REVEAL THE MYSTERY OF DPO: THE CONNECTION BETWEEN DPO AND RL ALGORITHMS

Xuerui Su*
School of Mathematics and Statistics
Beijing Jiaotong University
{24110486}@bjtu.edu.cn

Yue Wang*
Independent Researcher
yuewang_yw@foxmail.com

Jinhua Zhu
University of Science and Technology of China
teslazhu@mail.ustc.edu.cn

Mingyang Yi
School of information
Renmin University of China
yimingyang@ruc.edu.cn

Feng Xu
School of Management
Fudan University
fxu23@m.fudan.edu.cn

Zhiming Ma
Academy of Mathematics and Systems Science
mazm@amt.ac.cn

Yuting Liu†
School of Mathematics and Statistics
Beijing Jiaotong University
{ytliu}@bjtu.edu.cn

ABSTRACT

With the rapid development of Large Language Models (LLMs), numerous Reinforcement Learning from Human Feedback (RLHF) algorithms have been introduced to improve model safety, and alignment with human preferences. These algorithms can be divided into two main frameworks based on whether they require an explicit reward (or value) function for training: actor-critic-based Proximal Policy Optimization (PPO) and alignment-based Direct Preference Optimization (DPO). The mismatch between DPO and PPO, such as DPO's use of a classification loss driven by human-preferred data, has raised confusion about whether DPO should be classified as a Reinforcement Learning (RL) algorithm. To address these ambiguities, we focus on three key aspects related to DPO, RL, and other RLHF algorithms: (1) the construction of the loss function; (2) the target distribution at which the algorithm converges; (3) the impact of key components within the loss function. Specifically, we first establish a unified framework named UDRRA connecting these algorithms based on the construction of their loss functions. Next, we uncover their target policy distributions within this framework. Finally, we investigate the critical components of DPO to understand their impact on the convergence rate. Our work provides a deeper understanding of the relationship between DPO, RL, and other RLHF algorithms, offering new insights for improving existing algorithms.

1 INTRODUCTION

Large Language Models (LLMs) are considered one of the most promising advancements toward Artificial General Intelligence (AGI) Zhao et al. (2023a); Ouyang et al. (2022); Minaee et al. (2024); Achiam et al. (2023). In practice, post-training the LLM with Reinforcement Learning from Human

* These authors contributed equally to this work.
† Corresponding author.

Feedback (RLHF) has significantly improved its safety, compliance, and the alignment with human preferences Arumugam et al. (2019); Singh et al. (2022); Bai et al. (2022); Dai et al. (2023).

While well-established RLHF approaches, such as the Proximal Policy Optimization (PPO-RLHF) Ziegler et al. (2019), have demonstrated significant success, Direct Preference Optimization (DPO) Rafailov et al. (2024) has also emerged and garnered widespread attention. DPO differs from PPO in that it does not require explicit modeling of the reward function, making it simpler and more efficient in practice. Instead, DPO directly optimizes the language model to align with human preferences by minimizing a simple classification loss function, which is based solely on the log probability of outputs generated by the LLMs. Since its inception, DPO has gained considerable traction, leading to the development of several derivatives, such as IPO Azar et al. (2024) and DRO Richemond et al. (2024), among others.

Despite its practical success, DPO differs significantly from typical Reinforcement Learning (RL) algorithms, particularly in its underlying algorithmic structure. While most RL algorithms, such as PPO-RL Schulman et al. (2017a), rely on an actor-critic framework and explicitly model value functions, DPO takes a distinct approach. Under a Bradly-Terry Bradley & Terry (1952) framework to modeling human preference, DPO transferred training policy model into minimizing a human-preference labeled loss function, without requiring an explicit reward model or training value function with policy gradient. This fundamental difference leads to limited theoretical understanding of how DPO connects with standard RL algorithms, particularly in terms of its theoretical advantages. As a result, more research is needed to clarify the specific scenarios where DPO might offer distinct benefits over typical RL algorithms.

However, addressing the relationship between DPO and RL algorithms is a non-trivial task. On one hand, DPO is rooted in the optimal solution analysis of the KL-constrained reward maximization problem, based on PPO-RLHF. On the other hand, the absence of a reward function in DPO complicates its classification within the typical RL framework. As a result, recent research has even suggested that DPO may not be a true RL algorithm Panchenko (2024), further highlighting the need for a deeper theoretical investigation. In this paper, we aim to bridge this gap by investigating the connections between DPO, RL, and other RLHF algorithms, and focus on three key aspects:

- 1. What are the distinctions and connections of the construction of the loss function between DPO, RL and other representative RLHF algorithms?
- 2. What is the target distribution of these loss function?
- 3. How do the key components within these algorithms affect the algorithm performance?

To do so, we fist construct a unified framework (Figure 1) that uniformly covers the standard RL algorithms (i.e., PPO and Soft Actor Critic (SAC) Haarnoja et al. (2018)) and the DPO-based algorithms i.e., (IPO, DRO, DPO). Based on the construction of loss functions of these algorithms, we clearly reveal the connections between them in our framework. Besides, the target distribution under this framework is also clearly revealed. Notably, though some existing literature Xu et al. (2024); Ivison et al. (2024); Yan et al. (2024) have also explored the connection between DPO and PPO. However, they mainly focus on the technical details of training LLM and the comparison of experimental results between DPO and PPO. In contrast, our framework is more general, and is not restricted to any specific model scenario.

Then we analyze the target distributions of the series of methods mentioned in our framework. Although these methods have different requirements for the reward function, the same versions (e.g. the posterior version) all share the same target distribution. Furthermore, by analyzing the relationship between DPO and the PRA-P method in our framework, we proved that the target distribution of DPO is not $\bar{\pi}^{\tau}$ introduced in the original DPO's paper. Finally, considering the importance of DPO in training RLHF models, we further investigate the convergence rate of the DPO algorithm. Building on our theoretical results, we analyze how the hyper-parameter τ and the offline dataset of the algorithm influence its convergence performance.

To the best of our knowledge, we are the first to establish such a unified framework that provides a cohesive perspective on DPO Algorithm and a variety of RLHF algorithms. Our explorations provide us deeper understanding and new insights in improving the existing algorithms for RLHF. The contributions in this paper can be summarized as follows:

- We established a framework to bridge the theoretical gap between DPO and RL algorithms.
- We analyzed the target distribution of the loss function for algorithm in our Framework.
- We explored the impact of the hyper parameter τ and the preference dataset on algorithm performance.

2 PRELIMINARY

2.1 PROBLEM SETUP

Consider a set of state or prompt $\mathbb{X}$. Denote the response space as $\mathbb{Y}$, and the reward function as $r : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$. The optimization problem that most RLHF algorithm aim to solve is defined below where $\mathcal{D}$ is arbitrary distribution and define the target distribution $\pi^\delta = \arg\min_{\pi_\theta} J$.

$$\min_{\pi_\theta} J \triangleq \min_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(y|x)} \left[-r(x,y)\right], \tag{1}$$

Definition 2.1. Boltzmann Distribution Approximation Problem (BDAP). Motivated by the idea of Soft Actor Critic algorithm Haarnoja et al. (2018) that they update the policy towards the exponential of the new Q-function, another important problem we consider is the Boltzmann Distribution Approximation Problem:

$$\min_{\pi' \in \Pi} \mathbb{E}_{x\sim\mathcal{D}} \left[\mathrm{D_{KL}}\left(\pi'(\cdot \mid x) \| \pi^\tau(\cdot \mid x)\right)\right], \quad \pi^\tau(\cdot \mid x) \triangleq \frac{\exp\left(\tau r(x,\cdot)\right)}{Z(x)}. \tag{2}$$

Problem (2) describes the objective of optimizing π' towards the Boltzmann distribution of reward function by the KL divergence, though in principle any distribution distance is suitable. The partition function $Z(x) = \sum_{y\in\mathbb{Y}} \exp\left(\tau r(x,y)\right)$ normalizes the distribution. τ is the temperature parameter. Π is the set of policies. π^τ is essentially a soft approximation of the optimal solution π^δ of Problem (1), while π^δ is sharp. This approximation is referred to as the Boltzmann approximation in physics. Proposition 2.2 shows that π^τ converges to π^δ.

Proposition 2.2. *Without loss of generality, assume that the reward function $r(x,y)$ has a unique maximum for any given x. Then we have:* $\lim_{\tau\to\infty} \pi^\tau(y|x) = \pi^\delta(y|x)$.

2.2 RELATED WORKS

PPO-RL: The Proximal Policy Optimization (PPO) algorithm Schulman et al. (2017a) in RL is based on the Trust Region Policy Optimization (TRPO) algorithm Schulman (2015). The KL penalty version of PPO in RL is below:

$$\mathcal{L}_{\text{PPO}}(\pi_\theta) = \hat{\mathbb{E}}_t[-\frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}\hat{A}_t + \frac{1}{\tau}\mathrm{D_{KL}}\left(\pi_{\theta_{\text{old}}}(\cdot|s_t)\|\pi_\theta(\cdot|s_t)\right)]. \tag{3}$$

where θ_{old} is the vector of policy parameters before the update. $\hat{A}_t$ is an estimator of the advantage function at timestep t. For one step policy optimization problem, $\hat{A}_t = r(s_0, a_0)$ where s_0, a_0 are taken as the prompt and response separately. Based on the theory of TRPO, PPO-RL will converge to the optimal solution π^δ of Problem (1).

PPO-RLHF: The PPO based RLHF Ziegler et al. (2019) typically consists of two major stages: reward modeling and RL optimization. In the first stage, human annotators select the preferred answer y_w over the less preferred one y_l, forming the preference pair (y_w, y_l, x) based on a given input x. A reward model $r_\phi(x,y)$ is optimized by:

$$\mathcal{L}_R(r_\phi) = -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_R}\left[\log\sigma\left(r_\phi(x,y_w) - r_\phi(x,y_l)\right)\right]. \tag{4}$$

where $\sigma(\cdot)$ is the sigmoid function, $\mathcal{D}_R \triangleq \{(x,y_w,y_l)|x \sim \mathcal{D}, y_w, y_l \sim \pi_0(\cdot|x), (y_w \succ y_l) \sim p^*(1|y_w,y_l,x)\}$, π_0 is the offline data sampling distribution. $\{z = 1|y_1,y_2,x\} \triangleq \{r(x,y_1) \geq r(x,y_2)\}$ and p^* is modeled by the BT model. In the RL optimization stage, the trained reward model evaluates the outputs of the language model, guiding policy optimization with loss function of the classic PPO algorithm (Typically the reference policy $\pi_{ref}(\cdot|x)$ is a pretrained LLM):

$$\mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[-r_\phi(x,y) + \frac{1}{\tau}\mathrm{D_{KL}}\left(\pi_\theta(\cdot|x)\|\pi_{ref}(\cdot|x)\right)\right]. \tag{5}$$

DPO: The Direct Preference Optimization (DPO) Rafailov et al. (2024) leverages the optimal policy form in Eq.5 as theoretical support by representing the comparison probability (based on BT model) of human preferences through the ratio between the policy π_θ and the reference policy π_{ref}. This approach eliminates the need for explicitly modeling the reward function. Thus DPO directly optimizes the policy by maximizing the log-likelihood function based on human preference feedback:

$$\begin{aligned} \bar{h}_\theta\left(x, y_w, y_l\right) &= \frac{1}{\tau} \log \frac{\pi_\theta\left(y_w \mid x\right)}{\pi_{\text{ref}}\left(y_w \mid x\right)} - \frac{1}{\tau} \log \frac{\pi_\theta\left(y_l \mid x\right)}{\pi_{\text{ref}}\left(y_l \mid x\right)}, \\ \mathcal{L}_{\text{DPO}}\left(\pi_\theta ; \pi_{\text{ref}}\right) &= -\mathbb{E}_{\left(x, y_w, y_l\right) \sim \mathcal{D}_R}\left[\log \sigma\left(\bar{h}_\theta\left(x, y_w, y_l\right)\right)\right]. \end{aligned} \tag{6}$$

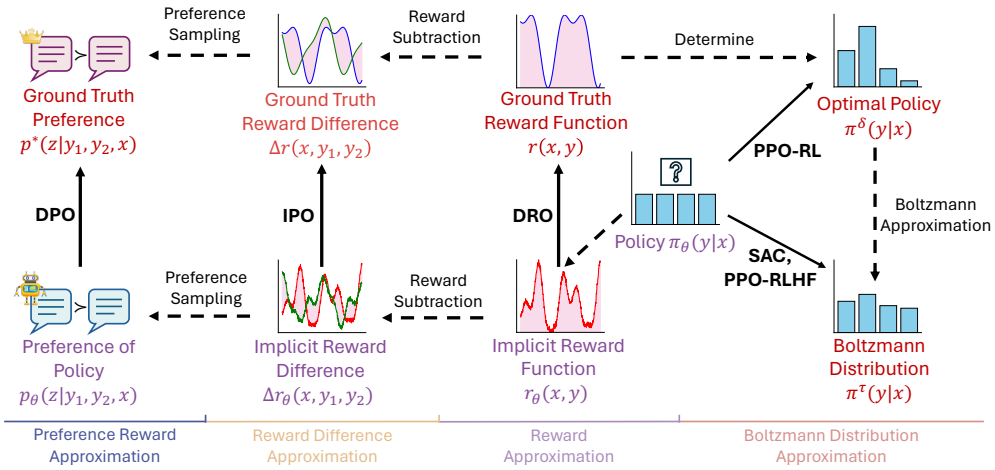

Figure 1: The unified UDRRA framework to connect DPO and other RLHF algorithms with typical RL methods. The algorithms in our framework can construct loss functions for four different scenarios (from right to left): (1) between $\pi_\theta(y|x)$ and $\pi^\delta(y|x)$ (or $\pi^\tau(y|x)$); (2) between $r_\theta(x,y)$ and $r(x,y)$; (3) between $\Delta r_\theta(x,y_1,y_2)$ and $\Delta r(x,y_1,y_2)$; and (4) between $p_\theta(z|y_1,y_2,x)$ and $p^*(z|y_1,y_2,x)$. From scenario (1) to scenario (4), the requirements of the reward function $r(x,y)$ are progressively relaxed. The notations in the figure will be elaborated in Section 2 and Section 3. Notably, we utilize Proposition 2.2 to illustrate the relationship between π^δ and π^τ.

3 The Framework to Unify DPO, RL and Other Representative RLHF Algorithms

In this section, we unify DPO, RL, and other representative RLHF algorithms within a unified framework (see Figure 1) based on the construction of loss functions. The distinctions of loss functions are mainly because how loss functions are defined across various reward function scenarios. Our framework clarifies these relationships, highlighting each method's strengths and use cases while bridging RL principles with RLHF techniques to improve methods like DPO.

3.1 The UDRRA Framework

Most RL algorithms aim to solve Problem (1): given a reward function, find the optimal policy π^δ. To balance exploration and exploitation, the target distribution π^δ is often approximated by a Boltzmann distribution π^τ, turning the optimization into a distribution approximation problem. The critical challenge lies in designing a loss function that enables the parameterized policy π_θ to effectively approximate the target distribution π^τ through gradient-based methods. Our framework emphasizes the differences in loss function construction across various algorithms. Specifically, the construction of loss functions in current algorithms primarily addresses four types of scenarios involving policy distributions, reward functions, reward differences, and preferences, corresponding to scenarios (1)-(4) in Figure 1.

From scenario (1) to scenario (4), the requirements on the reward function $r(x,y)$ are progressively relaxed. Our paper will illustrate the characteristics of loss function construction in different scenarios

and prove the equivalence of different loss designs in the sense of target distribution. Furthermore, we will analyze the position of DPO within these scenarios and its relationship with our proposed method PRA-P for offering a new perspective for the theoretical investigation of DPO.

3.1.1 SCENARIO (1): BOLTZMANN DISTRIBUTION APPROXIMATION

Proposition 2.2 supports the Boltzmann approximation, allowing π^τ to relax π^δ. This relaxation is validated by RL algorithms like SAC. Therefore, in scenario (1) there is a natural relaxation method to approximate the solution to Problem (1) by solving π^τ. Here, we summarize the method based on Definition 2.1 using the KL divergence as the loss function. For the Forward-KL and Reverse-KL, we have Eq.7 and Eq.8 (See derivation in Appendix E.1, E.2):

$$\begin{aligned}\mathcal{L}_{\text{Forward}-\text{BDA}}(\pi_\theta) &= \mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{D}_{\mathrm{KL}}\left(\pi_\theta\left(\cdot \mid x\right) \| \pi^\tau\left(\cdot \mid x\right)\right)\right] \\ &= -\mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[\left(\tau r(x,y) - \log(\pi_\theta(y|x))\right)\right], \end{aligned} \tag{7}$$

$$\begin{aligned}\mathcal{L}_{\text{Reverse}-\text{BDA}}(\pi_\theta) &= \mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{D}_{\mathrm{KL}}\left(\pi^\tau\left(\cdot \mid x\right) \| \pi_\theta\left(\cdot \mid x\right)\right)\right] \\ &= -\mathbb{E}_{x\sim\mathcal{D}, y\sim\pi^\tau(\cdot|x)}[\log \pi_\theta(y|x)]. \end{aligned} \tag{8}$$

We name the method using Eq.7 or Eq.8 as the loss function as the **Boltzmann Distribution Approximation (BDA)** method. Eq.7 is exactly the loss function of SAC.

3.1.2 SCENARIO (2): REWARD APPROXIMATION

Denote the implicit reward function $r_\theta(x, y)$:

$$r_\theta(x,y) \triangleq \frac{1}{\tau}\log(Z(x)\pi_\theta(y|x)). \tag{9}$$

Scenario (2) focuses on the intuition of using $r_\theta(x, y)$ to approximate the ground truth $r(x, y)$. Considering using $r_\theta(x, y)$ to approximate $r(x, y)$ is a typical regression problem, the Mean Square Error (MSE) is natural to be the loss function (See derivation in Appendix E.4):

$$\mathcal{L}_{\mathrm{RA}}(\pi_\theta) = \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[(r_\theta(x,y) - r(x,y))^2\right] = \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[\frac{1}{\tau^2}\left(\log(\frac{\pi_\theta(y|x)}{\pi^\tau(y|x)})\right)^2\right]. \tag{10}$$

We name the method using Eq.10 as the loss function as the **Reward Approximation (RA)** method. Due to the special design of $r_\theta(x, y)$, although the forms are different, Eq.7, Eq.8 and Eq.10 all effectively share the same target distribution. We will formally prove it later in Theorem 4.1.

In RLHF, the alignment task builds upon a reference model $\pi_{ref}(\cdot|x)$ rather than an initialized language model policy. Thus we incorporate $\pi_{ref}(y|x)$ via the posterior implicit reward function $\bar{r}_\theta(x, y)$. Treating π_{ref} as the prior, the posterior distribution becomes $\bar{\pi}^\tau$:

$$\bar{\pi}^\tau(y|x) \triangleq \frac{\pi_{ref}(y|x)\exp(\tau r(x,y))}{\sum_{y'\in Y}\pi_{ref}(y'|x)\exp(\tau r(x,y'))}. \tag{11}$$

The corresponding posterior implicit reward function is $\bar{r}_\theta(x, y) = \frac{1}{\tau}\log(Z'(x)\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)})$ where $Z'(x) = \sum_{y'\in Y}\pi_{ref}(y'|x)\exp(\tau r(x, y'))$. Then Eq.10 will be changed into:

$$\begin{aligned}\mathcal{L}_{\mathrm{RA}-\mathrm{P}}(\pi_\theta) &= \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[(\bar{r}_\theta(x,y) - r(x,y))^2\right] \\ &= \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[\left(\frac{1}{\tau}\log(Z'(x)\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}) - r(x,y)\right)^2\right]. \end{aligned} \tag{12}$$

We call the method using Eq.12 as loss function as the **Reward Approximation-Posterior (RA-P)** method.

3.1.3 SCENARIO (3): REWARD DIFFERENCE APPROXIMATION

Define the general difference function $\Delta f(x, y_1, y_2) \triangleq f(x, y_1) - f(x, y_2)$. In scenario (3), the loss function is constructed to align the reward difference function, $\Delta r_\theta(x, y_1, y_2) = r_\theta(x, y_1) - r_\theta(x, y_2)$ and $\Delta r(x, y_1, y_2) = r(x, y_1) - r(x, y_2)$. Denote $\mathcal{D}_{\text{pw}} \triangleq \{(x, y_1, y_2) \mid x \in \mathcal{D}, y_1, y_2 \sim \pi_\theta(y|x)\}$. Similar to the RA method, using MSE, we have the loss function:

$$\mathcal{L}_{\mathrm{RDA}}(\pi_\theta) = \mathbb{E}_{\mathcal{D}_{\text{pw}}}\left[\left(\frac{1}{\tau}\log\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)} - \Delta r(x, y_1, y_2)\right)^2\right]. \tag{13}$$

The method using Eq.13 as the loss function is named **Reward Difference Approximation (RDA)**. Its posterior version, RDA-P, replaces r_θ with $\bar{r}_\theta$. All our methods have posterior versions, this paper only formulate RA-P and PRA-P as example. The RDA method simplifies computation by canceling the partition function $Z(x)$, which requires multiple reward function queries. Additionally, RDA only needs reward differences, not absolute values, while maintaining the same target distribution as BDA and RA. This is formally proven in Theorem 4.2.

3.1.4 Scenario (4): Preference Reward Approximation

In scenario (4), we consider using $p^*(z \mid y_1, y_2, x) = \omega(r(x, y_{2-z}), r(x, y_{z+1}))$, where $z = 1$ if $r(x, y_1) > r(x, y_2)$, and $z = 0$ otherwise, to model the comparison probability between $r(x, y_1)$ and $r(x, y_2)$. Only (x, y_1, y_2, z) sampled with $p^*(z \mid y_1, y_2, x)$ is available, while $r(x, y)$ is unobservable. The BDA, RA, and RDA methods are less effective in scenario (4). These methods rely on inverting ω to estimate $r(x, y)$. Table 3 and Appendix C discuss various ω formulations. In general, most ω^{-1} are difficult to compute because p^* is not analytic. Thus developing methods that rely only on the forward evaluation of ω instead of ω^{-1} would greatly improve practicality in this scenario.

Therefore we propose the **Preference Reward Approximation (PRA)** method, assuming ω is known and data is sampled as (x, y_w, y_l) where $\{r(x, y_w) > r(x, y_l)\}$ is true with probability $p^*(1 \mid y_w, y_l, x)$. Define the parameterized comparison distribution as $p_\theta(z|y_1, y_2, x) = \omega(r_\theta(x, y_{2-z}), r_\theta(x, y_{z+1}))$. Assuming $\omega(x, y)$ is injective and satisfies the symmetric complementarity property $\omega(x, y) = 1 - \omega(y, x)$, the winning probability of x against y equals y's losing probability against x, then we have:

$$\begin{aligned}\mathcal{L}_{\text{PRA}}(\pi_\theta) =& \mathbb{E}_{\mathcal{D}_\theta}\left[\mathrm{D}_{\mathrm{KL}}(p^*(z|y_1, y_2, x)||p_\theta(z|y_1, y_2, x))\right] \\ =& -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_\theta}\left[\log\left(p_\theta(1|y_w, y_l, x)\right)\right] + \mathbb{E}_{x\sim\mathcal{D}, y_1, y_2\sim\pi_\theta(y|x)}\left[M(x, y_1, y_2)\right].\end{aligned} \tag{14}$$

Defer the proof in Appendix E.3. $\mathcal{D}_\theta \triangleq \{(x, y_w, y_l)|x \sim \mathcal{D}, y_w, y_l \sim \pi_\theta(\cdot|x), (y_w \succ y_l) \sim p^*(1|y_w, y_l, x)\}$ and $M(x, y_1, y_2) = \sum_{z=0,1} p^*(z|y_1, y_2, x) \log p^*(z|y_1, y_2, x)$. PRA uses the KL divergence between p^* and p_θ as the loss function, combining a cross-entropy term that avoids computing gradients of the reward function or p^*, and a regularization term involving integration over p^* and π_θ. Similar to the RA-P method, we also propose a posterior version of PRA. Define $\bar{p}_\theta(z|y_1, y_2, x) = \omega(\bar{r}_\theta(x, y_{2-z}), \bar{r}_\theta(x, y_{z+1}))$. Based on Eq.14, we have:

$$\begin{aligned}\mathcal{L}_{\text{PRA-P}}(\pi_\theta) =& \mathbb{E}_{\mathcal{D}_\theta}\left[\mathrm{D}_{\mathrm{KL}}(p^*(z|y_1, y_2, x)||\bar{p}_\theta(z|y_1, y_2, x))\right] \\ =& -\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_\theta}\left[\log \bar{p}_\theta\left(1|y_w, y_l, x\right)\right] + \mathbb{E}_{x\sim\mathcal{D}, y_1, y_2\sim\pi_\theta(y|x)}\left[M(x, y_1, y_2)\right].\end{aligned} \tag{15}$$

The method using Eq.15 as the loss function is name **Preference Reward Approximation-Posterior (PRA-P)** method. Firstly, PRA and PRA-P avoid calculating ω^{-1} while maintaining the same target distribution as previous methods. Theorem 4.3 shows that PRA shares its target distribution with methods like BDA, while Theorem 4.4 demonstrates that PRA-P and RA-P share the same target distribution. Secondly, the DPO algorithm aligns with scenario (4) and is closely related to PRA-P. In fact, the first term of Eq.15 matches the DPO loss Rafailov et al. (2024) if $\mathcal{D}_\theta$ is replaced with an offline dataset, though this introduces a distribution shift issue, which is discussed in Theorem 4.5.

3.2 Summary

Our framework is both logically reasonable and empirically valid. After developing the solution method for Problem 2.1 based on different reward function scenarios, we found that previous studies have explored specific methodologies for the some scenarios and proposed efficient algorithms. For example, as shown in Section 3.1.1, the loss function of SAC (Eq.10 in Haarnoja et al. (2018)) aligns with the Forward-BDA method. To highlight these relationships, we summarize the correspondence between these algorithms and our methods in Table 2.

For scenario (1), the Reverse-BDA method, using reverse-KL (Eq.8) as a loss function for policy improvement, has been applied in finite state-action space control problems (see Eq.2 in RERPI Abdolmaleki et al. (2018)). However, in continuous state-action spaces, the term $Z(x)$ is hard to compute, limiting the method's applicability. For scenario (2), the loss function of the Direct Reward Optimization (DRO) method matches the RA-P method's loss function (Eq.12), as shown the Eq.4 in Richemond et al. (2024). For scenario (3), the Step-level Value Preference Optimization (SVPO) algorithm Chen et al. (2024) uses $\Delta r_\pi\left(\mathbf{s}_{t+1}^w, \mathbf{s}_{t+1}^l\right)$ to learn towards $\mathrm{sg}\left[\Delta r_\phi\left(\mathbf{s}_{t+1}^w, \mathbf{s}_{t+1}^l\right)\right]$

(Eq.10 in Chen et al. (2024)), which corresponds to our RDA-P method. The RDA-P method is also a generalized version of Identity-PO (IPO) Azar et al. (2024), relaxing the data requirement from querying reward differences to only ranking queries (i.e., $\mathbb{I}(r(x, y_1) - r(x, y_2))$, where $\mathbb{I}$ is the indicator function). Though $\mathbb{I}(r(x, y_1) - r(x, y_2))$ is more like a preference, we classify IPO as a relaxed version of RDA due to its similar form as RDA-P. Finally, for scenario (4), we will show in Theorem 4.5 that DPO is simply an offline version of PRA-P.

4 Target Distribution Analysis

In this section, we analyze the target distribution of the loss function mentioned in our framework. Although these methods have different requirements for the reward function, the same versions (e.g. the posterior version) all share the same target distribution. Additionally, as noted in Section 3.1, we highlight how using an offline dataset in DPO introduces distribution shift issues, distinguishing it from online methods like SAC and PPO-RLHF.

4.1 The Target Distribution Equivalence of Our Framework

Here we state the approximation equivalence among the BDA, RA, RDA and PRA methods.

Theorem 4.1. *Define* $\pi^*_{\text{Forward-BDA}} = \arg\min_{\pi_\theta} \mathcal{L}_{\text{Forward-BDA}}(\pi_\theta)$, $\pi^*_{\text{Reverse-BDA}} = \arg\min_{\pi_\theta} \mathcal{L}_{\text{Reverse-BDA}}(\pi_\theta)$, $\pi^*_{\text{RA}} = \arg\min_{\pi_\theta} \mathcal{L}_{\text{RA}}(\pi_\theta)$. *The following property holds:*

$$\pi^*_{\text{Forward-BDA}} = \pi^*_{\text{Reverse-BDA}} = \pi^*_{\text{RA}} = \pi^\tau. \tag{16}$$

Theorem 4.2. *Recall* $\mathcal{L}_{\text{RDA}}(\pi_\theta)$ *in Eq.13. Define* $\pi^*_{\text{RDA}} = \arg\min_{\pi_\theta} \mathcal{L}_{\text{RDA}}(\pi_\theta)$. *Then* $\pi^*_{\text{RDA}} = \pi^\tau$.

Theorem 4.3. *Recall* $\mathcal{L}_{\text{PRA}}(\pi_\theta)$ *in Eq.14. Define* $\pi^*_{\text{PRA}} = \arg\min_{\pi_\theta} \mathcal{L}_{\text{PRA}}(\pi_\theta)$. *Then* $\pi^*_{\text{PRA}} = \pi^\tau$.

Theorem 4.1 demonstrate that the "arg min" of Eq.7, 8, 10 are equivalent and equal to π^τ. Theorem 4.2 proves that the target distribution corresponding to the loss function of the RDA method is π^τ, which is consistent with the target distribution of the BDA and RA methods. Theorem 4.3 shows that although the loss function of the PRA method is designed by minimizing the distance between the two distributions p^*, p_θ, its target distribution remains π^τ. Similarly, we use Theorem 4.4 to demonstrate the relationship between the target distributions of the posterior versions of the RA and PRA methods.

Theorem 4.4. *Recall* $\mathcal{L}_{\text{RA-P}}(\pi_\theta)$ *in Eq.12 and* $\mathcal{L}_{\text{PRA-P}}(\pi_\theta)$ *in Eq.15. Define*

$$\pi^*_{\text{RA-P}} = \arg\min_{\pi_\theta} \mathcal{L}_{\text{RA-P}}(\pi_\theta),\ \pi^*_{\text{PRA-P}} = \arg\min_{\pi_\theta} \mathcal{L}_{\text{PRA-P}}(\pi_\theta). \tag{17}$$

The following property holds:

$$\pi^*_{\text{RA-P}} = \pi^*_{\text{PRA-P}} = \bar{\pi}^\tau. \tag{18}$$

In summary, while the BDA, RA, RDA, and PRA methods are designed for different scenarios, they all share the same target distribution and aim to solve the same problem (BDAP) under varying conditions of reward function access. Specifically, BDA and RA require exact reward values, RDA only needs reward differences across variables y, and PRA relies on ordinal relationships of the reward function. Based on our analyzing about the methods in our UDRRA framework, we can more accurately select the appropriate method for different scenarios.

4.2 The Relationship between DPO and PRA-P (The Target Distribution of DPO)

This section analyzes the target distribution of DPO. DPO mentions that the DPO target distribution is $\bar{\pi}^\tau$ which is also the the target distribution of the PRA-P method (Eq.4 in Rafailov et al. (2024)). However, we will show that there is a distribution shift between the the target distribution $\bar{\pi}^\tau$ of the PRA-P method and the the target distribution of DPO. We use Theorem 4.5 to strictly analyze the differences and connections between DPO and PRA-P.

Theorem 4.5. *When* p^* *is modeled by the BT model:*

$$p^*(1|y_1, y_2, x) = \omega(r(x, y_{2-z}), r(x, y_{z+1})) = \sigma(r(x, y_1) - r(x, y_2)), \tag{19}$$

then we have the following equality:

$$\mathcal{L}_{\text{PRA-P}}(\pi_\theta) = \mathcal{L}_{\text{DPO}}(\pi_\theta) + \eta_1(\pi_\theta, \pi_0) + \eta_2(\pi_\theta). \tag{20}$$

where $\eta_1(\pi_\theta, \pi_0)$ *equals to 0 if and only if* $\pi_\theta = \pi_0$ *and* $\eta_2(\pi_\theta) \triangleq \mathbb{E}_{x\sim\mathcal{D}, y_1, y_2\sim\pi_\theta(y|x)}[M(x, y_1, y_2)]$.

Theorem 4.5 tells us that since $\mathcal{D}_R$ in DPO is an offline dataset, a shift term $\eta_1(\pi_\theta, \pi_0)$ is introduced between DPO and PRA-P. Because of the difference between π_θ and π_0, the existence of the non-zero term $\eta_1(\pi_\theta, \pi_0)$ makes the target distribution of DPO deviate from the target distribution $\bar{\pi}^\tau$ of the PRA-P method. We call this phenomenon as distribution shift. Defer the proofs to Appendix F.5. Moreover, $\eta_2(\pi_\theta)$ is a regularization term which serves to increase the information entropy $(-M(x, y_1, y_2))$ of the comparison probability p^* under the current policy distribution π_θ. The ignorance of this regularization term will make the DPO deviate further from the target distribution $\bar{\pi}^\tau$. Lastly, we emphasize that the ω function in the PRA method only needs to satisfy the symmetric complementarity property and is not limited to the BT model, which is a special case. We list more ω functions meeting this requirement in Table 3.

5 Components Influence in Algorithms within UDRRA

After analyzing the connections and distinctions between algorithms like DPO, the remaining question is how the common components in algorithms within UDRRA affect their performance. Here we focus on the impact of hyperparameters τ and preference datasets on algorithm performance.

5.1 Q1: The Influence of τ on Algorithm Performance

For Q1, Increasing hyper-parameter τ can speed up the convergence of DPO, which will formulate in Theorem 5.1. As τ increases, the optimal solution $\bar{\pi}^\tau$ of DPO tends to π^δ (See Proposition 2.2), while deviating from π_{ref}. Thus there exists a trade-off between the convergence rate and maintaining proximity to π_{ref}.

Theorem 5.1. *Assume π_θ is constructed by Definition D.1. Consider the DPO loss function $\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{ref})$. Given the learning rate α_t satisfying finite squared summability, suppose the parameters θ are updated by:*

$$\theta_{t+1} = \theta_t - \alpha_t g(x, y_w, y_l, \theta_t), \tag{21}$$

where $g(x, y_w, y_l, \theta_t)$ is a stochastic gradient of $\mathcal{L}_{\text{DPO}}(\pi_{\theta_t}; \pi_{ref})$. Assume $\|g(\cdot,\cdot,\cdot,\cdot)\|^2 \leq G^2$. Denote $\mathcal{L}^*_{\text{DPO}} = \min_{\pi_\theta} \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{ref})$, then:

$$\min_{1\leq i\leq T} ||\nabla_\theta \mathcal{L}_{\text{DPO}}(\pi_{\theta_i}; \pi_{ref})||_2^2 \leq \frac{2G^2 \sum_{t=1}^{T-1} \alpha_t^2}{\tau^2 \sum_{t=1}^{T-1} \alpha_t} + \frac{\mathcal{L}_{\text{DPO}}(\pi_{\theta_1}) - \mathcal{L}^*_{\text{DPO}}}{\sum_{t=1}^{T-1} \alpha_t}. \tag{22}$$

We assume the policy is a softmax policy (Definition D.1) and show that the gradient norm of the DPO loss function has an upper bound under non-convexity, which is negatively correlated with the hyper-parameter τ. Thus we have that τ is inversely related to the convergence rate of DPO.

This paper focuses on the impact of τ, rather than the design of the LLMs. Thus, we simplify Theorem 5.1 to apply the convergence theorem for SGD in non-convex settings. Our argument is that, as long as the policy distribution is a softmax policy, we show that the loss functions for all four methods in our framework, including DPO, are L-smooth with respect to the policy parameters θ leading to a sub-linear convergence rate of $O(\frac{1}{T})$.

Thus we summarize the convergence properties of DPO and related methods (BDA, RA, RDA, PRA) in Table 1, corresponding to Lemma G.8-G.12. The variations in L arise from differences in how their loss functions are constructed. First, the Forward-BDA method has a larger smooth coefficient $(6\epsilon_1+10)$ than the Reverse-BDA method, indicating slower convergence but a more sampling-friendly policy π_θ. Second, the RA method outperforms both BDA methods by setting ϵ_1 and τ to reduce the smooth coefficient. For instance, when ϵ_1 and τ satisfy $0 \leq \epsilon_1 \leq \frac{\sqrt{1+\tau}-1}{\tau} \leq \frac{\sqrt{6}-1}{5}$, the smooth coefficient of the RA method is bounded by 2, ensuring faster convergence. Third, the smooth coefficients of the RDA and PRA methods can be reduced by adjusting $\epsilon_2, \epsilon_3, \tau$, and the parameter domain diameter d, to speed up their convergence. Lastly, the DPO method has a higher smooth coefficient than PRA, with additional terms like $20\log(1+e^{\frac{d}{\tau}}) + \frac{16\epsilon_3}{\tau} + 16\log 2$, indicating that DPO converges faster than PRA-P due to the distribution shift bias.

Method	#Loss Functions	#Smooth Coefficient (L)	#$p(x, y; \pi_\theta)$
Forward-BDA	Eq.7	$6\epsilon_1 + 10$	$\pi_\theta(y\|x), x \sim \mathcal{D}$
Reverse-BDA	Eq.8	2	$\pi^\tau(y\|x), x \sim \mathcal{D}$
RA	Eq.10 and Eq.12	$3\epsilon_1^2 + \frac{18\epsilon_1}{\tau} + \frac{8}{\tau^2} + \max\left\{\epsilon_1^2 + \frac{2}{\tau}\epsilon_1, \frac{1}{\tau}\right\}$	$\pi_\theta(y\|x), x \sim \mathcal{D}$
RDA	Eq.13	$20\epsilon_2^2 + \frac{32\epsilon_2}{\tau} + \frac{8}{\tau^2}$	$\pi_\theta(y\|x), x \sim \mathcal{D}$
PRA	Eq.14 and Eq.15	$20\log(1 + e^{\frac{d}{\tau}}) + \frac{16\epsilon_3}{\tau} + \frac{4}{\tau^2} + 16\log 2$	$\pi_\theta(y\|x), x \sim \mathcal{D}$
DPO	Eq.6	$\frac{4}{\tau^2}$	$\pi_0(y\|x), x \sim \mathcal{D}$

Table 1: Smooth Coefficients for different methods. π_θ is the current policy to be optimized, and π_0 is the sampling distribution for DPO's offline dataset. Let $|\log(\pi_\theta(y|x)) - \log(\pi^\tau(y|x))| \leq \epsilon_1$, where ϵ_1 represents the maximum likelihood difference between π_θ and π^τ. Let $|(r_\theta(x, y_1) - r_\theta(x, y_2)) - (r(x, y_1) - r(x, y_2))| \leq \epsilon_2$, where ϵ_2 is the maximum deviation between r_θ and r. Define d as the diameter of the smallest manifold sphere containing the domain of θ (as per Definition D.1), i.e. $|\theta(x_1, y_1) - \theta(x_2, y_2)| \leq d$. Similarly, let $|p^*(z|y_1, y_2, x) - p_\theta(z|y_1, y_2, x)| \leq \epsilon_3$, where ϵ_3 bounds the difference between $\bar{p}_\theta(z|y_1, y_2, x)$ and $p^*(z|x, y_1, y_2)$.

5.2 Q2: The Influence of Preference Dataset on Algorithm Performance

Since DPO is a representative algorithm in RLHF, this paper uses it to answer Q2. Essentially, the design of preference dataset $\mathcal{D}_R$ depends on the design of the sampling strategy π_0, in other words, the relationship between the sampling strategy π_0 and the current strategy π_θ. Thus for Q2, we use Theorem 5.3 to show that data (x, y_1, y_2) should be selected where both the reward values and policy probabilities have a large and consistent margin between y_1 and y_2. Larger margins accelerate convergence, while consistent margins reduce distribution shift.

To prove Theorem 5.3, we first consider a simplified question: should π_0 prioritize samples with large or small reward differences or margins when identifying winners and losers? The challenge is that π_0 and π_θ are interdependent, as we optimize π_θ. To address this, Lemma 5.2 quantifies their relationship when π_0 is a uniform distribution. Define event sets $\Omega_1 = \{(y_1, y_2, x) \mid |\log \frac{p^*(1|y_1,y_2,x)}{p^*(0|y_1,y_2,x)}| \geq \epsilon_0\}$ and $\Omega_2 = \{(y_1, y_2, x) \mid |\log \frac{\pi_\theta(y_1|x)\pi_{\text{ref}}(y_2|x)}{\pi_\theta(y_2|x)\pi_{\text{ref}}(y_1|x)}| \geq \epsilon_0\}$. Let $\gamma(x) = \frac{|\Omega_1 \cap \Omega_2|}{K^2} \leq \gamma$, where $K = |\mathbb{Y}|$. ϵ_0 represents the log-probability significance threshold of p^*. Intuitively, the parameter γ reflects the alignment between π_0 and π_θ for large-margin data pairs (x, y_1, y_2).

Lemma 5.2. *Assume $\|g_t\|^2 \leq G^2$. See $\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{ref})$ on Eq.6. Given the learning rate α_t satisfying finite squared summability, for DPO, let $\gamma(x) \leq \gamma$, when π_0 is a uniform distribution, then:*

$$\min_{1 \leq i \leq T} ||\nabla_\theta \mathcal{L}_{\text{DPO}}(\pi_{\theta_i}; \pi_{ref})||_2^2 \leq \frac{\mathcal{L}_{\text{DPO}}(\pi_{\theta_1}) - \mathcal{L}^*_{\text{DPO}}}{\sum_{t=1}^{T-1} \alpha_t} + \frac{2(\gamma c_0 + 1)G^2 \sum_{t=1}^{T-1} \alpha_t^2}{\tau^2 \sum_{t=1}^{T-1} \alpha_t}. \tag{23}$$

where $\mathcal{L}^*_{\text{DPO}} = \min_{\pi_\theta} \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{ref})$ and $c_0 = \sigma(\frac{\epsilon_0}{\tau})\sigma(-\frac{\epsilon_0}{\tau}) - 1 \in (-1, 0)$.

Assuming π_0 is uniform, Lemma 5.2 prepares for analyzing how sampling certain types of data affects convergence. Since c_0 is negative, a higher γ value indicates better alignment and faster convergence, as shown in the second term of Eq.23.

Based on Lemma 5.2, we further demonstrate with Theorem 5.3 that when the data pairs (x, y_1, y_2) satisfy that $(x, y_1, y_2) \in \Omega_1 \cap \Omega_2$ and are sampled more, the convergence rate can be accelerated. In Theorem 5.3, μ represents a preference for selecting data pairs with large margins that are also consistent with π_θ. Compared to the upper bound of inequality 23, the upper bound of inequality 24 is smaller, indicating that the data selection distribution π_1 plays a beneficial role in accelerating the DPO optimization process. Thus we can sample data by π_1 to achieve faster convergence for DPO.

Theorem 5.3. *Define $\pi_1(y_1, y_2|x) = \frac{\mu}{K^2}$ if $(x, y_1, y_2) \in (\Omega_1 \cap \Omega_2)$; $\frac{1-\mu\gamma}{(1-\gamma)K^2}$ else where $\mu \in (0, 1)$, be the sample distribution of DPO. Assume $\|g_t\|^2 \leq G^2$. Given the learning rate α_t satisfying finite squared summability, we have:*

$$\min_{1 \leq i \leq T} ||\nabla_\theta \mathcal{L}_{\text{DPO}}(\pi_{\theta_i}; \pi_{ref})||_2^2 \leq \frac{\mathcal{L}_{\text{DPO}}(\pi_{\theta_1}) - \mathcal{L}^*_{\text{DPO}}}{\sum_{t=1}^{T-1} \alpha_t} + \frac{2(\mu\gamma c_0 + 1)G^2 \sum_{t=1}^{T-1} \alpha_t^2}{\tau^2 \sum_{t=1}^{T-1} \alpha_t}. \tag{24}$$

In summary, we conduct the answers for Q1 and Q2:

- About Q1: A larger value of τ leads to faster convergence of DPO. As τ increases, the optimal solution $\bar{\pi}^{\tau}$ of DPO converges to π^{δ}, while deviating from π_{ref}. Hence, there is a trade-off between the convergence speed and the preservation of proximity to π_{ref}.
- About Q2: Selecting data where both the current policy probability values ($\pi(y_1|x)$ and $\pi(y_2|x)$) and the reward values ($r(x, y_1)$ and $r(x, y_1)$) consistently exhibit a large distance (i.e. large margin) can accelerate the convergence of the DPO algorithm. And with the consistence, the distribution shift problem will reduce.

6 Conclusion

In this paper, we investigated the connections of DPO, RL and other RLHF algorithms. We proposed UDRRA to link these algorithms through their different reward function scenarios. Then we analyzed the target distributions corresponding to the methods in UDRRA and pointed out that the distribution shift problem of DPO comparing PRA-P. Furthermore, we examined the impact of key components of the DPO loss on algorithm performance. Our findings offer a deeper understanding of DPO's theoretical positioning and practical implications, providing a foundation for future developments in RLHF algorithms.

References

Abbas Abdolmaleki, Jost Tobias Springenberg, Jonas Degrave, Steven Bohez, Yuval Tassa, Dan Belov, Nicolas Heess, and Martin Riedmiller. Relative entropy regularized policy iteration. *arXiv preprint arXiv:1812.02256*, 2018.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Dilip Arumugam, Jun Ki Lee, Sophie Saskin, and Michael L Littman. Deep reinforcement learning from policy-dependent human feedback. *arXiv preprint arXiv:1902.04257*, 2019.

Mohammad Gheshlaghi Azar, Zhaohan Daniel Guo, Bilal Piot, Remi Munos, Mark Rowland, Michal Valko, and Daniele Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pp. 4447–4455. PMLR, 2024.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Ralph Allan Bradley and Milton E Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

Guoxin Chen, Minpeng Liao, Chengxi Li, and Kai Fan. Step-level value preference optimization for mathematical reasoning. *arXiv preprint arXiv:2406.10858*, 2024.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback. *arXiv preprint arXiv:2310.12773*, 2023.

Logan Engstrom, Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Firdaus Janoos, Larry Rudolph, and Aleksander Madry. Implementation matters in deep rl: A case study on ppo and trpo. In *International conference on learning representations*, 2019.

Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pp. 1861–1870. PMLR, 2018.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. A survey on large language models: Applications, challenges, limitations, and practical usage. *Authorea Preprints*, 2023.

Muhammad Usman Hadi, Qasem Al Tashi, Abbas Shah, Rizwan Qureshi, Amgad Muneer, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, et al. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*, 2024.

Jiwoo Hong, Noah Lee, and James Thorne. Reference-free monolithic preference optimization with odds ratio. *arXiv preprint arXiv:2403.07691*, 2024.

Xiangkun Hu, Tong He, and David Wipf. New desiderata for direct preference optimization. *arXiv preprint arXiv:2407.09072*, 2024.

Hamish Ivison, Yizhong Wang, Jiacheng Liu, Zeqiu Wu, Valentina Pyatkin, Nathan Lambert, Noah A Smith, Yejin Choi, and Hannaneh Hajishirzi. Unpacking dpo and ppo: Disentangling best practices for learning from preference feedback. *arXiv preprint arXiv:2406.09279*, 2024.

Ziniu Li, Tian Xu, and Yang Yu. Policy optimization in rlhf: The impact of out-of-preference data. *arXiv preprint arXiv:2312.10584*, 2023.

Yong Lin, Skyler Seto, Maartje ter Hoeve, Katherine Metcalf, Barry-John Theobald, Xuan Wang, Yizhe Zhang, Chen Huang, and Tong Zhang. On the limited generalization capability of the implicit reward model induced by direct preference optimization. *arXiv preprint arXiv:2409.03650*, 2024.

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International conference on machine learning*, pp. 6820–6829. PMLR, 2020.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. Large language models: A survey. *arXiv preprint arXiv:2402.06196*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Michael Panchenko. Rlhf without rl - direct preference optimization. In *ICLR Blogposts 2024*, 2024. URL https://iclr-blogposts.github.io/2024/blog/rlhf-without-rl/. https://iclr-blogposts.github.io/2024/blog/rlhf-without-rl/.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

Pierre Harvey Richemond, Yunhao Tang, Daniel Guo, Daniele Calandriello, Mohammad Gheshlaghi Azar, Rafael Rafailov, Bernardo Avila Pires, Eugene Tarassov, Lucas Spangher, Will Ellsworth, et al. Offline regularised reinforcement learning for large language models alignment. *arXiv preprint arXiv:2405.19107*, 2024.

John Schulman. Trust region policy optimization. *arXiv preprint arXiv:1502.05477*, 2015.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017a.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017b.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International conference on machine learning*, pp. 387–395. Pmlr, 2014.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15638–15650, 2022.

Yunhao Tang, Daniel Zhaohan Guo, Zeyu Zheng, Daniele Calandriello, Yuan Cao, Eugene Tarassov, Rémi Munos, Bernardo Ávila Pires, Michal Valko, Yong Cheng, et al. Understanding the performance gap between online and offline alignment algorithms. *arXiv preprint arXiv:2405.08448*, 2024.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.

Shusheng Xu, Wei Fu, Jiaxuan Gao, Wenjie Ye, Weilin Liu, Zhiyu Mei, Guangju Wang, Chao Yu, and Yi Wu. Is dpo superior to ppo for llm alignment? a comprehensive study. *arXiv preprint arXiv:2404.10719*, 2024.

Yuzi Yan, Yibo Miao, Jialian Li, Yipin Zhang, Jian Xie, Zhijie Deng, and Dong Yan. 3d-properties: Identifying challenges in dpo and charting a path forward. *arXiv preprint arXiv:2406.07327*, 2024.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023a.

Yao Zhao, Rishabh Joshi, Tianqi Liu, Misha Khalman, Mohammad Saleh, and Peter J Liu. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023b.

Han Zhong, Guhao Feng, Wei Xiong, Li Zhao, Di He, Jiang Bian, and Liwei Wang. Dpo meets ppo: Reinforced token optimization for rlhf. *arXiv preprint arXiv:2404.18922*, 2024.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*, 2019.

APPENDIX CONTENTS

- Appendix A: Related Work.
- Appendix B: Correspondence between the Series of Methods in Our Framework and Existing Algorithms.
- Appendix C: Discussion about $p^*(z \mid y_1, y_2, x) = \omega(r(x, y_{2-z}), r(x, y_{z+1}))$.
- Appendix D: Proof in Section 2.
- Appendix D.1: Proof of Proposition 2.2.
- Appendix D.2: The Softmax Transform of Policy π_θ.
- Appendix E: Proof in Section 3.
- Appendix E.1: Derivation of Equation 7.
- Appendix E.2: Derivation of Equation 8.
- Appendix E.3: Derivation of Equation 14.
- Appendix E.4: Derivation of Equation 10.
- Appendix E.5: Derivation of Equation 6.
- Appendix F: Proof in Section 4.
- Appendix F.1: The Target Distribution Equivalence of Eq.7, Eq.8 and Eq.10.
- Appendix F.2: The Target Distribution Equivalence of Eq.10 and Eq.13.
- Appendix F.3: The Target Distribution Equivalence of Eq.13 and Eq.14.
- Appendix F.4: Proof of The Target Distribution Equivalence of Eq.12 and Eq.15.
- Appendix F.5: Proof of Theorem 4.5.
- Appendix G: Proof in Section 5.
- Appendix G.1: Proof of Proposition G.1.
- Appendix G.2: Proof of Theorem 5.1.
- Appendix G.3: Proof of Lemma 5.2.
- Appendix G.4: Proof of Theorem 5.3.
- Appendix G.5: Proposition G.2.
- Appendix G.6: Sub Lemma for Lemma G.8-G.12.
- Appendix G.7: Proof of Lemma G.8.
- Appendix G.8: Proof of Lemma G.9.
- Appendix G.9: Proof of Lemma G.10.
- Appendix G.10: Proof of Lemma G.11.
- Appendix G.11: Proof of Lemma G.12.

A Related Work

Reinforcement Learning from Human Feedback (RLHF). The Large Language Models (LLMs) Zhao et al. (2023a); Chang et al. (2024); Hadi et al. (2024); Minaee et al. (2024); Hadi et al. (2023); Achiam et al. (2023); Bubeck et al. (2023) is one of the most promising evolutions towards Artificial General Intelligence (AGI). The success of this transformation lies a critical component: Reinforcement Learning from Human Feedback (RLHF) or Human Alignment (HA), which is the final and crucial step in LLMs' training Arumugam et al. (2019); Ouyang et al. (2022); Singh et al. (2022); Bai et al. (2022); Dai et al. (2023). Motivated by the instability, complexity, and incurring significant computational costs of the RLHF process, Rafailov et al. (2024) proposed an algorithmic framework for directly optimizing the language model to follow human preferences, namely the DPO algorithm, based on the analysis of the optimal solutions of classic RL algorithm PPO Schulman et al. (2017b); Engstrom et al. (2019). DPO has received widespread attention since its inception, and hundreds of algorithms for improving DPO have been derived within a year, e.g. IPOAzar et al. (2024), KTOEthayarajh et al. (2024), SimPOMeng et al. (2024), ORPOHong et al. (2024), etc.

DPO & PPO discussing. The success of DPO Rafailov et al. (2024) benefits from the optimal solution analysis of the KL-constrained reward maximization objective (Eq.3 in Rafailov et al. (2024)) come from the classical Reinforcement Learning (RL) algorithm, PPO Schulman et al. (2017a). Therefore, the performance difference between the DPO algorithm and the PPO based RLHF method Ziegler et al. (2019) has gradually attracted the attention of researchers in the RL community Ivison et al. (2024). In general, while the DPO algorithm can fit the static training dataset comparably, it generalizes less effectively than PPO based RLHF Lin et al. (2024). While PPO based RLHF usually performs better in the state-of-the-art production-level LLMs Yan et al. (2024), its correct fine-tuning usually requires more sophisticated techniques Xu et al. (2024).

There are many works that empirically discuss the relationship between the DPO algorithm and the PPO algorithm in RLHF. From the perspective of algorithm design, Xu et al. (2024); Ivison et al. (2024); Yan et al. (2024) pointed out that the PPO algorithm is generally better than the DPO algorithm in human preference alignment tasks, but the PPO algorithm requires various additional tricks, such as advantage normalization, large batch size, and exponential moving average update for the reference model. Zhong et al. (2024) combined the advantages of DPO and PPO to create a more effective algorithm named RTO at the token-level. From the perspective of data source, Tang et al. (2024) classified algorithms such as DPO as offline algorithms, while the PPO based RLHF is an online algorithm. This demonstration is consistent with our paper. Li et al. (2023); Lin et al. (2024) showed that compared with the DPO algorithm that only relies on static data sets, the PPO algorithm can use sufficient non-preferred data for policy optimization to significantly improve performance by relying on the generalization ability of its Reward model.

B Correspondence between the Series of Methods in Our Framework and Existing Algorithms

Method	#Corresponding Algorithms
Forward-BDA	SAC Haarnoja et al. (2018), DPG Silver et al. (2014)
Reverse-BDA	RERPI Abdolmaleki et al. (2018)
RA (RA-P)	DRO Richemond et al. (2024), Kimi k1.5 Team et al. (2025)
RDA (RDA-P)	IPO Azar et al. (2024), SVPO Chen et al. (2024)
PRA (PRA-P)	DPO Rafailov et al. (2024)

Table 2: Correspondence between the series of methods in our Framework and existing algorithms.

C Discussion about $p^*(z \mid y_1, y_2, x) = \omega(r(x, y_{2-z}), r(x, y_{z+1}))$

As for modeling the comparison probability $p^*(z \mid y_1, y_2, x)$ between $r(x, y_1)$ and $r(x, y_2)$, i.e. the pair (x, y_1, y_2, z) is sampled with probability $p^*(z \mid y_1, y_2, x) = \omega(r(x, y_{2-z}), r(x, y_{z+1}))$ where $z \in \{0, 1\}$, in general, the difficulty of estimating the reward depends on the complexity of ω^{-1}. Table 3 presents several special forms of the ω function, including cases based on the BT model assumptions Bradley & Terry (1952), along with an analysis of their relevant properties.

$\omega(r(x, y_{2-z}), r(x, y_{z+1}))$	#Requirement about $r(x, y)$	#Inverse Formula about $r(x, y_1) - r(x, y_2)$	#Corresponding Algorithm
[1] $\frac{\exp(\eta r(x,y_{2-z}))}{\exp(\eta r(x,y_{2-z}))+\exp(\eta r(x,y_{z+1}))}$	$[-\infty, \infty]$	$\frac{1}{\eta}\log(\frac{p^*(1\|y_1,y_2,x)}{1-p^*(1\|y_1,y_2,x)})$	DPORafailov et al. (2024)
$\frac{(\eta r(x,y_{2-z}))}{(\eta r(x,y_{2-z}))+(\eta r(x,y_{z+1}))}$	$[0, \infty]$	N/A	TYPOHu et al. (2024)
[2] $\frac{1}{2}+\frac{1}{2}\tanh(r(x, y_{2-z}) - r(x, y_{z+1}))$	$[-\infty, \infty]$	$\frac{1}{2}\log(\frac{p^*(1\|y_1,y_2,x)}{1-p^*(1\|y_1,y_2,x)})$	N/A
$\frac{1}{2}+\frac{1}{2}\sin(r(x, y_{2-z}) - r(x, y_{z+1}))$	$[-\frac{\pi}{2}, \frac{\pi}{2}]$	$\arcsin(2p^*(1 \mid y_1, y_2, x) - 1)$	N/A
$\mathbb{I}(r(x, y_{2-z}) - r(x, y_{z+1}))$	$[-\infty, \infty]$	1	IPOAzar et al. (2024)
[4] $\max\{0, 1 - \eta(r(x, y_{2-z}) - r(x, y_{z+1}))\}$	$[-\infty, \infty]$	N/A	SLiCZhao et al. (2023b)
[3][4] $\frac{\exp(\eta r(x,y_{2-z}))}{\exp(\eta r(x,y_{2-z}))+\exp(\eta r(x,y_{\text{average}}))}$	$[-\infty, \infty]$	$\frac{1}{\eta}\log(\frac{p^*(1\|y_1,y_2,x)(1-p^*(0\|y_1,y_2,x))}{p^*(0\|y_1,y_2,x)(1-p^*(1\|y_1,y_2,x))})$	KTOEthayarajh et al. (2024)
[4] $\frac{1}{(1+\exp(\eta(r(x,y_{2-z})-r(x,y_{z+1}))))^2}$	$[-\infty, \infty]$	$\frac{1}{\eta}\log(\frac{\sqrt{p^*(1\|y_1,y_2,x)}}{1-\sqrt{p^*(1\|y_1,y_2,x)}})$	N/A
[4] $\exp(\eta(r(x, y_{2-z}) - r(x, y_{z+1})))$	$[-\infty, \infty]$	$\frac{1}{\eta}\log(p^*(1 \mid y_1, y_2, x))$	N/A

[1] Here is the generalized BT model expression, like other work did, we simplified $\eta = 1$ in Eq.19.
[2] $\frac{1}{2} + \frac{1}{2}\tanh(r(x, y_1) - r(x, y_2))$ actually is a BT model expression when $\eta = 2$.
[3] $r(x, y_{\text{average}})$ is the reference point z_{ref} in KTOEthayarajh et al. (2024).
[4] These forms of ω may not strictly satisfy $p^*(1 \mid y_1, y_2, x) = (1 - p^*(0 \mid y_1, y_2, x))$.

Table 3: Various expressions of function $\omega(r(x, y_{2-z}), r(x, y_{z+1}))$.

D Proof in Section 2

D.1 Proof of Proposition 2.2

Proposition 2.2: Without loss of generality, assume that the reward function $r(x, y)$ has a unique maximum for any given x. Then the following limit holds: $\lim_{\tau\to\infty} \pi^\tau(y|x) = \pi^\delta(y|x)$.

Proof: Recall that

$$\pi^\delta(y|x) = \begin{cases} 1, & \text{if } y = \arg\max_{y'} r(x, y'), \\ 0, & \text{otherwise.} \end{cases}$$

Given x and defining $y^* = \arg\max\ r(x, y)$, we consider the limit of $\pi^\tau(y|x)$ at the point (x, y^*) as τ approaches infinity:

$$\begin{aligned} \lim_{\tau\to\infty} \pi^\tau(y^*|x) &= \lim_{\tau\to\infty} \frac{\exp(\tau r(x, y^*))}{\sum_{y\in Y}\exp(\tau r(x, y))} = \frac{1}{\lim_{\tau\to\infty}\sum_{y\in Y}\exp(\tau(r(x, y) - r(x, y^*)))} \\ &= \frac{1}{\sum_{y\in Y}\lim_{\tau\to\infty}\exp(\tau(r(x, y) - r(x, y^*)))} = \frac{1}{1 + 0 + \cdots + 0} = 1. \end{aligned} \tag{25}$$

Next, for $y \neq y^*$, we consider the limit of $\pi^\tau(y|x)$ at the point (x, y):

$$\begin{aligned} &\lim_{\tau\to\infty} \pi^\tau(y|x) = \lim_{\tau\to\infty} \frac{\exp(\tau r(x, y))}{\sum_{y'\in Y}\exp(\tau r(x, y'))} = \frac{1}{\lim_{\tau\to\infty}\sum_{y'\in Y}\exp(\tau(r(x, y') - r(x, y)))} \\ &= \frac{1}{\sum_{y'\in Y}\lim_{\tau\to\infty}\exp(\tau(r(x, y') - r(x, y)))} \\ &= \frac{1}{\sum_{y'\in Y,\ r(x,y')>r(x,y)}\lim_{\tau\to\infty}\exp(\tau(r(x, y') - r(x, y))) + \sum_{y'\in Y,\ r(x,y')\le r(x,y)}\lim_{\tau\to\infty}\exp(\tau(r(x, y') - r(x, y)))} \\ &= \frac{1}{\sum_{y'\in Y,\ r(x,y')>r(x,y)}\lim_{\tau\to\infty}\exp(\tau(r(x, y') - r(x, y)))} = 0. \end{aligned} \tag{26}$$

Thus, we conclude that $\lim_{\tau\to\infty} \pi^\tau(y|x) = \pi^\delta(y|x)$. Proof finished.

D.2 The Softmax Transform of Policy π_θ

Given a policy $\pi : \mathbb{X} \to \Delta(\mathbb{Y})$, the softmax transform of a vector exponentiates the components of the vector and normalizes it so that the result lies in the simplex. This can be used to transform vectors assigned to state-action pairs into policies:

Definition D.1. (Softmax transform.) Given the function $\theta : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}$, the softmax transform of θ is defined as $\pi_\theta(\cdot \mid x) \triangleq \text{softmax}(\theta(x, \cdot))$, where for all $y \in \mathbb{Y}$,

$$\pi_\theta(y \mid x) \triangleq \frac{\exp\{\theta(x, y)\}}{\sum_{y'}\exp\{\theta(x, y')\}}. \tag{27}$$

We call the values $\theta(x, y)$ the logit values and the function θ itself a logit function due to its origin in logistic regression. This paper assume the set is parameterized by a softmax function, i.e., $\Pi = \{\pi_\theta \mid \pi_\theta(\cdot \mid x) \triangleq \text{softmax}(\theta(x, \cdot)),\ \theta : \mathbb{X} \times \mathbb{Y} \to \mathbb{R}\}$, which is commonly used in the proof of the policy gradient theorem Mei et al. (2020).

E Proof in Section 3

E.1 Derivation of Equation 7

$$
\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{D}_{\mathrm{KL}}\left(\pi_\theta\left(\cdot \mid x\right) \| \pi^\tau\left(\cdot \mid x\right)\right)\right] = \mathbb{E}_{x\sim\mathcal{D}}[\mathrm{D}_{\mathrm{KL}}(\pi_\theta(y|x)||\frac{\exp(\tau r(x,y))}{Z(x)})] \\
=&\mathbb{E}_{x\sim\mathcal{D}}\left[\sum_{y\in Y}\pi_\theta(y|x)\log\left(\frac{\pi_\theta(y|x)}{\frac{\exp(\tau r(x,y))}{Z(x)}}\right)\right] = \mathbb{E}_{x\sim\mathcal{D}}\left[\sum_{y\in Y}\pi_\theta(y|x)\log\left(\frac{Z(x)\pi_\theta(y|x)}{\exp(\tau r(x,y))}\right)\right] \\
=&\mathbb{E}_{x\sim\mathcal{D}}\left[\sum_{y\in Y}-\pi_\theta(y|x)(\tau r(x,y)-\log(\pi_\theta(y|x)))+\log(Z(x))\right] \\
=&\mathbb{E}_{x\sim\mathcal{D}}\left[\sum_{y\in Y}-\pi_\theta(y|x)(\tau r(x,y)-\log(\pi_\theta(y|x)))\right]+\mathbb{E}_{x\sim\mathcal{D}}[\log(Z(x))] \\
\equiv&\mathbb{E}_{x\sim\mathcal{D}}\left[\sum_{y\in Y}-\pi_\theta(y|x)(\tau r(x,y)-\log(\pi_\theta(y|x)))\right] \\
=&\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(y|x)}\left[-(\tau r(x,y)-\log(\pi_\theta(y|x)))\right].
\end{aligned}
\tag{28}
$$

The "$\equiv$" means that the left and right ends differ by a term that is unrelated to θ.

E.2 Derivation of Equation 8

$$
\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{D}_{\mathrm{KL}}\left(\pi^\tau\left(\cdot \mid x\right) \| \pi_\theta\left(\cdot \mid x\right)\right)\right] \equiv \mathbb{E}_{x\sim\mathcal{D}}[\sum_{y\in Y}\frac{\exp(\tau r(x,y))}{Z(x)}(-\log\pi_\theta(y|x))] \\
=&-\mathbb{E}_{x\sim\mathcal{D},y\sim\pi^\tau(\cdot|x)}[\log\pi_\theta(y|x)] = \mathbb{E}_{x\sim\mathcal{D}}[\sum_{y\in Y}\pi_\theta(y|x)\frac{1}{\pi_\theta(y|x)}\frac{\exp(\tau r(x,y))}{Z(x)}(-\log\pi_\theta(y|x))] \\
=&-\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}[\frac{\exp(\tau r(x,y))}{Z(x)\pi_\theta(y|x)}\log\pi_\theta(y|x)].
\end{aligned}
\tag{29}
$$

E.3 Derivation of Equation 14

$$
\begin{aligned}
&\mathcal{L}_{\mathrm{PRA}}(\pi_\theta) = \mathbb{E}_{\mathcal{D}_\theta}\left[\mathrm{D}_{\mathrm{KL}}(p^*(z|y_1,y_2,x)||p_\theta(z|y_1,y_2,x))\right] \\
=&-\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_\theta(y|x)}\left[p^*(1|y_1,y_2,x)\log p_\theta\left(1|y_1,y_2,x\right)+p^*(0|y_1,y_2,x)\log p_\theta\left(0|y_1,y_2,x\right)\right] \\
&+\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_\theta(y|x)}\left[p^*(1|y_1,y_2,x)\log p^*(1|y_1,y_2,x)+p^*(0|y_1,y_2,x)\log p^*(0|y_1,y_2,x)\right] \\
=&-\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_\theta(y|x)}\left[p^*(1|y_1,y_2,x)\log p_\theta\left(1|y_1,y_2,x\right)+p^*(0|y_1,y_2,x)\log p_\theta\left(0|y_1,y_2,x\right)\right] \\
&+\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_\theta(y|x)}\left[M(x,y_1,y_2)\right] \\
=&-\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_\theta(y|x)}\left[p^*(1|y_1,y_2,x)\log p_\theta\left(1|y_1,y_2,x\right)+p^*(1|y_2,y_1,x)\log p_\theta\left(1|y_2,y_1,x\right)\right] \\
&+\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_\theta(y|x)}\left[M(x,y_1,y_2)\right] \\
=&-\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_\theta(y|x)}\left[p^*(1|y_w,y_l,x)\log p_\theta\left(1|y_w,y_l,x\right)+p^*(1|y_l,y_w,x)\log p_\theta\left(1|y_l,y_w,x\right)\right] \\
&+\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_\theta(y|x)}\left[M(x,y_1,y_2)\right] \\
=&-\mathbb{E}_{(x,y_w,y_l)\sim\mathcal{D}_\theta}\left[\log p_\theta\left(1|y_w,y_l,x\right)\right]+\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_\theta(y|x)}\left[M(x,y_1,y_2)\right].
\end{aligned}
\tag{30}
$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function, $M(x, y_1, y_2) = \sum_{z=0,1} p^*(z|y_1, y_2, x) \log p^*(z|y_1, y_2, x)$ and $(x, y_1, y_2) \sim \mathcal{D}_\theta \triangleq y_1 \succ y_2 \sim p^*(z = 1|y_1, y_2, x), y_1, y_2 \sim \pi_\theta(y|x), x \sim \mathcal{D}$. The third equality is because the symmetric complementarity property of $\omega(\cdot, \cdot)$:

$$p^*(1|y_2, y_1, x) = \omega(r(x, y_2), r(x, y_1)) = 1-\omega(r(x, y_1), r(x, y_2)) = 1-p^*(1|y_1, y_2, x) = p^*(0|y_1, y_2, x). \tag{31}$$

E.4 Derivation of Equation 10

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[(r_\theta(x, y) - r(x, y))^2\right] = \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(y|x)}\left(\frac{1}{\tau}\log(\pi_\theta(y|x)) + \frac{1}{\tau}\log(Z(x)) - r(x, y)\right)^2 \\
=&\frac{1}{\tau^2}\mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[\left(\log(\pi_\theta(y|x)) - \log(\frac{\exp(\tau r(x, y))}{Z(x)})\right)^2\right] = \frac{1}{\tau^2}\mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[\left(\log(\frac{\pi_\theta(y|x)}{\pi^\tau(y|x)})\right)^2\right].
\end{aligned} \tag{32}$$

E.5 Derivation of Equation 6

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{D}, y_1, y_2\sim\pi_0(y|x)}\left[\mathbb{D}_{\mathrm{KL}}(p^*(z|y_1, y_2, x)||\bar{p}_\theta(z|y_1, y_2, x))\right] \\
\equiv& -\mathbb{E}_{x\sim\mathcal{D}, y_1, y_2\sim\pi_0(y|x)}\left[p^*(1|y_1, y_2, x)\log\sigma\left(\bar{h}_\theta(x, y_1, y_2)\right) + p^*(0|y_1, y_2, x)\log\sigma\left(\bar{h}_\theta(x, y_2, y_1)\right)\right] \\
=& -\mathbb{E}_{x\sim\mathcal{D}, y_1, y_2\sim\pi_0(y|x)}\left[p^*(1|y_1, y_2, x)\log\sigma\left(\bar{h}_\theta(x, y_1, y_2)\right) + p^*(1|y_2, y_1, x)\log\sigma\left(\bar{h}_\theta(x, y_2, y_1)\right)\right] \\
=& -\mathbb{E}_{x\sim\mathcal{D}, y_1, y_2\sim\pi_0(y|x)}\left[p^*(1|y_w, y_l, x)\log\sigma\left(\bar{h}_\theta(x, y_w, y_l)\right) + p^*(1|y_l, y_w, x)\log\sigma\left(\bar{h}_\theta(x, y_l, y_w)\right)\right] \\
=& -\mathbb{E}_{(x, y_w, y_l)\sim\mathcal{D}_R}\left[\log\sigma\left(\bar{h}_\theta(x, y_w, y_l)\right)\right] \\
=& -\mathbb{E}_{(x, y_w, y_l)\sim\mathcal{D}_R}\left[\log\sigma\left(\frac{1}{\tau}\log\frac{\pi_\theta(y_w|x)}{\pi_{ref}(y_w|x)} - \frac{1}{\tau}\log\frac{\pi_\theta(y_l|x)}{\pi_{ref}(y_l|x)}\right)\right].
\end{aligned} \tag{33}$$

where $z \in \{0, 1\}$, $\bar{p}_\theta(0|y_1, y_2, x) = \sigma\left(\bar{h}_\theta(x, y_2, y_1)\right)$ and $\bar{p}_\theta(1|y_1, y_2, x) = \sigma\left(\bar{h}_\theta(x, y_1, y_2)\right)$.

F Proof in Section 4

F.1 The Target Distribution Equivalence of Eq.7, Eq.8 and Eq.10

Denote:

$$\begin{aligned}
\pi^*_{\mathrm{Forward-KL}} &= \arg\min_{\pi_\theta} -\mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[(\tau r(x, y) - \log(\pi_\theta(y|x)))\right], \\
\pi^*_{\mathrm{Reverse-KL}} &= \arg\min_{\pi_\theta} -\mathbb{E}_{x\sim\mathcal{D}, y\sim\pi^\tau(\cdot|x)}[\log\pi_\theta(y|x)], \\
\pi^*_{\mathrm{RA}} &= \arg\min_{\pi_\theta}\mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[\frac{1}{\tau^2}\left(\log(\frac{\pi_\theta(y|x)}{\pi^\tau(y|x)})\right)^2\right],
\end{aligned} \tag{34}$$

then $\pi^*_{\mathrm{Forward-KL}} = \pi^*_{\mathrm{Reverse-KL}} = \pi^*_{\mathrm{RA}} = \pi^\tau$.

Proof: First, we prove $\pi^*_{\text{Forward}-\text{KL}} = \pi^\tau$:

$$
\begin{aligned}
&-\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}\left[\tau r(x,y)-\log(\pi_\theta(y|x))\right]-(-\mathbb{E}_{x\sim\mathcal{D},y\sim\pi^\tau(\cdot|x)}\left[\tau r(x,y)-\log(\pi^\tau(y|x))\right])\\
=&-\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}\left[\tau r(x,y)-\log(\pi_\theta(y|x))\right]+\mathbb{E}_{x\sim\mathcal{D},y\sim\pi^\tau(\cdot|x)}\left[\tau r(x,y)-\log(\frac{\exp\left(\tau r\left(x,y\right)\right)}{Z(x)})\right]\\
=&-\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}\left[\tau r(x,y)-\log(\pi_\theta(y|x))\right]+\mathbb{E}_{x\sim\mathcal{D},y\sim\pi^\tau(\cdot|x)}\left[\log(Z(x))\right]\\
=&\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}\left[-\tau r(x,y)+\log(\pi_\theta(y|x))+\frac{\pi^\tau(y|x)}{\pi_\theta(y|x)}\log(Z(x))\right]\\
=&\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}\left[-(\log\exp(\tau r(x,y))-\log(Z(x))+\log(Z(x)))+\log(\pi_\theta(y|x))+\frac{\pi^\tau(y|x)}{\pi_\theta(y|x)}\log(Z(x))\right]\\
=&\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}\left[-\log(\pi^\tau(y|x))-\log(Z(x))+\log(\pi_\theta(y|x))+\frac{\pi^\tau(y|x)}{\pi_\theta(y|x)}\log(Z(x))\right]\\
=&\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}\left[\log(\frac{\pi_\theta(y|x)}{\pi^\tau(y|x)})+\frac{\pi^\tau(y|x)-\pi_\theta(y|x)}{\pi_\theta(y|x)}\log(Z(x))\right]\\
=&\mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{D}_{\mathrm{KL}}\left(\pi_\theta\left(\cdot\mid x\right)\|\pi^\tau\left(\cdot\mid x\right)\right)\right]\geq 0.
\end{aligned}
\tag{35}
$$

From the above equation, we can see that when π_θ is equal to π^τ, $-\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}\left[\tau r(x,y)-\log(\pi_\theta(y|x))\right]$ reaches the minimum value $-\mathbb{E}_{x\sim\mathcal{D},y\sim\pi^\tau(\cdot|x)}\left[\tau r(x,y)-\log(\pi^\tau(y|x))\right]$. Therefore $\pi^*_{\text{Forward}-\text{KL}} = \pi^\tau$.

Second, similarly we can prove $\pi^*_{\text{Reverse}-\text{KL}} = \pi^\tau$:

$$
-\mathbb{E}_{x\sim\mathcal{D},y\sim\pi^\tau(\cdot|x)}[\log\pi_\theta(y|x)]-(-\mathbb{E}_{x\sim\mathcal{D},y\sim\pi^\tau(\cdot|x)}[\log\pi^\tau(y|x)])=\mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{D}_{\mathrm{KL}}\left(\pi^\tau\left(\cdot\mid x\right)\|\pi_\theta\left(\cdot\mid x\right)\right)\right]\geq 0.
\tag{36}
$$

At last, we prove $\pi^*_{\text{RA}} = \pi^\tau$ like above:

$$
\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}\left[\frac{1}{\tau^2}\left(\log(\frac{\pi_\theta(y|x)}{\pi^\tau(y|x)})\right)^2\right]-\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}\left[\frac{1}{\tau^2}\left(\log(\frac{\pi^\tau(y|x)}{\pi^\tau(y|x)})\right)^2\right]\\
=&\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}\left[\frac{1}{\tau^2}\left(\log(\frac{\pi_\theta(y|x)}{\pi^\tau(y|x)})\right)^2\right]\geq 0.
\end{aligned}
\tag{37}
$$

Proof finished.

F.2 The Target Distribution Equivalence of Eq.10 and Eq.13

Here we prove the target distribution equivalence of Eq.10 and Eq.13. Denote

$$
\pi^*_{\text{RDA}} = \arg\min_{\pi_\theta}\mathbb{E}_{\mathcal{D}_{\text{pw}}}\left[\left(\frac{1}{\tau}\log\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)}-(r(x,y_1)-r(x,y_2))\right)^2\right],
\tag{38}
$$

then $\pi^*_{\text{RDA}} = \pi^*_{\text{RA}} = \pi^\tau$.

Proof:

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}_{\text{pw}}}\left[\left(\frac{1}{\tau}\log\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)}-(r(x,y_1)-r(x,y_2))\right)^2\right]-\mathbb{E}_{\mathcal{D}_{\text{pw}}}\left[\left(\frac{1}{\tau}\log\frac{\pi^\tau(y_1|x)}{\pi^\tau(y_2|x)}-(r(x,y_1)-r(x,y_2))\right)^2\right]\\
=&\mathbb{E}_{\mathcal{D}_{\text{pw}}}\left[\left(\frac{1}{\tau}\log\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)}-(r(x,y_1)-r(x,y_2))\right)^2\right]\geq 0.
\end{aligned}
\tag{39}
$$

When π_θ is equal to π^τ, $\mathbb{E}_{\mathcal{D}_{\text{pw}}}\left[\left(\frac{1}{\tau}\log\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)}-(r(x,y_1)-r(x,y_2))\right)^2\right]$ reaches the minimum value 0. Therefore $\pi^*_{\text{RDA}} = \pi^*_{\text{RA}} = \pi^\tau$. Proof finished.

F.3 THE TARGET DISTRIBUTION EQUIVALENCE OF EQ.13 AND EQ.14

Denote

$$\pi^*_{\text{PRA}} = \arg\min_{\pi_\theta} \mathbb{E}_{\mathcal{D}_\theta}\left[\mathrm{D}_{\mathrm{KL}}(p^*(z|y_1,y_2,x)||p_\theta(z|y_1,y_2,x))\right], \tag{40}$$

then $\pi^*_{\text{PRA}} = \pi^*_{\text{RDA}} = \pi^\tau$.

Proof: Because $r_\theta(x,y) = \frac{1}{\tau}\log(Z(x)\pi_\theta(y|x))$, $p^*(z \mid y_1,y_2,x) = \omega(r(x,y_{2-z}), r(x,y_{z+1}))$ and $p_\theta(z|y_1,y_2,x) = \omega(r_\theta(x,y_{2-z}), r_\theta(x,y_{z+1}))$, when π_θ is equal to π^τ, $r_\theta(x,y) = \frac{1}{\tau}\log(Z(x)\pi_\theta(y|x)) = r(x,y)$. Then $p_\theta(z|y_1,y_2,x) = p^*(z|y_1,y_2,x)$ and $\mathbb{E}_{\mathcal{D}_\theta}\left[\mathrm{D}_{\mathrm{KL}}(p^*(z|y_1,y_2,x)||p_\theta(z|y_1,y_2,x))\right] = 0$, then $\pi^*_{\text{PRA}} = \pi^\tau = \pi^*_{\text{RDA}}$. Proof finished.

F.4 PROOF OF THE TARGET DISTRIBUTION EQUIVALENCE OF EQ.12 AND EQ.15

Proof: See the definition of $\bar{\pi}^\tau$ in Eq.11. Denote

$$\begin{aligned}
\pi^*_{\text{RA}-\text{P}} &= \arg\min_{\pi_\theta} \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[\left(\frac{1}{\tau}\log(Z'(x)\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}) - r(x,y)\right)^2\right] \\
\pi^*_{\text{PRA}-\text{P}} &= \arg\min_{\pi_\theta} \mathbb{E}_{\mathcal{D}_\theta}\left[\mathrm{D}_{\mathrm{KL}}(p^*(z|y_1,y_2,x)||\bar{p}_\theta(z|y_1,y_2,x))\right].
\end{aligned} \tag{41}$$

First, when π_θ is equal to $\bar{\pi}^\tau$, then

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[\left(\frac{1}{\tau}\log(Z'(x)\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}) - r(x,y)\right)^2\right] \\
=&\mathbb{E}_{x\sim\mathcal{D}, y\sim\bar{\pi}^\tau(\cdot|x)}\left[\left(\frac{1}{\tau}\log(Z'(x)\frac{\bar{\pi}^\tau(y|x)}{\pi_{ref}(y|x)}) - r(x,y)\right)^2\right] \\
=&\mathbb{E}_{x\sim\mathcal{D}, y\sim\bar{\pi}^\tau(\cdot|x)}\left[\left(\frac{1}{\tau}\log(\exp(\tau r(x,y)) - r(x,y)\right)^2\right] = 0.
\end{aligned} \tag{42}$$

Because $\mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[\left(\frac{1}{\tau}\log(Z'(x)\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)}) - r(x,y)\right)^2\right] \geq 0$, then $\pi^*_{\text{RA}-\text{P}} = \bar{\pi}^\tau$.

Second, when π_θ is equal to $\bar{\pi}^\tau$, $r_\theta(x,y) = \frac{1}{\tau}\log(Z'(x)\frac{\pi_\theta(y|x)}{\pi_{ref}(y|x)})) = \frac{1}{\tau}\log(Z'(x)\frac{\bar{\pi}^\tau(y|x)}{\pi_{ref}(y|x)})) = r(x,y)$. Then $\bar{p}_\theta(z|y_1,y_2,x) = p^*(z|y_1,y_2,x)$ and $\mathbb{E}_{\mathcal{D}_\theta}\left[\mathrm{D}_{\mathrm{KL}}(p^*(z|y_1,y_2,x)||\bar{p}_\theta(z|y_1,y_2,x))\right] = 0$, then $\pi^*_{\text{PRA}-\text{P}} = \bar{\pi}^\tau$. Then $\pi^*_{\text{RA}-\text{P}} = \pi^*_{\text{PRA}-\text{P}} = \bar{\pi}^\tau$. Proof finished.

F.5 PROOF OF THEOREM 4.5

Theorem 4.5: Denote

$$\begin{aligned}
\mathcal{L}_{\text{DPO}}(\pi_\theta) &= \mathbb{E}_{\mathcal{D}_R}\left[-\log\sigma\left(\bar{h}_\theta(x,y_w,y_l)\right)\right], \\
\mathcal{L}_{\text{PRA}-\text{P}}(\pi_\theta) &= \mathbb{E}_{\mathcal{D}_\theta}\left[\mathrm{D}_{\mathrm{KL}}(p^*(z|y_1,y_2,x)||\bar{p}_\theta(z|y_1,y_2,x))\right].
\end{aligned}$$

where $\mathcal{D}_R \triangleq \{(x,y_w,y_l)|x\sim\mathcal{D}, y_w, y_l \sim \pi_0(\cdot|x), (y_w \succ y_l) \sim p^*(1|y_w,y_l,x)\}$ and $\mathcal{D}_\theta \triangleq \{(x,y_1,y_2)|x\sim\mathcal{D}, y_1,y_2\sim\pi_\theta(\cdot|x), (y_1\succ y_2)\sim p^*(1|y_1,y_2,x)\}$. When p^* is modeled by the BT model with Eq.19:

$$p^*(1|y_1,y_2,x) = \omega(r(x,y_{2-z}), r(x,y_{z+1})) = \sigma(r(x,y_1) - r(x,y_2)), \tag{43}$$

then we have the following equality:

$$\mathcal{L}_{\text{PRA}-\text{P}}(\pi_\theta) = \mathcal{L}_{\text{DPO}}(\pi_\theta) + \eta_1(\pi_\theta,\pi_0) + \eta_2(\pi_\theta). \tag{44}$$

where $\eta_1(\pi_\theta,\pi_0)$ equals to 0 if and only if $\pi_\theta = \pi_0$ and $\eta_2(\pi_\theta) \triangleq \mathbb{E}_{x\sim\mathcal{D}, y_1,y_2\sim\pi_\theta(y|x)}\left[M(x,y_1,y_2)\right]$.

Proof:

$$
\begin{aligned}
&\mathcal{L}_{\text{PRA-P}}\left(\pi_\theta\right)=\mathbb{E}_{\mathcal{D}_\theta}\left[\mathrm{D}_{\mathrm{KL}}(p^*(z|y_1,y_2,x)||\bar{p}_\theta\left(x,y_1,y_2\right))\right]\\
=&-\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_\theta(y|x)}\left[p^*(z=1|y_1,y_2,x)\log\bar{p}_\theta\left(x,y_1,y_2\right)+p^*(z=0|y_1,y_2,x)\log\bar{p}_\theta\left(x,y_2,y_1\right)\right]\\
&+\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_\theta(y|x)}\left[p^*(z=1|y_1,y_2,x)\log p^*(z=1|y_1,y_2,x)+p^*(z=0|y_1,y_2,x)\log p^*(z=0|y_1,y_2,x)\right]\\
=&-\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_\theta(y|x)}\left[p^*(z=1|y_1,y_2,x)\log\bar{p}_\theta\left(x,y_1,y_2\right)+p^*(z=0|y_1,y_2,x)\log\bar{p}_\theta\left(x,y_2,y_1\right)\right]\\
&+\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_\theta(y|x)}\left[M(x,y_1,y_2)\right].
\end{aligned}
\tag{45}
$$

where $\sigma(x)=\frac{1}{1+e^{-x}}$ is the sigmoid function, $M(x,y_1,y_2)=\sum_{z=0,1}p^*(z|y_1,y_2,x)\log p^*(z|y_1,y_2,x)$ and $(x,y_1,y_2)\sim\mathcal{D}_\theta\triangleq y_1\succ y_2\sim p^*(z=1|y_1,y_2,x),y_1,y_2\sim\pi_\theta(y|x),x\sim\mathcal{D}$.

Denote $\eta_2(\pi_\theta)=\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_\theta(y|x)}\left[M(x,y_1,y_2)\right]$, $\zeta(\bar{p}_\theta;x,y_1,y_2)=p^*(z=1|y_1,y_2,x)\log\bar{p}_\theta\left(x,y_1,y_2\right)+p^*(z=0|y_1,y_2,x)\log\bar{p}_\theta\left(x,y_2,y_1\right)$. Then we have:

$$
\begin{aligned}
&\mathcal{L}_{\text{PRA-P}}\left(\pi_\theta\right)=-\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_\theta(y|x)}\left[\zeta(\bar{p}_\theta;x,y_1,y_2)\right]+\eta_2(\pi_\theta)\\
=&-\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_0(y|x)}\left[\zeta(\bar{p}_\theta;x,y_1,y_2)\right]\\
&+\left(\sum_{x\in\mathbb{X},y_1,y_2\in\mathbb{Y}}(\pi_0(y|x)-\pi_\theta(y|x))\cdot\zeta(\bar{p}_\theta;x,y_1,y_2)\right)+\eta_2(\pi_\theta).
\end{aligned}
\tag{46}
$$

Denote $\sum_{x\in\mathbb{X},y_1,y_2\in\mathbb{Y}}(\pi_0(y|x)-\pi_\theta(y|x))\cdot\zeta(\bar{p}_\theta;x,y_1,y_2)$ as $\eta_1(\pi_\theta,\pi_0)$. Based on Eq.E.5 and $p^*(1|y_1,y_2,x)=\omega(r(x,y_{2-z}),r(x,y_{z+1}))=\sigma(r(x,y_1)-r(x,y_2))$ because of the BT model, then we have the final conclusion:

$$
\mathcal{L}_{\text{PRA-P}}\left(\pi_\theta\right)=\mathcal{L}_{\text{DPO}}\left(\pi_\theta\right)+\eta_1(\pi_\theta,\pi_0)+\eta_2(\pi_\theta).
\tag{47}
$$

Proof finished.

G Proof in Section 5

Proposition G.1. *If $F(\theta)$ is L-smooth for θ and $\|g_t\|^2\leq G^2$. Suppose the parameters θ are updated by:*

$$
\theta_{t+1}=\theta_t-\alpha_t g(x,y,\theta_t),
\tag{48}
$$

Then the following inequality holds:

$$
\min_{1\leq t\leq T-1}||\nabla F(\theta_t)||_2^2\leq\frac{LG^2\sum_{t=1}^{T-1}\alpha_t^2+2(F(\theta_1)-F^*)}{2\sum_{t=1}^{T-1}\alpha_t}.
\tag{49}
$$

where $F^*=\arg\min_\theta F(\theta)$.

Proof of Sketch: Before proof, firstly we need to introduce a formal setting. Assume that the parameterized policy π_θ is given by Definition D.1 and the parameters θ is optimized by the Stochastic Gradient Descent (SGD) method. Denote $F(\theta)=\mathbb{E}_{x,y\sim p(x,y;\pi_\theta)}[f(x,y,\theta)]$ where $p(x,y;\pi_\theta)$ is a distribution and $f(x,y,\theta)$ is the loss function on data (x,y) for policy parameters θ. Let $g(x,y,\theta_k)$ be the stochastic gradient of $\nabla F(\theta_k)$, i.e. $E_{x,y\sim p(x,y;\pi_\theta)}[g(x,y,\theta)]=\nabla_\theta F(\theta)$. Usually $g(x,y,\theta_t)$ is simplified to g_t when it does not cause ambiguity. With learning rate α_t, the SGD update rule is Eq.49. For Theorem G.1, since the L-smooth property cannot guarantee convex, thus for a non-convex optimization problem, the commonly used measure is the gradient norm. Therefore, with the properties of L-smooth and SGD, Proposition G.1 holds. See details in Appendix G.1.

Observing Eq.49, $||\nabla F(\theta_t)||_2^2$ will converge to 0 when $T\to\infty$ if $\sum_{t=1}^{\infty}\alpha_t^2\leq\infty$. The smaller the variance G of stochastic gradient g_t is, the faster the convergence rate of $||\nabla F(\theta_t)||_2^2$ is. For the methods in Figure 1, we prove that the loss functions of these methods all satisfy the L-smooth assumption with different coefficients under the softmax parametrization of policy π_θ (See Table 1, corresponding to Lemma G.8-G.12).

G.1 PROOF OF PROPOSITION G.1

Proof: Because $F(\theta)$ is L-smooth for θ, we have:

$$|F(\theta_{t+1}) - F(\theta_t) - \nabla_\theta F(\theta_t)(\theta_{t+1} - \theta_t)| \leq \frac{L}{2}\|\theta_{t+1} - \theta_t\|_2^2. \tag{50}$$

where $\theta_{t+1} = \theta_t - \alpha_t g_t$.

Substituting into $\theta_{t+1} = \theta_t - \alpha_t g_t$, we get:

$$\begin{aligned} &(F(\theta_{t+1}) - F(\theta_t) + \alpha_t \nabla_\theta F(\theta_t) g_t) \\ \leq &|F(\theta_{t+1}) - F(\theta_t) + \alpha_t \nabla_\theta F(\theta_t) g_t| \\ = &|F(\theta_{t+1}) - F(\theta_t) - \nabla_\theta F(\theta_t)(\theta_{t+1} - \theta_t)| \\ \leq &\frac{L}{2}\|\theta_{t+1} - \theta_t\|_2^2 = \frac{L}{2}\|\alpha_t g_t\|_2^2 \leq \frac{L}{2}\alpha_t^2 G^2. \end{aligned} \tag{51}$$

Base on Eq.51, we get:

$$\alpha_t \nabla_\theta F(\theta_t) g_t \leq \frac{L}{2}\alpha_t^2 G^2 + F(\theta_t) - F(\theta_{t+1}). \tag{52}$$

Summing both sides with respect to t yields:

$$\sum_{t=1}^{T-1} \alpha_t \nabla_\theta F(\theta_t) g_t \leq \frac{L}{2} G^2 \sum_{t=1}^{T-1} \alpha_t^2 + F(\theta_1) - F(\theta_T) \leq \frac{L}{2} G^2 \sum_{t=1}^{T-1} \alpha_t^2 + F(\theta_1) - F^*. \tag{53}$$

where $F^* = \arg\min_\theta F(\theta)$.

Expect the stochastic gradient g_t to get:

$$\min_{1 \leq t \leq T-1} ||\nabla_\theta F(\theta_t)||_2^2 \sum_{t=1}^{T-1} \alpha_t \leq \sum_{t=1}^{T-1} \alpha_t ||\nabla_\theta F(\theta_t)||_2^2 \leq \frac{L}{2} G^2 \sum_{t=1}^{T-1} \alpha_t^2 + F(\theta_1) - F^*. \tag{54}$$

Finally we have:

$$\min_{1 \leq t \leq T-1} ||\nabla F(\theta_t)||_2^2 \leq \frac{L G^2 \sum_{t=1}^{T-1} \alpha_t^2 + 2(F(\theta_1) - F^*)}{2 \sum_{t=1}^{T-1} \alpha_t}. \tag{55}$$

Proof finished.

G.2 PROOF OF THEOREM 5.1

Theorem 5.1: Assume $\|g_t\|^2 \leq G^2$. Given the Definition D.1 for policy π_θ and the learning rate α_t, for DPO algorithm, the following properties hold:

(0). $\forall r, \tau$, then $\theta \to \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}})$ is $\frac{4}{\tau^2}$-smooth.

(1). If there is π_θ^* such that $\mathbb{E}_{x \sim \mathcal{D}, y_1, y_2 \sim \pi_0(y|x)}[(p^*(1|y_1, y_2, x) - \sigma(\bar{h}_\theta(x, y_1, y_2)))] = 0$, then $\pi_\theta^* = \arg\min_{\pi_\theta} \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}, \pi_0)$. And $\bar{\pi}^\tau(y|x) = \frac{\pi_{ref}(y|x)\exp(\tau r(x,y))}{Z'(x)}$ is an example π_θ^* with the Bradley-Terry (BT) model.

(2). Denote $\mathcal{L}_{\text{DPO}}^* = \min_{\pi_\theta} \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}, \pi_0)$. $\min_{1 \leq i \leq T} ||\nabla_\theta \mathcal{L}_{\text{DPO}}(\pi_{\theta_i}; \pi_{\text{ref}})||_2^2 \leq \frac{2G^2 \sum_{t=1}^{T-1} \alpha_t^2}{\tau^2 \sum_{t=1}^{T-1} \alpha_t} + \frac{\mathcal{L}_{\text{DPO}}(\pi_{\theta_T}) - \mathcal{L}_{\text{DPO}}^*}{\sum_{t=1}^{T-1} \alpha_t}$.

See $\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}})$ on Eq.6. $\mathcal{D}_R \triangleq \{(x, y_w, y_l) | x \sim \mathcal{D}, y_w, y_l \sim \pi_0(\cdot|x), (y_w \succ y_l) \sim p^*(1|y_w, y_l, x)\}$ which π_0 is an unanalytical distribution, $\mathcal{D}$ is an arbitrary distribution and $\{z = 1|y_1, y_2, x\} \triangleq \{r(x, y_1) \geq r(x, y_2)\}$.

$$\begin{aligned} \bar{h}_\theta(x, y_w, y_l) &= \frac{1}{\tau} \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \frac{1}{\tau} \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}, \\ \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) &= -\mathbb{E}_{(x, y_w, y_l) \sim \mathcal{D}_R}\left[\log \sigma\left(\bar{h}_\theta(x, y_w, y_l)\right)\right]. \end{aligned} \tag{56}$$

Proof:

(0). $\forall r, \tau$, then $\theta \to \mathcal{L}_{\text{DPO}}\left(\pi_\theta; \pi_{\text{ref}}\right)$ is $\frac{8}{\tau^2}$-smooth.

Proof: By Lemma G.4, it suffices to show that the spectral radius of the hessian matrix of the second derivative of $\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}})$, i.e.

$$\left|\sum_{x,x'\in\mathbb{X}}\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\frac{\partial^2\mathcal{L}_{\text{DPO}}(\pi_\theta;\pi_{\text{ref}})}{\partial\theta(x,y_i)\partial\theta(x',y_j)}z(x',y_j)\right| \leq \left(\frac{4}{\tau^2}\right)||z(\cdot,\cdot)||_2^2. \tag{57}$$

Denote $h(\pi_\theta, x, y_1, y_2) = -p^*(1|y_1,y_2,x)\log\sigma\left(\bar{h}_\theta(x,y_1,y_2)\right) - p^*(0|y_1,y_2,x)\log\sigma\left(\bar{h}_\theta(x,y_2,y_1)\right)$, we have:

$$\begin{aligned}
&\left|\sum_{x,x'\in\mathbb{X}}\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\frac{\partial^2\mathcal{L}_{\text{DPO}}\left(\pi_\theta;\pi_{\text{ref}}\right)}{\partial\theta(x,y_i)\partial\theta(x',y_j)}z(x',y_j)\right| \\
=&\left|\sum_{x\in\mathbb{X}}\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\frac{\partial^2\sum_{x\in\mathbb{X}}\mathcal{D}(x)\sum_{y_1,y_2\in\mathbb{Y}}\pi_0(y_1|x)\pi_0(y_2|x)h(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}z(x,y_j)\right| \\
=&\left|\sum_{x\in\mathbb{X}}\mathcal{D}(x)\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\frac{\partial^2 f_{\text{DPO}}(x,\theta)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}z(x,y_j)\right| \\
\triangleq&|\sum_{x\in\mathbb{X}}\mathcal{D}(x)\psi(x)| \leq \|\mathcal{D}(\cdot)\|_1\|\psi(\cdot)\|_\infty = 1\cdot\|\psi(\cdot)\|_\infty.
\end{aligned} \tag{58}$$

where $f_{\text{DPO}}(x,\theta) = \sum_{y_1,y_2\in\mathbb{Y}}\pi_0(y_1|x)\pi_0(y_2|x)h(\pi_\theta,x,y_1,y_2)$ and $\psi(x) = \sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\frac{\partial^2 f_{\text{DPO}}(x,\theta)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}z(x,y_j)$.

The second derivative of $f_{\text{DPO}}(x,\theta)$ is:

$$\frac{\partial^2 f_{\text{DPO}}(x,\theta)}{\partial\theta(x,y_i)\partial\theta(x,y_j)} = \sum_{y_1,y_2\in\mathbb{Y}}\pi_0(y_1|x)\pi_0(y_2|x)\frac{\partial^2 h(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}. \tag{59}$$

Consider the first derivative of $h(\pi_\theta, x, y_1, y_2)$, denote $p^* = p^*(1|y_1,y_2,x)$:

$$\begin{aligned}
&\frac{\partial h(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_i)} \\
=&-p^*\frac{\partial}{\partial\theta(x,y_i)}(\log\sigma(\frac{1}{\tau}\log\frac{\pi_\theta(y_1|x)\pi_{\text{ref}}(y_2|x)}{\pi_\theta(y_2|x)\pi_{\text{ref}}(y_1|x)})) - (1-p^*)\frac{\partial}{\partial\theta(x,y_i)}(\log\sigma(\frac{1}{\tau}\log\frac{\pi_\theta(y_1|x)\pi_{\text{ref}}(y_2|x)}{\pi_\theta(y_2|x)\pi_{\text{ref}}(y_1|x)})) \\
=&-\left(p^*(1-\sigma(\frac{1}{\tau}\log\frac{\pi_\theta(y_1|x)\pi_{\text{ref}}(y_2|x)}{\pi_\theta(y_2|x)\pi_{\text{ref}}(y_1|x)})) - (1-p^*)\sigma(\frac{1}{\tau}\log\frac{\pi_\theta(y_1|x)\pi_{\text{ref}}(y_2|x)}{\pi_\theta(y_2|x)\pi_{\text{ref}}(y_1|x)})\right)\frac{1}{\tau}\frac{\partial}{\partial\theta(x,y_i)}(\log\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)}) \\
=&-\left(p^*-\sigma(\frac{1}{\tau}\log\frac{\pi_\theta(y_1|x)\pi_{\text{ref}}(y_2|x)}{\pi_\theta(y_2|x)\pi_{\text{ref}}(y_1|x)})\right)\frac{1}{\tau}(\delta_{y_1y_i}-\delta_{y_2y_i}).
\end{aligned} \tag{60}$$

Consider the second derivative of $h(\pi_\theta, x, y_1, y_2)$:

$$\begin{aligned}
&\frac{\partial^2 h(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_i)\partial\theta(x,y_j)} \\
=&-\frac{\partial}{\partial\theta(x,y_j)}(\sigma(\frac{1}{\tau}\log\frac{\pi_\theta(y_2|x)\pi_{\text{ref}}(y_1|x)}{\pi_\theta(y_1|x)\pi_{\text{ref}}(y_2|x)})\frac{1}{\tau}(\delta_{y_1y_i}-\delta_{y_2y_i})) \\
=&\frac{1}{\tau^2}(\delta_{y_1y_i}-\delta_{y_2y_i})(\delta_{y_1y_j}-\delta_{y_2y_j})\sigma(\frac{1}{\tau}\log\frac{\pi_\theta(y_1|x)\pi_{\text{ref}}(y_2|x)}{\pi_\theta(y_2|x)\pi_{\text{ref}}(y_1|x)})\sigma(\frac{1}{\tau}\log\frac{\pi_\theta(y_2|x)\pi_{\text{ref}}(y_1|x)}{\pi_\theta(y_1|x)\pi_{\text{ref}}(y_2|x)}).
\end{aligned} \tag{61}$$

Then:

$$
\begin{aligned}
&|\psi(x)| = |\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\frac{\partial^2 f_{\text{DPO}}(x,\theta)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}z(x,y_j)| \\
=&|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}}\pi_0(y_1|x)\pi_0(y_2|x)\frac{\partial^2 h(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}z(x,y_j)| \\
\leq&|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}}\pi_0(y_1|x)\pi_0(y_2|x)\frac{1}{\tau^2}(\delta_{y_1y_i}-\delta_{y_2y_i})(\delta_{y_1y_j}-\delta_{y_2y_j})z(x,y_j)| \\
=&|\sum_{y_i\in\mathbb{Y}} z(x,y_i)\pi_0(y_i|x)\frac{1}{\tau^2}z(x,y_i)-\sum_{y_i\in\mathbb{Y}} z(x,y_i)\pi_0(y_i|x)\sum_{y_j\in\mathbb{Y}}\pi_0(y_j|x)\frac{1}{\tau^2}z(x,y_j) \\
&-\sum_{y_i\in\mathbb{Y}} z(x,y_i)\pi_0(y_i|x)\sum_{y_j\in\mathbb{Y}}\pi_0(y_j|x)\frac{1}{\tau^2}z(x,y_j)+\sum_{y_i\in\mathbb{Y}} z(x,y_i)\pi_0(y_i|x)\frac{1}{\tau^2}z(x,y_i)| \\
\leq&\frac{2}{\tau^2}|\sum_{y_i\in\mathbb{Y}} z(x,y_i)\pi_0(y_i|x)z(x,y_i)|+|\sum_{y_i\in\mathbb{Y}} z(x,y_i)\pi_0(y_i|x)\sum_{y_j\in\mathbb{Y}}\pi_0(y_j|x)z(x,y_j)| \\
=&\frac{2}{\tau^2}(\pi_0^T(\cdot|x)z^2(x,\cdot)+(\pi_0^T(\cdot|x)z(x,\cdot))^2)\leq\frac{2}{\tau^2}(\|\pi_0(\cdot|x)\|_1\|z^2(x,\cdot)\|_\infty+(\|\pi_0(\cdot|x)\|_1\|z(x,\cdot)\|_\infty)^2) \\
\leq&\frac{4}{\tau^2}\|z(x,\cdot)\|_2^2.
\end{aligned}
\tag{62}
$$

The first inequality is because $\sigma(\cdot)\leq 1$. Therefore,

$$
\|\psi(\cdot)\|_\infty=\max_{x\in\mathbb{X}}|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\frac{\partial^2 f_{\text{DPO}}(x,\theta)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}z(x,y_j)|\leq\max_{x\in\mathbb{X}}\frac{4}{\tau^2}\|z(x,\cdot)\|_2^2\leq\frac{4}{\tau^2}\|z(\cdot,\cdot)\|_2^2. \tag{63}
$$

Then Eq.57 is proved. Proof finished.

(1). If π_θ such that $\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_0(y|x)}[(p^*(1|y_1,y_2,x)-\sigma(\bar{h}_\theta(x,y_1,y_2)))]=0$, then the DPO's loss function get a local optimal solution (zero gradient point) where π_0 is another unknown distribution. And $\bar{\pi}^\tau(y|x)=\frac{\pi_{ref}(y|x)\exp(\tau r(x,y))}{Z'(x)}$ is an example with the Bradley-Terry (BT) model.

Proof: From the above theorem, it is evident that the optimal solution for π_θ is influenced by the comparison distribution $p^*(z|y_1,y_2,x)$. Here, we provide an existence proof. We first propose a candidate solution $\bar{\pi}^\tau(y|x)=\frac{\pi_{ref}(y|x)\exp(\tau r(x,y))}{Z'(x)}$ where $Z'(x)=\sum_{y'\in Y}\pi_{ref}(y'|x)\exp(\tau r(x,y'))$, and subsequently demonstrate that if $p^*(z|y_1,y_2,x)$ is modeled by the Bradley-Terry (BT) model Bradley & Terry (1952), solution $\bar{\pi}^\tau(y|x)=\frac{\pi_{ref}(y|x)\exp(\tau r(x,y))}{Z'(x)}$ satisfies condition $\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_0(y|x)}[(p^*(1|y_1,y_2,x)-\sigma(\bar{h}_\theta(x,y_1,y_2)))]=0$ as a locally optimal solution. But it is not necessarily a global optimum, as we can only verify that $\bar{\pi}^\tau(y|x)=\frac{\pi_{ref}(y|x)\exp(\tau r(x,y))}{Z'(x)}$ is a point where the gradient of the loss function $\mathcal{L}_{\text{DPO}}(\pi_\theta;\pi_{\text{ref}})$ equals zero.

To compute the gradient of the loss function $\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_0(y|x)}\left[\mathrm{D}_{\mathrm{KL}}(p^*(z|y_1,y_2,x)||\bar{p}_\theta(z|y_1,y_2,x))\right]$ with respect to θ, we proceed as follows:

$$
\begin{aligned}
&\nabla_\theta\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_0(y|x)}\left[\mathrm{D}_{\mathrm{KL}}(p^*(z|y_1,y_2,x)||\bar{p}_\theta(z|y_1,y_2,x))\right]\\
=&-\nabla_\theta\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_0(y|x)}\left[p^*(1|y_1,y_2,x)\log\sigma(\bar{h}_\theta(x,y_1,y_2))+p^*(0|y_1,y_2,x)\log\sigma(\bar{h}_\theta(x,y_2,y_1))\right]\\
=&-\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_0(y|x)}\left[p^*(1|y_1,y_2,x)\nabla_\theta\log\sigma(\bar{h}_\theta(x,y_1,y_2))+p^*(0|y_1,y_2,x)\nabla_\theta\log\sigma(\bar{h}_\theta(x,y_2,y_1))\right]\\
=&-\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_0(y|x)}\left[\frac{p^*(1|y_1,y_2,x)}{\sigma(\bar{h}_\theta(x,y_1,y_2))}\nabla_\theta\sigma(\bar{h}_\theta(x,y_1,y_2))+\frac{p^*(0|y_1,y_2,x)}{\sigma(\bar{h}_\theta(x,y_2,y_1))}\nabla_\theta\sigma(\bar{h}_\theta(x,y_2,y_1))\right]\\
=&-\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_0(y|x)}\left[\frac{p^*(1|y_1,y_2,x)}{\sigma(\bar{h}_\theta(x,y_1,y_2))}\nabla_\theta\sigma(\bar{h}_\theta(x,y_1,y_2))-\frac{1-p^*(1|y_1,y_2,x)}{1-\sigma(\bar{h}_\theta(x,y_1,y_2))}\nabla_\theta\sigma(\bar{h}_\theta(x,y_1,y_2))\right]\\
=&-\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_0(y|x)}\left[\left(\frac{p^*(1|y_1,y_2,x)(1-\sigma(\bar{h}_\theta))-(1-p^*(1|y_1,y_2,x))\sigma(\bar{h}_\theta)}{\sigma(\bar{h}_\theta(x,y_1,y_2))(1-\sigma(\bar{h}_\theta)(x,y_1,y_2))}\right)\nabla_\theta\sigma(\bar{h}_\theta)\right]\\
=&-\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_0(y|x)}\left[\left(\frac{p^*(1|y_1,y_2,x)-\sigma(\bar{h}_\theta(x,y_1,y_2))}{\sigma(\bar{h}_\theta(x,y_1,y_2))(1-\sigma(\bar{h}_\theta(x,y_1,y_2)))}\right)\nabla_\theta\sigma(\bar{h}_\theta(x,y_1,y_2))\right]\\
=&-\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_0(y|x)}\left[\left(p^*(1|y_1,y_2,x)-\sigma(\bar{h}_\theta(x,y_1,y_2))\right)\nabla_\theta\bar{h}_\theta(x,y_1,y_2)\right].
\end{aligned}
\tag{64}
$$

where $z\in\{0,1\}$, $\bar{p}_\theta(0|y_1,y_2,x)=\sigma\left(\bar{h}_\theta(x,y_2,y_1)\right)$ and $\bar{p}_\theta(1|y_1,y_2,x)=\sigma\left(\bar{h}_\theta(x,y_1,y_2)\right)$.

Under the assumption of the Bradley-Terry (BT) model, which posits:

$$
p^*(1|y_1,y_2,x)=\frac{1}{1+e^{-(r(x,y_1)-r(x,y_2))}}=\sigma(h(x,y_1,y_2)),
\tag{65}
$$

substituting this into the gradient of the target function, we obtain:

$$
\begin{aligned}
&\nabla_\theta\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_0(y|x)}\left[\mathrm{D}_{\mathrm{KL}}(p^*(z|y_1,y_2,x)||\bar{p}_\theta(z|y_1,y_2,x))\right]\\
=&-\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_0(y|x)}\left[\left(p^*(1|y_1,y_2,x)-\sigma(\bar{h}_\theta(x,y_1,y_2))\right)\nabla_\theta\bar{h}_\theta(x,y_1,y_2)\right]\\
=&-\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_0(y|x)}\left[\left(\sigma(h(x,y_1,y_2))-\sigma(\bar{h}_\theta(x,y_1,y_2))\right)\nabla_\theta\bar{h}_\theta(x,y_1,y_2)\right].
\end{aligned}
\tag{66}
$$

where $\bar{h}_\theta\left(x,y_1,y_2\right)=\frac{1}{\tau}\log\frac{\pi_\theta(y_1|x)}{\pi_{\text{ref}}\left(y_1|x\right)}-\frac{1}{\tau}\log\frac{\pi_\theta(y_2|x)}{\pi_{\text{ref}}\left(y_2|x\right)}$ and $h\left(x,y_1,y_2\right)=r(x,y_1)-r(x,y_2)$.

Given that $\sigma(\cdot)$ is a strictly increasing function, it follows that when $\pi_\theta(y|x)=\bar{\pi}^\tau(y|x)=\frac{\pi_{ref}(y|x)\exp(\tau r(x,y))}{Z'(x)}$, we have $\bar{h}_\theta(x,y_1,y_2)=h(x,y_1,y_2)$ and $\nabla_\theta\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_0(y|x)}\left[\mathrm{D}_{\mathrm{KL}}(p^*(z|y_1,y_2,x)||\bar{p}_\theta(z|y_1,y_2,x))\right]=0$. Proof finished.

(2). Denote $\mathcal{L}^*_{\text{DPO}}=\min_{\pi_\theta}\mathcal{L}_{\text{DPO}}(\pi_\theta;\pi_{\text{ref}},\pi_0)$. $\min_{1\le i\le T}||\nabla_\theta\mathcal{L}_{\text{DPO}}(\pi_{\theta_i};\pi_{\text{ref}})||_2^2\le\frac{2G^2\sum_{t=1}^{T-1}\alpha_t^2}{\tau^2T}+\frac{\mathcal{L}_{\text{DPO}}(\pi_{\theta_T})-\mathcal{L}^*_{\text{DPO}}}{T}$.

This theorem can be obtained from Theorem 5.1.(1) and Theorem G.1.

G.3 Proof of Lemma 5.2

Lemma 5.2: Assume $\|g_t\|^2\le G^2$. Given the Definition D.1 for policy π_θ and the learning rate α_t, for DPO algorithm, let $\gamma(x)\le\gamma$, when π_0 is uniform distribution, then:

$$
\min_{1\le i\le T}||\nabla_\theta\mathcal{L}_{\text{DPO}}(\pi_{\theta_i};\pi_{\text{ref}})||_2^2\le\frac{\mathcal{L}_{\text{DPO}}(\pi_{\theta_T})-\mathcal{L}^*_{\text{DPO}}}{\sum_{t=1}^{T-1}\alpha_t}+\frac{2(\gamma c_0+1)G^2\sum_{t=1}^{T-1}\alpha_t^2}{\tau^2\sum_{t=1}^{T-1}\alpha_t}.
\tag{67}
$$

where $\mathcal{L}^*_{\text{DPO}}=\min_{\pi_\theta}\mathcal{L}_{\text{DPO}}(\pi_\theta;\pi_{\text{ref}},\pi_0)$ and $c_0=\sigma(\frac{\epsilon_0}{\tau})\sigma(-\frac{\epsilon_0}{\tau})-1\in(-1,0)$. See $\mathcal{L}_{\text{DPO}}(\pi_\theta;\pi_{\text{ref}})$ on Eq.6.

Proof: Based on Theorem G.1, we only need to know $\mathcal{L}_{\text{DPO}}(\pi_\theta;\pi_{\text{ref}})$ is $\frac{4(\gamma c_0+1)}{\tau^2}$-smooth.

Denote $h(\pi_\theta, x, y_1, y_2) = -p^*(1|y_1, y_2, x) \log \sigma\left(\bar{h}_\theta(x, y_1, y_2)\right) - p^*(0|y_1, y_2, x) \log \sigma\left(\bar{h}_\theta(x, y_2, y_1)\right)$, we have:

$$
\begin{aligned}
&\left| \sum_{x,x' \in \mathbb{X}} \sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i) \frac{\partial^2 \mathcal{L}_{\text{DPO}}\left(\pi_\theta; \pi_{\text{ref}}\right)}{\partial\theta(x, y_i)\partial\theta(x', y_j)} z(x', y_j) \right| \\
=&\left| \sum_{x \in \mathbb{X}} \sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i) \frac{\partial^2 \sum_{x \in \mathbb{X}} \mathcal{D}(x) \sum_{y_1, y_2 \in \mathbb{Y}} \pi_0(y_1|x)\pi_0(y_2|x) h(\pi_\theta, x, y_1, y_2)}{\partial\theta(x, y_i)\partial\theta(x, y_j)} z(x, y_j) \right| \\
=&\left| \sum_{x \in \mathbb{X}} \mathcal{D}(x) \sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i) \frac{\partial^2 f_{\text{DPO}}(x, \theta)}{\partial\theta(x, y_i)\partial\theta(x, y_j)} z(x, y_j) \right| \\
\triangleq&|\sum_{x \in \mathbb{X}} \mathcal{D}(x)\psi(x)| \leq \|\mathcal{D}(\cdot)\|_1 \|\psi(\cdot)\|_\infty = 1 \cdot \|\psi(\cdot)\|_\infty.
\end{aligned}
\tag{68}
$$

where $f_{\text{DPO}}(x, \theta) = \sum_{y_1, y_2 \in \mathbb{Y}} \pi_0(y_1|x)\pi_0(y_2|x) h(\pi_\theta, x, y_1, y_2)$ and $\psi(x) = \sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i) \frac{\partial^2 f_{\text{DPO}}(x,\theta)}{\partial\theta(x,y_i)\partial\theta(x,y_j)} z(x, y_j)$.

Because π_0 is uniform distribution, the second derivative of $f_{\text{DPO}}(x, \theta)$ is:

$$
\frac{\partial^2 f_{\text{DPO}}(x, \theta)}{\partial\theta(x, y_i)\partial\theta(x, y_j)} = \sum_{y_1, y_2 \in \mathbb{Y}} \frac{1}{K^2} \frac{\partial^2 h(\pi_\theta, x, y_1, y_2)}{\partial\theta(x, y_i)\partial\theta(x, y_j)}. \tag{69}
$$

Consider the second derivative of $h(\pi_\theta, x, y_1, y_2)$:

$$
\begin{aligned}
&\frac{\partial^2 h(\pi_\theta, x, y_1, y_2)}{\partial\theta(x, y_i)\partial\theta(x, y_j)} = -\frac{\partial}{\partial\theta(x, y_j)}(\sigma(\frac{1}{\tau} \log \frac{\pi_\theta(y_2|x)\pi_{\text{ref}}(y_1|x)}{\pi_\theta(y_1|x)\pi_{\text{ref}}(y_2|x)})\frac{1}{\tau}(\delta_{y_1 y_i} - \delta_{y_2 y_i})) \\
=&\frac{1}{\tau^2}(\delta_{y_1 y_i} - \delta_{y_2 y_i})(\delta_{y_1 y_j} - \delta_{y_2 y_j})\sigma(\frac{1}{\tau} \log \frac{\pi_\theta(y_1|x)\pi_{\text{ref}}(y_2|x)}{\pi_\theta(y_2|x)\pi_{\text{ref}}(y_1|x)})\sigma(\frac{1}{\tau} \log \frac{\pi_\theta(y_2|x)\pi_{\text{ref}}(y_1|x)}{\pi_\theta(y_1|x)\pi_{\text{ref}}(y_2|x)}) \\
=&\frac{1}{\tau^2}(\delta_{y_1 y_i} - \delta_{y_2 y_i})(\delta_{y_1 y_j} - \delta_{y_2 y_j})\zeta(x, y_1, y_2).
\end{aligned}
\tag{70}
$$

where $\zeta(x, y_1, y_2) = \sigma(\frac{1}{\tau} \log \frac{\pi_\theta(y_1|x)\pi_{\text{ref}}(y_2|x)}{\pi_\theta(y_2|x)\pi_{\text{ref}}(y_1|x)})\sigma(\frac{1}{\tau} \log \frac{\pi_\theta(y_2|x)\pi_{\text{ref}}(y_1|x)}{\pi_\theta(y_1|x)\pi_{\text{ref}}(y_2|x)})$.

Let $m_\theta(x, y_1, y_2) = \frac{\pi_\theta(y_1|x)\pi_{\text{ref}}(y_2|x)}{\pi_\theta(y_2|x)\pi_{\text{ref}}(y_1|x)}$ and . Because $\Omega_1(y_1, y_2, x) = \{|\log \frac{p^*(1|y_1,y_2,x)}{p^*(0|y_1,y_2,x)}| \geq \epsilon_0\}$, $\Omega_2(y_1, y_2, x) = \{|\log m_\theta(x, y_1, y_2)| \geq \epsilon_0\}$ and $\gamma(x) = \frac{\sum_{y_1, y_2 \in \mathbb{Y}} \mathbb{I}(\Omega_1 \cap \Omega_2)}{K^2}$.

$$
\begin{aligned}
&|\psi(x)| = |\sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i) \frac{\partial^2 f_{\text{DPO}}(x, \theta)}{\partial\theta(x, y_i)\partial\theta(x, y_j)} z(x, y_j)| \\
=&|\sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i) \sum_{y_1, y_2 \in \mathbb{Y}} \frac{1}{K^2} \frac{\partial^2 h(\pi_\theta, x, y_1, y_2)}{\partial\theta(x, y_i)\partial\theta(x, y_j)} z(x, y_j)| \\
=&|\sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i) \sum_{y_1, y_2 \in \mathbb{Y}} \frac{1}{K^2} \frac{1}{\tau^2}(\delta_{y_1 y_i} - \delta_{y_2 y_i})(\delta_{y_1 y_j} - \delta_{y_2 y_j})\zeta(x, y_1, y_2) z(x, y_j)| \\
=&|\sum_{y_i, y_2 \in \mathbb{Y}} z(x, y_i) \frac{1}{K^2} \zeta(x, y_i, y_2) \frac{1}{\tau^2} z(x, y_i) - \sum_{y_j, y_i \in \mathbb{Y}} z(x, y_i) \frac{1}{K^2} \zeta(x, y_i, y_j) \frac{1}{\tau^2} z(x, y_j) \\
&- \sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i) \frac{1}{K^2} \zeta(x, y_i, y_j) \frac{1}{\tau^2} z(x, y_j) + \sum_{y_1, y_j \in \mathbb{Y}} z(x, y_i) \frac{1}{K^2} \zeta(x, y_1, y_j) \frac{1}{\tau^2} z(x, y_i)| \\
\leq&\frac{2}{\tau^2}|\sum_{y_i, y_2 \in \mathbb{Y}} z(x, y_i) \frac{1}{K^2} \zeta(x, y_i, y_2) z(x, y_i)| + \frac{2}{\tau^2}|\sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i) \frac{1}{K^2} \zeta(x, y_i, y_j) z(x, y_j)| \\
=&\frac{2}{\tau^2}|\sum_{y_i \in \mathbb{Y}} z^2(x, y_i) \sum_{y_2 \in \mathbb{Y}} \frac{1}{K^2} \zeta(x, y_i, y_2)| + \frac{2}{\tau^2}|\sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i) \frac{1}{K^2} \zeta(x, y_i, y_j) z(x, y_j)|.
\end{aligned}
\tag{71}
$$

For the first term $\frac{2}{\tau^2}|\sum_{y_i\in\mathbb{Y}} z^2(x,y_i)\sum_{y_2\in\mathbb{Y}}\frac{1}{K^2}\zeta(x,y_i,y_2)|$, based on Hölder's inequality we have:

$$\begin{aligned}
&\frac{2}{\tau^2}|\sum_{y_i\in\mathbb{Y}} z^2(x,y_i)\sum_{y_2\in\mathbb{Y}}\frac{1}{K^2}\zeta(x,y_i,y_2)| \leq \frac{2}{\tau^2}\sum_{y_i,y_2\in\mathbb{Y}}|\frac{1}{K^2}\zeta(x,y_i,y_2)|\cdot\|z^2(x,\cdot)\|_\infty \\
\leq &\frac{2}{\tau^2}(\gamma(x)\sigma(\frac{\epsilon_0}{\tau})\sigma(-\frac{\epsilon_0}{\tau})+1-\gamma(x))\|z^2(x,\cdot)\|_\infty = \frac{2}{\tau^2}(\gamma(x)c_0+1)\|z(x,\cdot)\|_\infty^2 \leq \frac{2}{\tau^2}(\gamma(x)c_0+1)\|z(x,\cdot)\|_2^2.
\end{aligned} \tag{72}$$

For the second term $\frac{2}{\tau^2}|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\frac{1}{K^2}\zeta(x,y_i,y_j)z(x,y_j)|$, let $\mathbb{Y}_1(x) = \{(y_1,y_2)|\mathbb{I}(\Omega_1(x,y_1,y_2)\cap\Omega_2(x,y_1,y_2)) = 1\}$ and $\mathbb{Y}_2(x) = \{(y_1,y_2)|\mathbb{I}(\Omega_1(x,y_1,y_2)\cap\Omega_2(x,y_1,y_2)) = 0\}$, we have:

$$\begin{aligned}
&\frac{2}{\tau^2}|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\frac{1}{K^2}\zeta(x,y_i,y_j)z(x,y_j)| \\
\leq &\frac{2}{\tau^2}|\sum_{(y_i,y_j)\in\mathbb{Y}_1} z(x,y_i)\frac{1}{K^2}\zeta(x,y_i,y_j)z(x,y_j)| + \frac{2}{\tau^2}|\sum_{(y_i,y_j)\in\mathbb{Y}_2} z(x,y_i)\frac{1}{K^2}\zeta(x,y_i,y_j)z(x,y_j)| \\
\leq &\frac{2}{\tau^2}|\sum_{(y_i,y_j)\in\mathbb{Y}_1} z(x,y_i)\frac{1}{K^2}\sigma(\frac{\epsilon_0}{\tau})\sigma(-\frac{\epsilon_0}{\tau})z(x,y_j)| + \frac{2}{\tau^2}|\sum_{(y_i,y_j)\in\mathbb{Y}_2} z(x,y_i)\frac{1}{K^2}z(x,y_j)| \\
\leq &\frac{2}{\tau^2}\sigma(\frac{\epsilon_0}{\tau})\sigma(-\frac{\epsilon_0}{\tau})\sqrt{\sum_{(y_i,y_j)\in\mathbb{Y}_1} z^2(x,y_i)\frac{1}{K^2}}\sqrt{\sum_{(y_i,y_j)\in\mathbb{Y}_1}\frac{1}{K^2}z^2(x,y_j)} + \frac{2}{\tau^2}|\sum_{(y_i,y_j)\in\mathbb{Y}_2} z(x,y_i)\frac{1}{K^2}z(x,y_j)| \\
\leq &\frac{2}{\tau^2}\sigma(\frac{\epsilon_0}{\tau})\sigma(-\frac{\epsilon_0}{\tau})\gamma(x)\|z^2(x,\cdot)\|_\infty + \frac{2}{\tau^2}|\sum_{(y_i,y_j)\in\mathbb{Y}_2} z(x,y_i)\frac{1}{K^2}z(x,y_j)| \\
\leq &\frac{2}{\tau^2}\sigma(\frac{\epsilon_0}{\tau})\sigma(-\frac{\epsilon_0}{\tau})\gamma(x)\|z^2(x,\cdot)\|_\infty + \frac{2}{\tau^2}\sqrt{\sum_{(y_i,y_j)\in\mathbb{Y}_2} z^2(x,y_i)\frac{1}{K^2}}\sqrt{\sum_{(y_i,y_j)\in\mathbb{Y}_2}\frac{1}{K^2}z^2(x,y_j)} \\
\leq &\frac{2}{\tau^2}\sigma(\frac{\epsilon_0}{\tau})\sigma(-\frac{\epsilon_0}{\tau})\gamma(x)\|z^2(x,\cdot)\|_\infty + \frac{2}{\tau^2}(1-\gamma(x))\|z^2(x,\cdot)\|_\infty \leq \frac{2}{\tau^2}(\gamma(x)c_0+1)\|z(x,\cdot)\|_2^2.
\end{aligned} \tag{73}$$

Because $c_0 = \sigma(\frac{\epsilon_0}{\tau})\sigma(-\frac{\epsilon_0}{\tau}) - 1 \in (-1,0)$.

Therefore,

$$\|\psi(\cdot)\|_\infty \leq \max_{x\in\mathbb{X}}\frac{4(\gamma(x)c_0+1)}{\tau^2}\|z(x,\cdot)\|_2^2 \leq \frac{4(\gamma c_0+1)}{\tau^2}\|z(\cdot,\cdot)\|_2^2. \tag{74}$$

Then $\mathcal{L}_{\text{DPO}}(\pi_\theta;\pi_{\text{ref}})$ is $\frac{4(\gamma c_0+1)}{\tau^2}$-smooth is proved. Proof finished.

G.4 Proof of Theorem 5.3

Theorem 5.3: Define joint conditional probability distribution $\pi_1(y_1,y_2|x) = \frac{\mu}{K^2}$ if $\mathbb{I}(\Omega_1\cap\Omega_2) = 1$; $\frac{1-\mu\gamma}{(1-\gamma)K^2}$else where $\mu\in(0,1)$. Assume $\|g_t\|^2 \leq G^2$. Given the Definition D.1 for policy π_θ and the learning rate α_t, for DPO algorithm, let $\gamma(x)\leq\gamma$, then:

$$\min_{1\leq i\leq T}||\nabla_\theta\mathcal{L}_{\text{DPO}}(\pi_{\theta_i};\pi_{\text{ref}})||_2^2 \leq \frac{\mathcal{L}_{\text{DPO}}(\pi_{\theta_T})-\mathcal{L}^*_{\text{DPO}}}{\sum_{t=1}^{T-1}\alpha_t} + \frac{2(\mu\gamma c_0+1)G^2\sum_{t=1}^{T-1}\alpha_t^2}{\tau^2\sum_{t=1}^{T-1}\alpha_t}. \tag{75}$$

where $\mathcal{L}^*_{\text{DPO}} = \min_{\pi_\theta}\mathcal{L}_{\text{DPO}}(\pi_\theta;\pi_{\text{ref}},\pi_1)$ and $c_0 = \sigma(\frac{\epsilon_0}{\tau})\sigma(-\frac{\epsilon_0}{\tau}) - 1 \in (-1,0)$. See $\mathcal{L}_{\text{DPO}}(\pi_\theta;\pi_{\text{ref}})$ on Eq.6.

About $\Omega_1(y_1,y_2,x)$: As the optimization progresses, π_θ gradually approaches $\bar{\pi}^\tau$, causing the proportion of the event $\Omega_1\cap\Omega_2$ (denoted by γ) to increase (See Proposition G.2), thereby reducing the upper bound in inequality 24. This leads to an increasingly faster convergence rate for DPO. Conversely, if small-margin pairs are selected as event Ω_1 (i.e. $\Omega_1(y_1,y_2,x) = \{|\log\frac{p^*(1|y_1,y_2,x)}{p^*(0|y_1,y_2,x)}| \leq \epsilon_0\}$), the convergence rate of DPO will slow down as the optimization advances and intensify the distribution shift problem cause Ω_1 and Ω_2 will more and more insistent as the optimization advances.

Proof: Based on Theorem G.1, we only need to know $\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}})$ is $\frac{4(\mu\gamma c_0+1)}{\tau^2}$-smooth. This corollary shares the same proof process as Lemma 5.2. The only two differences are Eq.72 and Eq.73. Their $\gamma(x)$ will be changed into $\mu\gamma(x)$. Then we will get the smooth coefficient of $\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}})$ is $\frac{4(\mu\gamma c_0+1)}{\tau^2}$. Proof finished.

G.5 Proposition G.2

Proposition G.2.

$$\lim_{\epsilon_0 \to 0} P(|\log \frac{p^*(z=1)}{p^*(z=0)}| \geq \epsilon_0) - P(|\log \frac{\bar{\pi}^\tau(y_1|x)\pi_{ref}(y_2|x)}{\bar{\pi}^\tau(y_2|x)\pi_{ref}(y_1|x)}| \geq \epsilon_0) = 0. \tag{76}$$

Proof: comparison probability $p^*(1 \mid y_1, y_2, x) \propto r(x, y_1) - r(x, y_2)$, where it is identified that $r(x, y_1) > r(x, y_2)$, denoted as $y_1 \succ y_2$; similarly, (x, y_2, y_1) is sampled with probability $p^*(0 \mid y_1, y_2, x) \propto r(x, y_2) - r(x, y_1)$, and is identified as $r(x, y_1) < r(x, y_2)$, denoted as $y_2 \succ y_1$. $\bar{\pi}^\tau(y|x) = \frac{\pi_{ref}(y|x)\exp(\tau r(x,y))}{Z'(x)}$. Then

$$\begin{aligned} \{|\log \frac{p^*(z=1)}{p^*(z=0)}| \geq \epsilon_0\} &\xrightarrow{\epsilon_0 \to 0} \{r(x, y_1) \geq r(x, y_2)\}, \\ \{|\log \frac{\bar{\pi}^\tau(y_1|x)\pi_{\text{ref}}(y_2|x)}{\bar{\pi}^\tau(y_2|x)\pi_{\text{ref}}(y_1|x)}| \geq \epsilon_0\} &\xrightarrow{\epsilon_0 \to 0} \{r(x, y_1) \geq r(x, y_2)\}. \end{aligned} \tag{77}$$

Proof finished.

As the optimization progresses, π_θ gradually approaches $\bar{\pi}^\tau$, causing $\Omega_1(y_1, y_2, x)$ close to $\Omega_2(y_1, y_2, x)$, then the proportion of the event $\Omega_1 \cap \Omega_2$ (denoted by γ) increases. Conversely, if small-margin pairs are selected as event Ω_1 (i.e. $\Omega_1(y_1, y_2, x) = \{|\log \frac{p^*(1|y_1,y_2,x)}{p^*(0|y_1,y_2,x)}| \leq \epsilon_0\}$), the convergence rate of DPO will slow down as the optimization advances and intensify the distribution shift problem cause Ω_1 and Ω_2 will more and more insistent as the optimization advances.

G.6 Sub Lemma for Lemma G.8-G.12

Lemma G.3. *By the properties of the softmax derivative, we have:*

$$\frac{\partial \pi_\theta(y|x)}{\partial \theta(x, y')} = \pi_\theta(y|x)(\delta_{yy'} - \pi_\theta(y'|x)), \text{ where } \delta_{yy'} = \begin{cases} 1, & \text{if } y = y' \\ 0, & \text{otherwise} \end{cases}. \tag{78}$$

Lemma G.4. *$\forall x$, $x \to f(x)$ is L-smooth is equivalent to the following property:*

$$\left| \sum_{i,j=1}^{|\mathbb{X}|} x_i \frac{\partial^2 f(x)}{\partial x_i \partial x_j} x_j \right| \leq L||\vec{x}||_2^2. \tag{79}$$

By Taylor's theorem, it suffices to show that the spectral radius of the hessian matrix of the second derivative of $f(x)$ is bounded by L.

Lemma G.5. *$f(x, \theta) = \sum_{y_1, y_2 \in \mathbb{Y}} \pi_\theta(y_1|x)\pi_\theta(y_2|x)h(\pi_\theta, x, y_1, y_2)$. Then the second derivative of $f(x, \theta)$ is:*

$$\begin{aligned} \frac{\partial^2 f(x,\theta)}{\partial\theta(x,y_i)\partial\theta(x,y_j)} = \sum_{y_1,y_2\in\mathbb{Y}} & 2\frac{\partial^2\pi_\theta(y_1|x)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}\pi_\theta(y_2|x)h(\pi_\theta,x,y_1,y_2) + 2\frac{\partial\pi_\theta(y_1|x)}{\partial\theta(x,y_i)}\frac{\partial\pi_\theta(y_2|x)}{\partial\theta(x,y_j)}h(\pi_\theta,x,y_1,y_2) \\ & + 2\frac{\partial\pi_\theta(y_1|x)}{\partial\theta(x,y_i)}\pi_\theta(y_2|x)\frac{\partial h(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_j)} + 2\frac{\partial\pi_\theta(y_1|x)}{\partial\theta(x,y_j)}\pi_\theta(y_2|x)\frac{\partial h(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_i)} \\ & + \pi_\theta(y_1|x)\pi_\theta(y_2|x)\frac{\partial^2 h(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}. \end{aligned} \tag{80}$$

Lemma G.6. *For $f(p) = -p\log p - (1-p)\log(1-p)$, $f(p) \leq \log 2$.*

Proof: For $f(p) = -p\log p - (1-p)\log(1-p) \geq 0, p \in [0, 1]$, $f'(p) = \log(\frac{1}{p} - 1)$ is a monotonically decreasing function that reaches 0 when $p = 0.5$. Thus $f(p)$ increases first and then decreases, and reaches its maximum value at $p = 0.5$. So $f(p) \leq \log 2$.

Lemma G.7. *$D_{\text{TV}}(p||q)^2 \leq D_{\text{KL}}(p||q)$.*

G.7 PROOF OF LEMMA G.8

Lemma G.8. *(BDA(Reverse-KL) Smoothness) Given softmax parametrization of Definition D.1 for policy π_θ, $\forall r, \tau$, $\theta \to \mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{D}_{\mathrm{KL}}\left(\pi^\tau\left(\cdot \mid x\right) \| \pi_\theta\left(\cdot \mid x\right)\right)\right]$ is 2-smooth (see Eq.8).*

See proof in Appendix G.7.

Proof: By Lemma G.4, it suffices to show that the spectral radius of the hessian matrix of the second derivative of $\mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{D}_{\mathrm{KL}}\left(\pi^\tau\left(\cdot \mid x\right) \| \pi_\theta\left(\cdot \mid x\right)\right)\right]$ is bounded by 2, i.e.

$$\left|\sum_{x,x'\in\mathbb{X}}\sum_{y,y'\in\mathbb{Y}} z(x,y)\frac{\partial^2\mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{D}_{\mathrm{KL}}\left(\pi^\tau\left(\cdot \mid x\right) \| \pi_\theta\left(\cdot \mid x\right)\right)\right]}{\partial\theta(x,y)\partial\theta(x',y')}z(x',y')\right| \leq 2||z(\cdot,\cdot)||_2^2. \tag{81}$$

By Lemma G.3, the first derivative of $\mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{D}_{\mathrm{KL}}\left(\pi^\tau\left(\cdot \mid x\right) \| \pi_\theta\left(\cdot \mid x\right)\right)\right]$ is:

$$\begin{aligned}
&\frac{\partial\mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{D}_{\mathrm{KL}}\left(\pi^\tau\left(\cdot \mid x\right) \| \pi_\theta\left(\cdot \mid x\right)\right)\right]}{\partial\theta(x,y)}\\
=&-\frac{\partial\mathbb{E}_{x\sim\mathcal{D}}[\sum_{y'\in Y}\frac{\exp(\tau r(x,y'))}{Z(x)}(\log\pi_\theta(y'|x))]}{\partial\theta(x,y)}\\
=&-\mathcal{D}(x)\sum_{y'\in Y}\frac{\exp(\tau r(x,y'))}{Z(x)}\frac{\partial\log\pi_\theta(y'|x))}{\partial\theta(x,y)}\\
=&-\mathcal{D}(x)\sum_{y'\in Y}\frac{\exp(\tau r(x,y'))}{Z(x)}(\delta_{y'y}-\pi_\theta(y|x))\\
=&\mathcal{D}(x)\pi_\theta(y|x)-\mathcal{D}(x)\frac{\exp(\tau r(x,y))}{Z(x)}.
\end{aligned} \tag{82}$$

The second derivative of $\mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{D}_{\mathrm{KL}}\left(\pi^\tau\left(\cdot \mid x\right) \| \pi_\theta\left(\cdot \mid x\right)\right)\right]$ is:

$$\begin{aligned}
&\frac{\partial^2\mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{D}_{\mathrm{KL}}\left(\pi^\tau\left(\cdot \mid x\right) \| \pi_\theta\left(\cdot \mid x\right)\right)\right]}{\partial\theta(x,y)\partial\theta(x',y')}\\
=&\frac{\partial}{\partial\theta(x',y')}(\frac{\partial\mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{D}_{\mathrm{KL}}\left(\pi^\tau\left(\cdot \mid x\right) \| \pi_\theta\left(\cdot \mid x\right)\right)\right]}{\partial\theta(x,y)})\\
=&\frac{\partial\mathcal{D}(x)\pi_\theta(y|x)}{\partial\theta(x',y')}=\delta_{xx'}\mathcal{D}(x)\pi_\theta(y|x)(\delta_{yy'}-\pi_\theta(y'|x)).
\end{aligned} \tag{83}$$

Then the spectral radius of $\mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{D}_{\mathrm{KL}}\left(\pi^\tau\left(\cdot \mid x\right) \| \pi_\theta\left(\cdot \mid x\right)\right)\right]$ is:

$$\begin{aligned}
&\left|\sum_{x,x'\in\mathbb{X}}\sum_{y,y'\in\mathbb{Y}} z(x,y)\frac{\partial^2\mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{D}_{\mathrm{KL}}\left(\pi^\tau\left(\cdot \mid x\right) \| \pi_\theta\left(\cdot \mid x\right)\right)\right]}{\partial\theta(x,y)\partial\theta(x',y')}z(x',y')\right|\\
=&\left|\sum_{x\in\mathbb{X}}\sum_{y,y'\in\mathbb{Y}} z(x,y)\delta_{xx'}\mathcal{D}(x)\pi_\theta(y|x)(\delta_{yy'}-\pi_\theta(y'|x))z(x,y')\right|\\
=&\left|\sum_{x\in\mathbb{X}}\mathcal{D}(x)\sum_{y,y'\in\mathbb{Y}} z(x,y)\pi_\theta(y|x)(\delta_{yy'}-\pi_\theta(y'|x))z(x,y')\right|\\
\triangleq&|\sum_{x\in\mathbb{X}}\mathcal{D}(x)\psi(x)|\leq\|\mathcal{D}(\cdot)\|_1\|\psi(\cdot)\|_\infty=1\cdot\|\psi(\cdot)\|_\infty.
\end{aligned} \tag{84}$$

Then we have:

$$
\begin{aligned}
&\|\psi(\cdot)\|_\infty = \max_{x\in\mathbb{X}} \left| \sum_{y,y'\in\mathbb{Y}} z(x,y)\pi_\theta(y|x)(\delta_{yy'} - \pi_\theta(y'|x))z(x,y') \right| \\
&= \max_{x\in\mathbb{X}} \left| \sum_{y,y'\in\mathbb{Y}} z(x,y)\pi_\theta(y|x)\delta_{yy'}z(x,y') - z(x,y)\pi_\theta(y|x)\pi_\theta(y'|x)z(x,y') \right| \\
&\leq \max_{x\in\mathbb{X}} \left| \sum_{y,y'\in\mathbb{Y}} z(x,y)\pi_\theta(y|x)\delta_{yy'}z(x,y') \right| + \max_{x\in\mathbb{X}} \left| \sum_{y,y'\in\mathbb{Y}} z(x,y)\pi_\theta(y|x)\pi_\theta(y'|x)z(x,y') \right| \\
&= \max_{x\in\mathbb{X}} \left| \sum_{y,y'\in\mathbb{Y}} z(x,y)\pi_\theta(y|x)z(x,y) \right| + \max_{x\in\mathbb{X}} (\sum_{y\in\mathbb{Y}} z(x,y)\pi_\theta(y|x))^2 \\
&\leq \max_{x\in\mathbb{X}} \sum_{y\in\mathbb{Y}} z(x,y)z(x,y) + \max_{x\in\mathbb{X}} (z(x,\cdot)^T\pi_\theta(\cdot|x))^2 \\
&\leq \max_{x\in\mathbb{X}} ||z(x,\cdot)||_2^2 + \max_{x\in\mathbb{X}} (||\pi_\theta(\cdot|x)||_1 ||z(x,\cdot)||_\infty)^2 \\
&\leq 2\max_{x\in\mathbb{X}} ||z(x,\cdot)||_2^2 = 2\|||z(x,\cdot)||_2^2\|_\infty \leq 2||z(\cdot,\cdot)||_2^2.
\end{aligned}
\tag{85}
$$

where $\|||z(x,\cdot)||_2^2\|_\infty$ is taking the infinite norm of $||z(x,\cdot)||_2^2$ with x as the coordinate axis.

Therefore, we have the 2-smooth conclusion:

$$
\left| \sum_{x,x'\in\mathbb{X}} \sum_{y,y'\in\mathbb{Y}} z(x,y) \frac{\partial^2 \mathbb{E}_{x\sim\mathcal{D}} \left[\mathrm{D}_{\mathrm{KL}}\left(\pi^\tau(\cdot \mid x) \| \pi_\theta(\cdot \mid x)\right)\right]}{\partial\theta(x,y)\partial\theta(x',y')} z(x',y') \right| \leq 2||z(\cdot,\cdot)||_2^2. \tag{86}
$$

Proof finished.

G.8 Proof of Lemma G.9

Lemma G.9. *(BDA(Forward-KL) Smoothness) Given softmax parametrization of Definition D.1 for policy π_θ, assume $|\log(\pi_\theta(y|x)) - \log(\pi^\tau(y|x))| \leq \epsilon_1$, $\forall r, \tau$, $\theta \to \mathbb{E}_{x\sim\mathcal{D}} \left[\mathrm{D}_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid x) \| \pi^\tau(\cdot \mid x)\right)\right]$ is $((4+K)\epsilon_1 + 6 + 2K)$-smooth (see Eq.7).*

See proof in Appendix G.8.

Proof: By Lemma G.4, it suffices to show that the spectral radius of the hessian matrix of the second derivative of $\mathbb{E}_{x\sim\mathcal{D}} \left[\mathrm{D}_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid x) \| \pi^\tau(\cdot \mid x)\right)\right]$ is bounded by $(6\epsilon_1 + 10)$, i.e.

$$
\left| \sum_{x,x'\in\mathbb{X}} \sum_{y,y'\in\mathbb{Y}} z(x,y) \frac{\partial^2 \mathbb{E}_{x\sim\mathcal{D}} \left[\mathrm{D}_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid x) \| \pi^\tau(\cdot \mid x)\right)\right]}{\partial\theta(x,y)\partial\theta(x',y')} z(x',y') \right| \leq (6\epsilon_1 + 10)\, ||z(\cdot,\cdot)||_2^2. \tag{87}
$$

For Equation 7, we have:

$$
\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{D}} \left[\mathrm{D}_{\mathrm{KL}}\left(\pi_\theta(\cdot \mid x) \| \pi^\tau(\cdot \mid x)\right)\right] \\
&= -\mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)} \left[(\log(\pi^\tau(y|x) - \log(\pi_\theta(y|x)))\right] \\
&= \sum_{x\in\mathbb{X}} \mathcal{D}(x)\pi_\theta(\cdot|x)^T h(\pi_\theta, x) \quad \text{where } h(\pi_\theta, x, y) = (-\log(\pi^\tau(y|x) + \log(\pi_\theta(y|x))),\ h(\pi_\theta, x) = h(\pi_\theta, x, \cdot).
\end{aligned}
\tag{88}
$$

Denote $f_{\text{fKL}}(x,\theta) = \pi_\theta(\cdot|x)^T h(\pi_\theta, x)$. We can calculate the spectral radius of $\frac{\partial^2 \mathbb{E}_{x\sim\mathcal{D}}[\mathrm{D}_{\mathrm{KL}}(\pi_\theta(\cdot|x)\|\pi^\tau(\cdot|x))]}{\partial\theta(x,y)\partial\theta(x',y')}$:

$$\begin{aligned}
&\left|\sum_{x,x'\in\mathbb{X}}\sum_{y,y'\in\mathbb{Y}} z(x,y)\frac{\partial^2 \mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{D}_{\mathrm{KL}}\left(\pi_\theta\left(\cdot\mid x\right)\|\pi^\tau\left(\cdot\mid x\right)\right)\right]}{\partial\theta(x,y)\partial\theta(x',y')} z(x',y')\right| \\
=&\left|\sum_{x\in\mathbb{X}}\sum_{y,y'\in\mathbb{Y}} z(x,y)\frac{\partial^2 \sum_{x\in\mathbb{X}}\mathcal{D}(x)\pi_\theta(\cdot|x)^T h(\pi_\theta,x)}{\partial\theta(x,y)\partial\theta(x,y')} z(x,y')\right| \\
=&\left|\sum_{x\in\mathbb{X}}\sum_{y,y'\in\mathbb{Y}} z(x,y)\mathcal{D}(x)\frac{\partial^2 f_{\text{fKL}}(x,\theta)}{\partial\theta(x,y)\partial\theta(x,y')} z(x,y')\right|.
\end{aligned} \tag{89}$$

The first derivative of $f_{\text{fKL}}(x,\theta)$ is:

$$\frac{\partial f_{\text{fKL}}(x,\theta)}{\partial\theta(x,y_i)} = \sum_y \frac{\partial\pi_\theta(y|x)}{\partial\theta(x,y_i)} h(\pi_\theta,x,y) + \pi_\theta(y|x)\frac{\partial h(\pi_\theta,x,y)}{\partial\theta(x,y_i)}. \tag{90}$$

The second derivative of $f_{\text{fKL}}(x,\theta)$ is:

$$\begin{aligned}
\frac{\partial^2 f_{\text{fKL}}(x,\theta)}{\partial\theta(x,y_i)\partial\theta(x,y_j)} = &\sum_y \frac{\partial^2\pi_\theta(y|x)}{\partial\theta(x,y_i)\partial\theta(x,y_j)} h(\pi_\theta,x,y) + \frac{\partial\pi_\theta(y|x)}{\partial\theta(x,y_i)}\frac{\partial h(\pi_\theta,x,y)}{\partial\theta(x,y_j)} \\
&+ \frac{\partial\pi_\theta(y|x)}{\partial\theta(x,y_j)}\frac{\partial h(\pi_\theta,x,y)}{\partial\theta(x,y_i)} + \pi_\theta(y|x)\frac{\partial^2 h(\pi_\theta,x,y)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}.
\end{aligned} \tag{91}$$

By Lemma G.3, we have:

$$\frac{\partial\pi_\theta(y|x)}{\partial\theta(x,y_i)} = \pi_\theta(y|x)(\delta_{yy_i} - \pi_\theta(y_i|x)). \tag{92}$$

$$\frac{\partial\left(\pi_\theta(y|x)(\delta_{yy_i} - \pi_\theta(y_i|x))\right)}{\partial\theta(x,y_j)} = \pi_\theta(y|x)(\delta_{yy_j} - \pi_\theta(y_j|x))(\delta_{yy_i} - \pi_\theta(y_i|x)) - \pi_\theta(y|x)\pi_\theta(y_i|x)(\delta_{y_iy_j} - \pi_\theta(y_j|x)). \tag{93}$$

And:

$$\frac{\partial h(\pi_\theta,x,y)}{\partial\theta(x,y_j)} = \frac{\partial(-\log(\pi^\tau(y|x) + \log(\pi_\theta(y|x)))}{\partial\theta(x,y_j)} = (\delta_{yy_j} - \pi_\theta(y_j|x)). \tag{94}$$

$$\frac{\partial^2 h(\pi_\theta,x,y)}{\partial\theta(x,y_i)\partial\theta(x,y_j)} = \frac{\partial}{\partial\theta(x,y_j)}\left(\frac{\partial h(\pi_\theta,x,y)}{\partial\theta(x,y_i)}\right) = \frac{\partial(\delta_{yy_i} - \pi_\theta(y_i|x))}{\partial\theta(x,y_j)} = -\pi_\theta(y_i|x)(\delta_{y_iy_j} - \pi_\theta(y_j|x)). \tag{95}$$

Thus the second derivative of $f_{\text{fKL}}(x,\theta)$ is changed into:

$$\begin{aligned}
&\frac{\partial^2 f_{\text{fKL}}(x,\theta)}{\partial\theta(x,y_i)\partial\theta(x,y_j)} \\
=&\sum_y \pi_\theta(y|x)(\delta_{yy_j} - \pi_\theta(y_j|x))(\delta_{yy_i} - \pi_\theta(y_i|x))(h(\pi_\theta,x,y) + 2) \\
&- \pi_\theta(y|x)\pi_\theta(y_i|x)(\delta_{y_iy_j} - \pi_\theta(y_j|x))(h(\pi_\theta,x,y) + 1).
\end{aligned} \tag{96}$$

Let $\vec{h}(\theta,x) = (h(\pi_\theta,x,y_1), h(\pi_\theta,x,y_2),\ldots,h(\pi_\theta,x,y_{|\mathbb{Y}|}))$, $\vec{H}_1(\theta,x) = -\vec{h}(\theta,x) - 1$, and $\vec{H}_2(\theta,x) = \vec{h}(\theta,x) + 2$. So $H_1(\pi_\theta,x,y) = -h(\pi_\theta,x,y) - 1$ and $H_2(\pi_\theta,x,y)$ is similar. Then, the Hessian matrix of $f_{\text{fKL}}(x,\theta) = \pi_\theta(\cdot|x)^T h(\pi_\theta,x)$ simplifies as follows:

$$\begin{aligned}
\nabla^2_\theta f_{\text{fKL}}(x,\theta)_{i,j} = \sum_{y\in\mathbb{Y}}[&\pi_\theta(y|x)\pi_\theta(y_i|x)(\delta_{y_iy_j} - \pi_\theta(y_j|x))H_1(\pi_\theta,x,y) \\
&+ \pi_\theta(y|x)(\delta_{yy_j} - \pi_\theta(y_j|x))(\delta_{yy_i} - \pi_\theta(y_i|x))H_2(\pi_\theta,x,y)].
\end{aligned} \tag{97}$$

Then we compute the quadratic form of the Hessian matrix:

$$
\begin{aligned}
&|\sum_{x\in\mathbb{X}}\sum_{y,y'\in\mathbb{Y}} z(x,y)\mathcal{D}(x)\frac{\partial^2 f_{\text{fKL}}(x,\theta)}{\partial\theta(x,y)\partial\theta(x,y')}z(x,y')| \\
=&|\sum_{x\in\mathbb{X}}\mathcal{D}(x)\sum_{i=1}^{|\mathbb{Y}|}\sum_{j=1}^{|\mathbb{Y}|} z(x,y_i)\left(\nabla_\theta^2 f_{\text{fKL}}(x,\theta)_{i,j}\right)z(x,y_j)| \\
=&|\sum_{x\in\mathbb{X}}\mathcal{D}(x)(\pi_\theta(\cdot|x)^\top\vec{H}_1(\theta,x)(\pi_\theta(\cdot|x)^\top z(x,\cdot))^2+\pi_\theta(\cdot|x)^\top(\vec{H}_2(\theta,x)-\vec{H}_1(\theta,x))(\pi_\theta(\cdot|x)^\top z(x,\cdot))^2 \\
&\quad+\sum_{i=1}^{K} z(x,y_i)^2\pi_\theta(y_i|x)H_2(\pi_\theta,x,y_i)-2\pi_\theta(\cdot|x)^\top z(x,\cdot)\pi_\theta(y_i|x)H_2(\pi_\theta,x,y_i)z(x,y_i))| \\
=&|\sum_{x\in\mathbb{X}}\mathcal{D}(x)(\pi_\theta(\cdot|x)^\top\vec{H}_1(\theta,x)(\pi_\theta(\cdot|x)^\top z(x,\cdot))^2+\pi_\theta(\cdot|x)^\top(\vec{H}_2(\theta,x)-\vec{H}_1(\theta,x))(\pi_\theta(\cdot|x)^\top z(x,\cdot))^2 \\
&\quad+\sum_{i=1}^{K}(z(x,y_i)-2\pi_\theta(\cdot|x)^\top z(x,\cdot))\pi_\theta(y_i|x)H_2(\pi_\theta,x,y_i)z(x,y_i))| \\
\triangleq&|\sum_{x\in\mathbb{X}}\mathcal{D}(x)\psi(x)|\leq\|\mathcal{D}(\cdot)\|_1\|\psi(\cdot)\|_\infty=1\cdot\|\psi(\cdot)\|_\infty.
\end{aligned}
\tag{98}
$$

To establish an upper bound for the spectral radius of the Hessian matrix, let $\vec{H}_3(\theta,x)=\vec{H}_2(\theta,x)-\vec{H}_1(\theta,x)$. We can then write:

$$
\begin{aligned}
&\|\psi(\cdot)\|_\infty \\
=&\max_{x\in\mathbb{X}}\Big|\pi_\theta(\cdot|x)^\top\vec{H}_1(\theta,x)\cdot(\pi_\theta(\cdot|x)^\top z(x,\cdot))^2+\pi_\theta(\cdot|x)^\top\vec{H}_3(\theta,x)\cdot(\pi_\theta(\cdot|x)^\top z(x,\cdot))^2 \\
&\quad+\sum_{i=1}^{K} z(x,y_i)^2\pi_\theta(y_i|x)H_2(\pi_\theta,x,y_i)+\sum_{i=1}^{K}2\pi_\theta(\cdot|x)^\top z(x,\cdot)\pi_\theta(y_i|x)H_2(\pi_\theta,x,y_i)z(x,y_i)\Big| \\
\leq&\max_{x\in\mathbb{X}}\left|\pi_\theta(\cdot|x)^\top\vec{H}_1(\theta,x)\cdot(\pi_\theta(\cdot|x)^\top z(x,\cdot))^2\right|+\left|\pi_\theta(\cdot|x)^\top\vec{H}_3(\theta,x)\cdot(\pi_\theta(\cdot|x)^\top z(x,\cdot))^2\right| \\
&\quad+\left|\sum_{i=1}^{K} z(x,y_i)^2\pi_\theta(y_i|x)H_2(\pi_\theta,x,y_i)\right|+\left|\sum_{i=1}^{K}2\pi_\theta(\cdot|x)^\top z(x,\cdot)\pi_\theta(y_i|x)H_2(\pi_\theta,x,y_i)z(x,y_i)\right|,
\end{aligned}
\tag{99}
$$

Continuing from the previous equation:

$$
\begin{aligned}
&\|\psi(\cdot)\|_\infty \\
\leq&\max_{x\in\mathbb{X}}\|\pi_\theta(\cdot|x)\|_1\|\vec{H}_1(\theta,x)\|_\infty\|z(x,\cdot)\|_2^2+\|\pi_\theta(\cdot|x)\|_1\|\vec{H}_3(\theta,x)\|_\infty\|z(x,\cdot)\|_2^2 \\
&\quad+\|\vec{H}_2(\theta,x)\|_\infty\|z(x,\cdot)\|_2^2+2\|\pi_\theta(\cdot|x)\odot\vec{H}_2(\theta,x)\|_1\|z(x,\cdot)\|_2^2 \\
=&\max_{x\in\mathbb{X}}(\|\vec{H}_1(\theta,x)\|_\infty+\|\vec{H}_2(\theta,x)\|_\infty+\|\vec{H}_3(\theta,x)\|_\infty)\|z(x,\cdot)\|_2^2+2\|\pi_\theta(\cdot|x)\odot\vec{H}_2(\theta,x)\|_1\|z(x,\cdot)\|_2^2 \\
\leq&\max_{x\in\mathbb{X}}(\|\vec{H}_1(\theta,x)\|_\infty+\|\vec{H}_2(\theta,x)\|_\infty+\|\vec{H}_3(\theta,x)\|_\infty)\|z(x,\cdot)\|_2^2+2\|\vec{H}_2(\theta,x)\|_\infty\|z(x,\cdot)\|_2^2 \\
=&\max_{x\in\mathbb{X}}(\|\vec{H}_1(\theta,x)\|_\infty+3\|\vec{H}_2(\theta,x)\|_\infty+\|\vec{H}_3(\theta,x)\|_\infty)\|z(x,\cdot)\|_2^2 \\
\leq&\max_{x\in\mathbb{X}}(\|\vec{H}_1(\theta,x)\|_\infty+3\|\vec{H}_2(\theta,x)\|_\infty+\|\vec{H}_3(\theta,x)\|_\infty)\max_{x\in\mathbb{X}}\|z(x,\cdot)\|_2^2 \\
\leq&\max_{x\in\mathbb{X}}(\|\vec{H}_1(\theta,x)\|_\infty+3\|\vec{H}_2(\theta,x)\|_\infty+\|\vec{H}_3(\theta,x)\|_\infty)\|z(\cdot,\cdot)\|_2^2.
\end{aligned}
\tag{100}
$$

In the second inequality, the first term arises from Hölder's inequality, which states that $\pi_\theta^\top z(x,\cdot)\leq\|\pi_\theta\|_1\|z(x,\cdot)\|_\infty=\|z(x,\cdot)\|_\infty$, where $\|\pi_\theta\|_1=1$ and $\|z^2(x,\cdot)\|_1=\|z(x,\cdot)\|_2^2$. For the second term, we have the bound on $|(\pi_\theta^\top z(x,\cdot))^2|$:

$$
|(\pi_\theta^\top z(x,\cdot))^2|=\pi_\theta^\top z(x,\cdot)\cdot\pi_\theta^\top z(x,\cdot)\leq\|\pi_\theta\|_1\|z(x,\cdot)\|_\infty\cdot\|\pi_\theta\|_1\|z(x,\cdot)\|_\infty\leq\|z(x,\cdot)\|_2^2. \tag{101}
$$

For the third term:

$$\left|\sum_{i=1}^{K} z(x,y_i)^2 \pi_\theta(y_i|x) H_2(\pi_\theta,x,y_i)\right| \leq \left|\sum_{i=1}^{K} z(x,y_i)^2 |H_2(\pi_\theta,x,y_i)|\right| \leq \|\vec{H}_2(\theta,x)\|_\infty \|z(x,\cdot)\|_2^2. \tag{102}$$

For the fourth term:

$$\begin{aligned}
&|\sum_{i=1}^{K} 2\pi_\theta^T z(x,\cdot)\pi_\theta(y_i|x) H_2(\pi_\theta,x,y_i) z(x,y_i)| \\
\leq& 2||\pi_\theta||_1 ||z(x,\cdot)||_\infty |\sum_{i=1}^{K} \pi_\theta(y_i|x) H_2(\pi_\theta,x,y_i) z(x,y_i)|| \\
\leq& 2||z(x,\cdot)||_2 ||\pi_\theta(\cdot|x) \odot \vec{H}_2(\theta,x)||_1 ||z(x,\cdot)||_\infty \\
\leq& 2||\vec{H}_2(\theta,x)||_\infty ||z(x,\cdot)||_2^2.
\end{aligned} \tag{103}$$

In this Lemma, we assume $|h(\pi_\theta,x,y)| = |\log(\pi_\theta(y|x)) - \log(\pi^\tau(y|x))| \leq \epsilon_1$ where ϵ_1 denotes the upper bound of the error of the current parameterized distribution π_θ.

For $||\vec{H}_1(\theta,x)||_\infty$, we have:

$$||\vec{H}_1(\theta,x)||_\infty = ||-\vec{h}(\theta,x) - 1||_\infty \leq \epsilon_1 + 1. \tag{104}$$

For $||\vec{H}_2(\theta,x)||_\infty$, it follows that:

$$||\vec{H}_2(\theta,x)||_\infty = ||\vec{h}(\theta,x) + 2||_\infty \leq \epsilon_1 + 2. \tag{105}$$

For $||\vec{H}_3(\theta,x)||_\infty$, we find:

$$||\vec{H}_3(\theta,x)||_\infty = ||\vec{H}_2(\theta,x) - \vec{H}_1(\theta)||_\infty = ||\vec{h}(\theta,x) + 2 + \vec{h}(\theta,x) + 1||_\infty \leq 2\epsilon_1 + 3. \tag{106}$$

Combining these results yields:

$$||\vec{H}_1(\theta)||_\infty + ||\vec{H}_2(\theta)||_\infty + ||\vec{H}_3(\theta)||_\infty \leq 6\epsilon_1 + 10. \tag{107}$$

Thus,

$$\left|\sum_{x,x'\in\mathbb{X}} \sum_{y,y'\in\mathbb{Y}} z(x,y) \frac{\partial^2 \mathbb{E}_{x\sim\mathcal{D}}\left[\mathrm{D}_{\mathrm{KL}}\left(\pi_\theta\left(\cdot \mid x\right) \| \pi^\tau\left(\cdot \mid x\right)\right)\right]}{\partial\theta(x,y)\partial\theta(x',y')} z(x',y')\right| \leq (6\epsilon_1 + 10)\, ||z(\cdot,\cdot)||_2^2. \tag{108}$$

Proof finished.

G.9 Proof of Lemma G.10

Lemma G.10. *(RA Smoothness) Given softmax parametrization of Definition D.1 for policy π_θ, assume $|r_\theta(x,y) - r(x,y)| \leq \epsilon_2$, $\forall r, \tau$, $\theta \to \mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}\left[(r_\theta(x,y) - r(x,y))^2\right]$ is $(3\epsilon_1^2 + \frac{18\epsilon_1}{\tau} + \frac{8}{\tau^2} + \max\left\{\epsilon_1^2 + \frac{2}{\tau}\epsilon_1, \frac{1}{\tau}\right\})$-smooth (see Eq.10). See proof in Appendix G.9.*

Proof: Denote $\mathcal{L}_{RA}(\pi_\theta) = \mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}\left[(r_\theta(x,y) - r(x,y))^2\right]$ Let $S \triangleq S(\theta) \in \mathbb{R}^{K\times K}$ be the second derivative of the value map $\theta \to \mathcal{L}_{RA}(\pi_\theta)$, i.e. $S(\theta,x_1,y_1,x_2,y_2) = \frac{\partial^2 \mathcal{L}_{RA}(\pi_\theta)}{\partial\theta(x_1,y_1)\partial\theta(x_2,y_2)}$. Denote $L_{\mathrm{RA}} = (3\epsilon_1^2 + \frac{18\epsilon_1}{\tau} + \frac{8}{\tau^2} + \max\left\{\epsilon_1^2 + \frac{2}{\tau}\epsilon_1, \frac{1}{\tau}\right\})$. By Lemma G.4, it suffices to show that the spectral radius of S is bounded by L_{RA}. Because $r_\theta(x,y) = \frac{1}{\tau}\log(Z(x)\pi_\theta(y|x))$, denote

$z(x,y) = \frac{1}{\tau}\log\sum_{y\in Y} e^{\tau r(x,y)} - r(x,y)$, then we have:

$$\begin{aligned}
&\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}\left[(r_\theta(x,y)-r(x,y))^2\right]\\
=&\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(y|x)}\left[\left(\frac{1}{\tau}\log\pi_\theta(y)+z(x,y)\right)^2\right]\\
=&\sum_{x\in\mathbb{X},y\in\mathbb{Y}}\mathcal{D}(x)\pi_\theta(y|x)\left(\frac{1}{\tau}\log\pi_\theta(y|x)+z(x,y)\right)^2\\
=&\sum_{x\in\mathbb{X}}\mathcal{D}(x)\pi_\theta(\cdot|x)^T h(\pi_\theta,x)\quad\text{where } h(\pi_\theta,x,y)=\left(\frac{1}{\tau}\log\pi_\theta(y|x)+z(x,y)\right)^2,\ h(\pi_\theta,x)=h(\pi_\theta,x,\cdot).
\end{aligned}\tag{109}$$

Now, by Definition D.1 we have:

$$\pi_\theta(y\mid x)=\frac{\exp\{\theta(x,y)\}}{\sum_{y'}\exp\{\theta(x,y')\}}.\tag{110}$$

By Lemma G.3, we have:

$$\frac{\partial\pi_\theta(y|x)}{\partial\theta(x,y')}=\pi_\theta(y|x)(\delta_{yy'}-\pi_\theta(y'|x)),\ \text{where } \delta_{yy'}=\begin{cases}1, & \text{if } y=y'\\ 0, & \text{otherwise}\end{cases}.\tag{111}$$

If $x\neq x'$, then the second derivative of $\pi_\theta(y|x)$ is 0:

$$\frac{\partial^2\pi_\theta(y|x)}{\partial\theta(x,y_1)\partial\theta(x',y_2)}=\frac{\partial\pi_\theta(y|x)(\delta_{yy_1}-\pi_\theta(y_1|x))}{\partial\theta(x',y_2)}=0.\tag{112}$$

Similarly, because $\pi_\theta(y|x)$ and $h(\pi_\theta,x)$ don't have relevant parameters to $\theta(x',\cdot)$, for $x\neq x'$, we have:

$$\frac{\partial^2\pi_\theta(\cdot|x)^T h(\pi_\theta,x)}{\partial\theta(x,y_1)\partial\theta(x',y_2)}=0.\tag{113}$$

Therefore if $x_1\neq x_2$,

$$\frac{\partial^2\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}\left[(r_\theta(x,y)-r(x,y))^2\right]}{\partial\theta(x_1,y_1)\partial\theta(x_2,y_2)}=\mathcal{D}(x_1)\frac{\partial}{\partial\theta(x_2,y_2)}\left(\frac{\partial\pi_\theta(\cdot|x_1)^T h(\pi_\theta,x_1)}{\partial\theta(x_1,y_1)}\right)=0.\tag{114}$$

So the only care about the term when $x_1=x_2$, i.e.

$$\frac{\partial^2\mathbb{E}_{x\sim\mathcal{D},y\sim\pi_\theta(\cdot|x)}\left[(r_\theta(x,y)-r(x,y))^2\right]}{\partial\theta(x,y_1)\partial\theta(x,y_2)}=\mathcal{D}(x)\frac{\partial}{\partial\theta(x,y_2)}\left(\frac{\partial\pi_\theta(\cdot|x)^T h(\pi_\theta,x)}{\partial\theta(x,y_1)}\right).\tag{115}$$

Denote $f_{RA}(x,\theta)=\pi_\theta(\cdot|x)^T h(\pi_\theta,x)=\sum_{y\in\mathbb{Y}}\pi_\theta(y|x)h(\pi_\theta,x,y)$ and $h(\pi_\theta,x,y)=g^2(\pi_\theta,x,y)$, $g(\pi_\theta,x,y)=\frac{1}{\tau}\log\pi_\theta(y|x)+z(x,y)=r_\theta(x,y)-r(x,y)$, the first derivative of $f_{RA}(x,\theta)$ is given by:

$$\frac{\partial f_{RA}(x,\theta)}{\partial\theta(x,y')}=\sum_{y\in\mathbb{Y}}\left[\frac{\partial\pi_\theta(y|x)}{\partial\theta(x,y')}g^2(\pi_\theta,x,y)+2\pi_\theta(y|x)g(\pi_\theta,x,y)\frac{\partial g(\pi_\theta,x,y)}{\partial\theta(x,y')}\right].\tag{116}$$

where, $\frac{\partial g(\pi_\theta,x,y)}{\partial\theta(x,y')}=\frac{1}{\tau}\frac{\partial\log\pi_\theta(y|x)}{\partial\theta(x,y')}=\frac{1}{\tau}(\delta_{yy'}-\pi_\theta(y'|x))$.

To compute the Hessian matrix, we take the derivative of this expression with respect to $\theta(y_2)$, yielding the (y_1,y_2)-th element as

$$\frac{\partial^2 f_{RA}(x,\theta)}{\partial\theta(x,y_1)\partial\theta(x,y_2)}=\frac{\partial}{\partial\theta(x,y_2)}\left[\sum_{y\in\mathbb{Y}}\left(\frac{\partial\pi_\theta(y|x)}{\partial\theta(x,y_1)}g^2(\pi_\theta,x,y)+2\pi_\theta(y|x)g(\pi_\theta,x,y)\frac{\partial g(\pi_\theta,x,y)}{\partial\theta(x,y_1)}\right)\right].\tag{117}$$

We proceed by decomposing this expression into two parts for further derivation. The first part of our derivation is given by:

$$\sum_{y\in\mathbb{Y}} \frac{\partial \pi_\theta(y|x)}{\partial \theta(x,y_1)} g^2(\pi_\theta,x,y) = \sum_{y\in\mathbb{Y}} \pi_\theta(y|x)(\delta_{yy_1} - \pi_\theta(y_1|x)) g^2(\pi_\theta,x,y). \tag{118}$$

Applying the chain rule, we first differentiate $\pi_\theta(y|x)(\delta_{yy_1} - \pi_\theta(y_1|x))$ with respect to $\theta(x,y_2)$:

$$\frac{\partial}{\partial \theta(x,y_2)} \left(\pi_\theta(y|x)(\delta_{yy_1} - \pi_\theta(y_1|x))\right) = \frac{\partial \pi_\theta(y|x)}{\partial \theta(x,y_2)}(\delta_{yy_1} - \pi_\theta(y_1|x)) - \pi_\theta(y|x)\frac{\partial \pi_\theta(y_1|x)}{\partial \theta(x,y_2)}. \tag{119}$$

Since $\frac{\partial \pi_\theta(y|x)}{\partial \theta(x,y_2)} = \pi_\theta(y|x)(\delta_{yy_2} - \pi_\theta(y_2|x))$ and $\frac{\partial \pi_\theta(y_1|x)}{\partial \theta(x,y_2)} = \pi_\theta(y_1|x)(\delta_{y_1y_2} - \pi_\theta(y_2|x))$, the derivative of the first part is

$$\frac{\partial \left(\pi_\theta(y|x)(\delta_{yy_1} - \pi_\theta(y_1|x))\right)}{\partial \theta(x,y_2)} = \pi_\theta(y|x)(\delta_{yy_2} - \pi_\theta(y_2|x))(\delta_{yy_1} - \pi_\theta(y_1|x)) - \pi_\theta(y|x)\pi_\theta(y_1|x)(\delta_{y_1y_2} - \pi_\theta(y_2|x)). \tag{120}$$

Next, we differentiate $g^2(\pi_\theta,x,y)$ with respect to $\theta(x,y_2)$. Noting that $\frac{\partial g(\pi_\theta,x,y)}{\partial \theta(x,y_2)} = \frac{1}{\tau}(\delta_{yy_2} - \pi_\theta(y_2|x))$, we obtain

$$\frac{\partial g^2(\pi_\theta,x,y)}{\partial \theta(x,y_2)} = 2g(\pi_\theta,x,y)\frac{\partial g(\pi_\theta,x,y)}{\partial \theta(x,y_2)} = \frac{2}{\tau} g(\pi_\theta,x,y)(\delta_{yy_2} - \pi_\theta(y_2|x)). \tag{121}$$

Combining all results, we obtain the derivative of the first part as follows:

$$\sum_{y\in\mathbb{Y}} \Big[\left(\pi_\theta(y|x)(\delta_{yy_2} - \pi_\theta(y_2|x))(\delta_{yy_1} - \pi_\theta(y_1|x)) - \pi_\theta(y|x)\pi_\theta(y_1|x)(\delta_{y_1y_2} - \pi_\theta(y_2|x))\right) g^2(\pi_\theta,x,y) \\ + \pi_\theta(y|x)(\delta_{yy_1} - \pi_\theta(y_1|x))\frac{2}{\tau} g(\pi_\theta,x,y)(\delta_{yy_2} - \pi_\theta(y_2|x)) \Big]. \tag{122}$$

For the second part, we begin with the expression:

$$2\sum_{y\in\mathbb{Y}} \pi_\theta(y|x) g(\pi_\theta,x,y)\frac{\partial g(\pi_\theta,x,y)}{\partial \theta(x,y_1)} = \frac{2}{\tau}\sum_{y\in\mathbb{Y}} \pi_\theta(y|x) g(\pi_\theta,x,y)(\delta_{yy_1} - \pi_\theta(y_1|x)). \tag{123}$$

Using the product rule, this derivative with respect to $\theta(x,y_2)$ consists of two components:

1. The derivative of $\pi_\theta(y|x)$: $\frac{2}{\tau}\sum_{y\in\mathbb{Y}} \frac{\partial \pi_\theta(y|x)}{\partial \theta(x,y_2)} g(\pi_\theta,x,y)(\delta_{yy_1} - \pi_\theta(y_1|x))$,
2. The derivative of $g(\pi_\theta,x,y)(\delta_{yy_1} - \pi_\theta(y_1|x))$:

$$\frac{2}{\tau}\sum_{y\in\mathbb{Y}} \pi_\theta(y|x) \left[\frac{\partial g(\pi_\theta,x,y)}{\partial \theta(x,y_2)}(\delta_{yy_1} - \pi_\theta(y_1|x)) - g(\pi_\theta,x,y)\frac{\partial \pi_\theta(y_1|x)}{\partial \theta(x,y_2)} \right]$$

.

For the derivative of $\pi_\theta(y|x)$, noting that $\frac{\partial \pi_\theta(y|x)}{\partial \theta(x,y_2)} = \pi_\theta(y|x)(\delta_{yy_2} - \pi_\theta(y_2|x))$, we obtain

$$\frac{2}{\tau}\sum_{y\in\mathbb{Y}} \pi_\theta(y|x)(\delta_{yy_2} - \pi_\theta(y_2|x)) g(\pi_\theta,x,y)(\delta_{yy_1} - \pi_\theta(y_1|x)). \tag{124}$$

For the derivative of $g(\pi_\theta,x,y)(\delta_{yy_1} - \pi_\theta(y_1|x))$, using $g(\pi_\theta,x,y) = \frac{1}{\tau}\log \pi_\theta(y|x) + z(y)$, we find

$$\frac{\partial g(\pi_\theta,x,y)}{\partial \theta(x,y_2)} = \frac{1}{\tau}(\delta_{yy_2} - \pi_\theta(y_2|x)), \tag{125}$$

and since $\frac{\partial \pi_\theta(y_1|x)}{\partial \theta(x,y_2)} = \pi_\theta(y_1|x)(\delta_{y_1 y_2} - \pi_\theta(y_2|x))$, the derivative of the second part simplifies to:

$$
\begin{aligned}
2\sum_{y\in\mathbb{Y}} \pi_\theta(y|x) g(\pi_\theta, x, y) \frac{\partial g(\pi_\theta, x, y)}{\partial \theta(x, y_1)} &= \frac{2}{\tau}\sum_{y\in\mathbb{Y}} \pi_\theta(y|x)(\delta_{yy_2} - \pi_\theta(y_2|x)) g(\pi_\theta, x, y)(\delta_{yy_1} - \pi_\theta(y_1|x)) \\
&+ \frac{2}{\tau^2}\sum_{y\in\mathbb{Y}} \pi_\theta(y|x)(\delta_{yy_2} - \pi_\theta(y_2|x))(\delta_{yy_1} - \pi_\theta(y_1|x)) \\
&- \frac{2}{\tau}\sum_{y\in\mathbb{Y}} \pi_\theta(y|x) g(\pi_\theta, x, y) \pi_\theta(y_1|x)(\delta_{y_1 y_2} - \pi_\theta(y_2|x)).
\end{aligned} \tag{126}
$$

Finally, combining terms, we derive the compact expression for the derivative of the second part with respect to $\theta(x, y_2)$:

$$
\begin{aligned}
&2\sum_{y\in\mathbb{Y}} \pi_\theta(y|x) g(\pi_\theta, x, y) \frac{\partial g(\pi_\theta, x, y)}{\partial \theta(x, y_1)} \\
=&\sum_{y\in\mathbb{Y}} \left(\frac{2}{\tau} g(\pi_\theta, x, y) + \frac{2}{\tau^2}\right) \pi_\theta(y|x)(\delta_{yy_2} - \pi_\theta(y_2|x))(\delta_{yy_1} - \pi_\theta(y_1|x)) \\
&- \frac{2}{\tau}\pi_\theta(y|x) g(\pi_\theta, x, y) \pi_\theta(y_1|x)(\delta_{y_1 y_2} - \pi_\theta(y_2|x)).
\end{aligned} \tag{127}
$$

By gathering all terms, we derive the full expression for the Hessian matrix:

$$
\begin{aligned}
&\frac{\partial}{\partial \theta(x, y_2)} \left(\frac{\partial \pi_\theta(\cdot|x)^T h(\pi_\theta, x)}{\partial \theta(x, y_1)}\right) = \frac{\partial^2 f_{RA}(x, \theta)}{\partial \theta(x, y_1) \partial \theta(x, y_2)} \\
=&\frac{\partial}{\partial \theta(x, y_2)} \left[\sum_{y\in\mathbb{Y}} \left(\frac{\partial \pi_\theta(y|x)}{\partial \theta(x, y_1)} g(\pi_\theta, x, y)^2 + 2\pi_\theta(y|x) g(\pi_\theta, x, y) \frac{\partial g(\pi_\theta, x, y)}{\partial \theta(x, y_1)}\right)\right] \\
=&\sum_{y\in\mathbb{Y}} \Big[(\pi_\theta(y|x)(\delta_{yy_2} - \pi_\theta(y_2|x))(\delta_{yy_1} - \pi_\theta(y_1|x)) - \pi_\theta(y|x)\pi_\theta(y_1|x)(\delta_{y_1 y_2} - \pi_\theta(y_2|x)))\, g(\pi_\theta, x, y)^2 \\
&\quad + \pi_\theta(y|x)(\delta_{yy_1} - \pi_\theta(y_1|x)) \frac{2}{\tau} g(\pi_\theta, x, y)(\delta_{yy_2} - \pi_\theta(y_2|x))\Big] \\
&+ \sum_{y\in\mathbb{Y}} \left(\frac{2}{\tau} g(\pi_\theta, x, y) + \frac{2}{\tau^2}\right) \pi_\theta(y|x)(\delta_{yy_2} - \pi_\theta(y_2|x))(\delta_{yy_1} - \pi_\theta(y_1|x)) \\
&\quad - \frac{2}{\tau}\pi_\theta(y|x) g(\pi_\theta, x, y) \pi_\theta(y_1|x)(\delta_{y_1 y_2} - \pi_\theta(y_2|x)) \\
=&\sum_{y\in\mathbb{Y}} \left(g(\pi_\theta, x, y)^2 + \frac{4}{\tau} g(\pi_\theta, x, y) + \frac{2}{\tau^2}\right) \pi_\theta(y|x)(\delta_{yy_2} - \pi_\theta(y_2|x))(\delta_{yy_1} - \pi_\theta(y_1|x)) \\
&\quad - \sum_{y\in\mathbb{Y}} \left(g(\pi_\theta, x, y)^2 + \frac{2}{\tau} g(\pi_\theta, x, y)\right) \pi_\theta(y|x)\pi_\theta(y_1|x)(\delta_{y_1 y_2} - \pi_\theta(y_2|x)).
\end{aligned} \tag{128}
$$

Let $\vec{g}(\theta, x) = (g(\pi_\theta, x, y_1), g(\pi_\theta, x, y_2), \ldots, g(\pi_\theta, x, y_{|\mathbb{Y}|}))$, $\vec{G}_1(\theta, x) = -\vec{g}^2(\theta, x) - \frac{2}{\tau}\vec{g}(\theta, x)$, and $\vec{G}_2(\theta, x) = (\vec{g}(\theta, x) + \frac{2}{\tau})^2 - \frac{2}{\tau^2}$. So $G_1(\pi_\theta, x, y) = -g^2(\pi_\theta, x, y) - \frac{2}{\tau} g(\pi_\theta, x, y)$ and $G_2(\pi_\theta, x, y)$ is similar. Then, the Hessian matrix of $f_{RA}(x, \theta) = \pi_\theta(\cdot|x)^T h(\pi_\theta, x)$ simplifies as follows:

$$
\begin{aligned}
\nabla^2_\theta f_{RA}(x, \theta)_{i,j} = \sum_{y\in\mathbb{Y}} [&\pi_\theta(y|x)\pi_\theta(y_i|x)(\delta_{y_i y_j} - \pi_\theta(y_j|x)) G_1(\pi_\theta, x, y) \\
&+ \pi_\theta(y|x)(\delta_{yy_j} - \pi_\theta(y_j|x))(\delta_{yy_i} - \pi_\theta(y_i|x)) G_2(\pi_\theta, x, y)].
\end{aligned} \tag{129}
$$

Then we can calculate the spectral radius of $S(\theta)$, because if $x_1 \neq x_2$, $\frac{\partial^2 \mathbb{E}_{x\sim\mathcal{D}, y\sim\pi_\theta(\cdot|x)}\left[(r_\theta(x,y)-r(x,y))^2\right]}{\partial\theta(x_1,y_1)\partial\theta(x_2,y_2)} = 0$. So

$$\left|\sum_{x,x'\in\mathbb{X}}\sum_{y,y'\in\mathbb{Y}} z(x,y)S(\theta,x,y,x',y')z(x',y')\right| = \left|\sum_{x,x'\in\mathbb{X}}\sum_{y,y'\in\mathbb{Y}} z(x,y)\frac{\partial^2\mathcal{L}_{RA}(\pi_\theta)}{\partial\theta(x,y)\partial\theta(x',y')}z(x',y')\right|$$
$$=\left|\sum_{x\in\mathbb{X}}\sum_{y,y'\in\mathbb{Y}} z(x,y)\frac{\partial^2\mathcal{L}_{RA}(\pi_\theta)}{\partial\theta(x,y)\partial\theta(x,y')}z(x,y')\right| = \left|\sum_{x\in\mathbb{X}}\sum_{y,y'\in\mathbb{Y}} z(x,y)\mathcal{D}(x)\frac{\partial^2 f_{RA}(x,\theta)}{\partial\theta(x,y)\partial\theta(x,y')}z(x,y')\right|. \tag{130}$$

To compute the quadratic form of the Hessian matrix, we proceed as follows:

$$\begin{aligned}
&|\sum_{x\in\mathbb{X}}\sum_{y,y'\in\mathbb{Y}} z(x,y)\mathcal{D}(x)\frac{\partial^2 f_{RA}(x,\theta)}{\partial\theta(x,y)\partial\theta(x,y')}z(x,y')| \\
=&|\sum_{x\in\mathbb{X}}\mathcal{D}(x)\sum_{i=1}^{|\mathbb{Y}|}\sum_{j=1}^{|\mathbb{Y}|} z(x,y_i)\left(\nabla_\theta^2 f_{RA}(x,\theta)_{i,j}\right)z(x,y_j)| \\
=&|\sum_{x\in\mathbb{X}}\mathcal{D}(x)(\pi_\theta(\cdot|x)^\top\vec{G}_1(\theta,x)(\pi_\theta(\cdot|x)^\top z(x,\cdot))^2+\pi_\theta(\cdot|x)^\top(\vec{G}_2(\theta,x)-\vec{G}_1(\theta,x))(\pi_\theta(\cdot|x)^\top z(x,\cdot))^2 \\
&\quad+\sum_{i=1}^{K} z(x,y_i)^2\pi_\theta(y_i|x)G_2(\pi_\theta,x,y_i)-2\pi_\theta(\cdot|x)^\top z(x,\cdot)\pi_\theta(y_i|x)G_2(\pi_\theta,x,y_i)z(x,y_i))| \\
=&|\sum_{x\in\mathbb{X}}\mathcal{D}(x)(\pi_\theta(\cdot|x)^\top\vec{G}_1(\theta,x)(\pi_\theta(\cdot|x)^\top z(x,\cdot))^2+\pi_\theta(\cdot|x)^\top(\vec{G}_2(\theta,x)-\vec{G}_1(\theta,x))(\pi_\theta(\cdot|x)^\top z(x,\cdot))^2 \\
&\quad+\sum_{i=1}^{K}(z(x,y_i)-2\pi_\theta(\cdot|x)^\top z(x,\cdot))\pi_\theta(y_i|x)G_2(\pi_\theta,x,y_i)z(x,y_i))| \\
\triangleq&|\sum_{x\in\mathbb{X}}\mathcal{D}(x)\psi(x)| \leq \|\mathcal{D}(\cdot)\|_1\|\psi(\cdot)\|_\infty = 1\cdot\|\psi(\cdot)\|_\infty.
\end{aligned} \tag{131}$$

To establish an upper bound for the spectral radius of the Hessian matrix, let $\vec{G}_3(\theta, x) = \vec{G}_2(\theta, x) - \vec{G}_1(\theta, x)$. We can then write:

$$
\begin{aligned}
&\|\psi(\cdot)\|_\infty \\
=&\max_{x\in\mathbb{X}} \Big| \pi_\theta(\cdot|x)^\top \vec{G}_1(\theta, x) \cdot (\pi_\theta(\cdot|x)^\top z(x,\cdot))^2 + \pi_\theta(\cdot|x)^\top \vec{G}_3(\theta, x) \cdot (\pi_\theta(\cdot|x)^\top z(x,\cdot))^2 \\
&\quad + \sum_{i=1}^{K} z(x, y_i)^2 \pi_\theta(y_i|x) G_2(\pi_\theta, x, y_i) + \sum_{i=1}^{K} 2\pi_\theta(\cdot|x)^\top z(x,\cdot)\pi_\theta(y_i|x) G_2(\pi_\theta, x, y_i) z(x, y_i) \Big| \\
\leq&\max_{x\in\mathbb{X}} \Big| \pi_\theta(\cdot|x)^\top \vec{G}_1(\theta, x) \cdot (\pi_\theta(\cdot|x)^\top z(x,\cdot))^2 \Big| + \Big| \pi_\theta(\cdot|x)^\top \vec{G}_3(\theta, x) \cdot (\pi_\theta(\cdot|x)^\top z(x,\cdot))^2 \Big| \\
&\quad + \Big| \sum_{i=1}^{K} z(x, y_i)^2 \pi_\theta(y_i|x) G_2(\pi_\theta, x, y_i) \Big| + \Big| \sum_{i=1}^{K} 2\pi_\theta(\cdot|x)^\top z(x,\cdot)\pi_\theta(y_i|x) G_2(\pi_\theta, x, y_i) z(x, y_i) \Big| \\
\leq&\max_{x\in\mathbb{X}} \|\pi_\theta(\cdot|x)\|_1 \|\vec{G}_1(\theta, x)\|_\infty \|z(x,\cdot)\|_2^2 + \|\pi_\theta(\cdot|x)\|_1 \|\vec{G}_3(\theta, x)\|_\infty \|z(x,\cdot)\|_2^2 \\
&\quad + \|\vec{G}_2(\theta, x)\|_\infty \|z(x,\cdot)\|_2^2 + 2\|\pi_\theta(\cdot|x) \odot \vec{G}_2(\theta, x)\|_1 \|z(x,\cdot)\|_2^2 \\
=&\max_{x\in\mathbb{X}} \|\vec{G}_1(\theta, x)\|_\infty \|z(x,\cdot)\|_2^2 + \|\vec{G}_3(\theta, x)\|_\infty \|z(x,\cdot)\|_2^2 \\
&\quad + \|\vec{G}_2(\theta, x)\|_\infty \|z(x,\cdot)\|_2^2 + 2\|\vec{G}_2(\theta, x)\|_\infty \|z(x,\cdot)\|_2^2 \\
=&\max_{x\in\mathbb{X}} (\|\vec{G}_1(\theta, x)\|_\infty + 3\|\vec{G}_2(\theta, x)\|_\infty + \|\vec{G}_3(\theta, x)\|_\infty) \|z(x,\cdot)\|_2^2 \\
\leq&\max_{x\in\mathbb{X}} (\|\vec{G}_1(\theta, x)\|_\infty + 3\|\vec{G}_2(\theta, x)\|_\infty + \|\vec{G}_3(\theta, x)\|_\infty) \max_{x\in\mathbb{X}} \|z(x,\cdot)\|_2^2 \\
\leq&\max_{x\in\mathbb{X}} (\|\vec{G}_1(\theta, x)\|_\infty + 3\|\vec{G}_2(\theta, x)\|_\infty + \|\vec{G}_3(\theta, x)\|_\infty) \|z(\cdot,\cdot)\|_2^2.
\end{aligned}
\tag{132}
$$

In the second inequality, the first term arises from Hölder's inequality, which states that $\pi_\theta^\top z(x,\cdot) \leq \|\pi_\theta\|_1 \|z(x,\cdot)\|_\infty = \|z(x,\cdot)\|_\infty$, where $\|\pi_\theta\|_\infty \leq 1$ and $\|z^2(x,\cdot)\|_1 = \|z(x,\cdot)\|_2^2$. For the second term, we have the bound on $|(\pi_\theta^\top z(x,\cdot))^2|$:

$$
|(\pi_\theta^\top z(x,\cdot))^2| = \pi_\theta^\top z(x,\cdot) \cdot \pi_\theta^\top z(x,\cdot) \leq \|\pi_\theta\|_1 \|z(x,\cdot)\|_\infty \cdot \|\pi_\theta\|_1 \|z(x,\cdot)\|_\infty \leq \|z(x,\cdot)\|_2^2. \tag{133}
$$

For the third term:

$$
\Big| \sum_{i=1}^{K} z(x, y_i)^2 \pi_\theta(y_i|x) G_2(y_i) \Big| \leq \Big| \sum_{i=1}^{K} z(x, y_i)^2 |G_2(y_i)| \Big| \leq \|\vec{G}_2(\theta)\|_\infty \|z^2(x,\cdot)\|_1 = \|\vec{G}_2(\theta)\|_\infty \|z(x,\cdot)\|_2^2. \tag{134}
$$

For the fourth term:

$$
\begin{aligned}
|\sum_{i=1}^{K} 2\pi_\theta^T z(x,\cdot)\pi_\theta(y_i) G_2(y_i) z(x, y_i)| &\leq 2||\pi_\theta||_1 ||z(x,\cdot)||_\infty |\sum_{i=1}^{K} \pi_\theta(y_i) G_2(y_i) z(x, y_i)|| \\
&\leq 2||z(x,\cdot)||_2 ||\pi_\theta(\cdot|x) \odot \vec{G}_2(\theta)||_1 ||z(x,\cdot)||_\infty \\
&\leq 2||\vec{G}_2(\theta)||_\infty ||z(x,\cdot)||_2^2.
\end{aligned}
\tag{135}
$$

Next, we analyze the coefficient $(\|\vec{G}_1(\theta, x)\|_\infty + \|\vec{G}_2(\theta, x)\|_\infty + \|\vec{G}_3(\theta, x)\|_\infty + \|\vec{G}_2(\theta, x)\|_1)$. The term $|g(\pi_\theta, x, y)|$ represents the error between the parameterized distribution π_θ and the true distribution for component (π_θ, x, y). In this Lemma, we assume $|r_\theta(x, y) - r(x, y)| \leq \epsilon_1$. Therefore $|g(\pi_\theta, x, y)| \leq \epsilon_1$ for all components (π_θ, x, y), where ϵ_1 denotes the upper bound of the error of the current parameterized distribution π_θ.

For $||\vec{G}_1(\theta, x)||_\infty$, we have:

$$
||\vec{G}_1(\theta, x)||_\infty = ||\vec{g}^2(\theta, x) + \frac{2}{\tau}\vec{g}(\theta, x)||_\infty \leq \max\{\epsilon_1^2 + \frac{2}{\tau}\epsilon_1, \frac{1}{\tau}\}. \tag{136}
$$

For $||\vec{G}_2(\theta, x)||_\infty$, it follows that:

$$
||\vec{G}_2(\theta, x)||_\infty = ||\vec{g}^2(\theta, x) + \frac{4}{\tau}\vec{g}(\theta, x) + \frac{2}{\tau^2}||_\infty \leq \epsilon_1^2 + \frac{4\epsilon_1}{\tau} + \frac{2}{\tau^2}. \tag{137}
$$

For $||\vec{G}_3(\theta, x)||_\infty$, we find:

$$||\vec{G}_3(\theta, x)||_\infty = ||\vec{G}_2(\theta, x) - \vec{G}_1(\theta)||_\infty = ||\frac{6}{\tau}\vec{g}(\theta, x) + \frac{2}{\tau^2}||_\infty \leq \frac{6\epsilon_1}{\tau} + \frac{2}{\tau^2}. \tag{138}$$

Combining these results yields:

$$\begin{aligned} &||\vec{G}_1(\theta)||_\infty + 3||\vec{G}_2(\theta)||_\infty + ||\vec{G}_3(\theta)||_\infty \\ \leq &3\epsilon_1^2 + \frac{18\epsilon_1}{\tau} + \frac{8}{\tau^2} + \max\left\{\epsilon_1^2 + \frac{2}{\tau}\epsilon_1, \frac{1}{\tau}\right\}. \end{aligned} \tag{139}$$

Thus, the Lipschitz coefficient of the gradient of the target function with respect to the parameter θ is given by:

$$\begin{aligned} &\left|\sum_{x,x'\in\mathbb{X}}\sum_{y,y'\in\mathbb{Y}} z(x,y)S(\theta,x,y,x',y')z(x',y')\right| \\ \leq &\max_{x\in\mathbb{X}}(\|\vec{G}_1(\theta,x)\|_\infty + 3\|\vec{G}_2(\theta,x)\|_\infty + \|\vec{G}_3(\theta,x)\|_\infty)\|z(\cdot,\cdot)\|_2^2 \\ \leq &(3\epsilon_1^2 + \frac{18\epsilon_1}{\tau} + \frac{8}{\tau^2} + \max\left\{\epsilon_1^2 + \frac{2}{\tau}\epsilon_1, \frac{1}{\tau}\right\})\|z(\cdot,\cdot)\|_2^2. \end{aligned} \tag{140}$$

Proof finished.

G.10 Proof of Lemma G.11

Lemma G.11. *(RDA Smoothness) Given softmax parametrization of Definition D.1 for policy π_θ, assume $|(r_\theta(x,y_1) - r_\theta(x,y_2)) - (r(x,y_1) - r(x,y_2))| \leq \epsilon_3$, $\forall r, \tau$, $\theta \to \mathbb{E}_{\mathcal{D}_{pw}}\left[\left(\frac{1}{\tau}\log\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)} - (r(x,y_1) - r(x,y_2))\right)^2\right]$ is $\left(20\epsilon_2^2 + \frac{32\epsilon_2}{\tau} + \frac{8}{\tau^2}\right)$-smooth where $\mathcal{D}_{pw} \triangleq \{(x,y_1,y_2) \mid x \in \mathcal{D}, y_1, y_2 \sim \pi_\theta(y|x)\}$ (see Eq.13). See proof in Appendix G.10.*

Proof: By Lemma G.4, it suffices to show that the spectral radius of the hessian matrix of the second derivative of $\mathcal{L}_{\text{RDA}}(\pi_\theta) = \mathbb{E}_{\mathcal{D}_{\text{pw}}}\left[\left(\frac{1}{\tau}\log\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)} - (r(x,y_1) - r(x,y_2))\right)^2\right]$ is bounded by $\left(20\epsilon_2^2 + \frac{32\epsilon_2}{\tau} + \frac{8}{\tau^2}\right)$, i.e.

$$\left|\sum_{x,x'\in\mathbb{X}}\sum_{y,y'\in\mathbb{Y}} z(x,y)\frac{\partial^2\mathcal{L}_{\text{RDA}}(\pi_\theta)}{\partial\theta(x,y)\partial\theta(x',y')}z(x',y')\right| \leq \left(20\epsilon_2^2 + \frac{32\epsilon_2}{\tau} + \frac{8}{\tau^2}\right)||z(\cdot,\cdot)||_2^2. \tag{141}$$

To show the bound, denote $h(\pi_\theta, x, y_1, y_2) = g^2(\pi_\theta, x, y_1, y_2)$ where $g(\pi_\theta, x, y_1, y_2) = \frac{1}{\tau}\log\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)} - (r(x,y_1) - r(x,y_2))$, we have:

$$\begin{aligned} &\left|\sum_{x,x'\in\mathbb{X}}\sum_{y,y'\in\mathbb{Y}} z(x,y)\frac{\partial^2\mathbb{E}_{\mathcal{D}_{\text{pw}}}\left[\left(\frac{1}{\tau}\log\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)} - (r(x,y_1) - r(x,y_2))\right)^2\right]}{\partial\theta(x,y)\partial\theta(x',y')}z(x',y')\right| \\ = &\left|\sum_{x\in\mathbb{X}}\sum_{y,y'\in\mathbb{Y}} z(x,y)\frac{\partial^2\sum_{x\in\mathbb{X}}\mathcal{D}(x)\sum_{y_1,y_2\in\mathbb{Y}}\pi_\theta(y_1|x)\pi_\theta(y_2|x)h(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y)\partial\theta(x,y')}z(x,y')\right| \\ = &\left|\sum_{x\in\mathbb{X}}\mathcal{D}(x)\sum_{y,y'\in\mathbb{Y}} z(x,y)\frac{\partial^2 f_{\text{RDA}}(x,\theta)}{\partial\theta(x,y)\partial\theta(x,y')}z(x,y')\right| \\ \triangleq &|\sum_{x\in\mathbb{X}}\mathcal{D}(x)\psi(x)| \leq \|\mathcal{D}(\cdot)\|_1\|\psi(\cdot)\|_\infty = 1\cdot\|\psi(\cdot)\|_\infty. \end{aligned} \tag{142}$$

where $f_{\text{RDA}}(x,\theta) = \sum_{y_1,y_2\in\mathbb{Y}}\pi_\theta(y_1|x)\pi_\theta(y_2|x)h(\pi_\theta,x,y_1,y_2)$.

Based on Lemma G.5, the second derivative of $f_{\text{RDA}}(x, \theta)$ is:

$$
\begin{aligned}
&\frac{\partial^2 f_{\text{RDA}}(x,\theta)}{\partial\theta(x,y_i)\partial\theta(x,y_j)} \\
=& \sum_{y_1,y_2\in\mathbb{Y}} 2\frac{\partial^2\pi_\theta(y_1|x)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}\pi_\theta(y_2|x)h(\pi_\theta,x,y_1,y_2) + 2\frac{\partial\pi_\theta(y_1|x)}{\partial\theta(x,y_i)}\frac{\partial\pi_\theta(y_2|x)}{\partial\theta(x,y_j)}h(\pi_\theta,x,y_1,y_2) \\
&+ 2\frac{\partial\pi_\theta(y_1|x)}{\partial\theta(x,y_i)}\pi_\theta(y_2|x)\frac{\partial h(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_j)} + 2\frac{\partial\pi_\theta(y_1|x)}{\partial\theta(x,y_j)}\pi_\theta(y_2|x)\frac{\partial h(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_i)} \\
&+ \pi_\theta(y_1|x)\pi_\theta(y_2|x)\frac{\partial^2 h(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}.
\end{aligned} \tag{143}
$$

Because $|(r_\theta(x,y_1) - r_\theta(x,y_2)) - (r(x,y_1) - r(x,y_2))| \le \epsilon_2$, $|g(\pi_\theta,x,y_1,y_2)| \le \epsilon_2$.

According to the absolute value inequality, $|\psi(x)| = \left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\frac{\partial^2 f_{\text{RDA}}(x,\theta)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}z(x,y_j)\right|$ can be decomposed into five parts based on Eq.143 for further analysis.

$$
\begin{aligned}
&\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\frac{\partial^2 f_{\text{RDA}}(x,\theta)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}z(x,y_j)\right| \\
\le& \left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} 2\frac{\partial^2\pi_\theta(y_1|x)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}\pi_\theta(y_2|x)h(\pi_\theta,x,y_1,y_2)z(x,y_j)\right| \\
&+ \left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} 2\frac{\partial\pi_\theta(y_1|x)}{\partial\theta(x,y_i)}\frac{\partial\pi_\theta(y_2|x)}{\partial\theta(x,y_j)}h(\pi_\theta,x,y_1,y_2)z(x,y_j)\right| \\
&+ \left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} 2\frac{\partial\pi_\theta(y_1|x)}{\partial\theta(x,y_i)}\pi_\theta(y_2|x)\frac{\partial h(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_j)}z(x,y_j)\right| \\
&+ \left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} 2\frac{\partial\pi_\theta(y_1|x)}{\partial\theta(x,y_j)}\pi_\theta(y_2|x)\frac{\partial h(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_i)}z(x,y_j)\right| \\
&+ \left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} \pi_\theta(y_1|x)\pi_\theta(y_2|x)\frac{\partial^2 h(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}z(x,y_j)\right|.
\end{aligned} \tag{144}
$$

For the first and second terms of Eq.144, we have:

$$
\begin{aligned}
&\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} 2\frac{\partial^2\pi_\theta(y_1|x)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}\pi_\theta(y_2|x)h(\pi_\theta,x,y_1,y_2)z(x,y_j)\right| \\
&+\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} 2\frac{\partial\pi_\theta(y_1|x)}{\partial\theta(x,y_i)}\frac{\partial\pi_\theta(y_2|x)}{\partial\theta(x,y_j)}h(\pi_\theta,x,y_1,y_2)z(x,y_j)\right| \\
=&\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} 4\pi_\theta(y_1|x)\pi_\theta(y_2|x)(\delta_{y_1y_i}-\pi_\theta(y_i|x))(\delta_{y_2y_j}-\pi_\theta(y_j|x))h(\pi_\theta,x,y_1,y_2)z(x,y_j)\right| \\
&+\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} 2\pi_\theta(y_1|x)\pi_\theta(y_2|x)\pi_\theta(y_i|x)(\delta_{y_iy_j}-\pi_\theta(y_j|x))h(\pi_\theta,x,y_1,y_2)z(x,y_j)\right| \\
\leq&\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} 4\pi_\theta(y_1|x)\pi_\theta(y_2|x)(\delta_{y_1y_i}-\pi_\theta(y_i|x))(\delta_{y_2y_j}-\pi_\theta(y_j|x))\epsilon_2^2 z(x,y_j)\right| \\
&+\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} 2\pi_\theta(y_1|x)\pi_\theta(y_2|x)\pi_\theta(y_i|x)(\delta_{y_iy_j}-\pi_\theta(y_j|x))\epsilon_2^2 z(x,y_j)\right|.
\end{aligned}
\tag{145}
$$

The first equality is because:

$$
\frac{\partial^2\pi_\theta(y|x)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}=\pi_\theta(y|x)(\delta_{yy_j}-\pi_\theta(y_j|x))(\delta_{yy_i}-\pi_\theta(y_i|x))-\pi_\theta(y|x)\pi_\theta(y_i|x)(\delta_{y_iy_j}-\pi_\theta(y_j|x)).
\tag{146}
$$

The second inequality is because:

$$
|h(\pi_\theta,x,y_1,y_2)|=|g^2(\pi_\theta,x,y_1,y_2)|\leq\epsilon_2^2.
\tag{147}
$$

Then:

$$
\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} 4\pi_\theta(y_1|x)\pi_\theta(y_2|x)(\delta_{y_1y_i}-\pi_\theta(y_i|x))(\delta_{y_2y_j}-\pi_\theta(y_j|x))\epsilon_2^2 z(x,y_j)\right|\leq 16\epsilon_2^2\|z(x,\cdot)\|_2^2.
\tag{148}
$$

The inequality is because we can expand the absolute value operation in the second-to-last line into four terms, and these four terms can be deduced to be less than $\|z(x,\cdot)\|_2^2$ using some existing conclusions. The relevant conclusions are: $\pi_\theta^\top z(x,\cdot)\leq\|\pi_\theta\|_1\|z(x,\cdot)\|_\infty=\|z(x,\cdot)\|_\infty\leq\|z(x,\cdot)\|_2$, $\pi_\theta^\top z^2(x,\cdot)\leq\|z(x,\cdot)\|_2^2$, $\|\pi_\theta\|_\infty\leq 1$, $\|z^2(x,\cdot)\|_1=\|z(x,\cdot)\|_2^2$ and $|(\pi_\theta^\top z(x,\cdot))^2|\leq\|z(x,\cdot)\|_2^2$.

And similarly, we have:

$$
\begin{aligned}
&\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} 2\pi_\theta(y_1|x)\pi_\theta(y_2|x)\pi_\theta(y_i|x)(\delta_{y_iy_j}-\pi_\theta(y_j|x))\epsilon_2^2 z(x,y_j)\right| \\
\leq&2\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\delta_{y_iy_j}\epsilon_2^2 z(x,y_j)\right|+2\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\pi_\theta(y_i|x)\pi_\theta(y_j|x)\epsilon_2^2 z(x,y_j)\right| \\
\leq&2\epsilon_2^2+2\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\pi_\theta(y_i|x)\pi_\theta(y_j|x)\epsilon_2^2 z(x,y_j)\right| \\
\leq&2\epsilon_2^2+2\epsilon_2^2\left|\sum_{y_i\in\mathbb{Y}} z(x,y_i)\pi_\theta(y_i|x)\sum_{y_j\in\mathbb{Y}}\pi_\theta(y_j|x)z(x,y_j)\right| \\
\leq&2\epsilon_2^2+2\epsilon_2^2\left|\|\pi(\cdot|x)\|_1\|z(x,\cdot)\|_\infty\|\pi(\cdot|x)\|_1\|z(x,\cdot)\|_\infty\right|\leq 4\epsilon_2^2\|z(x,\cdot)\|_2^2.
\end{aligned}
\tag{149}
$$

For the third and forth terms of Eq.144, we have:

$$
\begin{aligned}
&\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} 2\frac{\partial\pi_\theta(y_1|x)}{\partial\theta(x,y_i)}\pi_\theta(y_2|x)\frac{\partial h(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_j)}z(x,y_j)\right| \\
&+\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} 2\frac{\partial\pi_\theta(y_1|x)}{\partial\theta(x,y_j)}\pi_\theta(y_2|x)\frac{\partial h(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_i)}z(x,y_j)\right| \\
=&2\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} 2\frac{\partial\pi_\theta(y_1|x)}{\partial\theta(x,y_i)}\pi_\theta(y_2|x)\frac{\partial h(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_j)}z(x,y_j)\right| \\
=&8\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} \frac{\partial\pi_\theta(y_1|x)}{\partial\theta(x,y_i)}\pi_\theta(y_2|x)g(\pi_\theta,x,y_1,y_2)\frac{\partial g(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_j)}z(x,y_j)\right| \\
=&8\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} \pi_\theta(y_1|x)(\delta_{y_1y_i}-\pi_\theta(y_i|x))\pi_\theta(y_2|x)g(\pi_\theta,x,y_1,y_2)\frac{1}{\tau}(\delta_{y_1y_j}-\delta_{y_2y_j})z(x,y_j)\right| \\
\leq&\frac{8\epsilon_2}{\tau}\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} \pi_\theta(y_1|x)(\delta_{y_1y_i}-\pi_\theta(y_i|x))\pi_\theta(y_2|x)\frac{1}{\tau}(\delta_{y_1y_j}-\delta_{y_2y_j})z(x,y_j)\right| \\
\leq&\frac{32\epsilon_2}{\tau}\|z(x,\cdot)\|_2^2.
\end{aligned}
\tag{150}
$$

where the first derivative of $g(\pi_\theta,x,y_1,y_2)$ is:

$$
\frac{\partial g(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_i)}=\frac{\partial\frac{1}{\tau}(\log\pi_\theta(y_1|x)-\log\pi_\theta(y_2|x))}{\partial\theta(x,y_i)}=\frac{1}{\tau}(\delta_{y_1y_i}-\pi_\theta(y_i|x)-\delta_{y_2y_i}+\pi_\theta(y_i|x))=\frac{1}{\tau}(\delta_{y_1y_i}-\delta_{y_2y_i}).
\tag{151}
$$

And the last inequality is because we can expand the absolute value operation in the second-to-last line into four terms, and these four terms can be deduced to be less than $\|z(x,\cdot)\|_2^2$ using some existing conclusions. The relevant conclusions are: $\pi_\theta^\top z(x,\cdot)\leq\|\pi_\theta\|_1\|z(x,\cdot)\|_\infty=\|z(x,\cdot)\|_\infty\leq\|z(x,\cdot)\|_2$, $\pi_\theta^\top z^2(x,\cdot)\leq\|z(x,\cdot)\|_2^2$, $\|\pi_\theta\|_\infty\leq 1$, $\|z^2(x,\cdot)\|_1=\|z(x,\cdot)\|_2^2$ and $|(\pi_\theta^\top z(x,\cdot))^2|\leq\|z(x,\cdot)\|_2^2$.

For the fifth term of Eq.144, we have:

$$
\begin{aligned}
&\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} \pi_\theta(y_1|x)\pi_\theta(y_2|x)\frac{\partial^2 h(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}z(x,y_j)\right| \\
=&\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} \pi_\theta(y_1|x)\pi_\theta(y_2|x)2\frac{\partial g(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_i)}\frac{\partial g(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_j)}z(x,y_j)\right| \\
=&\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} \pi_\theta(y_1|x)\pi_\theta(y_2|x)\frac{2}{\tau^2}(\delta_{y_1y_i}-\delta_{y_2y_i})(\delta_{y_1y_j}-\delta_{y_2y_j})z(x,y_j)\right| \\
=&|\sum_{y_i\in\mathbb{Y}} z(x,y_i)\pi_\theta(y_i|x)\frac{2}{\tau^2}z(x,y_i)-\sum_{y_i\in\mathbb{Y}} z(x,y_i)\pi_\theta(y_i|x)\sum_{y_j\in\mathbb{Y}}\pi_\theta(y_j|x)\frac{2}{\tau^2}z(x,y_j) \\
&-\sum_{y_i\in\mathbb{Y}} z(x,y_i)\pi_\theta(y_i|x)\sum_{y_j\in\mathbb{Y}}\pi_\theta(y_j|x)\frac{2}{\tau^2}z(x,y_j)+\sum_{y_i\in\mathbb{Y}} z(x,y_i)\pi_\theta(y_i|x)\frac{2}{\tau^2}z(x,y_i)|,
\end{aligned}
\tag{152}
$$

$$
\begin{aligned}
&\leq \frac{4}{\tau^2}|\sum_{y_i\in\mathbb{Y}} z(x,y_i)\pi_\theta(y_i|x)z(x,y_i)| + |\sum_{y_i\in\mathbb{Y}} z(x,y_i)\pi_\theta(y_i|x)\sum_{y_j\in\mathbb{Y}}\pi_\theta(y_j|x)z(x,y_j)| \\
&= \frac{4}{\tau^2}(\pi_\theta^T(\cdot|x)z^2(x,\cdot) + (\pi_\theta^T(\cdot|x)z(x,\cdot))^2) \leq \frac{4}{\tau^2}(\|\pi_\theta(\cdot|x)\|_1\|z^2(x,\cdot)\|_\infty + (\|\pi_\theta(\cdot|x)\|_1\|z(x,\cdot)\|_\infty)^2) \\
&\leq \frac{8}{\tau^2}\|z(x,\cdot)\|_2^2.
\end{aligned}
$$

where the second derivative of $g(\pi_\theta, x, y_1, y_2)$ is:

$$
\frac{\partial^2 g(\pi_\theta, x, y_1, y_2)}{\partial\theta(x,y_i)\partial\theta(x,y_j)} = \frac{\partial}{\partial\theta(x,y_j)}(\frac{1}{\tau}(\delta_{y_1 y_i} - \delta_{y_2 y_i})) = 0. \tag{153}
$$

In summary, we have the upper bound of Eq.144:

$$
\begin{aligned}
\|\psi(x)\|_\infty &= \max_{x\in\mathbb{X}}|\psi(x)| = \max_{x\in\mathbb{X}}\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\frac{\partial^2 f_{\text{RDA}}(x,\theta)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}z(x,y_j)\right| \\
&\leq \left(20\epsilon_2^2 + \frac{32\epsilon_2}{\tau} + \frac{8}{\tau^2}\right)\max_{x\in\mathbb{X}}\|z(x,\cdot)\|_2^2 \\
&\leq \left(20\epsilon_2^2 + \frac{32\epsilon_2}{\tau} + \frac{8}{\tau^2}\right)\|z(\cdot,\cdot)\|_2^2
\end{aligned} \tag{154}
$$

Then Eq.141 is proved. Proof finished.

G.11 Proof of Lemma G.12

Lemma G.12. *(PRA Smoothness) Given softmax parametrization of Definition D.1 for policy π_θ, assume $|\theta(x_1,y_1) - \theta(x_2,y_2)| \leq d$ and $|p^* - \sigma(\frac{1}{\tau}\log\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)})| \leq \epsilon_4$, $\forall r, \tau$, $\theta \to \mathbb{E}_{\mathcal{D}_\theta}\left[\mathrm{D}_{\mathrm{KL}}(p^*(z|y_1,y_2,x)||p_\theta(z|y_1,y_2,x))\right]$ is $(20\log(1+e^{\frac{d}{\tau}}) + \frac{16\epsilon_3}{\tau} + \frac{4}{\tau^2} + 16\log 2)$-smooth where $\mathcal{D}_\theta \triangleq \{(x,y_1,y_2,z) \mid x\in\mathcal{D}, y_1,y_2\sim\pi_\theta(\cdot|x), z = 1$ if $r(x,y_1) > r(x,y_2); 0$ otherwise$\}$ (see Eq.14). See proof in Appendix G.11.*

Proof: Based on Appendix E.3, we have:

$$
\begin{aligned}
&\mathbb{E}_{\mathcal{D}_\theta}\left[\mathrm{D}_{\mathrm{KL}}(p^*(z|y_1,y_2,x)||\sigma\left(h_\theta(x,y_1,y_2)\right))\right] \\
=& -\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_\theta(y|x)}\left[p^*(z=1|y_1,y_2,x)\log\sigma\left(h_\theta(x,y_1,y_2)\right) + p^*(z=0|y_1,y_2,x)\log\sigma\left(h_\theta(x,y_2,y_1)\right)\right] \\
&+ \mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_\theta(y|x)}\left[p^*(z=1|y_1,y_2,x)\log p^*(z=1|y_1,y_2,x) + p^*(z=0|y_1,y_2,x)\log p^*(z=0|y_1,y_2,x)\right] \\
=& -\mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_\theta(y|x)}\left[p^*(z=1|y_1,y_2,x)\log\sigma\left(h_\theta(x,y_1,y_2)\right) + p^*(z=0|y_1,y_2,x)\log\sigma\left(h_\theta(x,y_2,y_1)\right)\right] \\
&+ \mathbb{E}_{x\sim\mathcal{D},y_1,y_2\sim\pi_\theta(y|x)}\left[M(x,y_1,y_2)\right] \\
\triangleq& \mathcal{L}_{\text{PRA}}(\pi_\theta).
\end{aligned} \tag{155}
$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function, $M(x,y_1,y_2) = \sum_{z=0,1}p^*(z|y_1,y_2,x)\log p^*(z|y_1,y_2,x)$ and $(x,y_1,y_2) \sim \mathcal{D}_\theta \triangleq y_1 \succ y_2 \sim p^*(z = 1|y_1,y_2,x), y_1,y_2\sim\pi_\theta(y|x), x\sim\mathcal{D}$.

By Lemma G.4, it suffices to show that the spectral radius of the hessian matrix of the second derivative of $\mathcal{L}_{\text{PRA}}(\pi_\theta)$, i.e.

$$
\begin{aligned}
&\left|\sum_{x,x'\in\mathbb{X}}\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\frac{\partial^2\mathcal{L}_{\text{PRA}}(\pi_\theta)}{\partial\theta(x,y_i)\partial\theta(x',y_j)}z(x',y_j)\right| \\
&\leq \left(20\log(1+e^{\frac{d}{\tau}}) + \frac{16\epsilon_3}{\tau} + \frac{4}{\tau^2} + 16\log 2\right)||z(\cdot,\cdot)||_2^2.
\end{aligned} \tag{156}
$$

Denote $h(\pi_\theta, x, y_1, y_2) = -p^*(z = 1|y_1, y_2, x) \log \sigma\left(h_\theta(x, y_1, y_2)\right) - p^*(z = 0|y_1, y_2, x) \log \sigma\left(h_\theta(x, y_2, y_1)\right)$, we have:

$$
\begin{aligned}
&\left| \sum_{x,x' \in \mathbb{X}} \sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i) \frac{\partial^2 \mathcal{L}_{\text{PRA}}(\pi_\theta)}{\partial\theta(x, y_i)\partial\theta(x', y_j)} z(x', y_j) \right| \\
=&\left| \sum_{x \in \mathbb{X}} \sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i) \frac{\partial^2 \sum_{x \in \mathbb{X}} \mathcal{D}(x) \sum_{y_1, y_2 \in \mathbb{Y}} \pi_\theta(y_1|x)\pi_\theta(y_2|x)(h(\pi_\theta, x, y_1, y_2) + M(x, y_1, y_2))}{\partial\theta(x, y_i)\partial\theta(x, y_j)} z(x, y_j) \right| \\
=&\left| \sum_{x \in \mathbb{X}} \mathcal{D}(x) \sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i) \frac{\partial^2 f_{\text{PRA}}(x, \theta)}{\partial\theta(x, y_i)\partial\theta(x, y_j)} z(x, y_j) \right| \\
&+ \left| \sum_{x \in \mathbb{X}} \mathcal{D}(x) \sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i) \frac{\partial^2 \sum_{y_1, y_2 \in \mathbb{Y}} \pi_\theta(y_1|x)\pi_\theta(y_2|x) M(x, y_1, y_2)}{\partial\theta(x, y_i)\partial\theta(x, y_j)} z(x, y_j) \right| \\
\triangleq&|\sum_{x \in \mathbb{X}} \mathcal{D}(x)\psi(x)| + \left| \sum_{x \in \mathbb{X}} \mathcal{D}(x) \sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i) \frac{\partial^2 \sum_{y_1, y_2 \in \mathbb{Y}} \pi_\theta(y_1|x)\pi_\theta(y_2|x) M(x, y_1, y_2)}{\partial\theta(x, y_i)\partial\theta(x, y_j)} z(x, y_j) \right| \\
\leq&\|\mathcal{D}(\cdot)\|_1 \|\psi(\cdot)\|_\infty + \left| \sum_{x \in \mathbb{X}} \mathcal{D}(x) \sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i) \frac{\partial^2 \sum_{y_1, y_2 \in \mathbb{Y}} \pi_\theta(y_1|x)\pi_\theta(y_2|x) M(x, y_1, y_2)}{\partial\theta(x, y_i)\partial\theta(x, y_j)} z(x, y_j) \right| \\
=&1 \cdot \|\psi(\cdot)\|_\infty + \left| \sum_{x \in \mathbb{X}} \mathcal{D}(x) \sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i) \frac{\partial^2 \sum_{y_1, y_2 \in \mathbb{Y}} \pi_\theta(y_1|x)\pi_\theta(y_2|x) M(x, y_1, y_2)}{\partial\theta(x, y_i)\partial\theta(x, y_j)} z(x, y_j) \right|.
\end{aligned} \tag{157}
$$

where $f_{\text{PRA}}(x, \theta) = \sum_{y_1, y_2 \in \mathbb{Y}} \pi_\theta(y_1|x)\pi_\theta(y_2|x) h(\pi_\theta, x, y_1, y_2)$ and $\psi(x) = \sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i) \frac{\partial^2 f_{\text{PRA}}(x,\theta)}{\partial\theta(x,y_i)\partial\theta(x,y_j)} z(x, y_j)$.

Because $|\theta(x_1, y_1) - \theta(x_2, y_2)| \leq d$ and the Definition D.1 for policy π_θ, consider the upper bound of $h(\pi_\theta, x, y_1, y_2)$, we have:

$$
\begin{aligned}
|\log \sigma\left(h_\theta(x, y_1, y_2)\right)| =& |\log \sigma\left(\frac{1}{\tau} \log \frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)}\right)| = |\log \sigma\left(\frac{1}{\tau} \log \frac{\exp(\theta(y_1, x))}{\exp(\theta(y_2, x))}\right)| \\
=& |\log \sigma\left(\frac{1}{\tau}(\theta(y_1, x) - \theta(y_2, x))\right)| \leq -\log \sigma\left(\frac{-d}{\tau}\right) \\
=& \log \sigma^{-1}\left(\frac{-d}{\tau}\right) = \log(1 + e^{\frac{d}{\tau}}).
\end{aligned} \tag{158}
$$

Then

$$
\begin{aligned}
&|h(\pi_\theta, x, y_1, y_2)| \\
=&|p^*(z = 1|y_1, y_2, x) \log \sigma\left(h_\theta(x, y_1, y_2)\right) + p^*(z = 0|y_1, y_2, x) \log \sigma\left(h_\theta(x, y_2, y_1)\right)| \\
\leq& p^*(z = 1|y_1, y_2, x) \log(1 + e^{\frac{d}{\tau}}) + p^*(z = 0|y_1, y_2, x) \log(1 + e^{\frac{d}{\tau}}) = \log(1 + e^{\frac{d}{\tau}}).
\end{aligned} \tag{159}
$$

Consider the first derivative of $h(\pi_\theta, x, y_1, y_2)$, denote $p^* = p^*(1|y_1, y_2, x)$:

$$
\begin{aligned}
&\frac{\partial h(\pi_\theta, x, y_1, y_2)}{\partial\theta(x, y_i)} \\
=& -p^* \frac{\partial}{\partial\theta(x, y_i)}(\log \sigma(\frac{1}{\tau} \log \frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)})) - (1 - p^*) \frac{\partial}{\partial\theta(x, y_i)}(\log \sigma(\frac{1}{\tau} \log \frac{\pi_\theta(y_2|x)}{\pi_\theta(y_1|x)})) \\
=& -\left(p^*(1 - \sigma(\frac{1}{\tau} \log \frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)})) - (1 - p^*)\sigma(\frac{1}{\tau} \log \frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)})\right) \frac{1}{\tau} \frac{\partial}{\partial\theta(x, y_i)}(\log \frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)}) \\
=& -\left(p^* - \sigma(\frac{1}{\tau} \log \frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)})\right) \frac{1}{\tau}(\delta_{y_1 y_i} - \delta_{y_2 y_i}).
\end{aligned} \tag{160}
$$

Consider the second derivative of $h(\pi_\theta, x, y_1, y_2)$:

$$\begin{aligned}\frac{\partial^2 h(\pi_\theta, x, y_1, y_2)}{\partial\theta(x, y_i)\partial\theta(x, y_j)} =& -\frac{\partial}{\partial\theta(x, y_j)}(\sigma(\frac{1}{\tau}\log\frac{\pi_\theta(y_2|x)}{\pi_\theta(y_1|x)})\frac{1}{\tau}(\delta_{y_1 y_i} - \delta_{y_2 y_i})) \\ =& \frac{1}{\tau^2}(\delta_{y_1 y_i} - \delta_{y_2 y_i})(\delta_{y_1 y_j} - \delta_{y_2 y_j})\sigma(\frac{1}{\tau}\log\frac{\pi_\theta(y_2|x)}{\pi_\theta(y_1|x)})\sigma(\frac{1}{\tau}\log\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)}).\end{aligned} \tag{161}$$

Based on Lemma G.5, the second derivative of $f_{\text{PRA}}(x, \theta)$ is:

$$\begin{aligned}&\frac{\partial^2 f_{\text{PRA}}(x, \theta)}{\partial\theta(x, y_i)\partial\theta(x, y_j)} \\ =& \sum_{y_1, y_2 \in \mathbb{Y}} 2\frac{\partial^2 \pi_\theta(y_1|x)}{\partial\theta(x, y_i)\partial\theta(x, y_j)}\pi_\theta(y_2|x)h(\pi_\theta, x, y_1, y_2) + 2\frac{\partial\pi_\theta(y_1|x)}{\partial\theta(x, y_i)}\frac{\partial\pi_\theta(y_2|x)}{\partial\theta(x, y_j)}h(\pi_\theta, x, y_1, y_2) \\ &+ 2\frac{\partial\pi_\theta(y_1|x)}{\partial\theta(x, y_i)}\pi_\theta(y_2|x)\frac{\partial h(\pi_\theta, x, y_1, y_2)}{\partial\theta(x, y_j)} + 2\frac{\partial\pi_\theta(y_1|x)}{\partial\theta(x, y_j)}\pi_\theta(y_2|x)\frac{\partial h(\pi_\theta, x, y_1, y_2)}{\partial\theta(x, y_i)} \\ &+ \pi_\theta(y_1|x)\pi_\theta(y_2|x)\frac{\partial^2 h(\pi_\theta, x, y_1, y_2)}{\partial\theta(x, y_i)\partial\theta(x, y_j)}.\end{aligned} \tag{162}$$

According to the absolute value inequality, $|\psi(x)| = \left|\sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i)\frac{\partial^2 f_{\text{PRA}}(x,\theta)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}z(x, y_j)\right|$ can be decomposed into five parts based on Eq.162 for further analysis.

$$\begin{aligned}&\left|\sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i)\frac{\partial^2 f_{\text{PRA}}(x, \theta)}{\partial\theta(x, y_i)\partial\theta(x, y_j)}z(x, y_j)\right| \\ \leq& \left|\sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i)\sum_{y_1, y_2 \in \mathbb{Y}} 2\frac{\partial^2 \pi_\theta(y_1|x)}{\partial\theta(x, y_i)\partial\theta(x, y_j)}\pi_\theta(y_2|x)h(\pi_\theta, x, y_1, y_2)z(x, y_j)\right| \\ &+ \left|\sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i)\sum_{y_1, y_2 \in \mathbb{Y}} 2\frac{\partial\pi_\theta(y_1|x)}{\partial\theta(x, y_i)}\frac{\partial\pi_\theta(y_2|x)}{\partial\theta(x, y_j)}h(\pi_\theta, x, y_1, y_2)z(x, y_j)\right| \\ &+ \left|\sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i)\sum_{y_1, y_2 \in \mathbb{Y}} 4\frac{\partial\pi_\theta(y_1|x)}{\partial\theta(x, y_i)}\pi_\theta(y_2|x)\frac{\partial h(\pi_\theta, x, y_1, y_2)}{\partial\theta(x, y_j)}z(x, y_j)\right| \\ &+ \left|\sum_{y_i, y_j \in \mathbb{Y}} z(x, y_i)\sum_{y_1, y_2 \in \mathbb{Y}} \pi_\theta(y_1|x)\pi_\theta(y_2|x)\frac{\partial^2 h(\pi_\theta, x, y_1, y_2)}{\partial\theta(x, y_i)\partial\theta(x, y_j)}z(x, y_j)\right|.\end{aligned} \tag{163}$$

Similar to Eq.145 and its following derivation, we have:

$$
\begin{aligned}
&\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} 2\frac{\partial^2\pi_\theta(y_1|x)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}\pi_\theta(y_2|x)h(\pi_\theta,x,y_1,y_2)z(x,y_j)\right| \\
&+\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} 2\frac{\partial\pi_\theta(y_1|x)}{\partial\theta(x,y_i)}\frac{\partial\pi_\theta(y_2|x)}{\partial\theta(x,y_j)}h(\pi_\theta,x,y_1,y_2)z(x,y_j)\right| \\
\leq&\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} 4\pi_\theta(y_1|x)\pi_\theta(y_2|x)(\delta_{y_1y_i}-\pi_\theta(y_i|x))(\delta_{y_2y_j}-\pi_\theta(y_j|x))\log(1+e^{\frac{d}{\tau}})z(x,y_j)\right| \\
&+\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} 2\pi_\theta(y_1|x)\pi_\theta(y_2|x)\pi_\theta(y_i|x)(\delta_{y_iy_j}-\pi_\theta(y_j|x))\log(1+e^{\frac{d}{\tau}})z(x,y_j)\right| \\
\leq&20\log(1+e^{\frac{d}{\tau}}).
\end{aligned}
\tag{164}
$$

Similar to Eq.150, for the third term of Eq.163, we have:

$$
\begin{aligned}
&\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} 4\frac{\partial\pi_\theta(y_1|x)}{\partial\theta(x,y_i)}\pi_\theta(y_2|x)\frac{\partial h(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_j)}z(x,y_j)\right| \\
=&4\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} \frac{\partial\pi_\theta(y_1|x)}{\partial\theta(x,y_i)}\pi_\theta(y_2|x)\left(p^*-\sigma(\frac{1}{\tau}\log\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)})\right)\frac{1}{\tau}(\delta_{y_1y_j}-\delta_{y_2y_j})z(x,y_j)\right| \\
\leq&4\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} \frac{\partial\pi_\theta(y_1|x)}{\partial\theta(x,y_i)}\pi_\theta(y_2|x)\epsilon_3\frac{1}{\tau}(\delta_{y_1y_j}-\delta_{y_2y_j})z(x,y_j)\right| \\
=&\frac{4\epsilon_3}{\tau}\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} \pi_\theta(y_1|x)(\delta_{y_1y_i}-\pi_\theta(y_i|x))\pi_\theta(y_2|x)\epsilon_3\frac{1}{\tau}(\delta_{y_1y_j}-\delta_{y_2y_j})z(x,y_j)\right| \\
\leq&\frac{16\epsilon_3}{\tau}\|z(x,\cdot)\|_2^2.
\end{aligned}
\tag{165}
$$

The first inequality is because $|p^*-\sigma(\frac{1}{\tau}\log\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)})|\leq\epsilon_3$ where $p^*=p^*(1|x,y_1,y_2)$. The last inequality is because we can expand the absolute value operation in the second-to-last line into four terms, and these four terms can be deduced to be less than $\|z(x,\cdot)\|_2^2$ using some existing conclusions. The relevant conclusions are: $\pi_\theta^\top z(x,\cdot)\leq\|\pi_\theta\|_1\|z(x,\cdot)\|_\infty=\|z(x,\cdot)\|_\infty\leq\|z(x,\cdot)\|_2$, $\pi_\theta^\top z^2(x,\cdot)\leq\|z(x,\cdot)\|_2^2$, $\|\pi_\theta\|_\infty\leq 1$, $\|z^2(x,\cdot)\|_1=\|z(x,\cdot)\|_2^2$ and $|(\pi_\theta^\top z(x,\cdot))^2|\leq\|z(x,\cdot)\|_2^2$.

Similar to Eq.152, for the fifth term of Eq.163, based on Eq.161 and Eq.62, we have:

$$
\begin{aligned}
&\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} \pi_\theta(y_1|x)\pi_\theta(y_2|x)\frac{\partial^2 h(\pi_\theta,x,y_1,y_2)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}z(x,y_j)\right| \\
=&|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} \pi_\theta(y_1|x)\pi_\theta(y_2|x)(\delta_{y_1y_i}-\delta_{y_2y_i})(\delta_{y_1y_j}-\delta_{y_2y_j}) \\
&\cdot(\frac{1}{\tau^2}\sigma(\frac{1}{\tau}\log\frac{\pi_\theta(y_2|x)}{\pi_\theta(y_1|x)})\sigma(\frac{1}{\tau}\log\frac{\pi_\theta(y_1|x)}{\pi_\theta(y_2|x)}))z(x,y_j)| \\
\leq&\left|\sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i)\sum_{y_1,y_2\in\mathbb{Y}} \pi_\theta(y_1|x)\pi_\theta(y_2|x)(\delta_{y_1y_i}-\delta_{y_2y_i})(\delta_{y_1y_j}-\delta_{y_2y_j})\frac{1}{\tau^2}z(x,y_j)\right| \\
\leq&\frac{4}{\tau^2}\|z(x,\cdot)\|_2^2.
\end{aligned}
\tag{166}
$$

In summary, we have the upper bound of Eq.163:

$$\begin{aligned}\|\psi(x)\|_\infty &= \max_{x\in\mathbb{X}} |\psi(x)| = \max_{x\in\mathbb{X}} \left| \sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i) \frac{\partial^2 f_{\text{PRA}}(x,\theta)}{\partial\theta(x,y_i)\partial\theta(x,y_j)} z(x,y_j) \right| \\ &\leq \left(20\log(1+e^{\frac{d}{\tau}}) + \frac{16\epsilon_3}{\tau} + \frac{4}{\tau^2}\right) \|z(\cdot,\cdot)\|_2^2\end{aligned} \tag{167}$$

As for the second term $\left|\sum_{x\in\mathbb{X}} \mathcal{D}(x) \sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i) \frac{\partial^2 \sum_{y_1,y_2\in\mathbb{Y}} \pi_\theta(y_1|x)\pi_\theta(y_2|x)M(x,y_1,y_2)}{\partial\theta(x,y_i)\partial\theta(x,y_j)} z(x,y_j)\right|$ of Eq.157, we have:

$$\begin{aligned}&|\frac{\partial^2 \sum_{y_1,y_2\in\mathbb{Y}} \pi_\theta(y_1|x)\pi_\theta(y_2|x)M(x,y_1,y_2)}{\partial\theta(x,y_i)\partial\theta(x,y_j)}| \\ =&|\frac{\partial}{\partial\theta(x,y_j)}(\frac{\partial \sum_{y_1,y_2\in\mathbb{Y}} \pi_\theta(y_1|x)\pi_\theta(y_2|x)M(x,y_1,y_2)}{\partial\theta(x,y_i)})| \\ =&|\frac{\partial}{\partial\theta(x,y_j)}\left(\sum_{y_1,y_2\in\mathbb{Y}} \pi_\theta(y_1|x)\pi_\theta(y_2|x)(\delta_{y_1y_i}+\delta_{y_2y_i}-2\pi_\theta(y_i|x))M(x,y_1,y_2)\right)| \\ =&|\sum_{y_1,y_2\in\mathbb{Y}} \pi_\theta(y_1|x)\pi_\theta(y_2|x)(\delta_{y_1y_i}+\delta_{y_2y_i}-2\pi_\theta(y_i|x))(\delta_{y_1y_j}+\delta_{y_2y_j}-2\pi_\theta(y_j|x))M(x,y_1,y_2).\end{aligned} \tag{168}$$

Based on Lemma G.6. We have:

$$\left|\sum_{x\in\mathbb{X}} \mathcal{D}(x) \sum_{y_i,y_j\in\mathbb{Y}} z(x,y_i) \frac{\partial^2 \sum_{y_1,y_2\in\mathbb{Y}} \pi_\theta(y_1|x)\pi_\theta(y_2|x)M(x,y_1,y_2)}{\partial\theta(x,y_i)\partial\theta(x,y_j)} z(x,y_j)\right| \leq 16\log 2 \cdot \|z(\cdot,\cdot)\|_2^2 \tag{169}$$

This inequality is because we can expand the absolute value operation in the second-to-last line into nine terms, and these nine terms can be deduced to be less than $\|z(x,\cdot)\|_2^2$ or $2\|z(x,\cdot)\|_2^2$ using some existing conclusions. The relevant conclusions are: $\pi_\theta^\top z(x,\cdot) \leq \|\pi_\theta\|_1 \|z(x,\cdot)\|_\infty = \|z(x,\cdot)\|_\infty \leq \|z(x,\cdot)\|_2$, $\pi_\theta^\top z^2(x,\cdot) \leq \|z(x,\cdot)\|_2^2$, $\|\pi_\theta\|_\infty \leq 1$, $\|z^2(x,\cdot)\|_1 = \|z(x,\cdot)\|_2^2$ and $|(\pi_\theta^\top z(x,\cdot))^2| \leq \|z(x,\cdot)\|_2^2$.

Then Eq.156 is proved. Proof finished.