# Generalizing Offline Alignment Theoretical Paradigm with Diverse Divergence Constraints

Haoyuan Sun [1] Yuxin Zheng [1] Yifei Zhao [1] Yongzhe Chang [1] Xueqian Wang [1]

## Abstract

The enhanced capabilities of large language models (LLMs) necessitate effective AI alignment. Learning from preference-based feedback has recently become popular as a promising approach to align large language models with human preference. Despite the impressive capabilities demonstrated by these aligned models across various tasks, they lack a unified theoretical framework for expression and deeper theoretical understanding. In this work, we propose the unified theoretical paradigm on human preference-based optimization, known as the Unified Preference Optimization (UPO), which can be proven as the generalization of ΨPO. Through understanding of Unified Preference Optimization (UPO), we can obtain a deeper theoretical comprehension of the practical algorithms, as UPO serves as a generalization for them. Furthermore, we explore a specific scenario of UPO by simply setting the mapping to the Identity. By employing this method, we develop a novel practical algorithm, with the name of Identity Unified Preference Optimization (IUPO). It can be demonstrated that IUPO serves as a generalization of IPO under diverse divergence constraints. Our experiments comparing JS-divergence based IUPO to IPO on the fine-tuning task of GPT2 demonstrate that IUPO, particularly JS-IUPO, outperforms IPO.

## 1. Introduction

The significant advancement in the capabilities of large language models (LLMs) provides a substantial step towards achieving artificial general intelligence (AGI). However, this development has also brought about a plethora of concurrent risks, which requires effective AI alignment.

Reinforcement learning from Human Feedback(Christiano et al., 2023) (RLHF) has been proven to be effective in aligning the behavior of LLMs with human preferences and following human instructions. The process can be summarized as three main steps: 1) supervised fine-tuning, 2) reward model training and 3) RL fine-tuning. The RLHF pipeline, while effective, is considerably more intricate than supervised learning. Specially, RLHF greatly benefits from training an individual reward model. However, researchers has found that LLMs may exploit errors learned through the reward model(Gao et al., 2022), which raises potential misuse of LLMs(Hendrycks et al., 2023)(Shevlane et al., 2023). Furthermore, reinforcement learning algorithms such as PPO often exhibit less stability and demand more memory.

Direct Preference Optimisation (DPO)(Rafailov et al., 2023) has been proposed as an approach that skips the reward model building stage and learns a policy directly from collected data. The essence of this method lies in leveraging the mapping between the reward function and the optimal policy. Ψ-preference optimization (ΨPO)(Azar et al., 2023) demonstrates the feasibility of characterizing the objective functions of Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO) as specific instances of a broader objective exclusively formulated based on pairwise preferences. Furthermore, it provides a straightforward solution to mitigate the issue of overfitting by setting Ψ to identity in the ΨPO, calling Identity-PO (IPO), whose construction circumvents the modeling assumption of the Bradley-Terry model(Bradley & Terry, 1952) for preferences.

Nonetheless, most current studies focus on solutions limited to the KL divergence, with a lack of exploration into the integration of other divergences. In addition, it has been noted that fine-tuning large language models with RLHF under the KL regularization only will lead to a narrow range of political perspectives(Santurkar et al., 2023). As a countermeasure, incorporating various divergences can lead to solutions. $f$-DPO(Wang et al., 2023) generalizes

[1] Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen, China. Correspondence to: Yongzhe Chang <changyongzhe@sz.tsinghua.edu.cn>, Xueqian Wang <wang.xq@sz.tsinghua.edu.cn>.

*Proceedings of the ICML 2024 Workshop on Models of Human Feedback for AI Alignment*, Vienna, Austria. PMLR 235, 2024.

the DPO framework to integrate a range of divergence constraints, which not only eliminates the need for estimating the normalizing constant in the Bradley-Terry model, but also enables a tractable mapping between the reward function and the optimal policy.

In this work, we generalize the ΨPO framework(Azar et al., 2023) to incorporate $f$-divergence constraints which is named **Unified Preference Optimization (UPO)**. By analyzing the generalized optimization target using the Karush–Kuhn–Tucker (KKT) conditions, it can be demonstrated that the normalizing constant can be eliminated in the Bradley-Terry model. Furthermore, we set $\Psi$ as a special function, it can be demonstrated that $f$-DPO is a specific form of UPO. By employing various mathematical methods, we can establish the relationship between UPO, $f$-DPO, ΨPO and DPO: **UPO is the generalization of both ΨPO and $f$-DPO, and at the same time, ΨPO and $f$-DPO are generalizations of DPO.** Furthermore, motivated by the work(Azar et al., 2023), we propose a novel practical solution, achieved by setting $\Psi$ straightforward to the Identity in the UPO with the name of **Identity-UPO (IUPO), which guides to the practical solution of UPO.** The experiments are conducted by comparing JS-divergence based IUPO to IPO on the GPT2 language model(Radford et al., 2019) under different penalty coefficients. The experimental results indicate that, across a lot scenarios, IUPO, especially IUPO with JS-divergence constrains, consistently outperforms IPO. Hence, the main contributions of this work can be summarized as follows:

- A unified theoretical paradigm on human preference-based optimization **UPO** is proposed, which is the generalization of ΨPO, $f$-DPO and DPO.
- **IUPO**, as a practical solution of UPO, is the generalization of IPO within the $f$-divergence constrains.
- Experiments comparing JS-divergence based IUPO to IPO demonstrate that IUPO does outperform IPO.

## 2. Preliminary

### 2.1. Bradley-Terry Model

Bradley-Terry Model(Bradley & Terry, 1952) has been widely employed for pairwise comparisons. In practice, it usually adopts a specific form, denoted as

$$\mathbb{P}\left(y_{win} \succeq y_{loss}\right) = \frac{\exp\left(\mathcal{R}\left(y_{win}\right)\right)}{\exp\left(\mathcal{R}\left(y_{win}\right)\right) + \exp\left(\mathcal{R}\left(y_{loss}\right)\right)} \tag{1}$$

, where $\mathcal{R}(y_{win})$ represents the rating of the item preferred by human, and in contrast, $\mathcal{R}(y_{loss})$ represents the item that are less preferred by human. Furthermore, (1) is often expressed in the logistic formula as follows:

$$\begin{aligned}\mathbb{P}\left(y_{win} \succeq y_{loss}\right) &= \frac{1}{1 + \exp\left(-\left(\mathcal{R}\left(y_{win}\right) - \mathcal{R}\left(y_{loss}\right)\right)\right)} \\ &= \sigma\left(\mathcal{R}\left(y_{win}\right) - \mathcal{R}\left(y_{loss}\right)\right)\end{aligned} \tag{2}$$

, where $\sigma(x)$ is the Sigmoid function.

### 2.2. RLHF: Reinforcement Learning from Human Feedbacks

The impact of RLHF on fine-tuning the behavior of LLMs to better align with human values, such as helpfulness and harmlessness, has been thoroughly investigated. The advantages of RLHF have also been demonstrated in specific tasks, such as summarization, where models are trained to condense extensive information into succinct representations. The technique has three steps: supervised fine-tuning, reward model training and RL fine-tuning. In the last step, it often maximizes the following objective:

$$\mathbb{E}_{x\sim\mathcal{D}_{prompt}, y\sim\pi_\theta(\cdot|x)}[r_\varphi(y|x)] - \beta D_{\mathrm{KL}}(\pi_\theta(\cdot|x)|\pi_{ref}(\cdot|x)) \tag{3}$$

, where $\mathcal{D}_{prompt} = \left\{x^i, y^i_{win}, y^i_{loss}\right\}_{i=1}^N$ is the human preference dataset sampled from P; $r_\varphi(y|x)$represents the reward function learned from the dataset; $\pi_{ref}(\cdot|x)$ is the reference model which is often the supervised fine-tuning model in step 1; and $\beta$ is the penalty coefficient of KL divergence between $\pi_\theta(\cdot|x)$ and $\pi_{ref}(\cdot|x)$, with the purpose of avoiding model degeneration.

### 2.3. DPO: Direct Preference Optimization

The initial DPO technique (Rafailov et al., 2023) establishes a functional mapping between the reward model and the optimum policy with the constraint of KL divergence. With the method of reparameterizing the reward function using the policy in a supervised manner, it makes directly optimize the policy possible. By setting reward function as

$$\mathcal{R}(\cdot|x) = \beta \log \frac{\pi_\theta(\cdot|x)}{\pi_{ref}(\cdot|x)} + \beta \log \mathcal{Z}\left(x\right)$$

, where $\mathcal{Z}\left(x\right)$ is the partition function or the normalizing constant. Substituting the reward function into the Bradley-Terry model, it can be found that the partition function can be cancelled out. Hence, the loss function can be given as follows:

$$-\mathbb{E}_{\mathcal{D}}\left[\log\sigma\left(\beta\log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \beta\log\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}\right)\right] \tag{4}$$

, where $\mathcal{D} = \left\{x^i, y^i_w, y^i_l\right\}_{i=1}^N$ is the human preference dataset, $y^i_w$ represents the item preferred by human, and

$y_l^i$ represents the item that are less preferred by human. In the subsequent discussion, we will continue such notation.

### 2.4. $f$-divergence and $f$-DPO

The $f$-divergence encompasses a wide range of frequently utilized divergences, including reverse KL divergence, forward KL divergence, Jensen-Shannon divergence, $\alpha$-divergence and so on. Its rigorous mathematical definition is as follows. For any convex function $f$: $\mathrm{R}^+ \rightarrow \mathrm{R}$, satisfying $f(1) = 0$ and $f$ is strictly convex around 1, the divergence between two distributions $p(x)$ and $q(x)$ can be defined as:

$$D_f\left(p, q\right) = E_{q(x)}\left[f\left(\frac{p\left(x\right)}{q\left(x\right)}\right)\right].$$

The DPO framework is generalized to $f$-DPO (Wang et al., 2023) by integrating a variety of $f$-divergence for regularization. Through addressing the Karush–Kuhn–Tucker (KKT) conditions of the optimization objective, the reward function can be rewritten as:

$$\mathcal{R}(y|x) = \beta f'\left(\frac{\pi^*(y|x)}{\pi_{ref}(y|x)}\right) + \text{const.}$$

By plugging the reward function into the Bradley-Terry model, the following expression can be given:

$$\begin{aligned}&\mathbb{P}\left(y_w \succeq y_l\right)\\ &= \sigma\left(\beta f'\left(\frac{\pi^*(y_w|x)}{\pi_{ref}(y_w|x)}\right) - \beta f'\left(\frac{\pi^*(y_l|x)}{\pi_{ref}(y_l|x)}\right)\right)\end{aligned}$$

, where $\pi^*$ is the optimal policy; $f'$ is the derivative of function $f$. Therefore, in order to train a better model $\pi_\theta$, we can minimize the following negative log-likelihood loss function:

$$-\mathbb{E}_{\mathcal{D}}\left[\log\sigma\left(\beta f'\left(\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)}\right) - \beta f'\left(\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}\right)\right)\right] \tag{5}$$

When $D_f$ is the reverse KL divergence, we have $f(x) = x\log x$, and $f'(x) = \log x + 1$. In this situation, by simplifying (5), we can obtain the formula of (4). This indicates that $f$-DPO is a generalization of DPO.

### 2.5. $\Psi$PO: $\Psi$-Preference Optimization

In the work (Azar et al., 2023), it is demonstrated that the objective functions of RLHF and DPO can be characterized as special cases of a more general objective exclusively expressed using pairwise preferences. The objective is called $\Psi$-preference optimization ($\Psi$PO) objective, where $\Psi$ is an arbitrary non-deceasing mapping. The objective can be described as follows. Considering a general non-decreasing mapping $\Psi$ : $[0, 1] \rightarrow \mathrm{R}$ and a real positive regularisation parameter $\beta \in \mathrm{R}_+^*$, the $\Psi$PO objective can be defined as:

$$\max_\pi \mathop{\mathbb{E}}_{\substack{x\sim\rho\\ y\sim\pi(.|x)\\ y'\sim\mu(.|x)}}\left[\Psi\left(p^*\left(y \succeq y'|x\right)\right) - \beta D_{\mathrm{KL}}\left(\pi||\pi_{ref}\right)\right] \tag{6}$$

, where $x$ is the given context and follows the context distribution $\rho$; $y$ is the action generated by the policy $\pi$ that is an discrete probability distribution and associates to each context; $y'$ is the action generated by the reference policy $\pi_{ref}$ and $\mu(\cdot|x)$ is the behaviour policy. The true human preference, denoted as $p^*\left(y \succeq y'|x\right)$, is defined as the probability of action generated by the policy $\pi$ being preferred to action generated by the reference policy $\pi_{ref}$ given the context $x$.

By supposing that $\Psi(x) = \log(x/(1-x))$, the first term of (6) holds the following formula:

$$\begin{aligned}&\mathbb{E}_{y'\sim\mu(\cdot|x)}\left[\Psi\left(p^*\left(y \succeq y'|x\right)\right)\right]\\ =&\mathbb{E}_{y'\sim\mu(\cdot|x)}\left[\Psi\left(\frac{e^{r(y)}}{e^{r(y)} + e^{r(y')}}\right)\right]\\ =&\mathbb{E}_{y'\sim\mu(\cdot|x)}\left[\log\frac{e^{r(y)}}{e^{r(y')}}\right] = \mathrm{E}_{y'\sim\mu(\cdot|x)}\left[r\left(y\right) - r\left(y'\right)\right]\\ =&r\left(y\right) + \text{const}\end{aligned} \tag{7}$$

The result is equivalent to the reward in (3), and in the statement of DPO, we know that the optimal policy for the DPO objective in (4) is identical to the RLHF objective in (3). Hence, with a special form of $\Psi(x)$, it can be proved that $\Psi$PO is the generalization of RLHF and DPO.

### 2.6. IPO: Identity Preference Optimization

The Identity Preference Optimization (IPO) can be viewed as a practical application of the $\Psi$-Preference Optimization ($\Psi$PO). Specifically, this is achieved by setting $\Psi$ as the identity mapping in (6), leading to the direct regularized optimization of total preferences, just as the following formula:

$$\max_\pi \mathbb{P}^*\left(\pi \succeq \pi_{ref}\right) - \beta D_{\mathrm{KL}}\left(\pi||\pi_{ref}\right) \tag{8}$$

Through appropriate practical simplifications, the loss function for the IPO is ultimately derived. Firstly, we define:

$$\begin{aligned}h_\pi\left(x, y_w, y_l\right) &= \log\left(\frac{\pi\left(y_w\left|x\right.\right)\pi_{ref}\left(y_l\left|x\right.\right)}{\pi\left(y_l\left|x\right.\right)\pi_{ref}\left(y_w\left|x\right.\right)}\right)\\ &= \log\left(\frac{\pi\left(y_w\left|x\right.\right)}{\pi\left(y_l\left|x\right.\right)}\right) - \log\left(\frac{\pi_{ref}\left(y_w\left|x\right.\right)}{\pi_{ref}\left(y_l\left|x\right.\right)}\right).\end{aligned} \tag{9}$$

Then, the loss function of IPO can be written as :

$$\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[h_\pi\left(x, y_w, y_l\right) - \frac{1}{2\beta}\right]^2 \tag{10}$$

Comparing the loss function in (10) and (4), we can know that IPO, unlike DPO, is consistently regularized towards $\pi_{ref}$ by managing the discrepancy between the log-likelihood ratios $\log\left(\pi\left(y_w\,|x\right)/\pi\left(y_l\,|x\right)\right)$ and $\log\left(\pi_{ref}\left(y_w\,|x\right)/\pi_{ref}\left(y_l\,|x\right)\right)$, thereby mitigating overfitting to the preference dataset.

## 3. Methodology

### 3.1. The Unified Preference Optimization (UPO)

In the previous human preference-based algorithms (especially ΨPO: a general theoretical paradigm on learning from human preference), it is customary to apply regularization to the fine-tuned model in order to ensure its proximity to the original or reference model, as measured by KL divergence. However, this form of regularization is excessively limiting. Our goal is to build a wider range of regularization, with the help of $f$-divergence, which encompasses many commonly employed divergences such as forward KL, reverse KL, JS divergence and so on.

Starting with the objective of ΨPO in (6), we formulate the Unified Preference Optimization (UPO) objective as follows:

$$\max_{\pi} \mathop{\mathbb{E}}_{\substack{x\sim\rho \\ y\sim\pi(.|x) \\ y'\sim\mu(.|x)}} \left[\Psi\left(p^*\left(y \succeq y'|x\right)\right) - \beta D_f\left(\pi, \pi_{ref}\right)\right] \quad (11)$$

In order to get a general solution form of the optimization objective mentioned above, we propose the following theorem:

**Theorem 3.1.** *For all valid $x$, $\pi_{ref} \succ 0$ and $f'(x)$ is invertible with $0 \notin dom(f'(x))$, the expectation of the mapping $\Psi$ can be reparameterized using the policy model $\pi$ and the reference model $\pi_{ref}$ as following formula:*

$$\mathop{\mathbb{E}}_{\substack{x\sim\rho \\ y\sim\pi(.|x) \\ y'\sim\mu(.|x)}} \Psi\left(p^*\left(y \succeq y'|x\right)\right) = \beta f'\left(\frac{\pi^*\left(y\,|x\,\right)}{\pi_{ref}\left(y\,|x\,\right)}\right) + \lambda \quad (12)$$

*Proof.* We consider this problem from a optimization perspective, and describe it as a constrained optimization problem:

$$\max_{\pi} \mathop{\mathbb{E}}_{\substack{x\sim\rho \\ y\sim\pi(.|x) \\ y'\sim\mu(.|x)}} \left[\Psi\left(p^*\left(y \succeq y'|x\right)\right) - \beta D_f\left(\pi, \pi_{ref}\right)\right]$$

$$s.t. \quad \forall y \quad \sum_y \pi\left(y|x\right) = 1 \quad \text{and} \quad \pi\left(y|x\right) \geq 0$$

In order to solve the constrained problem, we apply the Lagrange Multiplier Method, which the following formula can be given:

$$\mathcal{L} = \mathop{\mathbb{E}}_{\substack{x\sim\rho \\ y\sim\pi(.|x) \\ y'\sim\mu(.|x)}} \left[\Psi\left(p^*\left(y \succeq y'|x\right)\right) - \beta f\left(\frac{\pi\left(y\,|x\,\right)}{\pi_{ref}\left(y\,|x\,\right)}\right)\right] - \lambda\left[\sum_y \pi\left(y|x\right) - 1\right] + \sum_y \eta(y)\pi\left(y|x\right) \quad (13)$$

Examining the Karush-Kuhn-Tucker (KKT) conditions for this optimization problem:

**1. Stationarity Condition:**

It is required that the gradient of the Equation with respect to the primal variables ($\pi(y|x)$) be zero:

$$\forall y \quad \nabla_{\pi(y|x)}\mathcal{L}(\pi(y|x), \lambda, \eta(y)) = 0$$

Hence, we can obtain:

$$\begin{aligned} &\nabla_{\pi(y|x)}\mathcal{L}(\pi(y|x), \lambda, \eta(y)) \\ &= \mathop{\mathbb{E}}_{\substack{x\sim\rho \\ y\sim\pi(.|x) \\ y'\sim\mu(.|x)}} \Psi\left(p^*\left(y \succeq y'|x\right)\right) - \beta\pi_{ref}\left(y|x\right) f'\left(\frac{\pi\left(y\,|x\right)}{\pi_{ref}\left(y\,|x\right)}\right) \\ &\cdot \frac{1}{\pi_{ref}\left(y\,|x\right)} - \lambda + \eta\left(y\right) \\ &= \mathop{\mathbb{E}}_{\substack{x\sim\rho \\ y\sim\pi(.|x) \\ y'\sim\mu(.|x)}} \Psi\left(p^*\left(y \succeq y'|x\right)\right) - \beta f'\left(\frac{\pi\left(y\,|x\,\right)}{\pi_{ref}(y\,|x)}\right) - \lambda + \eta\left(y\right) \\ &= 0 \end{aligned} \quad (14)$$

Therefore, we can derive the following equation:

$$\mathop{\mathbb{E}}_{\substack{x\sim\rho \\ y\sim\pi(.|x) \\ y'\sim\mu(.|x)}} \Psi\left(p^*\left(y \succeq y'|x\right)\right) = \beta f'\left(\frac{\pi\left(y\,|x\right)}{\pi_{ref}(y\,|x)}\right) + \lambda - \eta\left(y\right) \quad (15)$$

**2. Primal Feasibility:**

The ultimate solution must meet the initial constraints of the problem.

$$\forall y \quad \sum_y \pi\left(y|x\right) = 1 \quad \text{and} \quad \pi\left(y|x\right) \geq 0$$

**3. Dual Feasibility:**

It dictates that the Lagrange multipliers for the inequality constraints must be non-negative.

$$\forall y \quad \eta\left(y\right) \geq 0$$

**4. Complementary Slackness:**

It requires that each inequality constraint is either fulfilled with equality or has a corresponding Lagrange multiplier of zero.

$$\forall y \quad \pi\left(y|x\right)\eta\left(y\right)=0$$

Upon this condition, certain solutions can be disregarded, as $\pi\left(y|x\right)=0$ or $\eta\left(y\right)$ must hold for each $y$.

Therefore, for functions $f$ in which $0 \notin dom(f'(x))$ and under the assumption that $\pi_{ref} \succ 0$ almost everywhere, it can be deduced that $\pi\left(y|x\right) \succ 0$ almost everywhere. Thus, we must have: $\forall y \quad \eta\left(y\right) \equiv 0$. This implies that (15) can be simplified, as shown below:

$$\underset{\substack{x\sim\rho \\ y\sim\pi(.|x) \\ y'\sim\mu(.|x)}}{\mathbb{E}} \Psi\left(p^*\left(y \succeq y'|x\right)\right) = \beta f'\left(\frac{\pi\left(y\,|x\right)}{\pi_{ref}(y\,|x)}\right)+\lambda \tag{16}$$

□

By integrating human preferred action $y_w$ and less preferred action $y_l$ into (16), we can establish the following general form:

$$\begin{aligned}&\underset{\substack{x\sim\rho \\ y_w,y_l\sim\pi(.|x) \\ y'_w,y'_l\sim\mu(.|x)}}{\mathbb{E}} \left[\Psi\left(p^*\left(y_w \succeq y'_w|x\right)\right)-\Psi\left(p^*\left(y_l \succeq y'_l|x\right)\right)\right] \\ &= \beta\left[f'\left(\frac{\pi\left(y_w|x\right)}{\pi_{ref}\left(y_w|x\right)}\right)-f'\left(\frac{\pi\left(y_l\,|x\right)}{\pi_{ref}\left(y_l|x\right)}\right)\right]\end{aligned} \tag{17}$$

### 3.2. The Relationship between UPO and $f$-DPO

Motivated by the form of (7), we observe that **UPO serves as a generalization of $f$-DPO**. Actually, the following proposition establishes the connection between UPO and $f$-DPO.

**Proposition 3.2.** *Setting $\Psi(x) = \log(x/(1-x))$, and the true human preference $p^*$ satisfies to the Bradley-Terry model, then the optimal policy for UPO in (17) is equivalent to formula for $f$-DPO in (5).*

*Proof.* As shown in (7), for action $y_w$, we can obtain the conclusion as follows:

$$\underset{\substack{x\sim\rho \\ y_w\sim\pi(.|x) \\ y'_w\sim\mu(.|x)}}{\mathbb{E}} \Psi\left(p^*\left(y_w \succeq y'_w|x\right)\right)=\mathcal{R}\left(y_w|x\right)+\lambda$$

Combining the (12) derived in Theorem 3.1, we can get:

$$\mathcal{R}\left(y_w|x\right)=\beta f'\left(\frac{\pi\left(y_w|x\right)}{\pi_{ref}\left(y_w|x\right)}\right)+\lambda^{'} \tag{18}$$

Similarly, in the case of action $y_l$, we can also derive the following formula:

$$\mathcal{R}\left(y_l|x\right)=\beta f'\left(\frac{\pi\left(y_l|x\right)}{\pi_{ref}\left(y_l|x\right)}\right)+\lambda^{'} \tag{19}$$

According to the Bradley-Terry model, the following formula can be derived:

$$\begin{aligned}&p^*\left(y_w \succeq y_l|x\right)=\sigma\left(\mathcal{R}\left(y_w|x\right)-\mathcal{R}\left(y_l|x\right)\right) \\ =&\sigma\left(\beta f'\left(\frac{\pi\left(y_w|x\right)}{\pi_{ref}\left(y_w|x\right)}\right)-\beta f'\left(\frac{\pi\left(y_l|x\right)}{\pi_{ref}\left(y_l|x\right)}\right)\right)\end{aligned} \tag{20}$$

Therefore, in this instance, the loss function can be formulated as follows, which takes the same formula as (5):

$$\begin{aligned}&\mathcal{L}=-\mathbb{E}_{\mathcal{D}\sim(x,y_w,y_l)}\left[\log p^*\left(y_w \succeq y_l|x\right)\right] \\ &=-\mathbb{E}_{\mathcal{D}}\log\sigma\left(\beta f'\left(\frac{\pi(y_w|x)}{\pi_{ref}(y_w|x)}\right)-\beta f'\left(\frac{\pi(y_l|x)}{\pi_{ref}(y_l|x)}\right)\right)\end{aligned} \tag{21}$$

□

## 4. Practical Algorithm

### 4.1. Identity Unified Preference Optimization (IUPO)

Similarly to IPO(Azar et al., 2023), here we regard $\Psi$ to be an identity mapping, which is a particularly natural form to consider, and then we can derive the optimization objective of Identity Unified Preference Optimization (IUPO) as follows:

$$\max_{\pi}\mathbb{P}_{x\sim\rho}\left(\pi \succeq \pi_{ref}\right)-\beta D_f\left(\pi,\pi_{ref}\right) \tag{22}$$

Motivated by the processing in(Azar et al., 2023), where $h_\pi\left(x,y_w,y_l\right)$ is defined as the human preference function, we define the unified human preference function here. In transitioning from DPO to $f$-DPO and from $\Psi$PO to UPO, the equation's form changes, particularly through the generalization of $\log\left(\pi\left(y\,|x\right)/\pi_{ref}\left(y|x\right)\right)$ to $f'\left(\pi\left(y\,|x\right)/\pi_{ref}\left(y|x\right)\right)$. Therefore, we can obtain unified human preference function by means of induction from original $h_\pi\left(x,y_w,y_l\right)=\log\left(\pi\left(y_w\,|x\right)/\pi\left(y_l\,|x\right)\right)-\log\left(\pi_{ref}\left(y_w\,|x\right)/\pi_{ref}\left(y_l\,|x\right)\right)$ as follows:

$$\mathcal{U}h_\pi\left(x,y_w,y_l\right)=f'\left(\frac{\pi\left(y_w|x\right)}{\pi\left(y_l|x\right)}\right)-f'\left(\frac{\pi_{ref}\left(y_w\,|x\right)}{\pi_{ref}\left(y_l|x\right)}\right) \tag{23}$$

Subsequently, the loss function of IUPO can be expressed as follows:

$$\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}}\left[\mathcal{U}h_\pi\left(x,y_w,y_l\right)-\frac{1}{2\beta}\right]^2 \tag{24}$$

```
**Algorithm 1** Identity Unified Preference Optimization (IUPO)
**Require:** Preference dataset $\mathcal{D}(x, y_w, y_l)$, batch size $b$, constraint coefficient $\beta$, $f$-divergence function $f$, learning rate $lr$, reference policy $\pi_{ref}$.
Initialize the model $\pi_{\theta_0}$ with supervised fine-tuned on dataset $\mathcal{D}(x, y_w, y_l)$
Using $f$-divergence define
    $\mathcal{U}h_\pi(x, y_w, y_l) = f'\left(\frac{\pi(y_w|x)}{\pi(y_l|x)}\right) - f'\left(\frac{\pi_{ref}(y_w|x)}{\pi_{ref}(y_l|x)}\right)$
**for** *n = 1, 2 ......N iterations* **do**
    Sample a batch $\mathcal{B} = \left\{\left(x_i, y_w^i, y_l^i\right)\right\}_{i=1}^{b}$ from dataset $\mathcal{D}$;
    Compute the loss using (24) with the chosen function $f$;
    Compute the gradient loss $\nabla_\theta \mathcal{L}(\theta_{t-1}, \mathcal{B})$;
    Update the model $\theta_t \leftarrow \theta_{t-1} - \nabla_\theta \mathcal{L}(\theta_{t-1}, \mathcal{B})$
**end**
**Return:** Final model $\pi_{\theta_N}$
```

Actually, the simplified form of loss function provides valuable insights into IUPO's policy optimization: the model learns from the preferences dataset by regressing the gap between $f'(\pi(y_w|x)/\pi(y_l|x))$ and $f'(\pi_{ref}(y_w|x)/\pi_{ref}(y_l|x))$ to $1/(2\beta)$. We summarize the full algorithm of IUPO in Algorithm 1.

### 4.2. The relationship between IUPO and IPO

Having derived the IUPO loss function form from the inductive approach, it is evident to understand the relationship between IUPO and IPO: **IUPO is the generalization of IPO**. We illustrate their connection in detail through the following proposition.

**Proposition 4.1.** *Upon setting the $f$-divergence to reverse KL divergence, specifically by defining $f(x) = x \log x$, it can be observed that the $\mathcal{U}h_\pi(x, y_w, y_l)$ in (23) degenerates into the $h_\pi(x, y_w, y_l)$ in (9).*

*Proof.* Choosing the function $f(x)$ as the reverse KL divergence, we can calculate its derivative as follows:

$$f'(x) = (x \log x)' = \log x + x \cdot 1/x = \log x + 1$$

Thus, we can obtain that:

$$\begin{aligned}
&\mathcal{U}h_\pi(x, y_w, y_l) = f'\left(\frac{\pi(y_w|x)}{\pi(y_l|x)}\right) - f'\left(\frac{\pi_{ref}(y_w|x)}{\pi_{ref}(y_l|x)}\right) \\
&= \log\left(\frac{\pi(y_w|x)}{\pi(y_l|x)}\right) + 1 - \log\left(\frac{\pi_{ref}(y_w|x)}{\pi_{ref}(y_l|x)}\right) - 1 \\
&= \log\left(\frac{\pi(y_w|x)}{\pi(y_l|x)}\right) - \log\left(\frac{\pi_{ref}(y_w|x)}{\pi_{ref}(y_l|x)}\right) \\
&= h_\pi(x, y_w, y_l)
\end{aligned} \tag{25}$$

At this point, the loss function of IUPO in (24) is equivalent to the loss function of IPO in (10). □

## 5. Experiments

### 5.1. Experimental Setup

For our experiments, we adopt two datasets, including Anthropic HH dataset(Bai et al., 2022) and Stanford Human Preferences (SHP) dataset(Ethayarajh et al., 2022). Our primary baseline approach is IPO(Azar et al., 2023), and we will compare it with the performance of IUPO under JS-divergence (JS-IUPO). The expression for JS-divergence is given by the function $f(x) = x \log x - (x+1)\log((x+1)/2)$, and its derivative can be expressed as $f'(x) = \log(2x/(1+x))$. It is evident that $0 \notin dom(f'(x))$, and we can derive the loss function for **JS-IUPO** with the following formula:

$$\mathbb{E}_{\mathcal{D}}\left[\log\left(\frac{2 \cdot \frac{\pi(y_w|x)}{\pi(y_l|x)}}{\frac{\pi(y_w|x)}{\pi(y_l|x)} + 1}\right) - \log\left(\frac{2 \cdot \frac{\pi_{ref}(y_w|x)}{\pi_{ref}(y_l|x)}}{\frac{\pi_{ref}(y_w|x)}{\pi_{ref}(y_l|x)} + 1}\right) - \frac{1}{2\beta}\right]^2 \tag{26}$$

In the experiment, we select GPT-2(Radford et al., 2019) as our benchmark model, which has a parameter count of 137M. We initially carry out supervised fine-tuning towards the GPT-2 model using the SHP and HH datasets, with the purpose of alleviating the distribution shift between the true reference distribution which is unavailable and $\pi_{ref}$ utilized by IPO and JS-IUPO. Subsequently, we train the fine-tuned GPT-2 model on SHP and HH datasets with IPO and JS-IUPO. We set the penalty coefficients $\beta$ to be 0.1, 0.2 and 0.5, in order to compare the training outcomes under different penalty coefficients. By comparing the train loss curves, the evaluation loss curves, the evaluation accuracy curves and the evaluation margin reward curves, we can assess the effectiveness of the algorithms.

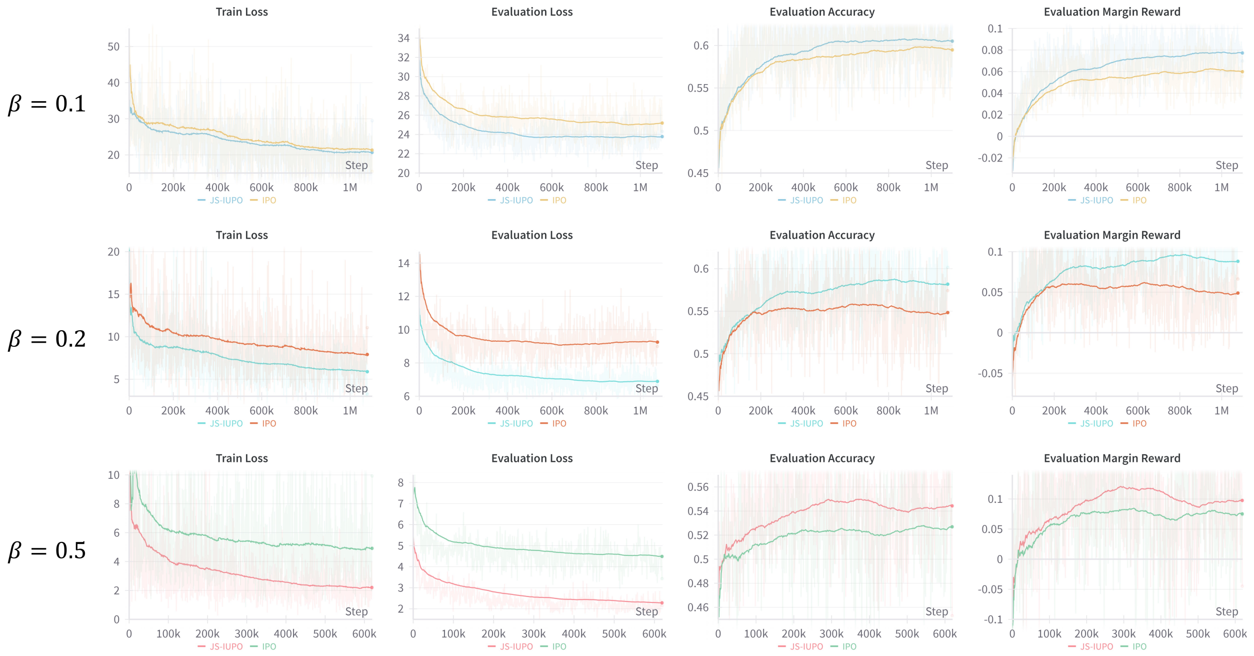

*Figure 1.* Setting the penalty coefficient $\beta$ separately as 0.1, 0.2 and 0.5, the train loss curves, the evaluation loss curves, the evaluation accuracy curves and the evaluation margin reward curves for IPO and JS-IUPO.

### 5.2. Experiments on the HH and SHP datasets

The result curves are depicted in Figure 1. In the analysis of the experimental results, we are dedicated to addressing the following two questions.

**Question 5.1.** *Have the effectiveness of IPO and IUPO been confirmed, and what are their performance on language models?*

Upon analysis of the result curves, it is apparent that the train loss curves and evaluation loss curves display a distinct decreasing trend, and additionally, the evaluation accuracy curves and the evaluation margin reward curves exhibit a significant increasing trend, which means that it is effective for IPO and IUPO to fine-tune language models.

As depicted in the figures, IPO consistently remains above JS-IUPO in curves of train loss and evaluation loss, while in curves of evaluation accuracy and evaluation margin reward, IPO consistently lies below JS-IUPO. It indicates that IUPO, particularly JS-divergence based IUPO, achieves better performance than IPO.

**Question 5.2.** *Are IPO and IUPO sensitive to the penalty coefficient $\beta$? If so, what is a good choice for $\beta$?*

Comparing the curves when penalty coefficient $\beta$ are set as 0.1, 0.2, and 0.5. When setting $\beta$ as 0.1, both the train loss curves and evaluation loss curves exhibit stable decreases, and the evaluation accuracy of JS-IUPO and IPO reach 0.605 and 0.595, respectively, after 1.1M steps. When setting $\beta$ as 0.2, the evaluation accuracy curves and the evaluation margin reward curves show slight fluctuations, with minor performance degradation after approximately 800k steps, and after 1.1M steps, the evaluation accuracy of JS-IUPO and IPO are respectively 0.582 and 0.548. Setting $\beta$ as 0.5 results in significant fluctuations in the evaluation accuracy curves and the evaluation margin reward curves, with model performance degradation occurring after approximately 300k steps, and we therefore implement the early stopping strategy to 610k steps during model training in order to prevent more severe degradation.

The analysis above indicates that setting the penalty coefficient $\beta = 0.1$ is an appropriate choice for both IPO and JS-divergence based IUPO.

## 6. Conclusion, Limitation and Future Work

In this paper, the unified theoretical paradigm on human preference-based optimization is proposed, namely Unified Preference Optimization (UPO), which can be theoretically proved to be the generalization of ΨPO, $f$-DPO and DPO. Furthermore, from a practical perspective, we develop the Identity-UPO (IUPO) algorithm, acting as the generalization of IPO within the $f$-divergence constrains.

The limitations of this paper are mainly as follows: Firstly, the experiments are conducted exclusively on GPT-2, a model with a limited number of parameters, the impact on language models with larger parameter sizes remains unexplored. Secondly, the ablation experiments on model hyperparameter $\beta$ is relatively limited. Finally, experiments on more tasks could be conducted to verify the effectiveness of the proposed method.

Hence, future work will focus on more sophisticated models, more ablation experiments on hyperparameter $\beta$ and more tasks. Furthermore, we would like to further research whether any potential ethical risks associates with this method.

## References

Azar, M. G., Rowland, M., Piot, B., Guo, D., Calandriello, D., Valko, M., and Munos, R. A general theoretical paradigm to understand learning from human preferences, 2023.

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., Johnston, S., Kravec, S., Lovitt, L., Nanda, N., Olsson, C., Amodei, D., Brown, T., Clark, J., McCandlish, S., Olah, C., Mann, B., and Kaplan, J. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.

Bradley, R. A. and Terry, M. E. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444.

Christiano, P., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences, 2023.

Ethayarajh, K., Choi, Y., and Swayamdipta, S. Understanding dataset difficulty with $\mathcal{V}$-usable information, 2022.

Gao, L., Schulman, J., and Hilton, J. Scaling laws for reward model overoptimization, 2022.

Hendrycks, D., Mazeika, M., and Woodside, T. An overview of catastrophic ai risks, 2023.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners, 2019.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model, 2023.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect?, 2023.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., Ho, L., Siddarth, D., Avin, S., Hawkins, W., Kim, B., Gabriel, I., Bolina, V., Clark, J., Bengio, Y., Christiano, P., and Dafoe, A. Model evaluation for extreme risks, 2023.

Wang, C., Jiang, Y., Yang, C., Liu, H., and Chen, Y. Beyond reverse kl: Generalizing direct preference optimization with diverse divergence constraints, 2023.