{\theta^*}(s_{\text{flip}}) = \frac{1-c}{1-2c} \sigma(-g_\theta(s_{\text{flip}})) + \frac{c}{1-2c} \sigma(g_\theta(s_{\text{flip}})).$$

We now evaluate the limit of the IF weight $\xi_{\theta^*}(s_{\text{flip}})$:

$$\lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \xi_{\theta^*}(s_{\text{flip}}) = \left(\frac{1-c}{1-2c} \cdot 1 \right) + \left(\frac{c}{1-2c} \cdot 0 \right) = \frac{1-c}{1-2c}.$$

By the bounded convergence theorem, the limit of the expectation is the expectation of the limit. Thus, the IF limit is upper bounded by:

$$\lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \|\text{IF}_{\text{rDPO}}\| \leq (1/L') \cdot \left(\frac{1-c}{1-2c} \right) \cdot 2C = \frac{2C(1-c)}{L'(1-2c)}.$$

The fact $0 < \frac{2C(1-c)}{L'(1-2c)} < \infty$ (according to $0 < C < \infty$, $0 \leq c < 1/2$, and $0 < L' < \infty$) completes the proof. \square

G.2 cDPO do not satisfy the redescending property

The objective of cDPO [76] is as follows:

$$\tilde{\mathcal{L}}_{\text{cDPO}}(\pi_\theta; \pi_{\text{ref}}) := (1-c) \mathbb{E}_{p_{\mathcal{D}}^{(\epsilon)}} [-\log \sigma(g_\theta(\tilde{s}))] - c \mathbb{E}_{p_{\mathcal{D}}^{(\epsilon)}} [-\log \sigma(-g_\theta(\tilde{s}))]. \quad (31)$$

We first show the IF for the cDPO.

Theorem 7. Suppose θ^* denotes the optimal parameters learned from the clean dataset $p_{\mathcal{D}}$, and $\theta^*(\epsilon)$ denotes those learned from the ϵ -contaminated dataset $p_{\mathcal{D}}^{(\epsilon)}$. Let the Hessian $H_{\theta^*}^{(\text{cDPO})}(s) := \nabla_\theta^2 \mathcal{L}_{\text{cDPO}}(s, \pi_\theta)|_{\theta=\theta^*}$ is positive definite. Then, the IF for the rDPO is given by:

$$\text{IF}_{\text{cDPO}}(x, \theta, p_{\mathcal{D}}) = - \left(\mathbb{E}_{p_{\mathcal{D}}} \left[H_{\theta^*}^{(\text{cDPO})}(s) \right] \right)^{-1} \mathbb{E}_{p(s_{\text{flip}})} [F_{\theta^*}^{(\text{cDPO})}(s_{\text{flip}})], \quad (32)$$

where $F_{\theta^*}^{(\text{cDPO})}(s_{\text{flip}}) := \xi_{\theta^*}(s_{\text{flip}}) \left(\nabla_\theta \log \pi_{\theta^*}(y_{\text{win}}^{\text{flip}} \mid x) - \nabla_\theta \log \pi_{\theta^*}(y_{\text{lose}}^{\text{flip}} \mid x) \right)$ and $\xi_{\theta^*}(s_{\text{flip}}) := (1-c) \sigma(-g_{\theta^*}(s_{\text{flip}})) + c \sigma(g_{\theta^*}(s_{\text{flip}}))$.

Proof. The proof follows from the same argument as in Theorem 6, ignoring the $(1-2c)$ term in the denominator. \square

Corollary 4 (cDPO is not robust). Suppose that the policy gradient $\nabla_\theta \log \pi_\theta(y \mid x)$ is bounded by C and satisfies L -Lipchitz in θ , where $0 < C < \infty$ and $0 < L < \infty$. Let $0 \leq c < 1$. Then, under Theorem 7, the IF of cDPO do not satisfy the robustness condition in Definition 2, i.e., $\lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \|\text{IF}_{\text{cDPO}}(x, \theta, p_{\mathcal{D}})\| \neq 0$.

Proof. By following the proof in Corollary 3 and ignoring the $(1 - 2c)$ term in the denominator, we have $0 < \mathbb{E}_{p_{\mathcal{D}}}[H_{\theta^*}^{(\text{cDPO})}(s)] \leq \xi_{\theta^*}(s) \cdot L < \infty$ and thus $\lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \|\text{IF}_{\text{cDPO}}(x, \theta, p_{\mathcal{D}})\|_2 \leq 2C(1 - c)/L'$, where $0 < L' \leq \mathbb{E}_{p_{\mathcal{D}}}[H_{\theta^*}(s)]$ and we use the fact that $\lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \xi_{\theta^*}(s_{\text{flip}}) = 1 - c$. The fact $0 < 2C(1 - c)/L' < \infty$ according to $0 < C < \infty$, $0 \leq c < 1$, and $0 < L' < \infty$ completes the proof. \square

G.3 IPO do not satisfy the redescending property

The objective of IPO [37] is as follows:

$$\tilde{\mathcal{L}}_{\text{IPO}}(\pi_{\theta}; \pi_{\text{ref}}) := \mathbb{E}_{p_{\mathcal{D}}^{(\epsilon)}} \left[\left(\frac{g_{\theta}(\tilde{s})}{\beta} - \frac{1}{2\beta} \right)^2 \right]. \quad (33)$$

We first show the IF for the IPO.

Theorem 8. Suppose θ^* denotes the optimal parameters learned from the clean dataset $p_{\mathcal{D}}$, and $\theta^*(\epsilon)$ denotes those learned from the ϵ -contaminated dataset $p_{\mathcal{D}}^{(\epsilon)}$. Let the Hessian $H_{\theta^*}^{(\text{IPO})}(s) := \nabla_{\theta}^2 \mathcal{L}_{\text{IPO}}(s, \pi_{\theta})|_{\theta=\theta^*}$ is positive definite. Then, the IF for the IPO is given by:

$$\text{IF}_{\text{IPO}}(x, \theta, p_{\mathcal{D}}) = - \left(\mathbb{E}_{p_{\mathcal{D}}} [H_{\theta^*}^{(\text{IPO})}(s)] \right)^{-1} \mathbb{E}_{p(s_{\text{flip}})} [F_{\theta^*}^{(\text{IPO})}(s_{\text{flip}})], \quad (34)$$

where $F_{\theta^*}^{(\text{IPO})}(s_{\text{flip}}) := 2 \left(\frac{g_{\theta}(s_{\text{flip}})}{\beta} - \frac{1}{2\beta} \right) \left(\nabla_{\theta} \log \pi_{\theta^*}(y_{\text{win}}^{\text{flip}} | x) - \nabla_{\theta} \log \pi_{\theta^*}(y_{\text{lose}}^{\text{flip}} | x) \right)$.

Proof. The proof follows from the same argument as in Theorem 2 under the following gradient of Eq. (33):

$$\nabla_{\theta} \tilde{\mathcal{L}}_{\text{IPO}}(\pi_{\theta}; \pi_{\text{ref}}) = \mathbb{E}_{p_{\mathcal{D}}^{(\epsilon)}} \left[2 \left(\frac{g_{\theta}(\tilde{s})}{\beta} - \frac{1}{2\beta} \right) \left(\nabla_{\theta} \log \pi_{\theta}(\tilde{y}_{\text{win}} | \tilde{x}) - \nabla_{\theta} \log \pi_{\theta}(\tilde{y}_{\text{lose}} | \tilde{x}) \right) \right].$$

\square

Corollary 5 (IPO is not robust). Suppose that the policy gradient $\nabla_{\theta} \log \pi_{\theta}(y | x)$ is bounded by C , where $0 < C < \infty$. Let $g_{\theta^*}(s)$ be bounded over $p_{\mathcal{D}}(s)$. Then, under Theorem 8, the IF of IPO do not satisfy the robustness condition in Definition 2, i.e., $\lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \|\text{IF}_{\text{IPO}}(x, \theta, p_{\mathcal{D}})\|_2 = \infty$.

Proof. From Theorem 8, the IF for IPO is $\text{IF}_{\text{IPO}} = -(\mathbb{E}_{p_{\mathcal{D}}}[H_{\theta^*}^{(\text{IPO})}(s)])^{-1} \mathbb{E}_{p(s_{\text{flip}})}[F_{\theta^*}^{(\text{IPO})}]$. From the positive definite assumption on the Hessian, let $L' = \lambda_{\min}(\mathbb{E}_{p_{\mathcal{D}}}[H_{\theta^*}^{(\text{IPO})}(s)]) > 0$. The norm of its inverse is bounded: $\|(\mathbb{E}_{p_{\mathcal{D}}}[H_{\theta^*}^{(\text{IPO})}(s)])^{-1}\| \leq 1/L'$. The gradient term $\|\nabla_{\theta} \log \pi_{\theta}(\dots)\|$ is also bounded by $2C$ from the assumption.

We analyze the limit of the IF:

$$\begin{aligned} & \lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \|\text{IF}_{\text{IPO}}\| \\ & \leq (1/L') \cdot \lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \left\| \mathbb{E}_{p(s_{\text{flip}})} [F_{\theta^*}^{(\text{IPO})}(s_{\text{flip}})] \right\| \\ & \leq (1/L') \cdot \mathbb{E}_{p(s_{\text{flip}})} \left[\lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \left\| \underbrace{2 \cdot \left(\frac{g_{\theta}(s_{\text{flip}})}{\beta} - \frac{1}{2\beta} \right)}_{\text{IF Weight}} \cdot \underbrace{\left(\nabla_{\theta} \log \pi_{\theta^*}(y_{\text{win}}^{\text{flip}} | x) - \nabla_{\theta} \log \pi_{\theta^*}(y_{\text{lose}}^{\text{flip}} | x) \right)}_{\text{Gradient Term}} \right\| \right], \end{aligned}$$

where we use Fatou's Lemma to exchange the limit.

In the limit, $\lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty}$, the IF weight term diverges:

$$\lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \left\| 2 \cdot \left(\frac{g_{\theta}}{\beta} - \frac{1}{2\beta} \right) \right\| = \infty.$$

Since the IF is proportional to the product of this diverging term ($\rightarrow \infty$) and a bounded, non-zero gradient term ($\leq 2C$), the IF itself diverges, that is,

$$\lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \|\text{IF}_{\text{IPO}}(x, \theta, p_{\mathcal{D}})\| = \infty.$$

This completes the proof. \square

G.4 Dr. DPO do not satisfy the redescending property

The objective of Dr. DPO [105] is as follows:

$$\tilde{\mathcal{L}}_{\text{Dr. DPO}}(\pi_{\theta}; \pi_{\text{ref}}) := -\beta' \log \mathbb{E}_{p_{\mathcal{D}}^{(\epsilon)}} \left[\exp \left(\frac{\log \sigma(g_{\theta}(\tilde{s}))}{\beta'} \right) \right]. \quad (35)$$

We first show the IF for the Dr. DPO.

Theorem 9. Suppose θ^* denotes the optimal parameters learned from the clean dataset $p_{\mathcal{D}}$, and $\theta^*(\epsilon)$ denotes those learned from the ϵ -contaminated dataset $p_{\mathcal{D}}^{(\epsilon)}$. Let the Hessian $H_{\theta^*}^{(\text{Dr. DPO})}(s) := \nabla_{\theta}^2 \tilde{\mathcal{L}}_{\text{Dr. DPO}}(s, \pi_{\theta})|_{\theta=\theta^*}$ is positive definite. Then, the IF for the Dr. DPO is given by:

$$\text{IF}_{\text{Dr. DPO}}(x, \theta, p_{\mathcal{D}}) = - \left(\mathbb{E}_{p_{\mathcal{D}}} [H_{\theta^*}^{(\text{Dr. DPO})}(s)] \right)^{-1} \mathbb{E}_{p(s_{\text{flip}})} [F_{\theta^*}^{(\text{Dr. DPO})}(s_{\text{flip}})], \quad (36)$$

where $F_{\theta^*}^{(\text{Dr. DPO})}(s_{\text{flip}}) := w_{\theta^*}(s_{\text{flip}}) \sigma(-g_{\theta^*}(s_{\text{flip}})) \left(\nabla_{\theta} \log \pi_{\theta^*}(y_{\text{win}}^{\text{flip}} | x) - \nabla_{\theta} \log \pi_{\theta^*}(y_{\text{lose}}^{\text{flip}} | x) \right)$ and $w_{\theta^*}(s_{\text{flip}}) := \exp \left(\frac{\log \sigma(g_{\theta^*}(s_{\text{flip}}))}{\beta'} \right) / \mathbb{E}_{p(s_{\text{flip}})} \left[\exp \left(\frac{\log \sigma(g_{\theta^*}(s_{\text{flip}}))}{\beta'} \right) \right]$.

Proof. The gradient of Eq. (35) under $p_{\mathcal{D}}^{(\epsilon)}$ is given by

$$\nabla_{\theta} \tilde{\mathcal{L}}_{\text{Dr. DPO}}(\pi_{\theta}; \pi_{\text{ref}}) = -\beta \mathbb{E}_{p_{\mathcal{D}}^{(\epsilon)}} \left[w_{\theta^*(\epsilon)}(\tilde{s}) \sigma(-g_{\theta^*(\epsilon)}(\tilde{s})) \left(\nabla_{\theta} \log \pi_{\theta}(\tilde{y}_{\text{win}} | \tilde{x}) - \nabla_{\theta} \log \pi_{\theta}(\tilde{y}_{\text{lose}} | \tilde{x}) \right) \right],$$

where

$$w_{\theta^*(\epsilon)}(\tilde{s}) := \frac{\exp \left(\frac{\log \sigma(g_{\theta^*(\epsilon)}(\tilde{s}))}{\beta'} \right)}{\mathbb{E}_{p_{\mathcal{D}}^{(\epsilon)}} \left[\exp \left(\frac{\log \sigma(g_{\theta^*(\epsilon)}(\tilde{s}))}{\beta'} \right) \right]}.$$

From the definition of $\theta^*(\epsilon)$, we have $0 = \nabla_{\theta} \tilde{\mathcal{L}}_{\text{Dr. DPO}}(\pi_{\theta}; \pi_{\text{ref}})|_{\theta=\theta^*(\epsilon)}$. By taking the derivation of this term w.r.t. ϵ , we obtain

$$\begin{aligned} 0 &= \frac{\partial}{\partial \epsilon} \nabla_{\theta} \tilde{\mathcal{L}}_{\text{Dr. DPO}}(\pi_{\theta}; \pi_{\text{ref}}) \Big|_{\theta=\theta^*(\epsilon)} \\ &= -\beta \frac{\partial}{\partial \epsilon} \mathbb{E}_{p_{\mathcal{D}}^{(\epsilon)}} \left[w_{\theta^*(\epsilon)}(\tilde{s}) \sigma(-g_{\theta^*(\epsilon)}(\tilde{s})) \left(\nabla_{\theta} \log \pi_{\theta^*(\epsilon)}(\tilde{y}_{\text{win}} | \tilde{x}) - \nabla_{\theta} \log \pi_{\theta^*(\epsilon)}(\tilde{y}_{\text{lose}} | \tilde{x}) \right) \right] \\ &\quad \underbrace{=: F_{\theta^*(\epsilon)}^{(\text{Dr. DPO})}(\tilde{s})} \\ &= -\beta \left\{ \int \left\{ \frac{\partial}{\partial \epsilon} p_{\mathcal{D}}^{(\epsilon)}(\tilde{s}) \right\} F_{\theta^*(\epsilon)}^{(\text{Dr. DPO})}(\tilde{s}) d\tilde{s} + \mathbb{E}_{p_{\mathcal{D}}^{(\epsilon)}} \left[\frac{\partial}{\partial \epsilon} F_{\theta^*(\epsilon)}^{(\text{Dr. DPO})}(\tilde{s}) \right] \right\} \\ &= -\beta \left\{ \int \left\{ \frac{\partial}{\partial \epsilon} p_{\mathcal{D}}^{(\epsilon)}(\tilde{s}) \right\} F_{\theta^*(\epsilon)}^{(\text{Dr. DPO})}(\tilde{s}) d\tilde{s} + \mathbb{E}_{p_{\mathcal{D}}^{(\epsilon)}} \left[\frac{\partial \theta^*(\epsilon)}{\partial \epsilon} \frac{\partial F_{\theta^*(\epsilon)}^{(\text{Dr. DPO})}(\tilde{s})}{\partial \theta^*(\epsilon)} \right] \right\} \\ &= -\beta \left\{ \int \left\{ \frac{\partial}{\partial \epsilon} p_{\mathcal{D}}^{(\epsilon)}(\tilde{s}) \right\} F_{\theta^*(\epsilon)}^{(\text{Dr. DPO})}(\tilde{s}) d\tilde{s} + \mathbb{E}_{p_{\mathcal{D}}^{(\epsilon)}} \left[\frac{\partial \theta^*(\epsilon)}{\partial \epsilon} H_{\theta^*(\epsilon)}^{(\text{Dr. DPO})}(\tilde{s}) \right] \right\}, \end{aligned} \quad (37)$$

where $H_{\theta^*(\epsilon)}^{(\text{Dr. DPO})}(\tilde{s}) := \frac{\partial F_{\theta^*(\epsilon)}^{(\text{Dr. DPO})}(\tilde{s})}{\partial \theta^*(\epsilon)}$.

From Definition 1, we obtain

$$\int \left\{ \frac{\partial}{\partial \epsilon} p_{\tilde{\mathcal{D}}}^{(\epsilon)}(\tilde{s}) \right\} F_{\theta^*(\epsilon)}^{(\text{Dr. DPO})}(\tilde{s}) d\tilde{s} = \mathbb{E}_{p(s_{\text{flip}})}[F_{\theta^*(\epsilon)}^{(\text{Dr. DPO})}(s_{\text{flip}})] - \mathbb{E}_{p_{\mathcal{D}}}[F_{\theta^*(\epsilon)}^{(\text{Dr. DPO})}(s)],$$

where $F_{\theta^*}(s_{\text{flip}}) := w_{\theta^*}(s_{\text{flip}})\sigma(-g_{\theta^*}(s_{\text{flip}}))(\nabla_{\theta} \log \pi_{\theta^*}(y_{\text{win}}^{\text{flip}} | x) - \nabla_{\theta} \log \pi_{\theta^*}(y_{\text{lose}}^{\text{flip}} | x))$. By taking $\epsilon \rightarrow 0$, we have

$$\left(\int \left\{ \frac{\partial}{\partial \epsilon} p_{\tilde{\mathcal{D}}}^{(\epsilon)}(\tilde{s}) \right\} F_{\theta^*(\epsilon)}^{(\text{Dr. DPO})}(\tilde{s}) d\tilde{s} \right) \Big|_{\epsilon=0} = \mathbb{E}_{p(s_{\text{flip}})}[F_{\theta^*}^{(\text{Dr. DPO})}(s_{\text{flip}})],$$

since $\theta^{(*)}(\epsilon) \rightarrow \theta^{(*)}$ and thus $\mathbb{E}_{p_{\mathcal{D}}}[F_{\theta^*}^{(\text{Dr. DPO})}(s)] = \nabla_{\theta} \mathcal{L}_{\text{Dr. DPO}}(\pi_{\theta}; \pi_{\text{ref}})|_{\theta=\theta^*} = 0$ from the first-order optimal condition.

Furthermore, we also obtain

$$\mathbb{E}_{p_{\tilde{\mathcal{D}}}^{(\epsilon)}} \left[\frac{\partial \theta^*(\epsilon)}{\partial \epsilon} H_{\theta^*(\epsilon)}^{(\text{Dr. DPO})}(\tilde{s}) \right] \Big|_{\epsilon=0} = \mathbb{E}_{p_{\mathcal{D}}} \left[\frac{\partial \theta^*(\epsilon)}{\partial \epsilon} H_{\theta^*}^{(\text{Dr. DPO})}(s) \right],$$

where $H_{\theta^*}^{(\text{Dr. DPO})}(s) := \frac{\partial F_{\theta^*}^{(\text{Dr. DPO})}(s)}{\partial \theta^*}$.

Then, Eq. (37) under $\epsilon \rightarrow 0$ can be rewritten as

$$\begin{aligned} 0 &= \left(\frac{\partial}{\partial \epsilon} \nabla_{\theta} \tilde{\mathcal{L}}_{\text{Dr. DPO}}(\pi_{\theta}; \pi_{\text{ref}}) \Big|_{\theta=\theta^*(\epsilon)} \right) \Big|_{\epsilon=0} \\ &= -\beta \left\{ \mathbb{E}_{p(s_{\text{flip}})}[F_{\theta^*}^{(\text{Dr. DPO})}(s_{\text{flip}})] + \mathbb{E}_{p_{\mathcal{D}}} \left[\frac{\partial \theta^*(\epsilon)}{\partial \epsilon} \Big|_{\epsilon=0} H_{\theta^*}^{(\text{Dr. DPO})}(s) \right] \right\}. \end{aligned}$$

By solving the above equality w.r.t. $\frac{\partial \theta^*(\epsilon)}{\partial \epsilon}$, we obtain

$$\frac{\partial \theta^*(\epsilon)}{\partial \epsilon} \Big|_{\epsilon=0} = - \left(\mathbb{E}_{p_{\mathcal{D}}} [H_{\theta^*}^{(\text{Dr. DPO})}(s)] \right)^{-1} \mathbb{E}_{p(s_{\text{flip}})}[F_{\theta^*}^{(\text{Dr. DPO})}(s_{\text{flip}})].$$

This completes the proof. \square

The following lemma is crucial to show the fact that Dr. DPO does not satisfy the redescending property.

Lemma 6 (Limit of Dr. DPO IF Weight). *Let $w_{\theta^*}(s_{\text{flip}})$ be the IF weight for Dr. DPO as defined in Theorem 9. Then, the limit of the total IF weight is 1, that is,*

$$\lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \mathbb{E}_{p(s_{\text{flip}})} \left[w_{\theta^*}(s_{\text{flip}}) \cdot \sigma(-g_{\theta^*}(s_{\text{flip}})) \right] = 1.$$

Proof. We first analyze the case where $p(s_{\text{flip}}) = \delta(s_{\text{flip}})$ (a single point mass). Here, the expectation in the denominator of w_{θ^*} is equal to the numerator, thus $w_{\theta^*}(s_{\text{flip}}) = 1$. Since $\lim \sigma(-g_{\theta^*}) = 1$, the total weight $\mathbb{E}[1 \cdot 1] = 1$.

We next analyze the case where $p(s_{\text{flip}})$ is a non-degenerate distribution. We track how fast $g_{\theta}(s_{\text{flip}})$ goes to $-\infty$ across the support of $p(s_{\text{flip}})$. Let us define:

$$S := \sup_{s_{\text{flip}}} g_{\theta}(s_{\text{flip}}), \quad r(s_{\text{flip}}) := g_{\theta}(s_{\text{flip}}) - S (\leq 0), \quad G := \{s_{\text{flip}} \mid r(s_{\text{flip}}) = 0\}$$

G is the non-empty set of “worst-case” label-flip samples. Using the bound $\sigma(z) \approx e^z$ for $z \rightarrow -\infty$, $\sigma(g_{\theta}(s_{\text{flip}})) \approx e^S e^{r(s_{\text{flip}})}$. The term $\exp(\log \sigma(g_{\theta})/\beta')$ simplifies to $\sigma(g_{\theta})^{1/\beta'}$. Thus,

$$w_{\theta^*}(s_{\text{flip}}) \approx \frac{(e^S e^{r(s_{\text{flip}})})^{1/\beta'}}{\mathbb{E}_{p(s_{\text{flip}})} [(e^S e^{r(s_{\text{flip}})})^{1/\beta'}]} = \frac{\exp(r(s_{\text{flip}})/\beta')}{\mathbb{E}_{p(s_{\text{flip}})} [\exp(r(s_{\text{flip}})/\beta')]}.$$

As $S \rightarrow -\infty$, the term $w_{\theta^*}(s_{\text{flip}})$ converges to $1/p(G)$ for $s_{\text{flip}} \in G$, and to 0 for $s_{\text{flip}} \notin G$.

The total IF weight is $W_{\text{total}} = w_{\theta^*}(s) \cdot \sigma(-g_{\theta^*}(s))$. We take the limit of its expectation (using the bounded convergence theorem):

$$\begin{aligned} \lim_{S \rightarrow -\infty} \mathbb{E}_{p(s_{\text{flip}})}[W_{\text{total}}] &= \mathbb{E}_{p(s_{\text{flip}})} \left[\lim_{S \rightarrow -\infty} w_{\theta^*}(s) \cdot \lim_{g \rightarrow -\infty} \sigma(-g_{\theta^*}(s)) \right] \\ &= \int_G \left(\lim w_{\theta^*}(s) \right) \cdot (1) \cdot p(s) ds + \int_{G^c} (0) \cdot (1) \cdot p(s) ds \\ &= \int_G \left(\frac{1}{p(G)} \right) \cdot p(s) ds = \frac{1}{p(G)} \int_G p(s) ds = \frac{p(G)}{p(G)} = 1. \end{aligned}$$

Thus, the limit of the total IF weight is 1 in all cases. \square

Now we can show the following corollary.

Corollary 6. *Suppose that the policy gradient $\nabla_{\theta} \log \pi_{\theta}(y \mid x)$ is bounded by C and satisfies L -Lipchitz in θ , where $0 < C < \infty$ and $0 < L < \infty$. Let the number of the label-flip data be $\lfloor N\epsilon \rfloor = M (< \infty)$, and $0 < \beta' < \infty$. Let the weight term in the gradient of Dr. DPO: $w_{\theta^*}(s)$ is bounded on $p_{\mathcal{D}}$. Then, under Theorem 9, the IF of Dr. DPO do not satisfy the robustness condition in Definition 2, i.e., $\lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \|\text{IF}_{\text{Dr. DPO}}(x, \theta, p_{\mathcal{D}})\| \neq 0$.*

Proof. The IF for Dr. DPO is

$$\text{IF}_{\text{Dr. DPO}} = - \left(\mathbb{E}_{p_{\mathcal{D}}} \left[H_{\theta^*}^{(\text{Dr. DPO})}(s) \right] \right)^{-1} \mathbb{E}_{p(s_{\text{flip}})} [F_{\theta^*}^{(\text{Dr. DPO})}(s_{\text{flip}})].$$

From the positive definite assumption on the Hessian, let $L' = \lambda_{\min}(\mathbb{E}_{p_{\mathcal{D}}} [H_{\theta^*}^{(\text{Dr. DPO})}(s)]) > 0$. The norm of its inverse is bounded: $\|(\mathbb{E}_{p_{\mathcal{D}}} [H_{\theta^*}^{(\text{Dr. DPO})}(s)])^{-1}\| \leq 1/L'$. The gradient term is bounded by $2C$.

We analyze the limit of the IF:

$$\begin{aligned} \lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \|\text{IF}_{\text{Dr. DPO}}\| &\leq (1/L') \cdot \lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \left\| \mathbb{E}_{p(s_{\text{flip}})} [F_{\theta^*}^{(\text{Dr. DPO})}(s_{\text{flip}})] \right\| \\ &\leq (1/L') \cdot \lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \mathbb{E}_{p(s_{\text{flip}})} \left[\underbrace{w_{\theta^*}(s_{\text{flip}}) \cdot \sigma(-g_{\theta^*}(s_{\text{flip}}))}_{\text{Total IF Weight}} \cdot \underbrace{\|\nabla_{\theta} \log \pi_{\theta}(\dots)\|}_{\leq 2C} \right] \\ &\leq (1/L') \cdot \lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \mathbb{E}_{p(s_{\text{flip}})} [w_{\theta^*}(s_{\text{flip}}) \cdot \sigma(-g_{\theta^*}(s_{\text{flip}}))] \cdot 2C \end{aligned}$$

As shown by Lemma 6, the limit of the total IF weight $\lim \mathbb{E}_{p(s_{\text{flip}})} [w_{\theta^*}(s) \cdot \sigma(-g_{\theta^*}(s))] = 1$.

Therefore, the IF limit is upper bounded by:

$$\lim_{\hat{r}_{\theta^*}(x, y_{\text{lose}}^{\text{flip}}) \rightarrow \infty} \|\text{IF}_{\text{Dr. DPO}}\| \leq (1/L') \cdot 1 \cdot 2C = 2C/L'.$$

The fact $0 < 2C/L' < \infty$ completes the proof. \square

H Additional Experimental details

```

1 import torch.nn.functional as F
2
3 # pi_logps : policy logprobs, shape (B,)
4 # ref_logps : reference model logprobs, shape (B,)
5 # yw_idxes : preferred completion indices, shape (T,)
6 # yl_idxes : dispreferred indices, shape (T,)
7 # beta, beta_1 : regularization coefficients
8
9 pi_yw_logps = pi_logps[yw_idxes]
10 pi_yl_logps = pi_logps[yl_idxes]
11 ref_yw_logps = ref_logps[yw_idxes]
12 ref_yl_logps = ref_logps[yl_idxes]
13
14 reward_win = pi_yw_logps - ref_yw_logps
15 reward_lose = pi_yl_logps - ref_yl_logps
16 g_theta = reward_win - reward_lose
17
18 if self.method == "dpo":
19     loss = -F.logsigmoid(self.beta * g_theta).mean()
20 elif self.method == "holder_dpo":
21     p = F.sigmoid(self.beta * g_theta)
22     loss = - (1.0 + self.gamma) * p.pow(self.gamma).mean() \
23           + self.gamma * (p.pow(self.gamma + 1)).mean()
24 return loss

```

Figure 6: Pseudocode for Hölder-DPO and DPO objectives

Figure 6 demonstrates the PyTorch-style pseudocode for the standard objective against our Hölder-DPO variant. Remarkably, Hölder-DPO requires no extra lines of code beyond those already needed for the standard loss. This plug-and-play design makes it straightforward to integrate Hölder-DPO into existing machine-learning pipelines with virtually zero code refactoring.

H.1 Dataset and model details

Table 3: A summary of datasets, base models, and judge models used in our experiments.

type	name	Hugging Face URL
dataset	IMDb [71]	https://huggingface.co/datasets/stanfordnlp/imdb
	Golden HH dataset [15]	https://huggingface.co/datasets/Unified-Language-Model-Alignment/Anthropic_HH_Golden
	OASST1-tasksource [89]	https://huggingface.co/datasets/tasksource/oasst1_pairwise_rlhf_reward
base model	GPT2-large [83]	https://huggingface.co/openai-community/gpt2-large
	Qwen2.5-1.5B [106]	https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct
	Phi-2 [50]	https://huggingface.co/microsoft/phi-2
	Minstral-8B [93]	https://huggingface.co/mistralai/Minstral-8B-Instruct-2410
	NeMo-12B [94]	https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407
judge models	SIEBERT [69, 46]	https://huggingface.co/siebert/sentiment-roberta-large-english
	GPT-4 [1]	gpt-4.1-nano-2025-04-14 from https://openai.com/api/

Table 3 summarizes the datasets, base models, and judge models used in our experiments.

H.2 Training and hyperparameter details

Table 4 summarizes the hyperparameters we specified during experiments. We basically used TRL default hyperparameters and existing works setting [84, 105] otherwise specified in Table 4. Each training takes 24 hours in the wall-clock time.