Faster WIND: Accelerating Iterative Best-of-N Distillation for LLM Alignment

Tong Yang* Jincheng Mei† Hanjun Dai† Zixin Wen* Shicong Cen*
Dale Schuurmans† Yuejie Chi* Bo Dai†

*Carnegie Mellon University
†Google Deepmind

October 2024

Abstract

Recent advances in aligning large language models with human preferences have corroborated the growing importance of best-of-N distillation (`BOND`). However, the iterative `BOND` algorithm is prohibitively expensive in practice due to the sample and computation inefficiency. This paper addresses the problem by revealing a unified game-theoretic connection between iterative `BOND` and self-play alignment, which unifies seemingly disparate algorithmic paradigms. Based on the connection, we establish a novel framework, **WIN** rate **D**ominance (`WIND`), with a series of efficient algorithms for regularized win rate dominance optimization that approximates iterative `BOND` in the parameter space. We provides provable sample efficiency guarantee for one of the `WIND` variant with the square loss objective. The experimental results confirm that our algorithm not only accelerates the computation, but also achieves superior sample efficiency compared to existing methods.

Keywords: Reinforcement learning from human feedback (RLHF), preference optimization, matrix game, sample efficiency

Contents

*Carnegie Mellon University; emails: `{tongyang,zixinw,shicongc,yuejiec}@andrew.cmu.edu`.
†Google DeepMind; emails: `{jcmei,hadai,schuurmans,bodai}@google.com`.

1 Introduction

Fine-tuning large language models (LLMs) to align with human preferences has become a critical challenge in artificial intelligence to ensure the safety of their deployment. Reinforcement Learning from Human Feedback (RLHF) has emerged as a dominant approach, significantly improving LLM performance as demonstrated by InstructGPT [Ouyang et al., 2022] and subsequent works. RLHF combines reward modeling to quantify human preferences and RL fine-tuning to adjust the LLM's output distribution, enhancing desired responses while suppressing unfavorable ones. While RLHF has shown promising results, it comes with significant extra post-training cost, and the aligned LLM may exhibit performance degeneration due to the alignment tax [Askell et al., 2021, OpenAI, 2023].

Alternatively, best-of-N (BoN) sampling has emerged as a simple and surprisingly effective technique to obtain high-quality outputs from an LLM [Stiennon et al., 2020]. In BoN sampling, multiple samples are drawn from an LLM, ranked according to a specific attribute, and the best one is selected. This simple approach can improve model outputs without the need for extensive fine-tuning, offering a potentially more efficient path to alignment. Building upon the success of BoN sampling, a few works explore the iterative variants of this approach [Dong et al., 2023, Sessa et al., 2024]. Iterative BoN applies the sampling and selection process repeatedly, potentially leading to even better alignments with human preferences.

However, BoN incurs significant computational overhead due to making n inference calls to generate one output, especially when n is high. To mitigate the high inference cost of (iterative) BoN, Sessa et al. [2024] proposed a distillation algorithm, best-of-N distillation (`BOND`), to train a new model emulating the output of iterative BoN. However, this approach also has a high training cost, due to the need of collecting multiple samples in each round of distillation, leading to a major bottleneck for wider adoption.

Given the growing importance and significance of the iterative BoN approach, it raises new questions about its theoretical properties, practical implementation, and relationship to established methods like RLHF. In this paper, we delve into the theoretical foundations and practical applications of iterative BoN sampling for LLM alignment. We address the following question:

What are the limiting points of iterative BoN, and can we design faster algorithms to find them?

1.1 Contributions

We provide comprehensive answers to these questions through the following key contributions:

- We introduce a general algorithmic framework for iterative BoN distillation, possibly with a slow moving anchor, and uncover its limiting point corresponds to the Nash equilibrium of a (regularized) two-player min-max game optimizing the logarithm of the expected win rate. This offers a fresh game-theoretic interpretation that is unavailable before.
- We show that the **WIN** rate **D**ominance (`WIND`) policy, which has a higher chance of winning against any other policy, solves the minmax game of win rate introduced in RLHF [Swamy et al., 2024, Munos et al., 2023], and approximates the iterative BoN's limiting point.

- We propose a novel algorithm framework, `WIND`, to find the win rate dominance policy with flexible loss configurations, and demonstrate it exhibits improved sample and computation efficiency, compared to prior work while maintaining provable convergence guarantees.
- We conduct extensive experiments to evaluate the performance of `WIND`, demonstrating competitive or better performance against state-of-the-art alignment methods such as `J-BOND` across various benchmarks, highlighting its efficiency especially in the sampling process and training cost.

1.2 Related work

RLHF and LLM alignment. Reinforcement Learning from human feedback (RLHF) is a effective approach to train AI models to produce outputs that aligns to human value and preference [Christiano et al., 2017, Stiennon et al., 2020, Nakano et al., 2021]. Recently, RLHF has become the most effective approach to align language models [Ouyang et al., 2022, Bai et al., 2022]. The famous InstructGPT [Ouyang et al., 2022] approach eventually led to the groundbreaking ChatGPT and GPT-4 [OpenAI, 2023]. A variety of RLHF methods have been proposed, including the direct preference optimization [Rafailov et al., 2024] and many other variants [Zhao et al., 2023, Yuan et al., 2023b, Azar et al., 2024, Meng et al., 2024, Xu et al., 2024, Ethayarajh et al., 2024, Tang et al., 2024], to name a few, which directly learns from the preference data without RL finetuning. Furthermore, value-incentive preference optimization [Cen et al., 2024] has been proposed to implements the provably optimistic principle and pessimistic principle for exploration-exploitation tradeoff in a practical way.

RLHF via self-play. One line of RLHF methods investigate self-play optimization for unregularized and regularized two-player win rate games, respectively [Swamy et al., 2024, Munos et al., 2023]. Wu et al. [2024b] introduced a scalable self-play algorithm for win rate games, enabling efficient fine-tuning of LLMs, see also Rosset et al. [2024], Zhang et al. [2024] among others.

Best-of-N and `BOND`. BoN has empirically shown impressive reward-KL trade-off [Nakano et al., 2021, Gao et al., 2023], which has been theoretically investigated by Gui et al. [2024] from the win rate maximization perspective. Beirami et al. [2024] analyzed the KL divergence between the BoN policy and the base policy, and Yang et al. [2024a] studied the asymptotic properties of the BoN policy. The recent work Gui et al. [2024] also proposed a method to use both best-of-N and worst-of-N to train language models. Sessa et al. [2024] introduced `BOND` and `J-BOND` to train language models to learn BoN policies. Amini et al. [2024] proposed vBoN which is equivalent to `BOND`. However, there is no existing work for characterizing the properties of iterative BoN yet.

Notation. We let $[n]$ denote the index set $\{1,\ldots,n\}$. Let I_n denote the $n\times n$ identity matrix, and inner product in Euclidean space $\mathbb{R}^n$ by $\langle\cdot,\cdot\rangle$. We let $\mathrm{supp}(\rho)$ denote the support set of the distribution ρ, and $\mathrm{relint}(\mathcal{C})$ represents the relative interior of set $\mathcal{C}$. We defer all the proofs to the appendix.

2 Preliminaries

2.1 RLHF: reward versus win rate

We consider the language model $\pi_\theta(\cdot)$ as a policy, where $\theta\in\Theta$ denotes its parameters, and Θ the compact parameter space. Given a prompt $x\in\mathcal{X}$, the policy generates an answer $y\in\mathcal{Y}$ according to the conditional distribution $\pi_\theta(\cdot|x)$. For notation simplicity, we drop the subscript θ when it is clear from the context. We let $\Delta_{\mathcal{Y}}$ denote the simplex over $\mathcal{Y}$. We let $\Delta_{\mathcal{Y}}^{\mathcal{X}}$ denote the space of policies as follows:

$$\Delta_{\mathcal{Y}}^{\mathcal{X}} \coloneqq \left\{\pi=[\pi(\cdot|x)]_{x\in\mathcal{X}} \mid \pi(\cdot|x)\in\Delta_{\mathcal{Y}}, \forall x\in\mathcal{X}\right\}.$$

In practice, RLHF optimize the policy model against the reward model while staying close to a reference model π_{ref}. There are two metrics being considered: reward and win rate.

Reward maximization. Suppose there is a reward model $r(x, y) : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$, which produce a scalar reward given a prompt x and a response y. RLHF aims to maximize the KL-regularized value function, given a reference model π_{ref}:

$$V_{\text{rm}}(\pi) = \mathop{\mathbb{E}}_{x \sim \rho, y \sim \pi(\cdot|x)} [r(x, y)] - \beta D_{\text{KL}}\left(\pi \| \pi_{\text{ref}}\right), \tag{1}$$

where

$$D_{\text{KL}}\left(\pi_1 \| \pi_2\right) = \mathbb{E}_{x \sim \rho}\left[\text{KL}\left(\pi_1(\cdot|x) \| \pi_2(\cdot|x)\right)\right]$$

is the KL divergence between policies π_1 and π_2, with ρ being the distribution of prompts. Here, $\beta \geq 0$ is a hyperparameter that balances the reward and the KL divergence. Without loss of generality, we assume $\text{supp}(\rho)$ is $\mathcal{X}$ throughout the paper.

Win rate maximization. Another scheme of RLHF aims to maximize the KL-regularized win rate against the reference model [Gui et al., 2024]. Given a reward model r, a preference model $P_x : \mathcal{Y} \times \mathcal{Y} \to \{0, 1/2, 1\}$ can be defined as:

$$P_x(y, y') \coloneqq \begin{cases} 1, & \text{if } r(x, y) > r(x, y'), \\ 1/2, & \text{if } r(x, y) = r(x, y'), \\ 0, & \text{if } r(x, y) < r(x, y'). \end{cases} \tag{2}$$

Given a policy pair π, π', the *win rate* of π over π' is thus [Swamy et al., 2024]

$$\begin{aligned} P(\pi \succ \pi') &\coloneqq \mathop{\mathbb{E}}_{\substack{x \sim \rho, y \sim \pi(\cdot|x), \\ y' \sim \pi'(\cdot|x)}} P_x(y, y') \\ &= \mathbb{E}_{x \sim \rho} \pi^{\top}(\cdot|x) P_x(\cdot, \cdot) \pi'(\cdot|x). \end{aligned} \tag{3}$$

The KL-regularized win rate maximization objective is defined as [Gui et al., 2024]:

$$V_{\text{wr}}(\pi) \coloneqq P(\pi \succ \pi_{\text{ref}}) - \beta D_{\text{KL}}\left(\pi \| \pi_{\text{ref}}\right). \tag{4}$$

The win rate maximization is more aligned with evaluation metric adopted in common benchmarks, and further, can be carried out without explicit reward models as long as the preference model P_x is well-defined.

2.2 Best-of-N distillation

Best-of-N (BoN) is a simple yet strong baseline in RLHF. Given a reward model r and a prompt x, BoN samples n i.i.d. responses $y_1, y_2, ..., y_n$ from the policy $\pi(\cdot|x)$ and select the response

$$y = \arg \max_{y \in y_1, y_2, \ldots, y_n} r(x, y), \quad y_1, \ldots, y_n \sim \pi(\cdot|x)$$

with the highest reward. We call $\pi^{(n)}$ the BoN policy which selects the sample with the highest reward given n samples i.i.d. drawn from π. Gui et al. [2024] shows that for any fixed small $\beta > 0$, $\pi_{\text{ref}}^{(n)}$ (approximately) maximizes $V_{\text{wr}}(\cdot)$ for properly chosen n. While BoN is widely used in practice [Beirami et al., 2024, Gao et al., 2023, Wang et al., 2024], yet can be quite expensive in terms of the inference cost for drawing n samples. Hence, BoN distillation (`BOND`) [Sessa et al., 2024] is developed to approximate the BoN policy $\pi^{(n)}$ through fine-tuning from some reference policy $\pi_{\text{ref}}(\cdot|x)$, which can be updated iteratively via an exponential moving average [Sessa et al., 2024].

3 A Unified Game-Theoretic View

In this section, we present a game-theoretic understanding of iterative BoN, which allows us to connect it to existing game-theoretic RLHF approaches under a *win rate maximization* framework.

3.1 Iterative BoN as game solving

Iterative BoN. Due to the success of BoN sampling, its iterative version has also been studied [Dong et al., 2023, Sessa et al., 2024], where BoN is performed iteratively by using a moving anchor as the reference policy. To understand its property in generality, we introduce the iterative BoN method in Algorithm 1 that encapsulates iterative BoN methods with or without moving reference model, which we call the *mixing* and *no-mixing* case.

Algorithm 1 Iterative BoN

```
Input: reference policy π_ref, iterate number T, Best-of-N parameter n, boolean value Mixing.
Optional: mixing rates α_1 > 0, α_2 ≥ 0 such that α_1 + α_2 ≤ 1.
Initialization: π_0 ← π_ref.
for t = 0, 1, ···, T − 1 do
  π_t^(n) ← Best-of-N(π_t, n).
  if Mixing then
    π_{t+1} ∝ (π_t^(n))^{α_1} π_t^{α_2} π_ref^{1−α_1−α_2};
  else if not Mixing then
    π_{t+1} ← π_t^(n).
  end if
end for
Return π_T.
```

Algorithm 1 demonstrates these two cases. In the *mixing* case, we obtain new policies by mixing the BoN policy $\pi_t^{(n)}$, π_t and π_{ref} at each iteration with mixing rates α_1, α_2. In the *no-mixing* case, the algorithm simply returns the BoN policy $\pi_t^{(n)}$ as π_{t+1} for the next iteration. We will provide some theoretical guarantees for both cases, using the following game-theoretic framework.

Game-theoretic perspective. We show that iterative BoN is implicitly solving the following game. Define a preference matrix $\overline{P}_x$ at $x \in \mathcal{X}$ of size $|\mathcal{Y}| \times |\mathcal{Y}|$ by

$$\overline{P}_x(y, y') := \begin{cases} 1, & \text{if } r(x,y) \geq r(x,y'), \\ 0, & \text{if } r(x,y) < r(x,y'). \end{cases} \tag{5}$$

Define further $f_\beta : \Delta_{\mathcal{X}}^{\mathcal{Y}} \times \Delta_{\mathcal{X}}^{\mathcal{Y}} \to \mathbb{R}$ as

$$f_\beta(\pi, \pi') := \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\in\pi(\cdot|x)}} \left[\log \mathop{\mathbb{E}}_{y'\in\pi'(\cdot|x)} \overline{P}_x(y \succeq y')\right] - \beta D_{\text{KL}}\left(\pi \| \pi_{\text{ref}}\right). \tag{6}$$

We introduce the following symmetric two-player *log-win-rate* game:

$$\begin{cases} \pi_1 = \arg\max_\pi f_\beta(\pi, \pi_2), \\ \pi_2 = \arg\max_\pi f_\beta(\pi, \pi_1). \end{cases} \tag{7}$$

Let $\overline{\pi}_\beta^\star$ be a Nash equilibrium of the log-win-rate game (7), which satisfies the fixed-point characterization:

$$\overline{\pi}_\beta^\star \in \arg\max_\pi \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\in\pi(\cdot|x)}} \left[\log \mathop{\mathbb{E}}_{y'\in\overline{\pi}_\beta^\star(\cdot|x)} \overline{P}_x(y \succeq y')\right] - \beta D_{\text{KL}}\left(\pi \| \pi_{\text{ref}}\right). \tag{8}$$

Now we present our Theorem 1, which guarantees the convergence to solutions for the above game under Algorithm 1.

Theorem 1 (Iterative BoN solves game (7))**.** *Let $\pi_{\text{ref}} \in$ relint$\left(\Delta_{\mathcal{Y}}^{\mathcal{X}}\right)$ and $n \geq 2$ in Algorithm 1. Then $\pi_\infty := \lim_{T\to\infty} \pi_T$ exists, and (π_∞, π_∞) is a Nash equilibrium of the log-win-rate game (7) when:*

1. *(no-mixing) $\alpha_1 = 1, \alpha_2 = 0$ for $\beta = 0$;*

2. *(mixing)* $\alpha_1 = \frac{\eta}{(1+\beta\eta)(n-1)}$, $\alpha_2 = \frac{n-1-\eta}{(1+\beta\eta)(n-1)}$ *for any* $\beta, \eta > 0$.

In the no-mixing case, we can show that π_T obtained by Algorithm 1 converges to the equilibrium of the unregularized log-win-rate game. In the mixing case, we show that with proper choice of mixing rates, Algorithm 1 solves the regularized log-win-rate game. To the best of our knowledge, this is the first game-theoretic understanding of *iterative* BoN using a general preference model.

3.2 Self-play and win rate dominance

The log-win-rate game (7) is a non-zero-sum game that may be challenging to optimize: the function f_β is not convex-concave, the Nash equilibria may not be unique, and the log term introduces nonlinearity, which induces difficulty in estimation. Therefore, we seek a good alternative to the log-win-rate game that maintains its core properties while being more amenable to optimization.

Specifically, we now consider the following alternative two-player *win-rate* game:

$$\max_{\pi} \min_{\pi'} P(\pi \succ \pi') - \beta D_{\mathrm{KL}}\left(\pi \| \pi_{\mathrm{ref}}\right) + \beta D_{\mathrm{KL}}\left(\pi' \| \pi_{\mathrm{ref}}\right), \tag{9}$$

which eliminates the nonlinearity in reward, and has been recently studied by Swamy et al. [2024], Wu et al. [2024b] for $\beta = 0$ and Munos et al. [2023] for $\beta > 0$.

The following proposition guarantees the game (9) is well-defined and is equivalent to the following fixed point problem:

$$\begin{aligned} \pi^\star_\beta &\in \arg\max_{\pi} P(\pi \succ \pi^\star_\beta) - \beta D_{\mathrm{KL}}\left(\pi \| \pi_{\mathrm{ref}}\right) \\ &= \mathop{\mathbb{E}}_{\substack{x\sim\rho, y\sim\pi(\cdot|x),\\ y'\sim\pi^\star_\beta(\cdot|x)}} \left[P_x(y, y') - \beta \log \frac{\pi(y|x)}{\pi_{\mathrm{ref}}(y|x)} \right]. \end{aligned} \tag{10}$$

Proposition 1 (existence of $\pi^\star_\beta$)**.** *$\pi^\star_\beta$ exists and $(\pi^\star_\beta, \pi^\star_\beta)$ is the Nash equilibrium of the minmax game (9). Moreover, when $\beta > 0$, $(\pi^\star_\beta, \pi^\star_\beta)$ is the unique Nash equilibrium.*

Win rate dominance. The fixed-point equation (10) identifies a policy with a higher winning probability against any other policy. For $\beta = 0$, $\pi^\star_0$ satisfies $P(\pi \succ \pi^\star_0) \le 1/2$ for any π, ensuring it outperforms other policies. When $\beta > 0$, the KL divergence term encourages $\pi^\star_\beta$ to remain close to π_{ref} while maintaining a high win rate. We term (10) the ***Win** rate **D**ominance* (`WIND`) optimization problem.

3.3 Connecting iterative BoN with `WIND`

Due to the monotonicity of $\log(\cdot)$, it is natural to believe the win rate game and the log-win-rate game beneath iterative BoN are connected. We establish the novel relationship rigorously, which allows a unifying game-theoretic view for many existing algorithms. We define a constant $c_\beta \in (0, +\infty]$ related to π_{ref}:

$$c_\beta := \min_{\substack{x\in\mathcal{X},\\ y\in\mathcal{Y}\setminus\mathcal{Y}^\star(x)}} \left\{ \frac{\sum_{y^\star\in\mathcal{Y}^\star(x)} \pi_{\mathrm{ref}}(y^\star|x)}{4\max\left\{\log \frac{\pi_{\mathrm{ref}}(y|x)}{\max_{y^\star\in\mathcal{Y}^\star(x)} \pi_{\mathrm{ref}}(y^\star|x)}, 0\right\}} \right\}, \tag{11}$$

where $\mathcal{Y}^\star(x) := \arg\max_{y\in\mathcal{Y}} r(x, y)$ is the set of optimal responses for each $x \in \mathcal{X}$. We now demonstrate the relationship between the equilibria set of the log-win-rate game $\overline{\pi}^\star_\beta$ and the win-rate game $\pi^\star_\beta$.

Theorem 2 (relationship between two games (informal))**.** *Let $\pi_{\mathrm{ref}} \in relint\left(\Delta^{\mathcal{X}}_{\mathcal{Y}}\right)$ and $n \ge 2$ in Algorithm 1. Then*

- *When $\beta = 0$, $\overline{\pi}^\star_\beta$ is also a solution to (10);*
- *When $\beta \in (0, c_\beta)$ where c_β is defined in (11), for all $x \in \mathcal{X}$, $\overline{\pi}^\star_\beta$ satisfies*

$$\left\|\overline{\pi}^\star_{\beta,x} - \pi^\star_{\beta,x}\right\|_1 \le 4(|\mathcal{Y}| - |\mathcal{Y}^\star(x)|) \exp\left(\frac{-\sum_{y^\star\in\mathcal{Y}^\star(x)} \pi_{\mathrm{ref}}(y^\star|x)}{4\beta}\right) \to 0 \text{ as } \beta \to 0. \tag{12}$$

Theorem 2 shows when $\beta = 0$, both games have the solution $\overline{\pi}^\star_0$. For small positive β, the ℓ_1 distance between the solutions of the two games is bounded by a term that decreases exponentially with $1/\beta$. We verify Theorem 2 empirically on contextual bandits in Section 5.1.

This result provides theoretical justification for using iterative BoN as an approximation to `WIND`, especially when β is small. More importantly, it paves a way for efficient algorithm to `WIND`, bypassing the log operator in the win-rate game.

4 Faster WIND

Based on the understanding of the connection between log-win-rate game and win-rate game, in this section, we propose a new sample-efficient algorithm for finding the `WIND` solution in (9), which includes two ingredients: **(i)** identifying a memory-efficient, exact policy optimization algorithm with linear last-iterate convergence [Sokota et al., 2023], and **(ii)** developing a series of sample-efficient algorithms with flexible loss functions and finite-time convergence guarantee. With slight abuse of terminology, we shall refer to our algorithmic framework `WIND`.

4.1 Exact policy optimization with last-iterate linear convergence

Recognizing that (9) is an KL-regularized matrix game, there are many existing algorithms that can be applied to find $\pi^\star_\beta$. Nonetheless, it is desirable to achieve fast last-iterate convergence with a small memory footprint. This is especially important in LLM optimization, for the memory efficiency. For example, extra-gradient algorithms (e.g., Korpelevich [1976], Popov [1980], Cen et al. [2021])—although fast-convergent—are expected to be expensive in terms of memory usage due to the need of storing an additional extrapolation point (i.e., the LLM) in each iteration.

It turns out that the magnetic mirror descent algorithm in Sokota et al. [2023], which is proposed to solve an equivalent variational inequality formulation to (9), meets our consideration. We present a tailored version of this algorithm in Algorithm 2, and state its linear last-iterate convergence in Theorem 3.

Algorithm 2 `WIND` (exact gradient, adapted from Sokota et al. [2023] tailored for our setting)

1: **Input:** reference policy π_{ref}, initial policy $\pi^{(0)}$, regularization coefficient $\beta > 0$, learning rate $\eta > 0$.
2: **for** $t = 0, 1, \cdots$ **do**
3: Update $\pi(\cdot|x)$ for all $x \in \mathcal{X}$:

$$\pi^{(t+1)}(y|x) \propto (\pi^{(t)}(y|x))^{\frac{1}{1+\beta\eta}} (\pi_{\text{ref}}(y|x))^{\frac{\beta\eta}{1+\beta\eta}} \exp\left(\frac{\eta}{1+\beta\eta}\mathbb{E}_{y'\sim\pi^{(t)}(\cdot|x)} P_x(y, y')\right) \tag{13}$$

4: **end for**

Theorem 3 (Linear last-iterate convergence of Algorithm 2, Sokota et al. [2023])**.** *Assume $\beta > 0$ and $\pi^{(0)}, \pi_{\text{ref}} \in relint(\Delta^{\mathcal{X}}_{\mathcal{Y}})$. When the learning rate $\eta \in (0, \beta]$, $\pi^{(t)}$ in Algorithm 2 satisfies:*

$$D_{\text{KL}}\big(\pi^\star_\beta||\pi^{(t)}\big) \le \left(\frac{1}{1+\eta\beta}\right)^t D_{\text{KL}}\left(\pi^\star_\beta\|\pi^{(0)}\right). \tag{14}$$

Remark 1. *We note that when $\beta = 0$, the update rule (13) recovers [Swamy et al., 2024, Algorithm 1]. When $\beta > 0$, the update rule in (13) is different from that of Munos et al. [2023], which is*

$$\pi^{(t+1)}(y|x) \propto \widetilde{\pi}^{(t)}(y|x) \cdot \exp\left(\eta \mathbb{E}_{y'\sim\widetilde{\pi}^{(t)}(\cdot|x)} P_x(y, y')\right),$$

where $\widetilde{\pi}^{(t)}$ is a mixed policy defined as

$$\widetilde{\pi}^{(t)}(y|x) \propto \left(\pi^{(t)}(y|x)\right)^{1-\eta\beta} (\pi_{\text{ref}}(y|x))^{\eta\beta}.$$

As such, it requires extra memory to store $\widetilde{\pi}^{(t)}$. Moreover, Munos et al. [2023] shows a slower rate of $\mathcal{O}(1/T)$, whereas Algorithm 2 admits linear convergence.

4.2 Sample-efficient algorithm

We now derive practical sample-efficient methods for approximating the exact update (13) of `WIND` in the parameter space Θ of the policy π_θ, $\theta \in \Theta$. For exposition, we use ϕ_θ to denote the logits before softmax, i.e.,

$$\pi_\theta = \mathsf{softmax} \circ \phi_\theta,$$

where $\mathsf{softmax}(x)_i \coloneqq e^{x_i} / \sum_j e^{x_j}$ is the softmax function.

We consider the existence of reward model approximation error, i.e., we may use an inaccurate judger $\widehat{P}_x$, which is an approximation of P_x. For example, instead of training a reward model, we could use an LLM $\widehat{P}$ as a judger to directly judge if response y is better than y' or not given a prompt x, and use $\widehat{P}_x$ as an approximation of P_x.

Algorithm derivation with the squared risk. Let $\theta_t, \theta_{\text{ref}}$ denote the parameters of $\pi^{(t)}$ and π_{ref} in Algorithm 2, respectively. We rewrite the update rule (13) as

$$\phi_{\theta_{t+1}}(y|x) = \frac{1}{1+\beta\eta}\phi_{\theta_t}(y|x) + \frac{\beta\eta}{1+\beta\eta}\phi_{\theta_{\text{ref}}}(y|x) + \frac{\eta}{1+\beta\eta}\mathbb{E}_{y'\sim\pi_{\theta_t}(\cdot|x)}P_x(y,y') + Z_t(x) \tag{15}$$

for some function $Z_t : \mathcal{X} \to \mathbb{R}$. We define a proxy $\varphi_t : \mathcal{X} \times \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ using the empirical win-rate as

$$\varphi_t(x,y,y') \coloneqq \frac{1}{1+\beta\eta}\phi_{\theta_t}(y|x) + \frac{\beta\eta}{1+\beta\eta}\phi_{\theta_{\text{ref}}}(y|x) + \frac{\eta}{1+\beta\eta}\widehat{P}_x(y,y') + Z_t(x), \tag{16}$$

Our observation is that the update (15) of $\phi_{\theta_t}(y|x)$ is approximating the conditional expectation of φ_t, which is

$$\psi_t(x,y) \coloneqq \mathbb{E}_{y'\sim\pi_{\theta_t}(\cdot|x)}[\varphi_t(x,y,y')|x,y], \ \forall(x,y).$$

Furthermore, this conditional expectation has the lowest risk, due to the following lemma:

Lemma 1 (Conditional mean minimizes the square loss)**.** *For any two random variables u, v, we have*

$$\mathbb{E}_{u,v}\left[(v - \mathbb{E}_v(v|u))^2\right] \le \mathbb{E}_{u,v}\left[(v - g(u))^2\right] \tag{17}$$

for any function g. In particular, the equality holds if and only if $g(u) = \mathbb{E}_v(v|u)$ almost everywhere on the support of the distribution of u.

To invoke Lemma 1, we assume the LLM is expressive enough, such that ψ_t can be represented by ϕ_θ:

Assumption 1 (expressive power)**.** *For any $t \in \mathbb{N}$, there exists $\theta^\star_{t+1} \in \Theta$ such that*

$$\forall(x,y) \in \mathcal{X}\times\mathcal{Y}: \quad \phi_{\theta^\star_{t+1}}(y|x) = \psi_t(x,y). \tag{18}$$

Note that $\text{supp}(\rho) = \mathcal{X}$, $\text{supp}(\pi^{(t)}(\cdot|x)) = \mathcal{Y}$ for all $x \in \mathcal{X}$, $t \in \mathbb{N}$. Thus under Assumption 1, by Lemma 1 we know that for all t, $\theta^\star_{t+1} \in \Theta$ satisfies (18) if and only if

$$\theta^\star_{t+1} \in \arg\min_{\theta\in\Theta} R_t(\theta), \tag{19}$$

where we define *the squared expected risk* at the t-th iteration $R^{\text{sq}}_t(\theta)$ as

$$R^{\text{sq}}_t(\theta) \coloneqq \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y,y'\sim\pi_{\theta_t}(\cdot|x)}}\left[\left(\varphi_t(x,y,y') - \phi_\theta(y|x)\right)^2\right].$$

In implementation, at each iteration t, we shall approximate $\theta^\star_{t+1}$ by minimizing the *empirical* risk: we sample $x_i^{(t)} \sim \rho$, $y_i^{(t)}, y_i'^{(t)} \sim \pi_{\theta_t}(\cdot|x_i^{(t)})$, $i \in [M]$, and compute θ_{t+1} by minimizing the empirical risk $R^{\text{sq}}_{t,M}(\theta)$ defined as

$$R^{\text{sq}}_{t,M}(\theta) \coloneqq \frac{1}{M}\sum_{i=1}^{M}\left(\varphi_t(x_i^{(t)}, y_i^{(t)}, y_i'^{(t)}) - \phi_\theta(y_i^{(t)}|x_i^{(t)})\right)^2. \tag{SQ}$$

We summarize the update procedure in Algorithm 3.

Algorithm 3 `WIND` (sample-efficient version)

Input: reference parameter θ_{ref}, initial parameter θ_0, regularization coefficient $\beta > 0$, learning rate $\eta > 0$, training set $\mathcal{D}$, iteration number $T \in \mathbb{N}_+$, sampling number $M \in \mathbb{N}_+$.
for $t = 0, 1, \cdots, T-1$ **do**
    Sample $x_i^{(t)} \sim \rho$, $y_i^{(t)}, y_i'^{(t)} \sim \pi_{\theta_t}(\cdot|x_i^{(t)})$, $i \in [M]$.
    $\theta_{t+1} \leftarrow \arg\min_{\theta\in\Theta} R_{t,M}(\theta)$.   ▷ $R_{t,M}$ could be $R_{t,M}^{\text{sq}}$, $R_{t,M}^{\text{kl}}$, $R_{t,M}^{\text{nce}}$, etc.
end for
Return θ_T.

Alternative risk functions. By utilizing different variational forms, we could derive objectives different from (SQ). For illustration, we provide two alternatives of $R_{t,M}^{\text{sq}}(\theta)$ by using the KL divergence and NCE loss, respectively (see Appendix A for derivations):

$$R_{t,M}^{\text{kl}}(\theta) := -\frac{1}{M}\sum_{i=1}^{M}\left[\mathbb{1}_{\{v_i=1\}}\log\zeta_\theta(x,y) + \mathbb{1}_{\{v_i=0\}}\log(1-\zeta_\theta(x,y))\right], \tag{KL}$$

and

$$R_{t,M}^{\text{nce}}(\theta) := -\frac{1}{M}\sum_{i=1}^{M}\Big[\left(\mathbb{1}_{\{v_i=1\}} + \mathbb{1}_{\{v_i'=0\}}\right)\log\frac{\zeta_\theta(x_i,y_i)}{\zeta_\theta(x_i,y_i)+p} + \left(\mathbb{1}_{\{v_i=0\}} + \mathbb{1}_{\{v_i'=1\}}\right)\log\frac{p}{\zeta_\theta(x_i,y_i)+p}\Big], \tag{NCE}$$

where $v_i \sim \mathsf{Ber}(\widehat{P}_{x_i}(y_i, y_i'))$, $v_i' \sim \mathsf{Ber}(p)$, $p \in (0,1)$ is a hyperparameter, ζ_θ is defined as

$$\zeta_\theta(x,y) = \frac{1+\beta\eta}{\eta}\phi_\theta(y|x) - \frac{1}{\eta}\phi_{\theta_t}(y|x) - \beta\phi_{\theta_{\text{ref}}}(y|x) - \frac{1+\beta\eta}{\eta}Z_t(x),$$

and $\mathbb{1}_{\{A\}}$ is the indicator function that equals 1 if A is true and 0 otherwise.

When we use the regression objective (SQ), our `WIND` algorithm shares a similar form to SPPO [Wu et al., 2024b]. However, `WIND` differs from SPPO in the following aspects: (i) we solve the regularized game with the KL regularization term $\beta D_{\text{KL}}(\pi' \| \pi_{\text{ref}})$. This term is crucial in practice and is also considered in other iterative `BOND` methods [Dong et al., 2023, Sessa et al., 2024]; (ii) our sampling scheme is more sample-efficient: in SPPO, for each x_i, they sample K responses $\{y_{i,j}\}_{j\in[K]}$ to estimate $\mathbb{E}_{y'\sim\pi_{\theta_t}(\cdot|x_i)}[P_{x_i}(y_{i,j}, y')]$ by $\frac{1}{K}\sum_{k=1}^{K} P_{x_i}(y_{i,j}, y_{i,k})$ for each $j \in [K]$. On the other hand, Lemma 1 implies estimating the conditional mean with multiple samples is unnecessary and for each x_i, sampling two responses y_i and y_i' is enough; (iii) we allow different risk functions beyond the squared loss.

4.3 Convergence analysis

We provide a finite-sample complexity guarantee for Algorithm 3 when the risk $R_{t,M} = R_t^{\text{sq}}$. Our results could be easily extended to other risks. Here we consider the existence of reward model approximation error, i.e., we may use an inaccurate judger $\widehat{P}_x$ as an approximation of P_x. For example, instead of training a reward model, we could use an LLM $\widehat{P}$ as a judger to directly judge if response y is better than y' or not given a prompt x, and use $\widehat{P}_x$ as an approximation of P_x.

We define the model approximation error δ_P as

$$\delta_P := \max_{x\in\mathcal{X}, y, y'\in\mathcal{Y}} \left| P_x(y,y') - \widehat{P}_x(y,y') \right|. \tag{20}$$

We require the following assumptions to prove the convergence of Algorithm 3. The first assumes ϕ_θ is differentiable and Θ, Z_t is bounded.

Assumption 2 (differentiability and boundedness)**.** *The parameter space Θ is compact, $\phi_\theta(y|x)$ is differentiable w.r.t. θ for any $(x,y) \in \mathcal{X}\times\mathcal{Y}$, and Z_t in (15) is uniformly bounded, i.e., $\exists Z \geq 0$ such that $|Z_t(x)| \leq Z$ for all $x \in \mathcal{X}$ and $t \in \mathbb{N}$.*

Assumption 2 guarantees the (uniform) boundedness of ϕ_θ. Especially, there exists $L_0 > 0$ such that for any $\theta, \theta' \in \Theta$ and $(x, y) \in \mathcal{X} \times \mathcal{Y}$, we have

$$|\phi_\theta(y|x) - \phi_{\theta'}(y|x)| \le L_0. \tag{21}$$

Assumption 2 also guarantees there exist $L, C > 0$ such that for all $x \in \mathcal{X}, y, y' \in \mathcal{Y}, \theta \in \Theta$ and t, we have

$$\left\| \nabla_\theta \left[\left(\varphi_t(x, y, y') - \phi_\theta(y|x) \right)^2 \right] \right\|_2 \le L. \tag{22}$$

and

$$\left(\varphi_t(x, y, y') - \phi_\theta(y|x) \right)^2 \le C. \tag{23}$$

The next assumption controls the concentrability coefficient, which is commonly used in the RL literature, see Yuan et al. [2023a], Munos [2003, 2005], Munos and Szepesvári [2008], Yang et al. [2023] for example.

Assumption 3 (concentrability coefficient)**.** *For Algorithm 3, there exists finite $C_\pi > 0$ such that for all $t \in \mathbb{N}$ and $x \in \mathcal{X}$, we have*

$$\mathbb{E}_{y \sim \pi_{\mathrm{ref}}(\cdot|x)} \left[\left(\frac{\pi^\star_\beta(y|x)}{\pi_{\mathrm{ref}}(y|x)} \right)^2 \right] \le C_\pi \qquad \textit{and} \qquad \mathbb{E}_{y \sim \pi_{\mathrm{ref}}(\cdot|x)} \left[\left(\frac{\pi_{\theta_{t+1}}(y|x)}{\pi_{\mathrm{ref}}(y|x)} \right)^2 \right] \le C_\pi.$$

We define

$$C_1 := \exp\left(\frac{2}{\beta} \left(\delta_P + \frac{1 + \beta\eta}{\eta} L_0 + 1 \right) \right) C_\pi. \tag{24}$$

We also assume for every t, the expected risk R_t and empirical risk $R_{t,N}$ both satisfy Polyak-Łojasiewicz (PL) condition, which has been proven to hold for over-parameterized neural networks including transformers [Liu et al., 2022, Wu et al., 2024a, Yang et al., 2024b].

Assumption 4 (Polyak-Łojasiewicz condition)**.** *For all $t \in \mathbb{N}$, risk R_t and empirical risk $R_{t,M}$ both satisfy Polyak-Łojasiewicz condition with parameter $\mu > 0$, i.e., for all $t \in \mathbb{N}$ and $\theta \in \Theta$, we have*

$$\frac{1}{2} \left\| \nabla_\theta R_t(\theta) \right\|_2^2 \ge \mu \left(R_t(\theta) - R_t(\theta^\star_{t+1}) \right)$$

and

$$\frac{1}{2} \left\| \nabla_\theta R_{t,N}(\theta) \right\|_2^2 \ge \mu \left(R_{t,M}(\theta) - R_{t,M}(\theta_{t+1}) \right).$$

Remark 2 (Assumption 4 is satisfied with linear function approximation)**.** *We consider a special case where $\phi_\theta(y|x) = \phi(x, y)^\top \theta$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, where $\phi(x, y)$ are the feature maps. If for all $t \in \mathbb{N}$, we have*

$$\mathbb{E}_{x \sim \rho, y \sim \pi_{\theta_t}(\cdot|x)} \left[\phi(x, y) \phi(x, y)^\top \right] \ge \frac{\mu}{2}$$

and

$$\frac{1}{M} \sum_{i=1}^{M} \phi(x_i^{(t)}, y_i^{(t)}) \phi(x_i^{(t)}, y_i^{(t)})^\top \ge \frac{\mu}{2},$$

then it's straightforward to verify that R_t and $R_{t,M}$ are both μ-strongly convex, which indicates Assumption 4 holds [Karimi et al., 2016].

The following theorem gives the convergence of Algorithm 3.

Theorem 4 (Convergence of Algorithm 3)**.** *Let $\theta_0 = \theta_{\mathrm{ref}}$ and $\eta \in (0, \beta]$ in Algorithm 3. Under Assumption 1,2,3,4, for any $T \in \mathbb{N}$ and $\delta \in (0, 1)$, with probability at least $1 - \delta$, Algorithm 3 satisfies:*

$$D_{\mathrm{KL}}\left(\pi^\star_\beta \| \pi_{\theta_T} \right) \le \left(\frac{1}{1 + \beta\eta} \right)^T D_{\mathrm{KL}}\left(\pi^\star_\beta \| \pi_{\theta_0} \right) + \frac{2}{\beta} \delta_P + \frac{2(1 + \beta\eta)}{\beta\eta} \sqrt{C_1 C_r \log\left(\frac{T}{\delta} \right)} \sqrt{\frac{2L^2 \log M}{\mu(M-1)} + \frac{C + 2L^2/\mu}{M}}, \tag{25}$$

where C_r is an absolute constant, C_1, L, δ_P, C, μ are defined in (24), (22), (20), (23), Assumption 4, respectively.

Theorem 4 indicates that, assuming no model approximation error, the total sample complexity for Algorithm 3 to reach ε-accuracy is

$$2MT = \widetilde{O}\left(\left(\frac{1+\beta\eta}{\beta\eta}\right)^2\left(\frac{L^2}{\mu}+C\right)C_1C_r\frac{1}{\varepsilon^2}\right).$$

In contrast with SPPO [Wu et al., 2024b], which only ensures average-iterate convergence without quantifying sample efficiency, our method has stronger theoretical guarantees, offering last-iterate convergence and explicit sample complexity bounds.

5 Experiments

We report our experiment results in this section.

5.1 Contextual bandits

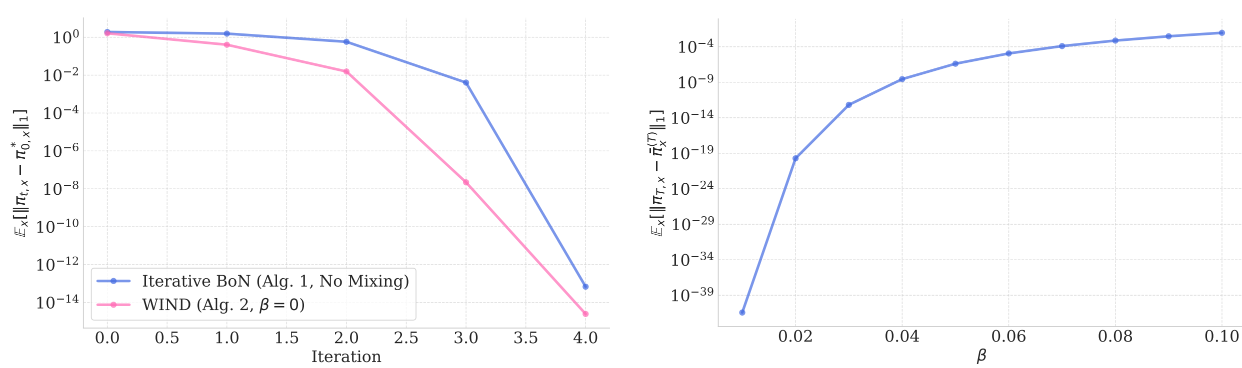
In this section we conduct contextual bandit experiments to validate Theorem 2.

Experiments setup. We set $|\mathcal{X}| = 20, |\mathcal{Y}| = 100$, and initialize $r(x_i, y_j) \overset{i.i.d}{\sim} \mathcal{N}(0,1)$, where $i \in [|\mathcal{X}|], j \in [|\mathcal{Y}|]$, and $\mathcal{N}(0,1)$ stands for the standard Gaussian distribution. We set π_{ref} and ρ to be uniform distributions and randomly initialized $\pi^{(0)}$ in Algorithm 2 using the Dirichlet distribution with parameters all set to be 1. For the distance metric, we use the average ℓ_1 distance $D_{\ell_1} : \Delta_{\mathcal{Y}}^{\mathcal{X}} \times \Delta_{\mathcal{Y}}^{\mathcal{X}} \to \mathbb{R}$ defined as

$$D_{\ell_1}(\pi, \pi') \coloneqq \mathbb{E}_{x\sim\rho}\left\|\pi_x - \pi'_x\right\|_1. \tag{26}$$

We conduct the following two experiments:

- For the no-mixing case where $\alpha_1 = 1, \alpha_2 = 0$, we show the convergence of both iterative BoN (c.f. Algorithm 1) and exact `WIND` (c.f. Algorithm 2) to $\overline{\pi}_0^\star$: we plot the average ℓ_1 distance between $\overline{\pi}_0^\star$ and the iterates for both algorithms. In this experiments we set learning rate η in Algorithm 2 to be 16.
- For the mixing case where $\alpha_1 = \frac{\eta}{(1+\beta\eta)(n-1)}$ and $\alpha_2 = \frac{n-1-\eta}{(1+\beta\eta)(n-1)}$, we verify that $\overline{\pi}_\beta^\star$ and $\pi_\beta^\star$ are very close to each other: we fix the iteration number $T = 5000$ for both Algorithm 1 and 2, and increase β from 0.01 to 0.1 to plot the change of average ℓ_1 distance between the final outputs of the two algorithms $D_{\ell_1}(\pi_T, \pi^{(T)})$ with respect to β. In this experiments we set $\eta = 1$.



(a) no-mixing

(b) mixing

Figure 1: Empirical validation of Theorem 2 on contextual bandit experiments. For (a) the no-mixing case, we show the convergence of both iterative BoN (c.f. Algorithm 1) and exact `WIND` (c.f. Algorithm 2) to $\overline{\pi}_0^\star$; for (b) the mixing case, we show $\overline{\pi}_\beta^\star$ and $\pi_\beta^\star$ are very close to each other.

Model	GSM8k	HellaSwag	MMLU	MT-Bench 1st Turn	MT-Bench 2nd Turn	MT-Bench Avg
Llama-3-8B-SPPO Iter1	75.44	79.80	65.65	8.3813	7.6709	8.0283
Llama-3-8B-SPPO Iter2	75.13	80.39	65.67	8.3875	7.4875	7.9375
Llama-3-8B-SPPO Iter3	74.91	**80.86**	65.60	8.0500	7.7625	7.9063
Llama-3-8B-JBOND Iter1	76.12	77.70	65.73	8.2875	7.4281	7.8578
Llama-3-8B-JBOND Iter2	75.74	77.47	65.85	8.2563	7.4403	7.8495
Llama-3-8B-JBOND Iter3	76.12	77.36	65.83	8.2750	7.2767	7.7774
Llama-3-8B-WIND Iter1 (Ours)	75.82	78.73	65.77	8.2875	7.6875	7.9875
Llama-3-8B-WIND Iter2 (Ours)	76.19	79.05	65.77	8.3625	7.7500	8.0563
Llama-3-8B-WIND Iter3 (Ours)	**77.18**	79.31	**65.87**	**8.5625**	**7.8354**	**8.2013**

Table 1: Results on GSM8k, HellaSwag, MMLU and MT-Bench.

Results. Our results are presented in Figure 1. Figure 1(a) indicates that for the no-mixing case, both algorithms converge to $\overline{\pi}_0^\star$ with `WIND` slightly faster than iterative BoN. From Figure 1(b), we can see that $\overline{\pi}_\beta^\star$ and $\pi_\beta^\star$ are very close to each other when β is small and their distance approaches 0 very quickly as $\beta \to 0$, which corroborates (12).

5.2 LLM alignment

We follow the experiment setup in Wu et al. [2024b] and Snorkel[1]. We use Llama-3-8B-Instruct[2] as the base pretrained model for baseline comparisons. For fair comparison, we chose the same prompt dataset UltraFeedback [Cui et al., 2023] and round splits, and the same Pair-RM framework [Jiang et al., 2023] for the preference model as in Wu et al. [2024b] and Snorkel. The learning rate is set to be 5×10^{-7}. In each iteration, we generate answers from 20000 prompts in the UltraFeedback dataset to train the model. The global training batch size is 64 (4 per device $\times$ 16 GPUs). Our experiments are run on 16 A100 GPUs, where each has 40 GB memory. We modify the per-device batch size and gradient accumulation steps in SPPO GitHub repository[3] while keeping the actual training batch size, to avoid out-of-memory error.

Baselines and Benchmarks. We consider two baselines: SPPO [Wu et al., 2024b] and a variant of J-`BOND` [Sessa et al., 2024]. Here we follow the exact same setting in their repository of the SPPO paper to reproduce SPPO results, with the only change being that we use different computation devices.

We consider 4 major evaluation benchmarks: GSM8k, HellaSwag, MMLU and MT-Bench. They evaluated the following capability:

- GSM8k [Cobbe et al., 2021] evaluates the mathematical reasoning at a grade school level.
- HellaSwag [Zellers et al., 2019] measures the commonsense reasoning by letting language models select a choice to finish a half-complete sentence.
- MMLU [Hendrycks et al., 2020] is a large-scale benchmark that encompasses a variety of tasks to measure the language models' knowledge.
- MT-Bench [Zheng et al., 2023] is also a LLM-as-a-judge benchmark that evaluates the LLM's multi-round chat capability. The scores given by GPT-4 is reported.

Results. For traditional benchmarks (GSM8k, HellaSwag and MMLU), which do not involve using LLMs as the judges, the results are shown in Table 1. The model Llama-3-8B-WIND of ours achieved optimal in the last iteration in GSM8k and MMLU, while performing better than the J-`BOND` variant and slightly

[1] https://huggingface.co/snorkelai/Snorkel-Mistral-PairRM-DPO
[2] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[3] https://github.com/uclaml/SPPO

worse than SPPO in HellaSwag. In fact, our method shows consistent improvement over iterations: for all three benchmarks, our method continues to improve with more iterations of training, while both SPPO and J-`BOND` variant show performance regressions with increasing number of iterations. For MT-Bench, Llama-3-8B-WIND achieves comparable results in comparison with SPPO, and outperforms J-`BOND`.

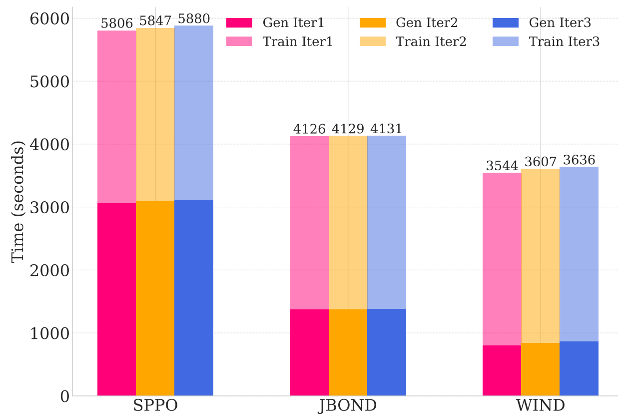


Figure 2: Run time (seconds) of different methods.

Runtime. We also report the running time used by different methods in our setting. Since we base our implementation on the SPPO GitHub Repository, we only modify the objectives and the sampling process to reflect the difference between these algorithms. Figure 2 shows that our method achieves much better sample efficiency during data generation. In sum, the proposed `WIND` achieves superior performance with less computation cost, making iterative `BOND` practice applicable.

6 Conclusion

This work establishes a unified game-theoretic framework that connects iterative BoN with existing game-theoretic RLHF approaches. We present `WIND`, a sample-efficient efficient algorithm for win rate dominance optimization with finite-sample guarantees, which provides an accelerated alternative to iterative `BOND`. Empirical validation on multiple public benchmarks demonstrates the effectiveness and efficiency of our approach compared to several state-of-the-art methods.

Acknowledgement

The work of T. Yang, Z. Wen, S. Cen and Y. Chi is supported in part by the grants NSF CIF-2106778, DMS-2134080 and ONR N00014-19-1-2404. In addition, the authors thank Xingyu Xu for the great discussion on the proof details.

References

A. Amini, T. Vieira, and R. Cotterell. Variational best-of-n alignment. *arXiv preprint arXiv:2407.06057*, 2024.

A. Askell, Y. Bai, A. Chen, D. Drain, D. Ganguli, T. Henighan, A. Jones, N. Joseph, B. Mann, N. DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

M. G. Azar, Z. D. Guo, B. Piot, R. Munos, M. Rowland, M. Valko, and D. Calandriello. A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*, pages 4447–4455. PMLR, 2024.

Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

H. H. Bauschke, J. M. Borwein, and P. L. Combettes. Bregman monotone optimization algorithms. *SIAM Journal on control and optimization*, 42(2):596–636, 2003.

A. Beck. *First-order methods in optimization.* SIAM, 2017.

A. Beirami, A. Agarwal, J. Berant, A. D'Amour, J. Eisenstein, C. Nagpal, and A. T. Suresh. Theoretical guarantees on the best-of-n alignment policy. *arXiv preprint arXiv:2401.01879*, 2024.

O. Bousquet and A. Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2: 499–526, 2002.

S. Cen, Y. Wei, and Y. Chi. Fast policy extragradient methods for competitive games with entropy regularization. *Advances in Neural Information Processing Systems*, 34:27952–27964, 2021.

S. Cen, J. Mei, K. Goshvadi, H. Dai, T. Yang, S. Yang, D. Schuurmans, Y. Chi, and B. Dai. Value-incentivized preference optimization: A unified approach to online and offline rlhf. *arXiv preprint arXiv:2405.19320*, 2024.

Z. Charles and D. Papailiopoulos. Stability and generalization of learning algorithms that converge to global optima. In *International conference on machine learning*, pages 745–754. PMLR, 2018.

P. F. Christiano, J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

G. Cui, L. Yuan, N. Ding, G. Yao, W. Zhu, Y. Ni, G. Xie, Z. Liu, and M. Sun. Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377*, 2023.

H. Dong, W. Xiong, D. Goyal, Y. Zhang, W. Chow, R. Pan, S. Diao, J. Zhang, K. Shum, and T. Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.

K. Ethayarajh, W. Xu, N. Muennighoff, D. Jurafsky, and D. Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

L. Gao, J. Schulman, and J. Hilton. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR, 2023.

L. Gui, C. Gârbacea, and V. Veitch. Bonbon alignment for large language models and the sweetness of best-of-n sampling. *arXiv preprint arXiv:2406.00832*, 2024.

M. Gutmann and A. Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 297–304. JMLR Workshop and Conference Proceedings, 2010.

D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

D. Jiang, X. Ren, and B. Y. Lin. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. *arXiv preprint arXiv:2306.02561*, 2023.

Y. Kang, Y. Liu, J. Li, and W. Wang. Sharper utility bounds for differentially private models: Smooth and non-smooth. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, pages 951–961, 2022.

H. Karimi, J. Nutini, and M. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2016, Riva del Garda, Italy, September 19-23, 2016, Proceedings, Part I 16*, pages 795–811. Springer, 2016.

Y. Klochkov and N. Zhivotovskiy. Stability and deviation optimal risk bounds with convergence rate $o(1/n)$. *Advances in Neural Information Processing Systems*, 34:5065–5076, 2021.

G. M. Korpelevich. The extragradient method for finding saddle points and other problems. *Matecon*, 12: 747–756, 1976.

C. Liu, L. Zhu, and M. Belkin. Loss landscapes and optimization in over-parameterized non-linear systems and neural networks. *Applied and Computational Harmonic Analysis*, 59:85–116, 2022.

Y. Meng, M. Xia, and D. Chen. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*, 2024.

R. Munos. Error bounds for approximate policy iteration. In *ICML*, volume 3, pages 560–567. Citeseer, 2003.

R. Munos. Error bounds for approximate value iteration. In *Proceedings of the National Conference on Artificial Intelligence*, volume 20, page 1006. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2005.

R. Munos and C. Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

R. Munos, M. Valko, D. Calandriello, M. G. Azar, M. Rowland, Z. D. Guo, Y. Tang, M. Geist, T. Mesnard, A. Michi, et al. Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*, 2023.

R. Nakano, J. Hilton, S. Balaji, J. Wu, L. Ouyang, C. Kim, C. Hesse, S. Jain, V. Kosaraju, W. Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021.

J. F. Nash et al. Non-cooperative games. 1950.

OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

S. Pattathil, K. Zhang, and A. Ozdaglar. Symmetric (optimistic) natural policy gradient for multi-agent learning with parameter convergence. In *International Conference on Artificial Intelligence and Statistics*, pages 5641–5685. PMLR, 2023.

L. D. Popov. A modification of the arrow–hurwicz method for search of saddle points. *Mathematical Notes of the Academy of Sciences of the USSR, 28(5):845–848*, 1980.

R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36, 2024.

C. Rosset, C.-A. Cheng, A. Mitra, M. Santacroce, A. Awadallah, and T. Xie. Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715*, 2024.

P. G. Sessa, R. Dadashi, L. Hussenot, J. Ferret, N. Vieillard, A. Ramé, B. Shariari, S. Perrin, A. Friesen, G. Cideron, et al. Bond: Aligning llms with best-of-n distillation. *arXiv preprint arXiv:2407.14622*, 2024.

S. Sokota, R. D'Orazio, J. Z. Kolter, N. Loizou, M. Lanctot, I. Mitliagkas, N. Brown, and C. Kroer. A unified approach to reinforcement learning, quantal response equilibria, and two-player zero-sum games. In *The Eleventh International Conference on Learning Representations*, 2023.

N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021, 2020.

G. Swamy, C. Dann, R. Kidambi, Z. S. Wu, and A. Agarwal. A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056*, 2024.

Y. Tang, Z. D. Guo, Z. Zheng, D. Calandriello, R. Munos, M. Rowland, P. H. Richemond, M. Valko, B. Á. Pires, and B. Piot. Generalized preference optimization: A unified approach to offline alignment. *arXiv preprint arXiv:2402.05749*, 2024.

Z. Wang, C. Nagpal, J. Berant, J. Eisenstein, A. D'Amour, S. Koyejo, and V. Veitch. Transforming and combining rewards for aligning large language models. *arXiv preprint arXiv:2402.00742*, 2024.

Y. Wu, F. Liu, G. Chrysos, and V. Cevher. On the convergence of encoder-only shallow transformers. *Advances in Neural Information Processing Systems*, 36, 2024a.

Y. Wu, Z. Sun, H. Yuan, K. Ji, Y. Yang, and Q. Gu. Self-play preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675*, 2024b.

H. Xu, A. Sharaf, Y. Chen, W. Tan, L. Shen, B. Van Durme, K. Murray, and Y. J. Kim. Contrastive preference optimization: Pushing the boundaries of llm performance in machine translation. *arXiv preprint arXiv:2401.08417*, 2024.

J. Q. Yang, S. Salamatian, Z. Sun, A. T. Suresh, and A. Beirami. Asymptotics of language model alignment. *arXiv preprint arXiv:2404.01730*, 2024a.

T. Yang, S. Cen, Y. Wei, Y. Chen, and Y. Chi. Federated natural policy gradient methods for multi-task reinforcement learning. *arXiv preprint arXiv:2311.00201*, 2023.

T. Yang, Y. Huang, Y. Liang, and Y. Chi. In-context learning with representations: Contextual generalization of trained transformers. *arXiv preprint arXiv:2408.10147*, 2024b.

R. Yuan, S. S. Du, R. M. Gower, A. Lazaric, and L. Xiao. Linear convergence of natural policy gradient methods with log-linear policies. In *International Conference on Learning Representations*, 2023a.

Z. Yuan, H. Yuan, C. Tan, W. Wang, S. Huang, and F. Huang. Rrhf: Rank responses to align language models with human feedback without tears. *arXiv preprint arXiv:2304.05302*, 2023b.

R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

Y. Zhang, D. Yu, B. Peng, L. Song, Y. Tian, M. Huo, N. Jiang, H. Mi, and D. Yu. Iterative nash policy optimization: Aligning llms with general preferences via no-regret learning. *arXiv preprint arXiv:2407.00617*, 2024.

Y. Zhao, R. Joshi, T. Liu, M. Khalman, M. Saleh, and P. J. Liu. Slic-hf: Sequence likelihood calibration with human feedback. *arXiv preprint arXiv:2305.10425*, 2023.

L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36: 46595–46623, 2023.

A Other Objectives for Algorithm 3

In this section we give some possible alternative objectives for Algorithm 3 by utilizing different variational forms.

f-divergence objectives: We could use f-divergence D_f as the objective function. For a convex function f, D_f is defined as

$$D_f(P||Q) \coloneqq \mathbb{E}_Q \left[f\left(\frac{P}{Q}\right)\right]. \tag{27}$$

Let

$$U = (x, y) \quad \text{and} \quad V = \begin{cases} 1, \text{ if } r(x,y) > r(x,y'), \\ 0, \text{ if } r(x,y) < r(x,y'), \\ z \sim Ber(1/2), \text{ if } r(x,y) = r(x,y'), \end{cases}$$

Then $V|U$ is a function of y' and $P_{V|U} = Ber(\mathbb{E}_{y'} P_x(y,y'))$. Further define

$$\zeta_\theta(x,y) = \frac{1+\beta\eta}{\eta}\phi_\theta(y|x) - \frac{1}{\eta}\phi_{\theta_t}(y|x) - \beta\phi_{\theta_{\text{ref}}}(y|x) - \frac{1+\beta\eta}{\eta} Z_t(x). \tag{28}$$

We let $Q_{V|U} = Ber(\zeta_\theta(y|x))$, then by solving

$$\theta_{t+1} = \arg\min_\theta \mathbb{E}_U D_f(P_{V|U}||Q_{V|U}), \tag{29}$$

we could approximate the update rule (15).

Especially, when $f(x) = x\log x$, we have $D_f = D_{\text{KL}}$, and (29) becomes

$$\begin{aligned} \theta_{t+1} &= \arg\min_\theta \mathbb{E}_{x\sim\rho, y\sim\pi_{\theta_t}(\cdot|x)} \mathbb{E}_{v\sim P_{V|U}} \log\frac{P_{V|U}(v)}{Q_{V|U}(v)} \\ &= \arg\min_\theta \mathbb{E}_{x\sim\rho, y\sim\pi_{\theta_t}(\cdot|x)} \mathbb{E}_{v\sim P_{V|U}} [-\log Q_{V|U}(v)]. \end{aligned} \tag{30}$$

We could approximate the above objective by sampling $x_i \sim \rho, y_i, y_i' \sim \pi_{\theta_t}(\cdot|x_i)$ $(i \in [M])$ and minimizing the empirical risk:

$$\theta_{t+1} = \arg\min_\theta R^{\text{kl}}_{t,M}(\theta) \coloneqq -\frac{1}{M}\sum_{i=1}^M \left[\mathbb{1}_{\{v_i=1\}} \log\zeta_\theta(x,y) + \mathbb{1}_{\{v_i=0\}}\log(1-\zeta_\theta(x,y))\right], \tag{31}$$

where

$$v_i = \begin{cases} 1, \text{ if } r(x_i,y_i) > r(x_i,y_i'), \\ 0, \text{ if } r(x_i,y_i) < r(x_i,y_i'), \\ z \sim Ber(1/2), \text{ if } r(x_i,y_i) = r(x_i,y_i'). \end{cases} \tag{32}$$

For other f-divergence objectives, we may not be able to get rid of the unknown $P_{V|U}$ on the RHS of (29), but (29) could provide a gradient estimator for the objective that allows us to optimize θ by stochastic gradient descent.

Noise contrastive estimation (NCE) objectives. We could use NCE [Gutmann and Hyvärinen, 2010] as objectives. NCE is a method to estimate the likelihood of a data point by contrasting it with noise samples. Let D_θ be the discriminator (parameterized by θ) that distinguishes the true data from noise samples. The NCE objective is

$$\min_\theta \mathbb{E}_{z\sim P_{\text{data}}}[-\log D_\theta(z)] + \mathbb{E}_{z\sim P_{\text{noise}}}[-\log(1-D_\theta(z))], \tag{33}$$

Then the uptimal solution of (33) is

$$D_{\theta^\star}(z) = \frac{P_{\text{data}}(z)}{P_{\text{data}}(z) + P_{\text{noise}}(z)}.$$

In our case, we let $P_{\text{data}} = P_{V|U}$ and $P_{\text{noise}} = Ber(p)$ where $p \in (0, 1)$ is a hyperparameter. We also let

$$D_\theta(1|x, y) = \frac{\zeta_\theta(x, y)}{\zeta_\theta(x, y) + p}, \quad \text{and} \quad D_\theta(0|x, y) = \frac{p}{\zeta_\theta(x, y) + p},$$

where ζ_θ is defined in (28). Then we could approximate the update rule (15) by solving

$$\begin{aligned}\theta_{t+1} = \arg\min_\theta \mathbb{E}_{x\sim\rho, y\sim\pi_{\theta_t}(\cdot|x)}\mathbb{E}_{v\sim P_{V|U}}\left[-\mathbb{1}_{\{v=1\}}\log\frac{\zeta_\theta(x, y)}{\zeta_\theta(x, y) + p} - \mathbb{1}_{\{v=0\}}\log\frac{p}{\zeta_\theta(x, y) + p}\right] \\ + \mathbb{E}_{x\sim\rho, y\sim\pi_{\theta_t}(\cdot|x)}\mathbb{E}_{v\sim Ber(p)}\left[-\mathbb{1}_{\{v=1\}}\log\frac{p}{\zeta_\theta(x, y) + p} - \mathbb{1}_{\{v=0\}}\log\frac{\zeta_\theta(x, y)}{\zeta_\theta(x, y) + p}\right].\end{aligned} \tag{34}$$

The sample version of (34) would be

$$\begin{aligned}\theta_{t+1} = \arg\min_\theta R^{\text{nce}}_{t,M}(\theta) := -\frac{1}{M}\sum_{i=1}^M\Big[\left(\mathbb{1}_{\{v_i=1\}} + \mathbb{1}_{\{v_i'=0\}}\right)\log\frac{\zeta_\theta(x_i, y_i)}{\zeta_\theta(x_i, y_i) + p} \\ + \left(\mathbb{1}_{\{v_i=0\}} + \mathbb{1}_{\{v_i'=1\}}\right)\log\frac{p}{\zeta_\theta(x_i, y_i) + p}\Big],\end{aligned} \tag{35}$$

where v_i is defined in (32) and $v_i' \sim Ber(p)$.

B Proofs

B.1 Proof of Proposition 1

We first prove the case when $\beta = 0$. This part of proof is inspired by Swamy et al. [2024, Lemma 2.1]. Let

$$\pi_1 := \arg\max_\pi \min_{\pi'} P(\pi \succ \pi'), \quad \pi_2 := \arg\min_{\pi'} \max_\pi P(\pi \succ \pi'),$$

i.e., (π_1, π_2) is a Nash equilibrium of (9) (which is guaranteed to exist since the policy space is compact). Then $\forall \pi, \pi', \forall x \in \mathcal{X}$, we have:

$$\pi_{1,x}^\top P_x \pi_x' \geq \pi_{1,x}^\top P_x \pi_{2,x} \geq \pi_x^\top P_x \pi_{2,x},$$

which is equivalent to

$$(\pi_x')^\top P_x^\top \pi_{1,x} \geq \pi_{2,x}^\top P_x^\top \pi_{1,x} \geq \pi_{2,x}^\top P_x^\top \pi_x.$$

Note that

$$P_x + P_x^\top = J, \tag{36}$$

where $J \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$ is the matrix of all ones. This gives

$$-(\pi_x')^\top P_x \pi_{1,x} \geq -\pi_{2,x}^\top P_x \pi_{1,x} \geq -\pi_{2,x}^\top P_x \pi_x,$$

i.e.,

$$(\pi_x')^\top P_x \pi_{1,x} \leq \pi_{2,x}^\top P_x \pi_{1,x} \leq \pi_{2,x}^\top P_x \pi_x, \ \forall \pi, \pi' \in \Delta_{\mathcal{Y}}^{\mathcal{X}}, \ \forall x \in \mathcal{X}.$$

This implies (π_2, π_1) is also a Nash equilibrium of (9). Then by the interchangeability of Nash equilibrium strategies for two-player zero-sum games [Nash et al., 1950], (π_1, π_1) and (π_2, π_2) are both the Nash equilibria of (9), which indicates that π_1, π_2 are both the solutions of (10).

Next we prove the case when $\beta > 0$. When $\beta > 0$, due to the strong concavity-convexity of the the minmax problem (9), there exists a unique Nash equilibrium $(\pi_1^\star, \pi_2^\star)$ of it. And it's straightforward to compute that $(\pi_1^\star, \pi_2^\star)$ satisfies the following relation:

$$\forall x \in \mathcal{X}: \quad \begin{cases}\pi_1^\star(\cdot|x) \propto \pi_{\text{ref}}(\cdot|x) \circ \exp\left(\frac{1}{\beta}P_x\pi_2^\star(\cdot|x)\right), \\ \pi_2^\star(\cdot|x) \propto \pi_{\text{ref}}(\cdot|x) \circ \exp\left(-\frac{1}{\beta}P_x^\top\pi_1^\star(\cdot|x)\right),\end{cases} \tag{37}$$

where we use $\circ$ to denote the element-wise product of two vectors.

Again, using (36), we have

$$\forall x \in \mathcal{X}: \quad \begin{cases} \pi_1^\star(\cdot|x) \propto \pi_{\text{ref}}(\cdot|x) \circ \exp\left(\frac{1}{\beta} P_x \pi_2^\star(\cdot|x)\right), \\ \pi_2^\star(\cdot|x) \propto \pi_{\text{ref}}(\cdot|x) \circ \exp\left(\frac{1}{\beta} P_x \pi_1^\star(\cdot|x)\right), \end{cases}$$

which implies $(\pi_2^\star, \pi_1^\star)$ is also a Nash equilibrium of (9). By the uniqueness of the Nash equilibrium we immediately know that $\pi_1^\star = \pi_2^\star$. Letting $\pi_\beta^\star = \pi_1^\star = \pi_2^\star$, we have $\pi_\beta^\star$ satisfies (10).

On the other hand, if $\pi_\beta^\star$ is the solution of (10), then $(\pi_\beta^\star, \pi_\beta^\star)$ satisfies (37) and thus is a Nash equilibrium of (9). In addition, by the uniqueness of (9), we deduce that (10) has a unique solution.

B.2 Proofs of Theorem 1 and Theorem 2

We merge Theorem 1 and Theorem 2 into the following theorem (recall we define $\overline{P}_x$ in (5)):

Theorem 5 (solution to iterative BoN (formal)). *Let $\pi_{\text{ref}} \in \text{relint}\left(\Delta_{\mathcal{Y}}^{\mathcal{X}}\right)$ and $n \geq 2$ in Algorithm 1. Then $\lim_{T\to\infty} \pi_T$ exists in the following two cases:*

- *(No-mixing) When $\alpha_1 = 1, \alpha_2 = 0$. In this case $\overline{\pi}_0^\star := \lim_{T\to\infty} \pi_T$ is a solution to both (8) and (10) with $\beta = 0$.*
- *(Mixing) When $\alpha_1 = \frac{\eta}{(1+\beta\eta)(n-1)}$, $\alpha_2 = \frac{n-1-\eta}{(1+\beta\eta)(n-1)}$ for any $\beta, \eta > 0$. In this case $\overline{\pi}_\beta^\star := \lim_{T\to\infty} \pi_T$ satisfies:*

$$\overline{\pi}_\beta^\star \in \arg\max_{\pi} \mathop{\mathbb{E}}_{\substack{x\sim\rho,\\ y\in\pi(\cdot|x)}} \log \mathop{\mathbb{E}}_{y'\in\overline{\pi}_\beta^\star(\cdot|x)} \overline{P}_x(y \succeq y') - \beta D_{\text{KL}}\left(\pi \| \pi_{\text{ref}}\right). \tag{38}$$

Moreover, if

$$\beta \leq \min_{\substack{x\in\mathcal{X},\\ y\in\mathcal{Y}\setminus\mathcal{Y}^\star(x)}} \left\{ \frac{\sum_{y^\star\in\mathcal{Y}^\star(x)} \pi_{\text{ref}}(y^\star|x)}{4\max\left\{\log\frac{\pi_{\text{ref}}(y|x)}{\max_{y^\star\in\mathcal{Y}^\star(x)} \pi_{\text{ref}}(y^\star|x)}, 0\right\}} \right\}, \tag{39}$$

then for all $x \in \mathcal{X}$, we have

$$\left\|\overline{\pi}_{\beta,x}^\star - \pi_{\beta,x}^\star\right\|_1 \leq 4(|\mathcal{Y}| - |\mathcal{Y}^\star(x)|) e^{\frac{-\sum_{y^\star\in\mathcal{Y}^\star(x)} \pi_{\text{ref}}(y^\star|x)}{4\beta}} \to 0 \text{ as } \beta \to 0. \tag{40}$$

Remark 3. *It's easy to see that $\overline{\pi}_\beta^\star$ is a solution to (38) (see also (8) in the main paper) if and only if $(\overline{\pi}_\beta^\star, \overline{\pi}_\beta^\star)$ is a nash equilibrium of the log-win-rate game (7).*

Now we give the proof of Theorem 5.

Step 1: show convergence for the no-mixing case. We first prove the convergence result for the no-mixing case. Note that for any policy π, $\pi^{(n)}$ has the expression

$$\begin{aligned} \forall (x,y) \in \mathcal{X}\times\mathcal{Y}: \quad \pi^{(n)}(y|x) &= \binom{n}{1}\pi(y|x)\mathbb{P}_{y_i\sim\pi(\cdot|x)}\left(r(x,y) \geq \max_{1\leq i\leq n-1} r(x,y_i)\right) \\ &= n\pi(y|x)\left(\overline{P}_x(y,:)\pi_x\right)^{n-1}, \end{aligned} \tag{41}$$

where $\overline{P}_x$ is defined in (5).

When $\alpha_1 = 1, \alpha_2 = 0$, Algorithm 1 could be simplified as

$$\forall t \in \mathbb{N}: \quad \pi_{t+1} = \pi_t^{(n)}.$$

Then π_T is equivalent to $\pi_{\text{ref}}^{(n^T)}$ — the best of-n^T policy of π_{ref}. For any x, define $\mathcal{Y}^\star(x)$ as the set of resonses that maximize the reward function $r(x,\cdot)$, i.e.,

$$\mathcal{Y}^\star(x) \coloneqq \arg\max_{y\in\mathcal{Y}} r(x,y).$$

Then for any $y \in \mathcal{Y}^\star(x)$ and $y' \in \mathcal{Y}$, we have $\overline{P}_x(y,y') = 1$ and $\overline{P}_x(y,:)\pi_{\text{ref},x} = 1$. And for any $y \notin \mathcal{Y}^\star(x)$, we have $\overline{P}_x(y,:)\pi_{\text{ref},x} < 1$ since $\pi_{\text{ref}} \in \text{relint}\left(\Delta_{\mathcal{Y}}^{\mathcal{X}}\right)$.

By the BoN expression (41) we deduce that

$$\lim_{T\to\infty} \pi_T(y|x) = \begin{cases} \frac{\pi_{\text{ref}}(y|x)}{\sum_{y\in\mathcal{Y}^\star(x)} \pi_{\text{ref}}(y|x)}, & \text{if } y \in \mathcal{Y}^\star(x), \\ 0, & \text{otherwise.} \end{cases} \tag{42}$$

We let $\overline{\pi}_0^\star \coloneqq \lim_{T\to\infty} \pi_T$. Now we show that $(\overline{\pi}_0^\star, \overline{\pi}_0^\star)$ is a nash equilibrium of (9) when $\beta = 0$, which implies $\overline{\pi}_0^\star$ is a solution to (10) when $\beta = 0$.

Note that for any $(x,y) \in \mathcal{X}\times\mathcal{Y}$, we have

$$(\overline{\pi}_{0,x}^\star)^\top P_x(:,y) = \begin{cases} \frac{1}{2}, & \text{if } y \in \mathcal{Y}^\star(x), \\ 1, & \text{otherwise,} \end{cases}$$

which gives

$$\forall \pi, x: \quad (\overline{\pi}_{0,x}^\star)^\top P_x \pi_x \geq \frac{1}{2} = (\overline{\pi}_{0,x}^\star)^\top P_x \overline{\pi}_{0,x}^\star. \tag{43}$$

On the other hand, for any $(x,y) \in \mathcal{X}\times\mathcal{Y}$, we have

$$P_x(y,:)\overline{\pi}_{0,x}^\star = \begin{cases} \frac{1}{2}, & \text{if } y \in \mathcal{Y}^\star(x), \\ 0, & \text{otherwise,} \end{cases}$$

which implies

$$\forall \pi, x: \quad (\pi_x)^\top P_x \overline{\pi}_{0,x}^\star \leq \frac{1}{2} = (\overline{\pi}_{0,x}^\star)^\top P_x \overline{\pi}_{0,x}^\star. \tag{44}$$

(43) and (44) together indicate that $(\overline{\pi}_0^\star, \overline{\pi}_0^\star)$ is a Nash equilibrium of (9) when $\beta = 0$, which also indicates that $\overline{\pi}_0^\star$ is a solution to (10). It's straightforward to verify with (42) that $\overline{\pi}_0^\star$ is also a solution to (8).

Step 2: show convergence of the mixing case. Recall that in the mixing case we set $\alpha_1 = \frac{\eta}{(1+\beta\eta)(n-1)}, \alpha_2 = \frac{n-1-\eta}{(1+\beta\eta)(n-1)}$. We take logarithm on both sides of the iteration in line 5 of Algorithm 1 and unroll it as follows:

$$\begin{aligned}
\log \pi_{t+1}(y|x) &= \alpha_1 \log \widetilde{\pi}_t(y|x) + \alpha_2 \log \pi_t(y|x) + (1-\alpha_1-\alpha_2)\log \pi_{\text{ref}}(y|x) + c_x \\
&= (\alpha_1+\alpha_2)\log \pi_t + (n-1)\alpha_1 \log\left(\overline{P}_x(y,:)\pi_{t,x}\right) + (1-\alpha_1-\alpha_2)\log \pi_{\text{ref}}(y|x) + c_x' \\
&= (\alpha_1+\alpha_2)^{t+1}\log \pi_0(y|x) + (n-1)\alpha_1 \sum_{i=0}^{t} (\alpha_1+\alpha_2)^i \log\left(\overline{P}_x(y,:)\pi_{t-i,x}\right) \\
&\quad + \left(1-(\alpha_1+\alpha_2)^{t+1}\right)\log \pi_{\text{ref}}(y|x) + c_x'' \\
&= \log \pi_{\text{ref}}(y|x) + (n-1)\alpha_1 \sum_{i=0}^{t} (\alpha_1+\alpha_2)^i \log\left(\overline{P}_x(y,:)\pi_{t-i,x}\right) + c_x'' \\
&= \log \pi_{\text{ref}}(y|x) + \frac{\eta}{1+\eta\beta} \sum_{i=0}^{t} \left(\frac{1}{1+\beta\eta}\right)^i \log\left(\overline{P}_x(y,:)\pi_{t-i,x}\right) + c_x'',
\end{aligned} \tag{45}$$

where c_x, c_x', c_x'' are constants that depend on x, and the second equality makes use of (41), and the last equality follows from our choice of α_1, α_2.

Note that for all $\pi \in \Delta_{\mathcal{Y}}^{\mathcal{X}}$, we have

$$\forall y \in \mathcal{Y}^\star(x): \quad \overline{P}_x(y,:)\pi_x = 1,$$
$$\forall y \in \mathcal{Y} \setminus \mathcal{Y}^\star(x): \quad \overline{P}_x(y,:)\pi_x \leq 1.$$

For each $x \in \mathcal{X}$, we let $y_1(x) \in \mathcal{Y}^\star(x)$ such that $\pi_{\mathrm{ref}}(y_1|x) = \max_{y\in\mathcal{Y}^\star(x)} \pi_{\mathrm{ref}}(y|x)$. For notation simplicity, when it does not cause confusion, we simply write $y_1(x)$ as y_1. (45) indicates that for all $y \in \mathcal{Y}$, we have

$$\log\left(\frac{\pi_{t+1}(y_1|x)}{\pi_{t+1}(y|x)}\right) = \log\left(\frac{\pi_{\mathrm{ref}}(y_1|x)}{\pi_{\mathrm{ref}}(y|x)}\right), \text{ if } y \in \mathcal{Y}^\star(x), \tag{46}$$

$$\log\left(\frac{\pi_{t+1}(y_1|x)}{\pi_{t+1}(y|x)}\right) = \log\left(\frac{\pi_{\mathrm{ref}}(y_1|x)}{\pi_{\mathrm{ref}}(y|x)}\right) + \frac{\eta}{1+\eta\beta}\sum_{i=0}^{t}\left(\frac{1}{1+\beta\eta}\right)^i \log\left(\frac{1}{\overline{P}_x(y,:)\pi_{t-i,x}}\right), \text{ if } y \notin \mathcal{Y}^\star(x). \tag{47}$$

Especially, (47) indicates that the ratio $\frac{\pi_{t+1}(y|x)}{\pi_{t+1}(y_1|x)}$ is decreasing with t for all $y \notin \mathcal{Y}^\star(x)$. Since it has a lower bound 0, we have that the ratio $\frac{\pi_{t+1}(y|x)}{\pi_{t+1}(y_1|x)}$ converges as $t \to \infty$ for all $y \notin \mathcal{Y}^\star(x)$. Therefore, (46) together with (47) implies that $\overline{\pi}_\beta^\star \coloneqq \lim_{t\to\infty} \pi_{t+1}$ exists. To see that $\overline{\pi}_\beta^\star$ is a solution to (38), we make use of the following lemma.

Lemma 2. *For any sequence $\{a_t\}_{t=0}^\infty$ in $\mathbb{R}$ where $a_t \leq 0$ for all t and $a \coloneqq \lim_{t\to\infty} a_t$ exists (a can be $-\infty$), for any $\alpha \in (0,1)$, we have*

$$\lim_{t\to\infty}\sum_{i=0}^{t}\alpha^i a_{t-i} = \frac{a}{1-\alpha}. \tag{48}$$

Proof of Lemma 2. If $a = -\infty$, then

$$\lim_{t\to\infty}\sum_{i=0}^{t}\alpha^i a_{t-i} \leq \lim_{t\to\infty} a_t = -\infty.$$

If $a > -\infty$, we have

$$\sum_{i=0}^{t}\alpha^i a_{t-i} = \sum_{i=0}^{t}\alpha^i a + \sum_{i=0}^{t}\alpha^i(\underbrace{a_{t-i}-a}_{e_{t-i}}) = \frac{1-\alpha^{t+1}}{1-\alpha}a + \sum_{i=0}^{t}\alpha^i e_{t-i},$$

thus we only need to verify that $\lim_{t\to\infty}\sum_{i=0}^{t}\alpha^i e_{t-i} = 0$.

For any $\epsilon > 0$, there exists $N \in \mathbb{N}$ such that

$$\forall t \geq N: \quad \left|\sum_{i=N}^{t}\alpha^i e_{t-i}\right| \leq \frac{\alpha^N}{1-\alpha}b \leq \epsilon/2,$$

where $b = \max_{i\geq N}|e_i|$. $b < \infty$ becuase e_t converges to 0. We fix N and choose T such that for all $t \geq T$, we have

$$\sum_{i=0}^{N}\alpha^i a_{t-i} \leq \epsilon/2.$$

Then for all $t \geq T$, we have

$$\left|\sum_{i=0}^{t}\alpha^i e_{t-i}\right| \leq \left|\sum_{i=0}^{N}\alpha^i e_{t-i}\right| + \left|\sum_{i=N+1}^{t}\alpha^i e_{t-i}\right| \leq \epsilon/2 + \epsilon/2 = \epsilon.$$

This completes the proof. □

Let $t \to \infty$ on both sides of (45), and by Lemma 2, we have

$$\overline{\pi}^\star_\beta(y|x) \propto \pi_{\text{ref}}(y|x) \frac{1}{\beta} \log \left(\overline{P}_x(y,:) \overline{\pi}^\star_{\beta,x} \right). \tag{49}$$

Note that for any x, the strongly concave problem

$$\max_{\pi_x \in \mathcal{Y}} \log \left(\mathop{\mathbb{E}}_{y' \sim \overline{\pi}^\star_\beta(\cdot|x)} \overline{P}_x(y,y') \right) - \beta \text{KL}(\pi_x || \pi_{\text{ref}}(\cdot|x))$$

has a unique solution $\overline{\pi}^\star_{\beta,x}$. Therefore, $\overline{\pi}^\star_\beta$ is a solution to (38). By Remark 3 we know that $(\overline{\pi}^\star_\beta, \overline{\pi}^\star_\beta)$ is a Nash equilibrium of the log-win-rate game (7).

Step 3: bound the distance between $\pi^\star_\beta$ and $\overline{\pi}^\star_\beta$. We let $\pi^{(0)} = \pi_{\text{ref}}$ in Algorithm 2 and unroll the iteration (13) similar to (45). We have

$$\log \pi^{(t+1)}(y|x) = \log \pi_{\text{ref}}(y|x) + \frac{\eta}{1+\eta\beta} \sum_{i=0}^{t} \left(\frac{1}{1+\beta\eta} \right)^i P_x(y,:) \pi_x^{(t-i)} + c_x''', \tag{50}$$

where $\pi^{(t)}$ is the policy at the t-th round of Algorithm 2. Furthermore, similar to (46) and (47), we have

$$\log \left(\frac{\pi^{(t+1)}(y_1|x)}{\pi^{(t+1)}(y|x)} \right) = \log \left(\frac{\pi_{\text{ref}}(y_1|x)}{\pi_{\text{ref}}(y|x)} \right), \text{ if } y \in \mathcal{Y}^\star(x), \tag{51}$$

$$\log \left(\frac{\pi^{(t+1)}(y_1|x)}{\pi^{(t+1)}(y|x)} \right) = \log \left(\frac{\pi_{\text{ref}}(y_1|x)}{\pi_{\text{ref}}(y|x)} \right) + \frac{\eta}{1+\eta\beta} \sum_{i=0}^{t} \left(\frac{1}{1+\beta\eta} \right)^i \left(P_x(y_1,:) - P_x(y,:) \right) \pi_x^{(t-i)}, \text{ if } y \notin \mathcal{Y}^\star(x), \tag{52}$$

For any $\pi \in \Delta_{\mathcal{Y}}^{\mathcal{X}}$, we have

$$\forall y \neq \mathcal{Y}^\star(x): \quad \left(P_x(y_1,:) - P_x(y,:) \right) \pi_x \geq \frac{1}{2} \sum_{y \in \mathcal{Y}^\star(x)} \pi(y|x) \geq 0, \tag{53}$$

we know that $\log \left(\frac{\pi^{(t+1)}(y_1|x)}{\pi^{(t+1)}(y|x)} \right)$ is increasing with t for all $y \in \mathcal{Y} \setminus \mathcal{Y}^\star(x)$, Thus $\pi^{(t)}(y|x)$ is decreasing with t for all $y \in \mathcal{Y} \setminus \mathcal{Y}^\star(x)$. Moreover, by a similar argument as in Step 1, we have that $\lim_{t \to \infty} \pi^{(t)}$ exists and is the solution to (10) (even when $\eta > \beta$).

Note that (52) is equivalent to

$$\log \left(\frac{\pi^{(t+1)}(y_1|x)}{\pi^{(t+1)}(y|x)} \right) = \frac{\eta}{1+\eta\beta} \sum_{i=0}^{t} \left(\frac{1}{1+\beta\eta} \right)^i \Bigg(\underbrace{\left(P_x(y_1,:) - P_x(y,:) \right) \pi_x^{(t-i)} + \beta \log \left(\frac{\pi_{\text{ref}}(y_1|x)}{\pi_{\text{ref}}(y|x)} \right)}_{\xi^{(t-i)}} \Bigg)$$
$$+ \left(\frac{1}{1+\beta\eta} \right)^{t+1} \log \left(\frac{\pi_{\text{ref}}(y_1|x)}{\pi_{\text{ref}}(y|x)} \right). \tag{54}$$

Also note that by (53) and the decreasing property of $\pi^{(t)}(y|x)$ for all $y \notin \mathcal{Y}^\star(x)$, we have

$$\forall x \in \mathcal{X}, \forall y \neq \mathcal{Y}^\star(x): \quad \left(P_x(y_1,:) - P_x(y,:) \right) \pi_x^{(t-i)} \geq \frac{1}{2} \sum_{y \in \mathcal{Y}^\star(x)} \pi_{\text{ref}}(y|x).$$

From the above expression and our choice of β we know that

$$\forall x \in \mathcal{X}, \forall y \neq \mathcal{Y}^\star(x): \quad \xi^{(t-i)} \geq \frac{1}{4} \sum_{y \in \mathcal{Y}^\star(x)} \pi_{\text{ref}}(y|x). \tag{55}$$

Then by (54) we know that

$$\forall x \in \mathcal{X}, \forall y \neq \mathcal{Y}^\star(x): \quad \log\left(\frac{\pi_\beta^\star(y_1|x)}{\pi_\beta^\star(y|x)}\right) \geq \frac{1}{4\beta}\sum_{y\in\mathcal{Y}^\star(x)} \pi_{\text{ref}}(y|x),$$

which indicates that for all $y \in \mathcal{Y} \setminus \mathcal{Y}^\star(x)$,

$$\pi_\beta^\star(y|x) \leq \pi_\beta^\star(y_1|x)\exp\left(-\frac{1}{4\beta}\sum_{y\in\mathcal{Y}^\star(x)} \pi_{\text{ref}}(y|x)\right) \leq \exp\left(-\frac{1}{4\beta}\sum_{y\in\mathcal{Y}^\star(x)} \pi_{\text{ref}}(y|x)\right), \tag{56}$$

which gives

$$\sum_{y\in\mathcal{Y}\setminus\mathcal{Y}^\star(x)} \pi_\beta^\star(y|x) \leq (|\mathcal{Y}| - |\mathcal{Y}^\star(x)|)\exp\left(-\frac{1}{4\beta}\sum_{y\in\mathcal{Y}^\star(x)} \pi_{\text{ref}}(y|x)\right).$$

Combining the above relation with (51), we obtain

$$\forall y \in \mathcal{Y}^\star(x): \quad \pi_\beta^\star(y|x) \geq \frac{\pi_{\text{ref}}(y|x)}{\sum_{y\in\mathcal{Y}^\star(x)}\pi_{\text{ref}}(y|x)}\cdot\left(1 - (|\mathcal{Y}| - |\mathcal{Y}^\star(x)|)\exp\left(-\frac{1}{4\beta}\sum_{y\in\mathcal{Y}^\star(x)} \pi_{\text{ref}}(y|x)\right)\right),$$

and

$$\forall y \in \mathcal{Y}^\star(x): \quad \pi_\beta^\star(y|x) \leq \frac{\pi_{\text{ref}}(y|x)}{\sum_{y\in\mathcal{Y}^\star(x)}\pi_{\text{ref}}(y|x)}.$$

Recall that we write the expression of $\overline{\pi}_0^\star$ in (42). Therefore, we have

$$\forall x \in \mathcal{X}: \quad \left\|\pi_{\beta,x}^\star - \overline{\pi}_{0,x}^\star\right\|_1 \leq 2(|\mathcal{Y}| - |\mathcal{Y}^\star(x)|)\exp\left(-\frac{1}{4\beta}\sum_{y\in\mathcal{Y}^\star(x)} \pi_{\text{ref}}(y|x)\right). \tag{57}$$

For the iteration in Algorithm 1, similar to (54) we have

$$\log\left(\frac{\pi^{(t+1)}(y_1|x)}{\pi^{(t+1)}(y|x)}\right) = \frac{\eta}{1+\eta\beta}\sum_{i=0}^{t}\left(\frac{1}{1+\beta\eta}\right)^i \underbrace{\left(\log\left(\frac{1}{\overline{P}_x(y,:)\pi_{t-i,x}}\right) + \beta\log\left(\frac{\pi_{\text{ref}}(y_1|x)}{\pi_{\text{ref}}(y|x)}\right)\right)}_{\delta^{(t-i)}}$$
$$+\left(\frac{1}{1+\beta\eta}\right)^{t+1}\log\left(\frac{\pi_{\text{ref}}(y_1|x)}{\pi_{\text{ref}}(y|x)}\right). \tag{58}$$

Note that for all $y \in \mathcal{Y} \setminus \mathcal{Y}^\star(x)$, we have

$$\log\left(\frac{1}{\overline{P}_x(y,:)\pi_{t-i,x}}\right) \geq \log\left(\frac{1}{1-\sum_{y\in\mathcal{Y}^\star(x)}\pi_{t-i}(y|x)}\right) \geq \log\left(\frac{1}{1-\sum_{y\in\mathcal{Y}^\star(x)}\pi_{\text{ref}}(y|x)}\right) \geq \sum_{y\in\mathcal{Y}^\star(x)}\pi_{\text{ref}}(y|x),$$

where in the second inequality we use the fact that $\pi_t(y|x)$ is decreasing with t for all $y \notin \mathcal{Y}^\star(x)$. Then by a similar argument as in (55), we have

$$\forall x \in \mathcal{X}, \forall y \neq \mathcal{Y}^\star(x): \quad \delta^{(t-i)} \geq \frac{3}{4}\sum_{y\in\mathcal{Y}^\star(x)} \pi_{\text{ref}}(y|x). \tag{59}$$

Therefore, analogous to (57), we have

$$\forall x \in \mathcal{X}: \quad \left\|\pi_{\beta,x}^\star - \overline{\pi}_{\beta,x}^\star\right\|_1 \leq 2(|\mathcal{Y}| - |\mathcal{Y}^\star(x)|)\exp\left(-\frac{3}{4\beta}\sum_{y\in\mathcal{Y}^\star(x)} \pi_{\text{ref}}(y|x)\right). \tag{60}$$

Combining (57) and (60), we have

$$\left\|\overline{\pi}^\star_{\beta,x} - \pi^\star_{\beta,x}\right\|_1 \le \left\|\overline{\pi}^\star_{\beta,x} - \overline{\pi}^\star_{0,x}\right\|_1 + \left\|\pi^\star_{\beta,x} - \overline{\pi}^\star_{0,x}\right\|_1$$
$$\le 2(|\mathcal{Y}| - |\mathcal{Y}^\star(x)|)\left(\exp\left(-\frac{3}{4\beta}\sum_{y\in\mathcal{Y}^\star(x)}\pi_{\mathrm{ref}}(y|x)\right) + \exp\left(-\frac{1}{4\beta}\sum_{y\in\mathcal{Y}^\star(x)}\pi_{\mathrm{ref}}(y|x)\right)\right),$$

from which we can see that (40) holds.

B.3 Proof of Theorem 3

To start with, we reformulate problem (10) as a monotone variational inequality (VI) problem.

We first define the operator $F_x : \Delta_{\mathcal{Y}} \to \mathbb{R}^{|\mathcal{Y}|}$ for all $x \in \mathcal{X}$ as

$$F_x(\pi_x) \coloneqq -P_x\pi_x - \beta \log \pi_{\mathrm{ref},x}, \ \forall \pi_x \in \Delta_{\mathcal{Y}}. \tag{61}$$

We also let

$$h : \Delta_{\mathcal{Y}} \to \mathbb{R}, \ h(p) \coloneqq \sum_i p_i \log p_i \tag{62}$$

denote the negative entropy, which is 1-strongly convex on $\Delta_{\mathcal{Y}}$ w.r.t. the l_1-norm [Beck, 2017].

The following lemma gives the VI form of `WIND`.

Lemma 3. *Assume $\beta > 0$ and $\pi^{(0)}, \pi_{\mathrm{ref}} \in relint(\Delta^{\mathcal{X}}_{\mathcal{Y}})$. Then (10) is equivalent to the following monotone VI problem:*

$$\mathop{\mathbb{E}}_{x\sim\rho}\left[\left\langle F_x(\pi^\star_{\beta,x}) + \beta\nabla h(\pi^\star_{\beta,x}), \pi_x - \pi^\star_{\beta,x}\right\rangle\right] \ge 0, \ \forall \pi \in \Delta^{\mathcal{X}}_{\mathcal{Y}}, \tag{63}$$

where for all $x \in \mathcal{X}$, F_x is monotone and 1-Lipschitz continuous w.r.t. the l_1-norm.

Proof of Lemma 3. By the proof of Proposition 1 we know that when $\beta > 0$ and $\pi_{\mathrm{ref}} \in \mathrm{relint}(\Delta^{\mathcal{X}}_{\mathcal{Y}})$, we have $\pi^\star_\beta \in \mathrm{relint}(\Delta^{\mathcal{X}}_{\mathcal{Y}})$.

By the optimality condition, $\pi^\star_\beta$ satisfies (10) if and only if

$$\left\langle \nabla f^\star(\pi^\star_\beta), \pi - \pi^\star_\beta \right\rangle \ge 0, \ \forall \pi, \tag{64}$$

where

$$f^\star(\pi) \coloneqq \mathop{\mathbb{E}}_{x\sim\rho}\left[\left\langle \pi_x, -P_x\pi^\star_{\beta,x} - \beta\log\pi_{\mathrm{ref},x} + \beta\log\pi_x\right\rangle\right].$$

By (61) and (62), we have (64) equivalent to (63). To see the monotonicity of F_x, we have

$$\begin{aligned}
&\langle F_x(\pi_x) - F_x(\pi'_x), \pi_x - \pi'_x\rangle \\
&= (\pi_x - \pi'_x)^\top P_x(\pi_x - \pi'_x) \\
&= (\pi_x - \pi'_x)^\top \frac{1}{2}(P_x + P_x^\top)(\pi_x - \pi'_x) \\
&= (\pi_x - \pi'_x)^\top \frac{1}{2} J(\pi_x - \pi'_x) = 0,
\end{aligned} \tag{65}$$

where $J \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$ is the matrix of all ones.

Furthermore, we have

$$\forall x \in \mathcal{X}, \forall p, q \in \Delta_{\mathcal{Y}} : \quad \|F_x(p) - F_x(q)\|_\infty \le \|P_x(p-q)\|_\infty \le \|p - q\|_1, \tag{66}$$

where the second inequality follows from the fact that each entry of P_x only take its value in $\{0, 1/2, 1\}$. (66) indicates that F_x is 1-Lipschitz with respect to ℓ_1-norm. □

We use the following *proximal mirror descent ascent* rule [Sokota et al., 2023, Pattathil et al., 2023] to solve the monotone VI problem (63):

$$\pi^{(t+1)} = \arg\min_{\pi} \mathop{\mathbb{E}}_{x\sim\rho}\left[\left\langle F_x(\pi_x^{(t)}), \pi_x\right\rangle + \beta h(\pi_x) + \frac{1}{\eta} B_h(\pi_x, \pi_x^{(t)})\right], \tag{67}$$

where $\eta > 0$ is the learning rate, and the Bregman distance $B_h : \Delta_{\mathcal{Y}} \times \Delta_{\mathcal{Y}} \to \mathbb{R}_+$ is generated from the negative entropy h:

$$B_h(p,q) \coloneqq h(p) - h(q) - \langle \nabla h(q), p-q\rangle = \mathrm{KL}(p||q).$$

It's straightforward to verify that the analytical solution of (67) is (13) in Algorithm 2.

Note that the negative entropy h (c.f. (62)) is 1-strongly convex on $\Delta_{\mathcal{Y}}$ with respect to the ℓ_1-norm [Beck, 2017, Example 5.27]. Furthermore, Lemma 3 shows F_x is 1-Lipschitz with respect to l_1-norm. With these facts, the theorem follows directly from Sokota et al. [2023, Theorem 3.4]:

$$\forall x \in \mathcal{X}: \quad \mathrm{KL}(\pi_\beta^\star(\cdot|x)||\pi^{(t)}(\cdot|x)) \le \left(\frac{1}{1+\eta\beta}\right)^t \mathrm{KL}(\pi_\beta^\star(\cdot|x)||\pi^{(0)}(\cdot|x)).$$

(14) can be deduced from the above relation by taking the expectation over $x \sim \rho$ on both sides.

B.4 Proof of Lemma 1

To start with, we have

$$\begin{aligned}
&\mathbb{E}_{u,v}\left[(v-g(u))^2\right] \\
&= \mathbb{E}_{u,v}\left[\left((v-\mathbb{E}_v(v|u)) + (\mathbb{E}_v(v|u)-g(u))\right)^2\right] \\
&= \mathbb{E}_{u,v}\left[(v-\mathbb{E}_v(v|u))^2\right] + 2\mathbb{E}_{u,v}\left[(v-\mathbb{E}_v(v|u))(\mathbb{E}_v(v|u)-g(u))\right] + \mathbb{E}_{u,v}\left[(\mathbb{E}_v(v|u)-g(u))^2\right].
\end{aligned} \tag{68}$$

We use $F(u)$, $F(u,v)$ and $F(v|u)$ to denote the distribution of u, the joint distribution of u, v and the distribution of v conditioned on u, resp. Then the cross term

$$\begin{aligned}
\mathbb{E}_{u,v}\left[(v-\mathbb{E}_v(v|u))(\mathbb{E}_v(v|u)-g(u))\right] &= \int_{(u,v)} (v-\mathbb{E}_v(v|u))(\mathbb{E}_v(v|u)-g(u))dF(u,v) \\
&= \int_u \left(\int_v v-\mathbb{E}_v(v|u)dF(v|u)\right)(\mathbb{E}_v(v|u)-g(u))dF(u) \\
&= 0,
\end{aligned} \tag{69}$$

where the last relation follows from the fact that

$$\int_v v - \mathbb{E}_v(v|u)dF(v|u) = \mathbb{E}_v(v|u) - \mathbb{E}_v(v|u) = 0.$$

Combining (69) and (68), we have that

$$\mathbb{E}_{u,v}\left[(v-g(u))^2\right] = \mathbb{E}_{u,v}\left[(v-\mathbb{E}_v(v|u))^2\right] + \mathbb{E}_u\left[(\mathbb{E}_v(v|u)-g(u))^2\right] \ge \mathbb{E}_{u,v}\left[(v-\mathbb{E}_v(v|u))^2\right],$$

and the equality holds if and only if $g(u) = \mathbb{E}_v(v|u)$ almost everywhere on the support set of $F(u)$.

B.5 Proof of Theorem 4

We first introduce the three-point property of the Bregman divergence [Sokota et al., 2023, Proposition D.1], [Bauschke et al., 2003, Proposition 2.3]:

Lemma 4 (three-point property of the Bregman divergence)**.** *Let $\psi : \Delta_{\mathcal{Y}} \to \mathbb{R}$ be a function that's differentiable on int$(\Delta_{\mathcal{Y}})$. Let $p, q \in \Delta_{\mathcal{Y}}$ and $r, s \in$ int$(\Delta_{\mathcal{Y}})$. Then the following equality holds:*

$$B_\psi(r,s) + B_\psi(s,r) = \langle \nabla\psi(r) - \nabla\psi(s), r-s\rangle. \tag{70}$$

$$B_\psi(p,r) = B_\psi(p,s) + B_\psi(s,r) + \langle \nabla\psi(s) - \nabla\psi(r), p-s\rangle. \tag{71}$$

$$B_\psi(p,s) + B_\psi(q,r) = B_\psi(p,r) + B_\psi(q,s) + \langle \nabla\psi(r) - \nabla\psi(s), p-q\rangle. \tag{72}$$

To start with, we rewrite the update rule in Algorithm 3

$$\theta_{t+1} \leftarrow \arg\min_{\theta\in\Theta} \frac{1}{M}\sum_{i=1}^{M}\left(\varphi_t(x_i^{(t)}, y_i^{(t)}, {y'}_i^{(t)}) - \phi_\theta(y_i^{(t)}|x_i^{(t)})\right)^2. \tag{73}$$

to a similar form as the PMDA rule (67).

Step 1: reformulate the update rule (73). We let $\delta_S^{(t)}, \delta_P^{(t)} \in \mathbb{R}^{|\mathcal{X}|\times|\mathcal{Y}|}$ denote the statistical error and model approximation error at the t-th round, respectively:

$$\forall (x,y)\in\mathcal{X}\times\mathcal{Y}: \quad \delta_S^{(t)}(x,y) \coloneqq \phi_{\theta_{t+1}}(y|x) - \phi_{\theta^\star_{t+1}}(y|x), \tag{74}$$

$$\delta_P^{(t)}(x,y) \coloneqq \mathbb{E}_{y'\sim\pi_{\theta_t}(\cdot|x)}\widehat{P}_x(y,y') - \mathbb{E}_{y'\sim\pi_{\theta_t}(\cdot|x)}P_x(y,y'). \tag{75}$$

We write $\delta_{S,x}^{(t)} \in \mathbb{R}^{|\mathcal{X}|}, \delta_{P,x}^{(t)} \in \mathbb{R}^{|\mathcal{X}|}$ as the shorthand of $\left(\delta_S^{(t)}(x,y)\right)_{y\in\mathcal{Y}}, \left(\delta_P^{(t)}(x,y)\right)_{y\in\mathcal{Y}}$, resp.

The above expression (74) combined with (18) gives

$$\pi_{\theta_{t+1}}(y|x) \propto (\pi_{\theta_t}(y|x))^{\frac{1}{1+\beta\eta}}(\pi_{\text{ref}}(y|x))^{\frac{\beta\eta}{1+\beta\eta}}\exp\left(\frac{\eta}{1+\beta\eta}\left(\mathbb{E}_{y'\sim\pi_{\theta_t}(\cdot|x)}\widehat{P}_x(y,y') + \frac{1+\beta\eta}{\eta}\delta_S^{(t)}(x,y)\right)\right). \tag{76}$$

For notation simplicity we let $\Pi \coloneqq \Delta_{\mathcal{Y}}^{\mathcal{X}}$ denote the whole policy space. Note that the above relation is equivalent to

$$\pi_{\theta_{t+1}} = \arg\min_{\pi\in\Pi}\mathbb{E}_{x\sim\rho}\left[\left\langle \widehat{F}_x^{(t)}, \pi_x\right\rangle + \beta h(\pi_x) + \frac{1}{\eta}B_h(\pi_x, \pi_{\theta_t,x})\right], \tag{77}$$

where $\widehat{F}_x^{(t)} \in \mathbb{R}^{|\mathcal{Y}|}$ is defined as

$$\forall x\in\mathcal{X}: \quad \widehat{F}_x^{(t)} \coloneqq \underbrace{-P_x\pi_{\theta_t,x} - \beta\log\pi_{\text{ref},x}}_{=F_x(\pi_{\theta_t,x}) \text{ by (61)}} - \frac{1+\beta\eta}{\eta}\delta_{S,x}^{(t)} - \delta_{P,x}^{(t)}, \tag{78}$$

which could be seen as an approximation of $F_x(\pi_{\theta_t,x})$. We let $\delta^{(t)} \in \mathbb{R}^{|\mathcal{X}||\mathcal{Y}|}$ denote

$$\forall x\in\mathcal{X}: \quad \delta_x^{(t)} \coloneqq \delta^{(t)}(x,\cdot) \coloneqq \widehat{F}_x^{(t)} - F_x(\pi_{\theta_t,x}) = -\frac{1+\beta\eta}{\eta}\delta_{S,x}^{(t)} - \delta_{P,x}^{(t)}. \tag{79}$$

The next step is to bound the distance between π_{θ_t} and $\pi_\beta^\star$ utilizing the reformulated update rule (77). This part of our proof is inspired by Sokota et al. [2023, Theorem 3.4].

Step 2: bound $D_{\text{KL}}\left(\pi_\beta^\star\|\pi_{\theta_t}\right)$. By the first-order optimality condition we know that (77) is equivalent to

$$\left\langle \widehat{F}_x^{(t)} + \beta\nabla h(\pi_{\theta_{t+1},x}) + \frac{1}{\eta}(\nabla h(\pi_{\theta_{t+1},x}) - \nabla h(\pi_{\theta_t,x})), \pi_x - \pi_{\theta_{t+1},x}\right\rangle \geq 0, \ \forall \pi\in\Pi, \ \forall x\in\mathcal{X}, \tag{80}$$

Reorganizing the terms in (80), we have

$$\begin{aligned}\left\langle \widehat{F}_x^{(t)} + \beta\nabla h(\pi_{\theta_{t+1},x}), \pi_x - \pi_{\theta_{t+1},x}\right\rangle &\geq \frac{1}{\eta}\left\langle \nabla h(\pi_{\theta_t,x}) - \nabla h(\pi_{\theta_{t+1},x}), \pi_x - \pi_{\theta_{t+1},x}\right\rangle \\ &\overset{(71)}{=} \frac{1}{\eta}\left(-B_h(\pi_x, \pi_{\theta_t,x}) + B_h(\pi_x, \pi_{\theta_{t+1},x}) + B_h(\pi_{\theta_{t+1},x}, \pi_{\theta_t,x})\right).\end{aligned} \tag{81}$$

Let $\pi = \pi_\beta^\star$ in (81) and reorganize the terms, we have

$$\begin{aligned}&B_h(\pi^\star_{\beta,x}, \pi_{\theta_{t+1},x}) \\ &\leq B_h(\pi^\star_{\beta,x}, \pi_{\theta_t,x}) - B_h(\pi_{\theta_{t+1},x}, \pi_{\theta_t,x}) + \eta\left\langle \widehat{F}_x^{(t)} + \beta\nabla h(\pi_{\theta_{t+1},x}), \pi^\star_{\beta,x} - \pi_{\theta_{t+1},x}\right\rangle\end{aligned}$$

$$
\begin{aligned}
&= B_h(\pi^\star_{\beta,x}, \pi_{\theta_t,x}) - B_h(\pi_{\theta_{t+1},x}, \pi_{\theta_t,x}) + \eta \left\langle F_x(\pi_{\theta_t,x}) + \beta \nabla h(\pi_{\theta_{t+1},x}), \pi^\star_{\beta,x} - \pi_{\theta_{t+1},x} \right\rangle \\
&\quad + \eta \left\langle \delta_x^{(t)}, \pi^\star_{\beta,x} - \pi_{\theta_{t+1},x} \right\rangle \\
&= B_h(\pi^\star_{\beta,x}, \pi_{\theta_t,x}) - B_h(\pi_{\theta_{t+1},x}, \pi_{\theta_t,x}) + \eta \left\langle F_x(\pi_{\theta_t,x}) - F_x(\pi_{\theta_{t+1},x}), \pi^\star_{\beta,x} - \pi_{\theta_{t+1},x} \right\rangle \\
&\quad + \eta \left\langle F_x(\pi_{\theta_{t+1},x}) + \beta \nabla h(\pi_{\theta_{t+1},x}), \pi^\star_{\beta,x} - \pi_{\theta_{t+1},x} \right\rangle + \eta \left\langle \delta_x^{(t)}, \pi^\star_{\beta,x} - \pi_{\theta_{t+1},x} \right\rangle, \ \forall x \in \mathcal{X}, \ \forall \pi \in \Pi. \qquad (82)
\end{aligned}
$$

Note that for any $\pi \in \Pi$ and $x \in \mathcal{X}$, we have

$$
\begin{aligned}
\left\langle F_x(\pi_x) + \beta \nabla h(\pi_x), \pi^\star_{\beta,x} - \pi_x \right\rangle &= \underbrace{\left\langle F_x(\pi_x) - F_x(\pi^\star_{\beta,x}), \pi^\star_{\beta,x} - \pi_x \right\rangle}_{=0 \text{ by } (65)} + \beta \left\langle \nabla h(\pi_x) - \nabla h(\pi^\star_{\beta,x}), \pi^\star_{\beta,x} - \pi_x \right\rangle \\
&\quad + \underbrace{\left\langle F_x(\pi^\star_{\beta,x}) + \beta \nabla h(\pi^\star_{\beta,x}), \pi^\star_{\beta,x} - \pi_x \right\rangle}_{\leq 0 \text{ by } (63)} \\
&\leq \beta \left\langle \nabla h(\pi_x) - \nabla h(\pi^\star_{\beta,x}), \pi^\star_{\beta,x} - \pi_x \right\rangle \\
&\overset{(70)}{=} -\beta \left(B_h(\pi_x, \pi^\star_{\beta,x}) + B_h(\pi^\star_{\beta,x}, \pi_x) \right). \qquad (83)
\end{aligned}
$$

Combining the above two expressions (82) and (83), we have

$$
\begin{aligned}
B_h(\pi^\star_{\beta,x}, \pi_{\theta_{t+1},x}) &\leq B_h(\pi^\star_{\beta,x}, \pi_{\theta_t,x}) - B_h(\pi_{\theta_{t+1},x}, \pi_{\theta_t,x}) + \eta \left\langle F_x(\pi_{\theta_t,x}) - F_x(\pi_{\theta_{t+1},x}), \pi^\star_{\beta,x} - \pi_{\theta_{t+1},x} \right\rangle \\
&\quad - \beta\eta \left(B_h(\pi_{\theta_{t+1},x}, \pi^\star_{\beta,x}) + B_h(\pi^\star_{\beta,x}, \pi_{\theta_{t+1},x}) \right) + \eta \left\langle \delta_x^{(t)}, \pi^\star_{\beta,x} - \pi_{\theta_{t+1},x} \right\rangle \\
&\overset{(21)}{\leq} B_h(\pi^\star_{\beta,x}, \pi_{\theta_t,x}) - B_h(\pi_{\theta_{t+1},x}, \pi_{\theta_t,x}) + \eta \left\| \pi_{\theta_t,x} - \pi_{\theta_{t+1},x} \right\|_1 \left\| \pi^\star_{\beta,x} - \pi_{\theta_{t+1},x} \right\|_1 \\
&\quad - \beta\eta \left(B_h(\pi_{\theta_{t+1},x}, \pi^\star_{\beta,x}) + B_h(\pi^\star_{\beta,x}, \pi_{\theta_{t+1},x}) \right) + \eta \left\langle \delta_x^{(t)}, \pi^\star_{\beta,x} - \pi_{\theta_{t+1},x} \right\rangle \\
&\leq B_h(\pi^\star_{\beta,x}, \pi_{\theta_t,x}) \underbrace{- B_h(\pi_{\theta_{t+1},x}, \pi_{\theta_t,x}) + \frac{1}{2} \left\| \pi_{\theta_t,x} - \pi_{\theta_{t+1},x} \right\|_1^2}_{\leq 0} + \underbrace{\frac{\eta^2}{2} \left\| \pi^\star_{\beta,x} - \pi_{\theta_{t+1},x} \right\|_1^2}_{\leq \eta^2 B_h(\pi_{\theta_{t+1},x}, \pi^\star_{\beta,x})} \\
&\quad - \beta\eta \left(B_h(\pi_{\theta_{t+1},x}, \pi^\star_{\beta,x}) + B_h(\pi^\star_{\beta,x}, \pi_{\theta_{t+1},x}) \right) + \eta \left\langle \delta_x^{(t)}, \pi^\star_{\beta,x} - \pi_{\theta_{t+1},x} \right\rangle \\
&\leq B_h(\pi^\star_{\beta,x}, \pi_{\theta_t,x}) + \eta \underbrace{(\eta - \beta)}_{\leq 0} B_h(\pi_{\theta_{t+1},x}, \pi^\star_{\beta,x}) - \beta\eta B_h(\pi^\star_{\beta,x}, \pi_{\theta_{t+1},x}) + \eta \left\langle \delta_x^{(t)}, \pi^\star_{\beta,x} - \pi_{\theta_{t+1},x} \right\rangle \\
&\leq B_h(\pi^\star_{\beta,x}, \pi_{\theta_t,x}) - \beta\eta B_h(\pi^\star_{\beta,x}, \pi_{\theta_{t+1},x}) + \eta \left\langle \delta_x^{(t)}, \pi^\star_{\beta,x} - \pi_{\theta_{t+1},x} \right\rangle, \qquad (84)
\end{aligned}
$$

where, in the third relation we use the 1-strong convexity of h w.r.t. the l_1-norm (see the proof of Theorem 3) to obtain that

$$
\begin{aligned}
&- B_h(\pi_{\theta_{t+1},x}, \pi_{\theta_t,x}) + \frac{1}{2} \left\| \pi_{\theta_t,x} - \pi_{\theta_{t+1},x} \right\|_1^2 \\
&= - \left(h(\pi_{\theta_{t+1},x}) - h(\pi_{\theta_t,x}) - \left\langle \nabla h(\pi_{\theta_t,x}), \pi_{\theta_{t+1},x} - \pi_{\theta_t,x} \right\rangle - \frac{1}{2} \left\| \pi_{\theta_t,x} - \pi_{\theta_{t+1},x} \right\|_1^2 \right) \leq 0.
\end{aligned}
$$

Note that

$$
\begin{aligned}
\left\langle \delta_x^{(t)}, \pi^\star_{\beta,x} - \pi_{\theta_{t+1},x} \right\rangle &\overset{(79)}{=} \left\langle -\frac{1+\beta\eta}{\eta} \delta_{S,x}^{(t)} - \delta_{P,x}^{(t)}, \pi^\star_{\beta,x} - \pi_{\theta_{t+1},x} \right\rangle \\
&\leq \frac{1+\beta\eta}{\eta} \underbrace{\left| \left\langle \delta_{S,x}^{(t)}, \pi^\star_{\beta,x} \right\rangle \right|}_{(i)} + \frac{1+\beta\eta}{\eta} \underbrace{\left| \left\langle \delta_{S,x}^{(t)}, \pi_{\theta_{t+1},x} \right\rangle \right|}_{(ii)} + \underbrace{\left\| \delta_{P,x}^{(t)} \right\|_\infty \left\| \pi^\star_{\beta,x} - \pi_{\theta_{t+1},x} \right\|_1}_{(iii)}. \qquad (85)
\end{aligned}
$$

To bound the error term, below we separately bound (i)-(iii).

To bound (i), we first unroll (76) similar as in (45) and obtain

$$
\log \pi_{\theta_{t+1}}(y|x) = \frac{1}{1+\beta\eta} \log \pi_{\theta_t}(y|x) + \frac{\beta\eta}{1+\beta\eta} \pi_{\text{ref}}(y|x)
$$

$$+\frac{\eta}{1+\beta\eta}\left(\mathbb{E}_{y'\sim\pi_{\theta_t}(\cdot|x)}P_x(y,y')+\frac{1+\beta\eta}{\eta}\delta_S^{(t)}(x,y)+\delta_P^{(t)}(x,y)\right)$$

$$=\log\pi_{\mathrm{ref}}(y|x)+\frac{\eta}{1+\eta\beta}\sum_{i=0}^{t}\left(\frac{1}{1+\beta\eta}\right)^i\left(\mathbb{E}_{y'\sim\pi_{\theta_{t-i}}(\cdot|x)}P_x(y,y')\right.$$
$$\left.+\frac{1+\beta\eta}{\eta}\delta_S^{(t-i)}(x,y)+\delta_P^{(t-i)}(x,y)\right)+z_x \tag{86}$$

for some z_x related to x. The above expression gives

$$\log\left(\frac{\pi_{\theta_{t+1}}(y'|x)}{\pi_{\theta_{t+1}}(y|x)}\right)=\log\left(\frac{\pi_{\mathrm{ref}}(y'|x)}{\pi_{\mathrm{ref}}(y|x)}\right)+\frac{\eta}{1+\eta\beta}\sum_{i=0}^{t}\left(\frac{1}{1+\beta\eta}\right)^i\left(\mathbb{E}_{y''\sim\pi_{\theta_{t-i}}(\cdot|x)}(P_x(y',y'')-P_x(y,y''))\right.$$
$$\left.+\frac{1+\beta\eta}{\eta}\delta_S^{(t-i)}(x,y')+\delta_P^{(t-i)}(x,y')-\frac{1+\beta\eta}{\eta}\delta_S^{(t-i)}(x,y)-\delta_P^{(t-i)}(x,y)\right),$$

for any $y,y'\in\mathcal{Y}$. This relation yields

$$\log\left(\frac{\pi_{\theta_{t+1}}(y'|x)}{\pi_{\theta_{t+1}}(y|x)}\right)\le\log\left(\frac{\pi_{\mathrm{ref}}(y'|x)}{\pi_{\mathrm{ref}}(y|x)}\right)+\frac{\eta}{1+\eta\beta}\sum_{i=0}^{t}\left(\frac{1}{1+\beta\eta}\right)^i\cdot 2\left(\delta_P+\frac{1+\beta\eta}{\eta}L_0+1\right),$$

where we use (21) and (74). The above expression indicates

$$\frac{\pi_{\theta_{t+1}}(y'|x)}{\pi_{\theta_{t+1}}(y|x)}\le\frac{\pi_{\mathrm{ref}}(y'|x)}{\pi_{\mathrm{ref}}(y|x)}\underbrace{\exp\left(\frac{2}{\beta}\left(\delta_P+\frac{1+\beta\eta}{\eta}L_0+1\right)\right)}_{C_2},$$

Summing over $y'\in\mathcal{Y}$ on both sides, we get

$$\forall y\in\mathcal{Y}:\quad\frac{1}{\pi_{\theta_{t+1}}(y|x)}\le\frac{1}{\pi_{\theta_{\mathrm{ref}}}(y|x)}C_2. \tag{87}$$

Therefore, we have

$$\begin{aligned}\left|\left\langle\delta_{S,x}^{(t)},\pi_{\beta,x}^\star\right\rangle\right|&=\sum_{y\in\mathcal{Y}}\frac{\pi_\beta^\star(y|x)}{\sqrt{\pi_{\theta_t}(y|x)}}\sqrt{\pi_{\theta_t}(y|x)\left(\delta_S^{(t)}(x,y)\right)^2}\\&\le\sqrt{\left(\sum_{y\in\mathcal{Y}}\frac{\left(\pi_\beta^\star(y|x)\right)^2}{\pi_{\theta_t}(y|x)}\right)\left(\sum_{y\in\mathcal{Y}}\pi_{\theta_t}(y|x)\left(\delta_S^{(t)}(x,y)\right)^2\right)}\\&=\sqrt{\mathbb{E}_{y\sim\pi_\beta^\star(\cdot|x)}\left[\frac{\pi_\beta^\star(y|x)}{\pi_{\theta_t}(y|x)}\right]\mathbb{E}_{y\sim\pi_{\theta_t}(\cdot|x)}\left[\left(\delta_S^{(t)}(x,y)\right)^2\right]}\\&\le\sqrt{C_2\mathbb{E}_{y\sim\pi_\beta^\star(\cdot|x)}\left[\frac{\pi_\beta^\star(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right]\mathbb{E}_{y\sim\pi_{\theta_t}(\cdot|x)}\left[\left(\delta_S^{(t)}(x,y)\right)^2\right]}\\&\le\sqrt{C_2\mathbb{E}_{y\sim\pi_{\mathrm{ref}}(\cdot|x)}\left[\frac{\pi_\beta^\star(y|x)}{\pi_{\mathrm{ref}}(y|x)}\right]^2\mathbb{E}_{y\sim\pi_{\theta_t}(\cdot|x)}\left[\left(\delta_S^{(t)}(x,y)\right)^2\right]}\\&\le\sqrt{C_1\mathbb{E}_{y\sim\pi_{\theta_t}(\cdot|x)}\left[\left(\delta_S^{(t)}(x,y)\right)^2\right]},\end{aligned} \tag{88}$$

where the second line follows from Cauchy-Schwartz inequality, and the last line uses Assumption 3.

By the same argument, we could also bound (ii):

$$\left|\left\langle\delta_{S,x}^{(t)},\pi_{\theta_{t+1},x}\right\rangle\right|\le\sqrt{C_1\mathbb{E}_{y\sim\pi_{\theta_t}(\cdot|x)}\left[\left(\delta_S^{(t)}(x,y)\right)^2\right]}. \tag{89}$$

For term (iii), note that $\left\|\delta_{P,x}^{(t)}\right\|_\infty \le \delta_P$, where δ_P is defined in (75), we have

$$\left\|\delta_{P,x}^{(t)}\right\|_\infty \left\|\pi^\star_{\beta,x} - \pi_{\theta_{t+1},x}\right\|_1 \le 2\delta_P. \tag{90}$$

Thus combining (88),(89),(90) with (85), we have

$$\left\langle \delta_x^{(t)}, \pi^\star_{\beta,x} - \pi_{\theta_{t+1},x} \right\rangle \le 2 \cdot \frac{1+\beta\eta}{\eta} \sqrt{C_1 \mathbb{E}_{y\sim\pi_{\theta_t}(\cdot|x)}\left[\left(\delta_S^{(t)}(x,y)\right)^2\right]} + 2\delta_P, \ \forall x \in \mathcal{X}. \tag{91}$$

Taking expectation w.r.t. x on both sides of (84) and making use of (91), we have

$$\begin{aligned}
&D_{\mathrm{KL}}\left(\pi^\star_\beta \| \pi_{\theta_{t+1}}\right) \\
&= \mathbb{E}_{x\sim\rho}\left[B_h(\pi^\star_{\beta,x}, \pi_{\theta_{t+1},x})\right] \\
&\le \frac{1}{1+\beta\eta}\mathbb{E}_{x\sim\rho}\left[B_h(\pi^\star_{\beta,x}, \pi_{\theta_t,x})\right] + 2\left(\mathbb{E}_{x\sim\rho}\sqrt{C_1\mathbb{E}_{y\sim\pi_{\theta_t}(\cdot|x)}\left[\left(\delta_S^{(t)}(x,y)\right)^2\right]} + \frac{\eta}{1+\beta\eta}\delta_P\right) \\
&\le \frac{1}{1+\beta\eta} D_{\mathrm{KL}}\left(\pi^\star_\beta \| \pi_{\theta_t}\right) + 2\left(\sqrt{C_1\mathbb{E}_{x\sim\rho, y\sim\pi_{\theta_t}(\cdot|x)}\left[\left(\delta_S^{(t)}(x,y)\right)^2\right]} + \frac{\eta}{1+\beta\eta}\delta_P\right) \\
&= \frac{1}{1+\beta\eta} D_{\mathrm{KL}}\left(\pi^\star_\beta \| \pi_{\theta_t}\right) + 2\left(\sqrt{C_1\mathbb{E}_{x\sim\rho, y\sim\pi_{\theta_t}(\cdot|x)}\left[\left(\delta_S^{(t)}(x,y)\right)^2\right]} + \frac{\eta}{1+\beta\eta}\delta_P\right)
\end{aligned} \tag{92}$$

where the second inequality follows from Jensen's inequality and $\delta_S^{(t)}(x,y)$.

Note that

$$\begin{aligned}
&\mathbb{E}_{x\sim\rho, y\sim\pi_{\theta_t}(\cdot|x)}\left[\left(\delta_S^{(t)}(x,y)\right)^2\right] \\
&\overset{(74)}{=} \mathbb{E}_{x\sim\rho, y\sim\pi_{\theta_t}(\cdot|x)}\left[\left(\phi_{\theta_{t+1}}(y|x) - \phi_{\theta^\star_{t+1}}(y|x)\right)^2\right] \\
&= \mathbb{E}_{x\sim\rho, y\sim\pi_{\theta_t}(\cdot|x)}\left[\left(\phi_{\theta_{t+1}}(y|x) + \phi_{\theta^\star_{t+1}}(y|x) - 2\mathbb{E}_{y'\sim\pi_{\theta_t}(\cdot|x)}[\varphi_t(x,y,y')|x,y]\right)\left(\phi_{\theta_{t+1}}(y|x) - \phi_{\theta^\star_{t+1}}(y|x)\right)\right] \\
&= \mathbb{E}_{\substack{x\sim\rho, y\sim\pi_{\theta_t}(\cdot|x), \\ y'\sim\pi_{\theta_t}(\cdot|x)}}\left[\left(\phi_{\theta_{t+1}}(y|x) - \varphi_t(x,y,y')\right)^2 - \left(\phi_{\theta^\star_{t+1}}(y|x) - \varphi_t(x,y,y')\right)^2\right] \\
&= R_t(\theta_{t+1}) - R_t(\theta^\star_{t+1}) \\
&= R_t(\theta_{t+1}) - R_t^\star,
\end{aligned} \tag{93}$$

where R_t is defined in (19), $R_t^\star \coloneqq \min_{\theta\in\Theta} R_t(\theta) = R_t(\theta^\star_{t+1})$, and the third line uses Assumption 1.

Combining the above expression (93) with (92), we obtain

$$D_{\mathrm{KL}}\left(\pi^\star_\beta \| \pi_{\theta_{t+1}}\right) \le \frac{1}{1+\beta\eta} D_{\mathrm{KL}}\left(\pi^\star_\beta \| \pi_{\theta_t}\right) + 2\Bigg(\underbrace{\sqrt{C_1(R_t(\theta_{t+1}) - R_t^\star)} + \frac{\eta}{1+\beta\eta}\delta_P}_{\coloneqq \xi_t}\Bigg). \tag{94}$$

The above expression implies we need to bound ξ_t. If for all t, ξ_t could be bounded by some finite ξ, then by (94) we have

$$\begin{aligned}
D_{\mathrm{KL}}\left(\pi^\star_\beta \| \pi_{\theta_t}\right) &\le \left(\frac{1}{1+\beta\eta}\right)^t D_{\mathrm{KL}}\left(\pi^\star_\beta \| \pi_{\theta_0}\right) + 2\sum_{s=0}^{t-1}\left(\frac{1}{1+\beta\eta}\right)^s \xi \\
&\le \left(\frac{1}{1+\beta\eta}\right)^t D_{\mathrm{KL}}\left(\pi^\star_\beta \| \pi_{\theta_0}\right) + \frac{2(1+\beta\eta)}{\beta\eta}\xi.
\end{aligned} \tag{95}$$

In the following, we bound ξ_t by bounding the *excess risk* $R_t(\theta_{t+1}) - R_t^\star$.

Step 3: bound the excess risk. To bound the excess risk, we first introduce the concept of *uniform stability* [Bousquet and Elisseeff, 2002]. Suppose we have a training dataset $\mathcal{D} = \{z_1, \cdots, z_M\}$ where each z_i is sampled i.i.d. from some unknown distribution P defined on some abstract set $\mathcal{Z}$. Given $\mathcal{D}$, a *learning algorithm* produces the decision rule $w_M = w_M(\mathcal{D}) = w_M(z_1, \cdots, z_M) \in \mathcal{W}$, where $\mathcal{W}$ is the set of all decision rules and is assumed to be a closed subset of a separable Hilbert space. We use w_M to refer to both the algorithm and the decision rule. For the loss function $\ell : \mathcal{Z} \times \mathcal{W} \to [0, \infty)$, we define the risk and the empirical risk of $w \in \mathcal{W}$ respectively as

$$R(w) = \mathbb{E}_{z\sim P}\ell(z, w) \quad \text{and} \quad R_M(w) = \frac{1}{M}\sum_{i=1}^{M} \ell(z_i, w). \tag{96}$$

Definition 1. *An algorithm w_M is uniformly γ-stable, if for any $z, z', z_1, \cdots, z_M \in \mathcal{Z}$ and any $i \in [M]$, it holds that*

$$|\ell(z, w_M(z_1, \cdots, z_M)) - \ell(z, w_M(z_1, \cdots, z_{i-1}, z', z_{i+1}, \cdots, x_M))| \le \gamma. \tag{97}$$

We will also use the *generalized Bernstein condition* defined as follows:

Definition 2 (Assumption 1.1 in Klochkov and Zhivotovskiy [2021])**.** *Define $\mathcal{W}^\star := \arg\min_{w\in\mathcal{W}} R(w)$ where $\mathcal{W}$ is a closed set. We say that $(\mathcal{W}, P, \ell)$ satisfies the generalized Bernstein condition if there exists some constant $B > 0$ such that for any $w \in \mathcal{W}$, there exists $w^\star \in \mathcal{W}^\star$ that satisfies*

$$\mathbb{E}_{z\in P}\left[(\ell(w, z) - \ell(w^\star, z))^2\right] \le B(R(w) - R(w^\star)). \tag{98}$$

With the above two lemmas, we now introduce the following important lemma that bounds the generalization error for uniformly stable algorithms:

Lemma 5 (Theorem 1.1 in Klochkov and Zhivotovskiy [2021])**.** *Assume loss ℓ is bounded by C on $\mathcal{Z} \times \mathcal{W}$, and $(\mathcal{W}, P, \ell)$ satisfies the generalized Bernstein condition with the parameter B (c.f. Definition 2). Let w be a γ-stable algorithm (c.f. Definition 1) that returns $w_M \in \arg\min_{w\in\mathcal{W}} R_M(w)$ given the training dataset $\mathcal{D}$. Then with probability at least $1 - \delta$, it holds that*

$$R(w_M) - \inf_{w\in\mathcal{W}} R(w) \le C_r\left(\gamma \log M + \frac{C + B}{M}\right)\log\left(\frac{1}{\delta}\right), \tag{99}$$

where $C_r > 0$ is an absolute constant.

To proceed, we analyze the generalization error at the t-th iterate of Algorithm 3 for a fixed arbitrary $t \in \mathbb{N}$. We'll let $\widehat{\theta}$ denote θ_{t+1} and drop superscript/subscript t when this causes no confusion. For example, we'll simply write the update rule (73) as

$$\widehat{\theta} \leftarrow \arg\min_{\theta\in\Theta} \frac{1}{M}\sum_{i=1}^{M} \left(\varphi(x_i, y_i, {y'}_i) - \phi_\theta(y_i|x_i)\right)^2.$$

For notation simplicity, we also let $u_i = (x_i, y_i)$, $v_i = \varphi(x_i, y_i, {y'}_i)$, $z_i = (u_i, v_i) \in \mathcal{Z} := \mathcal{X} \times \mathcal{Y} \times \mathbb{R}$, and let $\phi_\theta(u_i)$ denote $\phi_\theta(y_i|x_i)$ for all $i \in [M]$. Let $\mathcal{Z} := \mathcal{X}$. Then in our case, the loss function $\ell : \mathcal{Z} \times \Theta \to \mathbb{R}_+$ has the following form:

$$\ell(z, \theta) := (v - \phi_\theta(u))^2, \tag{100}$$

where $z = (u, v) \in \mathcal{Z}$, and similar as (96), our risk and empirical risk at the t-th itrerate satisfy:

$$\forall \theta \in \Theta: \quad R(\theta) = \mathbb{E}_{z\sim P}\ell(z, \theta) \quad \text{and} \quad R_M(\theta) = \frac{1}{M}\sum_{i=1}^{M} \ell(z_i, \theta), \tag{101}$$

where we let P denote the distribution of z_i ($i \in [M]$), and we have

$$\theta^\star = \arg\min_{\theta\in\Theta} R(\theta) \quad \text{and} \quad \widehat{\theta} = \arg\min_{\theta\in\Theta} R_M(\theta). \tag{102}$$

Mote that (22) implies that $L(z, \theta)$ is L-Lipschitz over θ for any $z \in \mathcal{Z}$. Then by Assumption 4 and Remark 3 in Kang et al. [2022], we have that (Θ, P, ℓ) satisfies the generalized Bernstein condition with

$$B = \frac{2L^2}{\mu}. \tag{103}$$

Furthermore, Corollary 4 in Charles and Papailiopoulos [2018] gives that, when Assumption 2,4 hold, the empirical risk R_M is γ-uniform stability (c.f. Definition 1) with

$$\gamma = \frac{2L^2}{\mu(M-1)}. \tag{104}$$

Substituting (103) and (104) into (99), we obtain that for any fixed t, with probability at least $1-\delta$, we have

$$R_t(\widehat{\theta}) - R_t^\star \le C_r \left(\frac{2L^2 \log M}{\mu(M-1)} + \frac{C + 2L^2/\mu}{M} \right) \log\left(\frac{1}{\delta}\right). \tag{105}$$

By the independence of the samples in different rounds we know that with probability at least $1-\delta$, we have

$$\begin{aligned} \forall t \le T-1: \quad R_t(\widehat{\theta}) - R_t^\star &\le C_r \left(\frac{2L^2 \log M}{\mu(M-1)} + \frac{C + 2L^2/\mu}{M} \right) \log\left(\frac{1}{1-(1-\delta)^{1/T}}\right) \\ &\le C_r \left(\frac{2L^2 \log M}{\mu(M-1)} + \frac{C + 2L^2/\mu}{M} \right) \log\left(\frac{T}{\delta}\right) \end{aligned} \tag{106}$$

Step 4: put everything together. Let

$$\xi \coloneqq \sqrt{C_1 C_r \left(\frac{2L^2 \log M}{\mu(M-1)} + \frac{C + 2L^2/\mu}{M} \right) \log\left(\frac{T}{\delta}\right)} + \frac{\eta}{1+\beta\eta}\delta_P. \tag{107}$$

Then (107), (106) and (95) together give the desired result.