
Accelerated Preference Optimization for Large Language Model Alignment

Jiafan He

Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
jiafanhe19@ucla.edu

Huizhuo Yuan

Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
hzyuan@cs.ucla.edu

Quanquan Gu

Department of Computer Science
University of California, Los Angeles
Los Angeles, CA 90095
qgu@cs.ucla.edu

Abstract

Reinforcement Learning from Human Feedback (RLHF) has emerged as a pivotal tool for aligning large language models (LLMs) with human preferences. Direct Preference Optimization (DPO), one of the most popular approaches, formulates RLHF as a policy optimization problem without explicitly estimating the reward function. It overcomes the stability and efficiency issues of two-step approaches, which typically involve first estimating the reward function and then optimizing the policy via proximal policy optimization (PPO). Since RLHF is essentially an optimization problem, and it is well-known that momentum techniques can accelerate optimization both theoretically and empirically, a natural question arises: Can RLHF be accelerated by momentum? This paper answers this question in the affirmative. In detail, we first show that the iterative preference optimization method can be viewed as a proximal point method. Based on this observation, we propose a general Accelerated Preference Optimization (APO) framework, which unifies many existing preference optimization algorithms and employs Nesterov’s momentum technique to speed up the alignment of LLMs. Theoretically, we demonstrate that APO can achieve a faster convergence rate than the standard iterative preference optimization methods, including DPO and SPPO. Empirically, we show the superiority of APO over DPO, iterative DPO, and other strong baselines for RLHF on the AlpacaEval 2.0 benchmark.

1 Introduction

Large Language Models (LLMs) have emerged as a pivotal technique in the era of artificial general intelligence and recently demonstrated impressive capabilities in tasks such as text generation (Bubeck et al., 2023; Anil et al., 2023; Touvron et al., 2023), coding (Chen et al., 2021; Austin et al., 2021), and problem solving (Cobbe et al., 2021; Wei et al., 2022). A key element contributing to these achievements is the alignment of LLMs with human preference data, utilizing reinforcement learning from human feedback (RLHF) (Ziegler et al., 2019; Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022; Munos et al., 2023).

The standard RLHF method (Ouyang et al., 2022) involves three main steps: feedback collection, reward modeling, and policy optimization. Specifically, the LLM receives human-generated prompts and produces several possible responses. Subsequently, human preferences for these responses are

collected and used to train a reward model that matches these preferences. Finally, the policy optimization process updates the large language model using optimization algorithms such as Proximal Policy Optimization (PPO) (Schulman et al., 2017) to produce responses with higher preference based on the trained reward model.

Recently, Rafailov et al. (2023) introduced the Direct Preference Optimization (DPO) method based on the Bradley-Terry (BT) model (Bradley and Terry, 1952). This method skips the reward modeling process and replaces it with a reparameterized reward with respect to the policy, allowing for direct optimization of the LLM. It simplifies the implementation, while maintaining a comparable or even better performance than standard RLHF. Later, several works (Munos et al., 2023; Azar et al., 2024; Wu et al., 2024; Rosset et al., 2024) considered the general preference model rather than the BT model, and designed algorithms (Nash-MD, IPO, SPPO, and DNO) that more flexibly represent human preferences while maintaining a simple implementation. Nonetheless, both the standard RLHF process and its variations rely on the policy optimization process. In fact, for general optimization problems, it is well-known that momentum techniques (Polyak, 1964; Nesterov, 2013; Bubeck et al., 2015b) can accelerate the convergence of the optimization algorithm in both theory and practice. Therefore, a natural question arises:

Can reinforcement learning from human feedback be accelerated?

In this paper, we affirmatively answer this question. In details, following Xu et al. (2023); Yuan et al. (2024b); Chen et al. (2024); Wu et al. (2024), we consider the iterative preference optimization framework. In this framework, a series of models is constructed, with each model being improved based on the previous one using preference optimization algorithm such as DPO and SPPO. The contributions of our work are highlighted as follows:

- We propose a general preference optimization framework, APO, based on Nesterov’s momentum to accelerate preference optimization. Our theoretical analysis shows that iterative DPO achieves an $\tilde{O}(1/t)$ sub-optimality gap from the optimal policy under the Bradley-Terry (BT) model. As a comparison, our algorithm APO achieves a smaller $\tilde{O}((1-\alpha)/t)$ sub-optimality gap, where α is the extrapolation parameter in the momentum. To the best of our knowledge, our work provides the first convergence analysis of iterative DPO and the first accelerated preference optimization algorithm with provable guarantees.
- With an additional minimal sub-optimality gap assumption in the BT model, we prove that APO will converge to the optimal policy in total variation (TV) distance at the rate of $\exp(-O(t/(1-\alpha)))$, improving the $\exp(-O(t))$ rate of iterative DPO. In addition, we extend our results to the general preference model and show that APO with the SPPO loss function can also accelerate SPPO under a similar minimal sub-optimality gap assumption.
- Empirically, we verify the performance of APO when applied to DPO method on fine-tuning Mistral-7B-Instruct-v0.2. In the AlpacaEval 2.0 evaluation tasks, APO with 3 iterations achieves a length-controlled win rate of 31.73%, demonstrating a 1.78% improvement over iterative DPO and a 5.34% improvement over Snorkel’s Mistral-PairRM-DPO. Moreover, APO with 2 iterations obtains a win rate of 37.53%, matching iterative DPO’s 37.65% with 3 iterations, with noticeably shorter response lengths. This is consistent with our theoretical analysis. In addition, the evaluation on five general instruction-following tasks from MT-Bench shows an average score of 9.57 out of 10, further demonstrating APO’s superior performance.

2 Preliminaries

In the setting of RLHF, we assume a finite context set \mathcal{X} , and possible response set \mathcal{Y} . For any prompts $x \in \mathcal{X}$, a policy $\pi : \mathcal{X} \rightarrow \Delta(\mathcal{Y})$ maps the prompt x to the discrete distributions over the response set \mathcal{Y} . For a given context $x \in \mathcal{X}$ collected from distribution ρ , we generate two responses y_1, y_2 with a reference policy μ and receive preferences from either humans or more advanced language models between these two responses ($y^w \succ y^l$), where y^w and y^l represent the preferred and dispreferred generated responses in $\{y_1, y_2\}$. Following Christiano et al. (2017); Ouyang et al. (2022); Rafailov et al. (2023), we assume the existence of a latent reward model $r^*(x, y)$, and the preference distribution satisfies the Bradley-Terry (BT) model (Bradley and Terry, 1952):

$$\mathbb{P}(y_1 \succ y_2 | x) = \frac{\exp(r^*(x, y_1))}{\exp(r^*(x, y_1)) + \exp(r^*(x, y_2))}. \quad (2.1)$$

3 Accelerated Preference Optimization

In this section, we present a general framework for language model alignment, namely accelerated preference optimization (APO), which is built upon the iterative preference optimization framework.

3.1 Iterative Preference Optimization Framework

Under the iterative Preference Optimization framework (Xu et al., 2023; Yuan et al., 2024b; Chen et al., 2024; Wu et al., 2024), the algorithm progressively updates the policy π_t , aiming to converge to the optimal policy. In detail, for each iteration $t \in [T]$, it designates the reference policy as the policy generated from the previous iteration, denoted by π_t . It estimates the reward model by minimizing the expected loss function $l(r, x, y^w, y^l)$ over the dataset \mathcal{D}_t :

$$r_t(\cdot, \cdot) \leftarrow \arg \max_{r(\cdot, \cdot)} \mathbb{E}_{(x, y^w, y^l) \sim \mathcal{D}_t} [\ell(r, x, y^w, y^l, \pi_t)], \quad (3.1)$$

Then, it updates the reference policy by solving the following KL-regularized optimization problem:

$$\hat{\pi}_{t+1} \leftarrow \arg \max_{\pi \in \Pi} \mathbb{E}_{x \sim \rho, y \sim \pi(\cdot|x)} [r_t(x, y)] - \beta \mathbb{E}_{x \sim \rho} [\text{KL}(\pi(\cdot|x) \| \pi_t(\cdot|x))]. \quad (3.2)$$

According to Rafailov et al. (2023), for each iteration $t \in [T]$, the optimization problem (3.2) has the following closed-form solution:

$$\hat{\pi}_{t+1}(y|x) \propto \pi_t(y|x) \cdot \exp(r_t(x, y)/\beta).$$

Thus, we can reparameterize the reward function for each policy π as follows:

$$r_\pi(x, y) = \beta \log \frac{\pi(y|x)}{\pi_t(y|x)}.$$

With this reparameterized reward function, the previous two-step optimization process in (3.1) and (3.2) can be integrated into the following one-step direct preference optimization:

$$\hat{\pi}_{t+1} \leftarrow \arg \min_{\pi \in \mathcal{R}_t} \mathbb{E}_{(x, y^w, y^l) \sim \mathcal{D}_t} [\ell(r_\pi, x, y^w, y^l, \pi_t)], \quad (3.3)$$

where \mathcal{D}_t represents the data collected at iteration t using the reference policy π_t .

For the vanilla iterative preference optimization framework, the updated policy $\hat{\pi}_{t+1}$ is directly used as the reference policy in the next iteration, where $\pi_{t+1} = \hat{\pi}_{t+1}$. In this situation, the iterative optimization of the policy resembles the (Bregman) Proximal Point Method (Bregman, 1967; Censor and Zenios, 1992; Kiwiel, 1997), which iteratively minimizes the following proximal subproblem:

$$\pi_{t+1} \leftarrow \arg \min_{\pi \in \Pi} \left\{ f_t(\pi) = L_t(\pi) + \beta D(\pi, \pi_t) \right\} \quad (3.4)$$

for a given regularization parameter β and Bregman divergence $D(\cdot, \cdot)$. In this reduction, the expected reward $-\mathbb{E}_{x \sim \rho, y \sim \pi(\cdot|x)} [r_t(x, y)]$ corresponds to the target function $L_t(\pi)$ in the proximal function $f_t(\pi)$, and the Bregman divergence $D(\pi, \pi_t)$ is chosen as the Kullback–Leibler (KL) divergence to the behavior policy π_t : $\mathbb{E}_{x \sim \rho} [\text{KL}(\pi(\cdot|x) \| \pi_t(\cdot|x))]$.

3.2 Accelerated Preference Optimization

So far, we have demonstrated that the iterative preference optimization framework resembles the proximal point method. For standard optimization problems, it is well known that Nesterov’s momentum method (Nesterov, 1983, 2008, 2013) can accelerate the optimization algorithm both theoretically and empirically. In particular, Lin et al. (2018) proposed a framework called Catalyst, which extends Nesterov’s momentum method to the proximal point method and has shown that it can accelerate it provably. In the Catalyst method, after solving the proximal operator

$$x_{t+1} = \arg \min_x \left\{ f_t(x) = f(x) + \kappa D(x, y_t) \right\},$$

where $D(x, y_t)$ is the Bregman divergence, an extrapolation step is introduced as follows:

$$y_{t+1} = x_{t+1} + \alpha_t (x_{t+1} - x_t),$$

where α_t is the Nesterov’s extrapolation parameter and $x_{t+1} - x_t$ denotes the momentum from the previous update. Following the above idea, we introduce an extrapolation step after solving (3.3) to obtain π_{t+1} :

$$\log \pi_{t+1}(y|x) = \log \hat{\pi}_{t+1} + \alpha (\log \hat{\pi}_{t+1} - \log \hat{\pi}_t), \quad (3.7)$$

where α is the Nesterov’s extrapolation parameter. After normalizing the policy $\pi_{t+1}(y|x)$, we obtain

$$\pi_{t+1}(y|x) \propto \hat{\pi}_{t+1}(y|x) \cdot (\hat{\pi}_{t+1}(y|x)/\hat{\pi}_t(y|x))^\alpha.$$

Putting together all key components discussed above, we present the APO framework in Algorithm 1.

Algorithm 1 APO (Accelerated Preference Optimization)

- 1: **input:** Reference policy π_{ref} , learning rate β , Extrapolation parameter α , number of iterations T
- 2: Initialize $\pi_0 = \hat{\pi}_0 = \pi_{\text{ref}}$
- 3: **for** iteration $t = 0, \dots, T$ **do**
- 4: Collect the dataset $|\mathcal{D}_t| = N$ with prompts distribution $x \sim \rho$ and current reference policy π_t
- 5: Set the reparameterized reward function class as following:

$$\mathcal{R}_t \leftarrow \left\{ r_\pi(x, y) = \beta \log \frac{\pi(y|x)}{\pi_t(y|x)} \mid \pi \in \Pi \right\}$$

- 6: Update the policy $\hat{\pi}_{t+1}$ as following:

$$\hat{\pi}_{t+1} \leftarrow \arg \min_{\pi \in \Pi} \mathbb{E}_{(x, y^w, y^l) \sim \mathcal{D}_t} [\ell(r_\pi, x, y^w, y^l, \pi_t)] \quad (3.5)$$

- 7: Compute the policy π_{t+1} with an extrapolation step

$$\pi_{t+1}(y|x) \propto \hat{\pi}_{t+1}(y|x) \cdot (\hat{\pi}_{t+1}(y|x) / \hat{\pi}_t(y|x))^\alpha \quad (3.6)$$

- 8: **end for**
-

4 Experiments

In this section, we detail the experimental settings used to validate the efficacy of APO algorithm. We evaluate the model’s performance across various benchmark tasks and explore the impact of different momentum schedules.

4.1 Experimental Setup

Model and Datasets. We use Mistral AI’s Mistral-7B-Instruct-v0.2 (Jiang et al., 2023a) as our base model, which is a fine-tuned version based on the pretrained Mistral-7B-v0.2 (Jiang et al., 2023a) on several publicly available datasets for instruction-following. For training, we employ the UltraFeedback dataset (Cui et al., 2023). Unlike traditional fine-tuning methods that depend on responses and preference labels generated by proprietary models like GPT-4, we utilize only the instruction set from UltraFeedback. All responses are autonomously generated by our model following an online principle, and the preference pairs are labeled using a separate reward model, PairRM (Jiang et al., 2023b). The instruction set used for both training and validation includes a total of 64k instructions that span a diverse range of tasks, such as UltraChat, ShareGPT, Evol-Instruct, TruthfulQA, FalseQA, and FLAN, as detailed in Cui et al. (2023). Over three training iterations, we divided the instruction set into three folds as in Hoang et al. (2023); Kawin et al. (2023), allocating approximately 2k instructions per iteration, with an additional 1k left out for validation. Overall, our training pipeline is independent of human or GPT inputs. We defer hyperparameters to Sec. F.3.

Evaluation. For evaluating the performance of our model, we utilize three common benchmarking tasks: the AlpacaEval 2.0 (Li et al., 2023b), the MT-Bench (Zheng et al., 2024), and the Open LLM Leaderboard (Beeching et al., 2023; Gao et al., 2023). Due to space limitations, only the results for AlpacaEval 2.0 are provided in the main text and the results for MT-Bench and Open LLM Leaderboard are deferred to Sec. F.4 and F.5.

AlpacaEval 2.0. As our primary evaluation metrics, AlpacaEval 2.0 incorporates an extensive set of 805 prompts. These prompts are simplified versions derived from the AlpacaFarm (Dubois et al., 2023) evaluation set, covering a wide range of topics such as Health, Linguistics, Entertainment, Technology, Coding, Gaming, Arts, Sports, and more (Yuan et al., 2024b). During the evaluation, we consult the help of GPT-4-Turbo to compare the responses generated by our model with those produced by GPT-4-Turbo. The final win rates are weighted based on the uncertainty of the judge.

Baselines. Starting from Mistral-7B-Instruct-v0.2, we compare our method with existing iterative training results, including iterative DPO (Hoang et al., 2023) and iterative KTO (Ethayarajh et al., 2024; Kawin et al., 2023). We also present results from our own iterative DPO training. All three baselines utilize the same base model, training dataset, dataset splits, and training pipelines, which effectively highlights the differences in methodology. Additionally, we include the Self-Rewarding

Table 1: **AlpacaEval 2.0 evaluation.** Comparison of APO with state-of-the-art iterative training algorithms. Results are reported in both raw win rate (%) and length-controlled (LC) win rate (%). Additionally, the average response character length (Avg. Len) of each model is provided. Improvements of LC win rate from the previous iteration to the current iteration are calculated and marked by a subscripted '(+)' in red.

Model	Epochs	AlpacaEval 2.0		
		LC Win Rate	Win Rate	Avg. Len
Mistral-7B-Instruct-v0.2	-	17.11	14.72	1676
Snorkel (Mistral-PairRM-DPO)	-	26.39	30.22	2736
Contextual AI (KTO-Mistral-PairRM)	-	29.71	33.23	2521
Self-Rewarding 70B Iter1	-	-	9.94	1092
Self-Rewarding 70B Iter2	-	-	15.38	1552
Self-Rewarding 70B Iter3	-	-	20.44	2552
DPO Iter1	Epoch 1	25.35	30.71	2369
DPO Iter2	Epoch 3	26.92	32.54	2529
DPO Iter3	Epoch 4	29.95	37.64	2736
APO Iter1 ($\hat{\pi}_1$)	Epoch 1	25.35 _(+8.24)	30.71	2369
APO Iter1 (π_1)	-	28.23	35.75	2640
APO Iter2 ($\hat{\pi}_2$)	Epoch 1	29.56 _(+4.21)	37.53	2636
APO Iter2 (π_2)	-	29.73	38.65	2799
APO Iter3 ($\hat{\pi}_3$)	Epoch 4	31.73_(+2.17)	39.38	2950

algorithm (Yuan et al., 2024b) as another baseline. Note that different from other baselines, the training for the Self-Rewarding algorithm incorporates additional self-instruct data augmentation and is based on the Llama2 70B model, but requires no external reward model.

4.2 Experimental Results

AlpacaEval 2.0 Evaluation. Table 1 summarizes the AlpacaEval 2.0 results of different models. As shown, APO surpasses all other state-of-the-art iterative training models under the same experimental setting, with a final length controlled (LC) win rate¹ of 31.73%, and a raw win rate of 39.38%. This demonstrates an increase of 1.78% in LC win rate and 1.74% in raw win rate over our implementation of iterative DPO. Delving deeper into the effects of extrapolation from one iteration to the next, we observe that starting from the same iteration 1 with an LC win rate of 25.35%, APO at iteration 2 achieves a 29.56% LC win rate, which is 4.21% higher than iteration 1, and 2.64% higher than the 26.92% attained by vanilla DPO. By iteration 3, APO exhibits a further 2.17% improvement over its previous iteration, maintaining a lead of 1.78% over vanilla DPO. This is significant, especially considering the increasing challenge of achieving gains at higher performance levels. Notably, the momentum acceleration step is both training-free and data-free. It maximizes the potential of a single iteration of training data, advancing further toward optimality without requiring additional inputs. While iterative training across three iterations with approximately 20k data points each often plateaus at around 30%, APO effectively surpasses this threshold by a large margin.

5 Conclusions and Future Work

In this work, we studied the iterative preference optimization framework for aligning large language models (LLMs) with human preferences and showed that it resembles the proximal point method. Based on this observation, we introduced a general framework, APO, incorporating Nesterov’s momentum technique. Theoretically, we show that our method achieves a faster convergence rate than the standard iterative DPO and SPPO methods. Our experimental results demonstrate the superiority of APO over iterative DPO on the AlpacaEval 2.0 benchmark and on the instruction-following tasks of MT-Bench, achieving both accelerated convergence rate, and better final performance.

¹The length-controlled learning rate is calculated using logistic regression, with inputs including the model weights, the instruction, and the length difference between the model and a baseline. The LC win rate is designed to correlate more strongly with human preferences and to be more robust against length attacks. Longer responses are penalized in a manner that closely simulates (by causal inference) human length preferences.

References

- ACHIAM, J., ADLER, S., AGARWAL, S., AHMAD, L., AKKAYA, I., ALEMAN, F. L., ALMEIDA, D., ALTENSCHMIDT, J., ALTMAN, S., ANADKAT, S. ET AL. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* .
- AGARWAL, A., KAKADE, S., KRISHNAMURTHY, A. and SUN, W. (2020). Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems* **33** 20095–20107.
- AHN, K. and SRA, S. (2022). Understanding nesterov’s acceleration via proximal point method. In *Symposium on Simplicity in Algorithms (SOSA)*. SIAM.
- ALLEN-ZHU, Z. and ORECCHIA, L. (2014). Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537* .
- ANIL, R., DAI, A. M., FIRAT, O., JOHNSON, M., LEPIKHIN, D., PASSOS, A., SHAKERI, S., TAROPA, E., BAILEY, P., CHEN, Z. ET AL. (2023). Palm 2 technical report. *arXiv preprint arXiv:2305.10403* .
- AUSTIN, J., ODENA, A., NYE, M., BOSMA, M., MICHALEWSKI, H., DOHAN, D., JIANG, E., CAI, C., TERRY, M., LE, Q. ET AL. (2021). Program synthesis with large language models. *arXiv preprint arXiv:2108.07732* .
- AZAR, M. G., GUO, Z. D., PIOT, B., MUNOS, R., ROWLAND, M., VALKO, M. and CALANDRIELLO, D. (2024). A general theoretical paradigm to understand learning from human preferences. In *International Conference on Artificial Intelligence and Statistics*. PMLR.
- BAI, Y., JONES, A., NDOUSSE, K., ASKELL, A., CHEN, A., DASSARMA, N., DRAIN, D., FORT, S., GANGULI, D., HENIGHAN, T. ET AL. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862* .
- BAUSCHKE, H. and COMBETTES, P. (????). Convex analysis and monotone operator theory in hilbert spaces, 2011. *CMS books in mathematics* **10** 978–1.
- BECK, A. and TEOULLE, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences* **2** 183–202.
- BEECHING, E., FOURRIER, C., HABIB, N., HAN, S., LAMBERT, N., RAJANI, N., SANSEVIERO, O., TUNSTALL, L. and WOLF, T. (2023). Open llm leaderboard. https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard.
- BRADLEY, R. A. and TERRY, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika* **39** 324–345.
- BREGMAN, L. M. (1967). The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics* **7** 200–217.
- BUBECK, S., CHANDRASEKARAN, V., ELKAN, R., GEHRKE, J., HORVITZ, E., KAMAR, E., LEE, P., LEE, Y. T., LI, Y., LUNDBERG, S. ET AL. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* .
- BUBECK, S., LEE, Y. T. and SINGH, M. (2015a). A geometric alternative to nesterov’s accelerated gradient descent. *arXiv preprint arXiv:1506.08187* .
- BUBECK, S. ET AL. (2015b). Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning* **8** 231–357.
- CENSOR, Y. and ZENIOS, S. A. (1992). Proximal minimization algorithm with d-functions. *Journal of Optimization Theory and Applications* **73** 451–464.
- CHEN, M., TWOREK, J., JUN, H., YUAN, Q., PINTO, H. P. D. O., KAPLAN, J., EDWARDS, H., BURDA, Y., JOSEPH, N., BROCKMAN, G. ET AL. (2021). Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374* .

- CHEN, S., SHI, B. and YUAN, Y.-X. (2022). Revisiting the acceleration phenomenon via high-resolution differential equations. *arXiv preprint arXiv:2212.05700* .
- CHEN, Z., DENG, Y., YUAN, H., JI, K. and GU, Q. (2024). Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335* .
- CHRISTIANO, P. F., LEIKE, J., BROWN, T., MARTIC, M., LEGG, S. and AMODEI, D. (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems* **30**.
- COBBE, K., KOSARAJU, V., BAVARIAN, M., CHEN, M., JUN, H., KAISER, L., PLAPPERT, M., TWOREK, J., HILTON, J., NAKANO, R. ET AL. (2021). Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168* .
- COHEN, M. B., SIDFORD, A. and TIAN, K. (2020). Relative lipschitzness in extragradient methods and a direct recipe for acceleration. *arXiv preprint arXiv:2011.06572* .
- CUI, G., YUAN, L., DING, N., YAO, G., ZHU, W., NI, Y., XIE, G., LIU, Z. and SUN, M. (2023). Ultrafeedback: Boosting language models with high-quality feedback. *arXiv preprint arXiv:2310.01377* .
- DIAKONIKOLAS, J. and ORECCHIA, L. (2017). Accelerated extra-gradient descent: A novel accelerated first-order method. *arXiv preprint arXiv:1706.04680* .
- DUBOIS, Y., LI, X., TAORI, R., ZHANG, T., GULRAJANI, I., BA, J., GUESTRIN, C., LIANG, P. and HASHIMOTO, T. B. (2023). AlpacaFarm: A simulation framework for methods that learn from human feedback.
- DUDÍK, M., HOFMANN, K., SCHAPIRE, R. E., SLIVKINS, A. and ZOGHI, M. (2015). Contextual dueling bandits. In *Conference on Learning Theory*. PMLR.
- ETHAYARAJH, K., XU, W., MUENNIGHOFF, N., JURAFSKY, D. and KIELA, D. (2024). Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306* .
- GAO, L., TOW, J., ABBASI, B., BIDERMAN, S., BLACK, S., DIPOFI, A., FOSTER, C., GOLDING, L., HSU, J., LE NOAC'H, A., LI, H., MCDONELL, K., MUENNIGHOFF, N., OCIEPA, C., PHANG, J., REYNOLDS, L., SCHOELKOPF, H., SKOWRON, A., SUTAWIKA, L., TANG, E., THITE, A., WANG, B., WANG, K. and ZOU, A. (2023). A framework for few-shot language model evaluation.
- GEER, S. A. (2000). *Empirical Processes in M-estimation*, vol. 6. Cambridge university press.
- GÜLER, O. (1992). New proximal point algorithms for convex minimization. *SIAM Journal on Optimization* **2** 649–664.
- HE, B. and YUAN, X. (2012). An accelerated inexact proximal point algorithm for convex minimization. *Journal of Optimization Theory and Applications* **154** 536–548.
- HOANG, T., CHRIS, G. and BRADEN, H. (2023). Iterative dpo alignment. Tech. rep., Snorkel AI.
- JIANG, A. Q., SABLAYROLLES, A., MENSCH, A., BAMFORD, C., CHAPLOT, D. S., CASAS, D. D. L., BRESSAND, F., LENGYEL, G., LAMPLE, G., SAULNIER, L. ET AL. (2023a). Mistral 7b. *arXiv preprint arXiv:2310.06825* .
- JIANG, D., REN, X. and LIN, B. Y. (2023b). Llm-blender: Ensembling large language models with pairwise comparison and generative fusion. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (ACL 2023)*.
- JIN, Y., SIDFORD, A. and TIAN, K. (2022). Sharper rates for separable minimax and finite sum optimization via primal-dual extragradient methods. In *Conference on Learning Theory*. PMLR.
- KAWIN, E., WINNIE, X., DAN, J. and DOUWE, K. (2023). Human-centered loss functions (halos). Tech. rep., Contextual AI. <https://github.com/ContextualAI/HALOs/blob/main/assets/report.pdf>.
- KIWIEL, K. C. (1997). Proximal minimization methods with generalized bregman functions. *SIAM journal on control and optimization* **35** 1142–1168.

- KNOX, W. B. and STONE, P. (2009). Interactively shaping agents via human reinforcement: The tamer framework. In *Proceedings of the fifth international conference on Knowledge capture*.
- KOVALEV, D., GASNIKOV, A. and RICHTÁRIK, P. (2022). Accelerated primal-dual gradient method for smooth and convex-concave saddle-point problems with bilinear coupling. *Advances in Neural Information Processing Systems* **35** 21725–21737.
- KWON, W., LI, Z., ZHUANG, S., SHENG, Y., ZHENG, L., YU, C. H., GONZALEZ, J. E., ZHANG, H. and STOICA, I. (2023). Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- LESSARD, L., RECHT, B. and PACKARD, A. (2016). Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization* **26** 57–95.
- LI, C. J., YUAN, H., GIDEL, G., GU, Q. and JORDAN, M. (2023a). Nesterov meets optimism: rate-optimal separable minimax optimization. In *International Conference on Machine Learning*. PMLR.
- LI, X., ZHANG, T., DUBOIS, Y., TAORI, R., GULRAJANI, I., GUESTRIN, C., LIANG, P. and HASHIMOTO, T. B. (2023b). AlpacaEval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.
- LIN, H., MAIRAL, J. and HARCHAOUI, Z. (2015). A universal catalyst for first-order optimization. *Advances in neural information processing systems* **28**.
- LIN, H., MAIRAL, J. and HARCHAOUI, Z. (2018). Catalyst acceleration for first-order convex optimization: from theory to practice. *Journal of Machine Learning Research* **18** 1–54.
- MARTINET, B. (1970). Régularisation d’inéquations variationnelles par approximations successives. rev. française informat. *Recherche Opérationnelle* **4** 154–158.
- MOREAU, J.-J. (1965). Proximité et dualité dans un espace hilbertien. *Bulletin de la Société mathématique de France* **93** 273–299.
- MUNOS, R., VALKO, M., CALANDRIELLO, D., AZAR, M. G., ROWLAND, M., GUO, Z. D., TANG, Y., GEIST, M., MESNARD, T., MICHI, A. ET AL. (2023). Nash learning from human feedback. *arXiv preprint arXiv:2312.00886*.
- NESTEROV, Y. (1983). A method for solving the convex programming problem with convergence rate $o(1/k^2)$. *Proceedings of the USSR Academy of Sciences* **269** 543–547.
- NESTEROV, Y. (2008). Accelerating the cubic regularization of newton’s method on convex problems. *Mathematical Programming* **112** 159–181.
- NESTEROV, Y. (2013). *Introductory lectures on convex optimization: A basic course*, vol. 87. Springer Science & Business Media.
- OPENAI, R. (2023). Gpt-4 technical report. arxiv 2303.08774. *View in Article* **2**.
- OUYANG, L., WU, J., JIANG, X., ALMEIDA, D., WAINWRIGHT, C., MISHKIN, P., ZHANG, C., AGARWAL, S., SLAMA, K., RAY, A. ET AL. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* **35** 27730–27744.
- POLYAK, B. T. (1964). Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics* **4** 1–17.
- RAFAILOV, R., SHARMA, A., MITCHELL, E., ERMON, S., MANNING, C. D. and FINN, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.
- ROCKAFELLAR, R. T. (1976). Monotone operators and the proximal point algorithm. *SIAM journal on control and optimization* **14** 877–898.

- ROSSET, C., CHENG, C.-A., MITRA, A., SANTACROCE, M., AWADALLAH, A. and XIE, T. (2024). Direct nash optimization: Teaching language models to self-improve with general preferences. *arXiv preprint arXiv:2404.03715* .
- SALZO, S., VILLA, S. ET AL. (2012). Inexact and accelerated proximal point algorithms. *Journal of Convex analysis* **19** 1167–1192.
- SCHULMAN, J., WOLSKI, F., DHARIWAL, P., RADFORD, A. and KLIMOV, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* .
- SHI, B. (2021). On the hyperparameters in stochastic gradient descent with momentum. *arXiv preprint arXiv:2108.03947* .
- SHI, B., DU, S. S., JORDAN, M. I. and SU, W. J. (2022). Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming* 1–70.
- SHI, B., DU, S. S., SU, W. and JORDAN, M. I. (2019). Acceleration via symplectic discretization of high-resolution differential equations. *Advances in Neural Information Processing Systems* **32**.
- SU, W., BOYD, S. and CANDES, E. J. (2016). A differential equation for modeling nesterov’s accelerated gradient method: Theory and insights. *Journal of Machine Learning Research* **17** 1–43.
- SWAMY, G., DANN, C., KIDAMBI, R., WU, Z. S. and AGARWAL, A. (2024). A minimaximalist approach to reinforcement learning from human feedback. *arXiv preprint arXiv:2401.04056* .
- TAYLOR, A. B., HENDRICKX, J. M. and GLINEUR, F. (2017). Smooth strongly convex interpolation and exact worst-case performance of first-order methods. *Mathematical Programming* **161** 307–345.
- THEKUMPARAMPIL, K. K., HE, N. and OH, S. (2022). Lifted primal-dual method for bilinearly coupled smooth minimax optimization. In *International Conference on Artificial Intelligence and Statistics*. PMLR.
- TOUVRON, H., MARTIN, L., STONE, K., ALBERT, P., ALMAHAIRI, A., BABAEI, Y., BASHLYKOV, N., BATRA, S., BHARGAVA, P., BHOSALE, S. ET AL. (2023). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288* .
- TSENG, P. (2009). On accelerated proximal gradient methods for convex-concave optimization. 2008. *Submitted to SIAM J. Optim* .
- TVERSKY, A. (1969). Intransitivity of preferences:’psychological review 76: 31-48.. 1972. *Elimination by aspects: A theory of choice* 281–299.
- WEI, J., WANG, X., SCHUURMANS, D., BOSMA, M., XIA, F., CHI, E., LE, Q. V., ZHOU, D. ET AL. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* **35** 24824–24837.
- WIRTH, C., AKROUR, R., NEUMANN, G., FÜRNKRANZ, J. ET AL. (2017). A survey of preference-based reinforcement learning methods. *Journal of Machine Learning Research* **18** 1–46.
- WU, Y., SUN, Z., YUAN, H., JI, K., YANG, Y. and GU, Q. (2024). Self-play preference optimization for language model alignment. *arXiv preprint arXiv:2405.00675* .
- XIONG, W., DONG, H., YE, C., ZHONG, H., JIANG, N. and ZHANG, T. (2023). Gibbs sampling from human feedback: A provable kl-constrained framework for rlhf. *arXiv preprint arXiv:2312.11456* .
- XU, J., LEE, A., SUKHBAATAR, S. and WESTON, J. (2023). Some things are more cringe than others: Preference optimization with the pairwise cringe loss. *arXiv preprint arXiv:2312.16682* .
- XU, Y. and ZEEVI, A. (2020). Upper counterfactual confidence bounds: a new optimism principle for contextual bandits. *arXiv preprint arXiv:2007.07876* .

- YUAN, A., LI, C. J., GIDEL, G., JORDAN, M., GU, Q. and DU, S. S. (2024a). Optimal extragradient-based algorithms for stochastic variational inequalities with separable structure. *Advances in Neural Information Processing Systems* **36**.
- YUAN, W., PANG, R. Y., CHO, K., SUKHBAATAR, S., XU, J. and WESTON, J. (2024b). Self-rewarding language models. *arXiv preprint arXiv:2401.10020*.
- ZHENG, L., CHIANG, W.-L., SHENG, Y., ZHUANG, S., WU, Z., ZHUANG, Y., LIN, Z., LI, Z., LI, D., XING, E. ET AL. (2024). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* **36**.
- ZIEGLER, D. M., STIENNON, N., WU, J., BROWN, T. B., RADFORD, A., AMODEI, D., CHRISTIANO, P. and IRVING, G. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Limitation

Due to limited computational resources, we do not evaluate APO with the SPPO (Wu et al., 2024) loss function in the current experiments, and we plan to investigate it in our future work. Additionally, while our model demonstrates consistent improvements on instruction-following tasks, it faces challenges in solving math problems. This limitation is largely due to the choice of dataset and the restriction of not utilizing additional information from GPT-4 or human sources. In the future, we aim to address this by incorporating larger datasets and leveraging GPT-4 supervision.

Broader Impact

This paper studies the alignment of LLMs with human preferences and presents a Nesterov’s method for the alignment process. We believe our work has the following social impacts. Firstly, Large Language Models (LLMs) have emerged as pivotal tools across several fields, including text generation, coding, and problem-solving, which are increasingly relied upon by individuals. Our method demonstrates superiority over existing benchmarks, which could be beneficial for constructing AI assistants. Secondly, aligning LLMs with human preferences requires material resources, and the accelerated method may expedite the learning process and reduce the required resources.

A Related Work

A.1 Reinforcement Learning from Human Feedback

Learning from human feedback in reinforcement learning can be traced back to Knox and Stone (2009); Wirth et al. (2017) and was later popularized by Christiano et al. (2017), which incorporated human preferences into deep reinforcement learning. Recently, RLHF has gained popularity in natural language processing and has become a paradigmatic method for fine-tuning large language models (LLMs) (Achiam et al., 2023; Touvron et al., 2023; OpenAI, 2023) to align with human objectives. The standard process for alignment with human feedback (Ouyang et al., 2022) requires both reward modeling and policy optimization. Recently, Rafailov et al. (2023) proposed the Direct Preference Optimization method, which replaces the reward modeling process with a reparameterized reward and directly performs policy optimization using preference data. Similarly, Kawin et al. (2023) proposed the Kahneman-Tversky Optimization method (KTO) with a reparameterized reward, while considering different human-aware loss functions.

More recently, iterative variants of policy optimization have garnered increasing attention. Xu et al. (2023) investigated the iterative preference optimization method and proposed the PAIRWISE CRINGE algorithm, which iteratively generates a new preference dataset using the current model and then updates the model with a combination of the original preference data and the newly labeled data. Additionally, they studied the iterative version of Direct Preference Optimization (DPO) and demonstrated that it achieves better performance than standard PPO or DPO. Later, Yuan et al. (2024b) studied Self-Rewarding Language Models, where the algorithm iteratively updates the reference policy in the DPO method and uses LLM-as-a-Judge prompting (Zheng et al., 2024) to provide its own rewards for the generated new preference data. Despite this work demonstrating the superiority of iterative optimization process, there is a lack of theoretical foundations for these practical frameworks. Recently, Xiong et al. (2023) proposed Gibbs Sampling from Human Feedback, offering theoretical analysis with the aid of linear function approximation and the incorporation of

an uncertainty bonus. Compared with previous methods, the reference policy remains fixed, the newly trained model is only used to generate new preference data, and this trained model will not be inherited in the subsequent iteration. On the contrary, our research focuses on studying the iterative DPO model without incorporating an uncertainty bonus, and it continually updates the reference policy throughout the process.

Most of the works motioned above rely on the assumption that the latent preference distribution $P(y_1 \succ y_2|x)$ follows the Bradley-Terry (BT) model (Bradley and Terry, 1952), and there exists a series of works focusing on general preference, where human preference may not strictly be transitive. Under the general preference assumption, iterative optimization (Munos et al., 2023; Swamy et al., 2024; Rosset et al., 2024; Wu et al., 2024) is also employed to find the Nash-equilibrium policy. We leave the extension of our algorithm to general preference model in the future work.

A.2 Accelerated Optimization with Momentum

The idea of accelerating gradient methods has been extensively explored over the decades. One of the earliest contributions to this field was the Polyak momentum (Polyak, 1964), which achieved faster convergence by leveraging the history of previous iterates. However, this early approach sometimes failed to converge even for strongly convex objectives (Lessard et al., 2016). This was further refined by Nesterov’s accelerated gradient (NAG) (Nesterov, 1983), with a guarantee for faster convergence rates for general smooth convex functions. Following these foundational works, other acceleration algorithms and analyses emerged for gradient descent and its proximal variants (Beck and Teboulle, 2009; Tseng, 2009; Taylor et al., 2017; Allen-Zhu and Orecchia, 2014; Bubeck et al., 2015a; Diakonikolas and Orecchia, 2017; Cohen et al., 2020). In addition, there has been another line of work in understanding Nesterov’s acceleration through continuous analysis and Lyapunov function analysis (Su et al., 2016; Shi et al., 2019; Shi, 2021; Shi et al., 2022; Chen et al., 2022). Momentum techniques have also proven effective in accelerating minimax game settings (Jin et al., 2022; Kovalev et al., 2022; Thekumparampil et al., 2022; Li et al., 2023a; Yuan et al., 2024a). However, the application of acceleration techniques to LLMs remains relatively unexplored.

The Proximal Point Method (PPM) (Moreau, 1965; Martinet, 1970; Rockafellar, 1976; Güler, 1992; Bauschke and Combettes) is a fundamental methodology that achieves optimality by iteratively solving auxiliary problems in a regularized form. Generally, these auxiliary problems cannot be solved directly, and different approximation methods yield various types of acceleration (He and Yuan, 2012; Salzo et al., 2012). One of the connections between PPM and NAG is established in Ahn and Sra (2022), where the authors interpret the updates of NAG as different approximations of the proximal method. The Catalyst approach (Lin et al., 2015, 2018) represents another variant of PPM that achieves accelerated convergence through practical inexact schemes (Lin et al., 2015; Salzo et al., 2012). This method effectively integrates Nesterov’s acceleration with the proximal point framework, making it applicable to a wide range of machine learning algorithms, including gradient descent and block coordinate descent. It can be applied to variance reduction algorithms (Lin et al., 2015, 2018), demonstrating its strong adaptability to various machine learning scenarios. Despite the widespread adoption of momentum acceleration in both theoretical and practical applications, we are the first to establish a connection between DPO and the proximal point method, linking iterative DPO with the Catalyst approach, and providing a comprehensive theoretical analysis of this interpretation.

B Theoretical Analysis

Notation. For any positive integer n , we employ $[n]$ to denote the set $\{0, \dots, n\}$. For two sequences $\{a_n\}$ and $\{b_n\}$, we write $a_n = O(b_n)$ if there exists an absolute constant C such that $a_n \leq Cb_n$. We use $\tilde{O}(\cdot)$ to further hide the logarithmic factors.

B.1 The Choice of Loss Function

By choosing different loss function $\ell(r_\pi, x, y^w, y^l, \pi_t)$ in (3.3), we can recover many existing iterative preference optimization algorithms. In detail, the loss function depends on the preference model \mathbb{P} , and we provide several concrete examples of the loss function and the corresponding preference optimization algorithms as follows.

Example B.1 (DPO). If we choose the loss function in (3.3) as:

$$\ell_{\text{DPO}}(r_\pi, x, y^w, y^l, \pi_t) = -\log \sigma(r_\pi(x, y^w) - r_\pi(x, y^l)),$$

it recovers DPO (Rafailov et al., 2023).

Example B.2 (SPPO). For the Self-Play Preference Optimization (SPPO) (Wu et al., 2024) algorithm, if we set the learning rate β in Algorithm 1 to be the inverse of the learning rate η in the SPPO algorithm, the preference optimization process can be expressed as follows:

$$\begin{aligned}
\pi_{t+1} &= \arg \min_{\pi} \mathbb{E}_{x \sim \rho, y \sim \pi_t(\cdot|x)} \left[\log \left(\frac{\pi(y|x)}{\pi_t(y|x)} \right) - \eta \left(\mathbb{P}(y \succ \pi_t|x) - \log Z_{\pi_t}(x) \right) \right]^2 \\
&= \arg \min_{\pi} \mathbb{E}_{x \sim \rho, y \sim \pi_t(\cdot|x)} \left[\beta \log \left(\frac{\pi(y|x)}{\pi_t(y|x)} \right) - \left(\mathbb{P}(y \succ \pi_t|x) - \log Z_{\pi_t}(x) \right) \right]^2 \\
&= \arg \min_{r_{\pi} \in \mathcal{R}_t} \mathbb{E}_{x \sim \rho, y \sim \pi_t(\cdot|x)} \left[r_{\pi}(x, y) - \mathbb{E}_{y' \sim \pi_t} [\mathbb{P}(y \succ y'|x)] + \log Z_{\pi_t}(x) \right]^2 \\
&= \arg \min_{r_{\pi} \in \mathcal{R}_t} \mathbb{E}_{x \sim \rho, y, y' \sim \pi_t(\cdot|x)} \left[r_{\pi}(x, y) - \mathbf{1}(y \succ y'|x) + \log Z_{\pi_t}(x) \right]^2 + C_{\pi_t} \\
&= \arg \min_{r_{\pi} \in \mathcal{R}_t} \mathbb{E}_{x \sim \rho, y, y' \sim \pi_t(\cdot|x)} \left[r_{\pi}(x, y) - \mathbf{1}(y \succ y'|x) + \log Z_{\pi_t}(x) \right]^2 / 2 \\
&\quad + \mathbb{E}_{x \sim \rho, y, y' \sim \pi_t(\cdot|x)} \left[r_{\pi}(x, y') - \mathbf{1}(y' \succ y|x) + \log Z_{\pi_t}(x) \right]^2 / 2 + C_{\pi_t},
\end{aligned}$$

where $Z_{\pi_t}(x) = \sum_{y \in \mathcal{Y}} \pi_t(y|x) \exp(\eta \mathbb{P}(y \succ \pi_t|x))$ represents the partition function for behavior policy π_t , $C_{\pi_t} = \mathbb{E}_{x \sim \rho, y, y' \sim \pi_t(\cdot|x)} [\mathbb{P}(y \succ y'|x) - \mathbf{1}(y \succ y'|x)]^2$ is the variance of behavior policy π_t , the second equation holds due to $\beta = \eta^{-1}$ and the last equation holds because y, y' collected under the same behavior policy. Therefore, if we choose the loss function in (3.3) as:

$$\ell_{\text{SPPO}}(r_{\pi}, x, y^w, y^l, \pi_t) = \frac{1}{2} (r_{\pi}(x, y^w) - 1 + \log Z_{\pi_t}(x))^2 + \frac{1}{2} (r_{\pi}(x, y^l) + \log Z_{\pi_t}(x))^2,$$

where $Z_{\pi_t}(x) = \sum_{y \in \mathcal{Y}} \pi_t(y|x) \exp(\eta \mathbb{P}(y \succ \pi_t|x))$ represents the partition function for behavior policy π_t , it recovers the Self-Play Preference Optimization (SPPO) algorithm (Wu et al., 2024).

Example B.3 (IPO). If we choose the loss function in (3.3) as

$$\ell_{\text{IPO}}(r_{\pi}, x, y^w, y^l, \pi_t) = (r_{\pi}(x, y^w) - r_{\pi}(x, y^l) - \tau^{-1})^2,$$

where τ is a regularization parameter, it recovers the Identity Preference Optimization (IPO) algorithm (Azar et al., 2024).

B.2 Accelerated Preference Optimization with Bradley-Terry model

In this section, we provide a theoretical analysis of APO in Algorithm 1.

We begin with the following theorem, which outlines the optimization dynamics of the policy $\hat{\pi}_{t+1}$ over different iterations $t \in [T]$.

Theorem B.4. Suppose that $\hat{\pi}_{t+1}(y|x) \cdot (\hat{\pi}_{t+1}(y|x)/\hat{\pi}_t(y|x))^{\alpha}$ belongs to the policy class Π for each iteration $t \in [T]$. Then, the updated policy $\hat{\pi}_t$ in Algorithm 1 satisfies

$$\hat{\pi}_{t+1}(y|x) = \frac{1}{Z_t(x)} \cdot \pi_{\text{ref}}(y|x) \cdot \exp \left(\frac{1}{\beta} \cdot \sum_{i=0}^t \left(\frac{1}{1-\alpha} - \frac{\alpha^{t+1-i}}{1-\alpha} \right) \cdot r_i(x, y) \right),$$

where $r_t(x, y) = \beta \log \hat{\pi}_{t+1}(y|x) - \beta \log \pi_t(y|x)$ represents the reparameterized reward at iteration t , and $Z_t(x) = \sum_y \pi_{\text{ref}}(y|x) \cdot \exp \left(\sum_{i=0}^t (1/(1-\alpha) - \alpha^{t+1-i}/(1-\alpha)) \cdot r_i(x, y) / \beta \right)$ is the partition function.

Theorem B.4 illustrates how the policy $\hat{\pi}_{t+1}$ evolves with respect to the reparameterized reward $r_t(x, y)$, which is highly dependent on the choice of the loss function ℓ in Algorithm 1.

For the Bradley-Terry (BT) model with the loss function ℓ_{BPPO} in Example B.1, Theorem 1 in Rafailov et al. (2023) demonstrates that all reward functions compatible with the BT model can be expressed by the reparameterized reward. In addition, we introduce the following two assumptions, which are required by our analysis.

Assumption B.5 (Realizability). For each iteration $t \in [T]$ and each policy $\pi \in \Pi$, the following updated policy belongs to the policy class Π :

$$\widehat{\pi}(\cdot|x) = \frac{1}{Z_\pi(x)} \cdot \pi(\cdot|x) \cdot \exp\left(\frac{r^*(x, \cdot)}{\beta}\right) \in \Pi,$$

where $Z_\pi(x) = \sum_y \pi(x) \cdot \exp(r^*(x, y)/\beta)$ is the partition function.

Assumption B.6 (Boundedness). For each iteration $t \in [T]$ and each policy $\pi, \pi_t \in \Pi$, we have

$$\beta \log \frac{\pi(y|x)}{\pi_t(y|x)} \in [-R, R]$$

for all $x \in \mathcal{X}, u \in \mathcal{Y}$.

Similar assumptions have been used in [Rosset et al. \(2024\)](#) to provide an analysis of the statistical error for the reparameterized reward. Equipped with these assumptions, we have the following performance guarantee for APO.

Theorem B.7 (APO with ℓ_{DPO}). For the Bradley-Terry (BT) model with loss function ℓ_{DPO} , under Assumptions [B.5](#) and [B.6](#), with probability at least $1 - \delta$, the sub-optimality gap between $\widehat{\pi}_{T+1}$ and the optimal policy $\pi^*(x) = \arg \max_{y \in \mathcal{Y}} r^*(x, y)$ is bounded by

$$\begin{aligned} & \mathbb{E}_{x \sim \rho, y \sim \pi^*(\cdot|x)} [r^*(x, y)] - \mathbb{E}_{x \sim \rho, y \sim \widehat{\pi}_{T+1}(\cdot|x)} [r^*(x, y)] \\ & \leq \widetilde{O}\left(\frac{(1-\alpha)\beta}{T}\right) + O\left(\sqrt{\frac{(T+1) \sum_{t=0}^T \kappa_t \cdot \log(T|\Pi|/\delta)}{N\beta^2(1-\alpha)^2}}\right), \end{aligned}$$

where the coverage coefficient κ_t is defined as:

$$\kappa_t = \max_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \frac{\widehat{\pi}_{T+1}(y|x) \pi_{T+1}^*(y|x)}{\pi_t^2(y|x)}. \quad (\text{B.1})$$

Remark B.8. The sub-optimality gap in [Theorem B.7](#) consists of two terms: the optimization error $\widetilde{O}((1-\alpha)\beta/T)$ and the statistical error $\widetilde{O}(\sqrt{T/(N\beta^2(1-\alpha)^2)})$, and there exists a tradeoff between these two errors. Specifically, a larger extrapolation parameter $\alpha \rightarrow 1$ will decrease the optimization error but increase the statistical error. Moreover, when the dataset size is sufficiently large ($N \rightarrow \infty$), [Theorem B.7](#) suggests a $\widetilde{O}((1-\alpha)\beta/T)$ sub-optimality gap for [Algorithm 1](#) with the loss function ℓ_{DPO} . Compared with the standard iterative DPO method with $\alpha = 0$, [Algorithm 1](#) improves the sub-optimality gap by a factor of $1/(1-\alpha)$ thanks to the Nesterov’s momentum technique.

[Theorem B.7](#) only provides theoretical guarantee on the sub-optimality gap for the policy $\widehat{\pi}_{T+1}$. In order to derive the convergence of policy $\widehat{\pi}_{T+1}$ to the optimal policy π^* , we need the following minimal sub-optimality gap assumption.

Assumption B.9 (Minimal sub-optimality gap). For each prompt $x \in \mathcal{X}$, let $y_x^* = \arg \max_{y \in \mathcal{Y}} r^*(x, y)$ be the optimal response. The sub-optimality gap between the optimal response and any other responses is strictly positive: $\min_{y \neq y_x^*} r^*(x, y_x^*) - r^*(x, y) \geq \Delta > 0$

Remark B.10. For a general Bradley-Terry (BT) model without the minimal sub-optimality gap, there may exist multiple responses that share the same maximum reward, i.e., $r(x, y_1) = r(x, y_2) = \max_{y \in \mathcal{Y}} r^*(x, y)$. In this case, the optimal policy is not unique, and it is impossible to show that the policy $\widehat{\pi}_{T+1}$ converges to a specific optimal policy.

Theorem B.11 (APO with ℓ_{DPO}). For the Bradley-Terry (BT) model with loss function ℓ_{DPO} , under Assumptions [B.5](#), [B.6](#) and [B.9](#), with probability at least $1 - \delta$, the TV-distance between $\widehat{\pi}_{T+1}$ and the optimal policy $\pi^*(x) = \arg \max_{y \in \mathcal{Y}} r^*(x, y)$ is bounded by

$$\begin{aligned} & \mathbb{E}_{x \sim \rho} \left[\text{D}_{\text{TV}}(\widehat{\pi}_{T+1}(\cdot|x), \pi^*(\cdot|x)) \right] \\ & \leq \exp\left(-O\left(\frac{T\Delta}{(1-\alpha)} \cdot \frac{1}{\beta}\right)\right) + O\left(\sqrt{\frac{(T+1) \sum_{t=0}^T \kappa_t \cdot \log(T|\Pi|/\delta)}{N\beta^2(1-\alpha)^2}}\right), \end{aligned}$$

where the coverage coefficient κ_t is defined in [\(B.1\)](#).

Remark B.12. The TV-distance between $\hat{\pi}_{t+1}$ and the optimal policy $\pi^*(x) = \arg \max_{y \in \mathcal{Y}} r^*(x, y)$ includes both the optimization error $\exp(-O(T/(1-\alpha)))$ and the statistical error $\tilde{O}(\sqrt{T/(N(1-\alpha)^2)})$. Similar to Theorem B.7, there is a tradeoff between these two errors based on the choice of α . In addition, when the dataset size is sufficiently large ($N \rightarrow \infty$), Theorem B.11 suggests that APO converges to the optimal policy at a rate of $\exp(-O(T/(1-\alpha)))$. In comparison, iterative DPO is a special case of our algorithm with $\alpha = 0$, which converges to the optimal policy at a slower rate of $\exp(-O(T))$.

B.3 Accelerated Preference Optimization with General Preferences

In the previous sections, we assumed the existence of a latent reward function and focused on the Bradley-Terry (BT) model. However, (Tversky, 1969) observed that the preferences across all possible responses may not be monotonous or transitive, which cannot be represented by a latent reward function. In this section, we extend the analysis to environments with general preferences and focus on the win probability between different responses.

In detail, for an environment with general preferences, we denote the probability that response y_1 is preferred over response y_2 given the context x as $\mathbb{P}(y_1 \succ y_2 | x)$, and we assume that the preference model is antisymmetric, such that $\mathbb{P}(y_2 \succ y_1 | x) = 1 - \mathbb{P}(y_1 \succ y_2 | x)$. Under this assumption, we define the reward $r_{\pi, \pi'} = \mathbb{E}_{y_1 \sim \pi, y_2 \sim \pi'} [\mathbb{P}(y_1 \succ y_2 | x)]$ as the win probability for policy π against policy π' . Our goal is to identify the **von Neumann winner** (Dudík et al., 2015) with the preference model. Specifically, the von Neumann winner corresponds to the Nash equilibrium of the following two-player zero-sum game:

$$\begin{aligned} (\pi^*, \pi^*) &= \arg \max_{\pi} \min_{\pi'} \mathbb{E}_{y_1 \sim \pi, y_2 \sim \pi'} [\mathbb{P}(y_1 \succ y_2 | x)] \\ &= \arg \max_{\pi} \min_{\pi'} r_{\pi, \pi'}. \end{aligned}$$

For a general preference model, Theorem B.14 provides a convergence rate for Algorithm 1 with the loss function ℓ_{SPP0} in Example B.2, under the assumption of a minimal sub-optimality gap.

Assumption B.13 (Minimal sub-optimality gap with general preferences). For each prompt $x \in \mathcal{X}$, we assume there exist a unique optimal response $y_x^* \in \mathcal{Y}$ such that for any other sub-optimal responses $y \neq y_x^*$, we have

$$\mathbb{P}(y_x^* \succ y' | x) - \mathbb{P}(y \succ y' | x) \geq \Delta, \forall y' \in \mathcal{Y}.$$

Theorem B.14 (APO with ℓ_{SPP0}). For general preference model with loss function ℓ_{SPP0} , under the mild assumptions of realizability (Assumption E.3), boundedness (Assumption E.4) and Assumption B.13, with probability at least $1 - \delta$, the TV-distance between $\hat{\pi}_{T+1}$ and the optimal policy $\pi^*(x) = y_x^*$ is bounded by

$$\begin{aligned} &\mathbb{E}_{x \sim \rho} \left[\text{D}_{\text{TV}}(\hat{\pi}_{T+1}(\cdot | x), \pi^*(\cdot | x)) \right] \\ &\leq O \left(\sqrt{\frac{(T+1) \sum_{t=0}^T \kappa_t \cdot \log(T|\Pi|/\delta)}{N\beta^2(1-\alpha)^2}} \right) + \exp \left(-O \left(\frac{t\Delta}{(1-\alpha)} \cdot \frac{1}{\beta} \right) \right), \end{aligned}$$

where the coverage coefficient κ_t is defined as:

$$\kappa_t = \max_{(x, y) \in \mathcal{X} \times \mathcal{Y}} \frac{\hat{\pi}_{T+1}(y | x) \pi_{T+1}^*(y | x)}{\pi_t^2(y | x)}.$$

Remark B.15. When the dataset size is sufficiently large ($N \rightarrow \infty$), Theorem B.14 suggests that APO converges to the optimal policy at a rate of $\exp(-O(T/(1-\alpha)))$. When $\alpha = 0$, our algorithm reduces to the standard SPPO algorithm in Wu et al. (2024), which only converges to the optimal policy at a slower rate of $\exp(-O(T))$. It is worth noting that Wu et al. (2024) provides a $\tilde{O}(1/\sqrt{T})$ sub-optimality gap guarantee for the average policy $\bar{\pi} = \frac{1}{T} \sum_{t=1}^T \pi_t$. In comparison, our result relies on a minimal sub-optimality gap assumption (Assumption B.13) and provides a faster convergence rate between the final policy $\hat{\pi}_{t+1}$ and the optimal policy π^* . As discussed in Remark B.10, this assumption is used to ensure the uniqueness of the optimal policy, which is necessary for the convergence guarantee.

C Proof of Main Results

C.1 Proof of Theorem B.4

In this section, we provide the proof of Theorem B.4, which is crucial for understanding the optimization dynamics of Algorithm 1.

Proof of Theorem B.4. Based on the definition of the reparameterized reward r_t , we have

$$r_t(x, y) = \beta \log \widehat{\pi}_{t+1}(y|x) - \beta \log \pi_t(y|x). \quad (\text{C.1})$$

Furthermore, the extrapolation step (3.6) satisfies

$$\pi_{t+1}(y|x) = \frac{1}{Z'_t(x)} \cdot \widehat{\pi}_{t+1}(y|x) \cdot (\widehat{\pi}_{t+1}(y|x)/\widehat{\pi}_t(y|x))^\alpha,$$

where $Z'_t(x) = \sum_y \widehat{\pi}_{t+1}(y|x) \cdot (\widehat{\pi}_{t+1}(y|x)/\widehat{\pi}_t(y|x))^\alpha$ represents the partition function. Taking the logarithm of both sides yields the following equation

$$\log \pi_{t+1}(y|x) = (1 + \alpha) \log \widehat{\pi}_{t+1}(y|x) - \alpha \log \widehat{\pi}_t(y|x) - \log Z'_t(x). \quad (\text{C.2})$$

For simplicity, we define $l_t(x, y) = \beta \log \widehat{\pi}_{t+1}(y|x) - \beta \log \widehat{\pi}_t(y|x)$, and thus we have

$$\begin{aligned} l_{t+1}(x, y) &= \beta \log \widehat{\pi}_{t+2}(y|x) - \beta \log \widehat{\pi}_{t+1}(y|x) \\ &= r_{t+1}(x, y) + \beta \log \pi_{t+1}(y|x) - \beta \log \widehat{\pi}_{t+1}(y|x) \\ &= r_{t+1}(x, y) + \alpha(\beta \log \widehat{\pi}_{t+1}(y|x) - \beta \log \widehat{\pi}_t(y|x)) - \beta \log Z'_t(x) \\ &= r_{t+1}(x, y) - \beta \log Z'_t(x) + \alpha l_t(x, y), \end{aligned} \quad (\text{C.3})$$

where the second equation holds due to (C.1) and the third equation holds due to (C.2). Recursively using (C.3) over all iterations, we derive the following equation:

$$\begin{aligned} l_t(x, y) &= r_t(x, y) - \beta \log Z'_{t-1}(x) + \alpha l_{t-1}(x, y) \\ &= r_t(x, y) - \beta \log Z'_{t-1}(x) + \alpha r_{t-1}(x, y) - \alpha \beta \log Z'_{t-2}(x) + \alpha^2 l_{t-2}(x, y) \\ &= \dots \\ &= \sum_{i=0}^t \alpha^{t-i} \cdot r_i(x, y) - \alpha^{t-i} \cdot \beta \log Z'_{i-1}(x), \end{aligned} \quad (\text{C.4})$$

where we assume $Z'_{-1}(x) = 1$ for simplicity. Finally, by summing (C.4) over all iterations $0 \leq j \leq t$, we have

$$\begin{aligned} &\beta \log \widehat{\pi}_{t+1}(y|x) - \beta \log \widehat{\pi}_1(y|x) \\ &= \sum_{j=0}^t \sum_{i=0}^j \alpha^{j-i} r_i(x, y) + \alpha^{j-i} \cdot \beta \log Z'_{i-1}(x) \\ &= \sum_{i=0}^t \sum_{j=i}^t \alpha^{j-i} r_i(x, y) + \alpha^{j-i} \cdot \beta \log Z'_{i-1}(x) \\ &= \sum_{i=0}^t \left(\frac{1}{1-\alpha} - \frac{\alpha^{t+1-i}}{1-\alpha} \right) \cdot r_i(x, y) + \left(\frac{1}{1-\alpha} - \frac{\alpha^{t+1-i}}{1-\alpha} \right) \cdot \beta \log Z'_{i-1}(x). \end{aligned} \quad (\text{C.5})$$

Given that $\widehat{\pi}_1 = \pi_{\text{ref}}$, we have

$$\begin{aligned} &\widehat{\pi}_{t+1}(y|x) \\ &= \pi_{\text{ref}}(y|x) \cdot \exp \left\{ \beta \cdot \sum_{i=0}^t \left(\frac{1}{1-\alpha} - \frac{\alpha^{t+1-i}}{1-\alpha} \right) \cdot r_t + \left(\frac{1}{1-\alpha} - \frac{\alpha^{t+1-i}}{1-\alpha} \right) \cdot \beta \log Z'_{i-1}(x) \right\} \\ &\propto \pi_{\text{ref}}(y|x) \cdot \exp \left\{ \beta \cdot \sum_{i=0}^t \left(\frac{1}{1-\alpha} - \frac{\alpha^{t+1-i}}{1-\alpha} \right) \cdot r_t \right\}. \end{aligned}$$

Thus, we complete the proof of Theorem B.4. \square

C.2 Proof of Theorem B.7

In this section, we provide the proof of Theorem B.7. To simplify the notation, we define the auxiliary policy, which is updated following the dynamics of $\hat{\pi}_{t+1}$ in Theorem D.1, but using the latent reward $r^*(x, y)$ instead of the reparameterized reward $r_t(x, y)$:

$$\pi_{t+1}^*(y|x) = \frac{1}{Z_t^*(x)} \cdot \pi_{\text{ref}}(y|x) \cdot \exp\left(\frac{1}{\beta} \cdot \sum_{i=0}^t \left(\frac{1}{1-\alpha} - \frac{\alpha^{t+1-i}}{1-\alpha}\right) \cdot r^*(x, y)\right), \quad (\text{C.6})$$

where $Z_t^*(x) = \sum_y \pi_{\text{ref}}(y|x) \cdot \exp\left(\sum_{i=0}^t (1/(1-\alpha) - \alpha^{t+1-i}/(1-\alpha)) \cdot r^*(x, y)/\beta\right)$ is the partition function for the auxiliary policy. With this notation, the following lemma provide an upper bound for the statistical errors arising from the gap between the reparameterized reward $r_t(x, y)$ and the latent reward $r^*(x, y)$.

Lemma C.1. Under the mild assumptions of realizability (Assumption B.5) and boundedness (Assumption B.6), with probability at least $1 - \delta$, the TV-distance between the policy $\hat{\pi}_{T+1}$ and auxiliary policy π_{T+1}^* is upper bounded by:

$$\mathbb{E}_{x \sim \rho} \left[\text{D}_{\text{TV}}(\hat{\pi}_{T+1}(\cdot|x), \pi_{T+1}^*(\cdot|x)) \right] \leq O\left(\sqrt{\frac{(T+1) \sum_{t=0}^T \kappa_t \cdot \log(T|\Pi|/\delta)}{N\beta^2(1-\alpha)^2}}\right),$$

where the coverage coefficient κ_t is defined as:

$$\kappa_t = \max_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \frac{\hat{\pi}_{T+1}(y|x) \pi_{T+1}^*(y|x)}{\pi_t^2(y|x)}.$$

With the help of Lemma C.1, we start the proof of Theorem B.7.

Proof of Theorem B.7. For each iteration $t \in [T]$ and prompt $x \in \mathcal{X}$, according to the definition of π_{t+1}^* in (C.6), we have

$$\begin{aligned} \pi_{t+1}^*(y|x) &= \frac{1}{Z_t^*(x)} \cdot \pi_{\text{ref}}(y|x) \cdot \exp\left(\frac{1}{\beta} \cdot \sum_{i=0}^t \left(\frac{1}{1-\alpha} - \frac{\alpha^{t+1-i}}{1-\alpha}\right) \cdot r^*(x, y)\right) \\ &\propto \pi_{\text{ref}}(y|x) \cdot \exp\left(\frac{1}{\beta} \cdot \sum_{i=0}^t \left(\frac{1}{1-\alpha} - \frac{\alpha^{t+1-i}}{1-\alpha}\right) \cdot r^*(x, y)\right) \\ &= \pi_{\text{ref}}(y|x) \cdot \exp\left(\frac{1}{\beta} \cdot \left(\frac{t+1}{1-\alpha} - \frac{\alpha}{(1-\alpha)^2} + \frac{\alpha^{t+2}}{(1-\alpha)^2}\right) \cdot r^*(x, y)\right). \end{aligned}$$

For simplicity, we set $\gamma = ((t+1)/(1-\alpha) - \alpha/(1-\alpha)^2 + \alpha^{t+2}/(1-\alpha)^2)/\beta$, and the sub-optimality gap for the policy $\hat{\pi}_{t+1}$ and a fixed prompt x can be denoted as:

$$\begin{aligned} &\mathbb{E}_{y \sim \pi^*(\cdot|x)} [r^*(x, y)] - \mathbb{E}_{y \sim \hat{\pi}_{t+1}(\cdot|x)} [r^*(x, y)] \\ &= r^*(x, y_x^*) - \frac{\sum_{y \in \mathcal{Y}} \pi_{\text{ref}}(y|x) \cdot \exp(\gamma \cdot r^*(x, y)) \cdot r^*(x, y)}{\sum_{y' \in \mathcal{Y}} \pi_{\text{ref}}(y'|x) \cdot \exp(\gamma \cdot r^*(x, y'))} \\ &= \frac{\sum_{y \neq y_x^*} \pi_{\text{ref}}(y|x) \cdot \exp(\gamma \cdot r^*(x, y) - \gamma \cdot r^*(x, y_x^*)) \cdot (r^*(x, y_x^*) - r^*(x, y))}{\sum_{y' \in \mathcal{Y}} \pi_{\text{ref}}(y'|x) \cdot \exp(\gamma \cdot r^*(x, y') - \gamma \cdot r^*(x, y_x^*))}, \end{aligned}$$

where $y_x^* = \arg \max_{y \in \mathcal{Y}} r^*(x, y)$ denotes the optimal response given the prompt x . For simplicity, we set $z(x, y) = \exp(\gamma \cdot r^*(x, y) - \gamma \cdot r^*(x, y_x^*))$ and $Z(x) = \sum_{y \neq y_x^*} \pi_{\text{ref}}(y|x) \cdot z(x, y)$. With this notation, we have

$$\begin{aligned} &\mathbb{E}_{y \sim \pi^*(\cdot|x)} [r^*(x, y)] - \mathbb{E}_{y \sim \hat{\pi}_{t+1}(\cdot|x)} [r^*(x, y)] \\ &= - \frac{\sum_{y \neq y_x^*} \pi_{\text{ref}}(y|x) \cdot z(x, y) \cdot \log z(x, y)}{\gamma \cdot \sum_{y' \in \mathcal{Y}} \pi_{\text{ref}}(y'|x) \cdot z(x, y')} \end{aligned}$$

$$\begin{aligned}
&= -\frac{(1 - \pi_{\text{ref}}(y_x^*|x)) \cdot \sum_{y \neq y_x^*} \pi_{\text{ref}}(y|x) / (1 - \pi_{\text{ref}}(y_x^*|x)) \cdot z(x, y) \cdot \log z(x, y)}{\gamma \cdot \sum_{y' \in \mathcal{Y}} \pi_{\text{ref}}(y'|x) \cdot z(x, y')} \\
&\leq -\frac{\sum_{y \neq y_x^*} \pi_{\text{ref}}(y|x) \cdot z(x, y) \cdot \log \left(\sum_{y \neq y_x^*} \pi_{\text{ref}}(y|x) \cdot z(x, y) / (1 - \pi_{\text{ref}}(y_x^*|x)) \right)}{\gamma \cdot \sum_{y' \in \mathcal{Y}} \pi_{\text{ref}}(y'|x) \cdot z(x, y')} \\
&= -\frac{Z(x) \cdot \log \left(Z(x) / (1 - \pi_{\text{ref}}(y_x^*|x)) \right)}{\gamma \cdot (Z(x) + \pi_{\text{ref}}(y_x^*|x))}, \tag{C.7}
\end{aligned}$$

where the inequality holds due to $f(x) = x \log x$ is a convex function with the fact that $\sum_{y \neq y_x^*} \pi_{\text{ref}}(y|x) / (1 - \pi_{\text{ref}}(y_x^*|x)) = 1$, and the last equation holds due to $z(x, y_x^*) = 1$. Now, we consider the following auxiliary function for $0 < Z \leq 1 - \pi_{\text{ref}}(y_x^*|x)$:

$$f(Z) = -Z \cdot \log \left(Z / (1 - \pi_{\text{ref}}(y_x^*|x)) \right) - (Z + \pi_{\text{ref}}(y_x^*|x)) \cdot \log \left((1 - \pi_{\text{ref}}(y_x^*|x)) / \pi_{\text{ref}}(y_x^*|x) \right).$$

With basic math calculation, we have

$$\begin{aligned}
f'(Z) &= \log \left(\frac{1 - \pi_{\text{ref}}(y_x^*|x)}{Z} \right) - \left(1 + \log \left((1 - \pi_{\text{ref}}(y_x^*|x)) / \pi_{\text{ref}}(y_x^*|x) \right) \right) \\
f(Z) &\leq f \left((1 - \pi_{\text{ref}}(y_x^*|x)) / \exp \left(1 + \log \left((1 - \pi_{\text{ref}}(y_x^*|x)) / \pi_{\text{ref}}(y_x^*|x) \right) \right) \right) \leq 0.
\end{aligned}$$

Substituting the result $f(Z) \leq 0$ into (C.7), we have

$$\begin{aligned}
&\mathbb{E}_{y \sim \pi^*(\cdot|x)} [r^*(x, y)] - \mathbb{E}_{y \sim \pi_{T+1}^*(\cdot|x)} [r^*(x, y)] \\
&\leq -\frac{Z(x) \cdot \log \left(Z(x) / (1 - \pi_{\text{ref}}(y_x^*|x)) \right)}{\gamma \cdot (Z(x) + \pi_{\text{ref}}(y_x^*|x))} \\
&\leq \frac{\log \left((1 - \pi_{\text{ref}}(y_x^*|x)) / \pi_{\text{ref}}(y_x^*|x) \right)}{\gamma} \\
&= \tilde{O} \left(\frac{(1 - \alpha)\beta}{t} \right). \tag{C.8}
\end{aligned}$$

Therefore, the sub-optimality gap between $\hat{\pi}_{T+1}$ and the optimal policy $\pi^*(x) = \arg \max_{y \in \mathcal{Y}} r^*(x, y)$ is bounded by

$$\begin{aligned}
&\mathbb{E}_{x \sim \rho, y \sim \pi^*(\cdot|x)} [r^*(x, y)] - \mathbb{E}_{x \sim \rho, y \sim \hat{\pi}_{T+1}(\cdot|x)} [r^*(x, y)] \\
&= \mathbb{E}_{x \sim \rho, y \sim \pi^*(\cdot|x)} [r^*(x, y)] - \mathbb{E}_{x \sim \rho, y \sim \pi_{T+1}^*(\cdot|x)} [r^*(x, y)] \\
&\quad + \mathbb{E}_{x \sim \rho, y \sim \pi_{T+1}^*(\cdot|x)} [r^*(x, y)] - \mathbb{E}_{x \sim \rho, y \sim \hat{\pi}_{T+1}(\cdot|x)} [r^*(x, y)] \\
&\leq \mathbb{E}_{x \sim \rho, y \sim \pi^*(\cdot|x)} [r^*(x, y)] - \mathbb{E}_{x \sim \rho, y \sim \pi_{T+1}^*(\cdot|x)} [r^*(x, y)] \\
&\quad + \mathbb{E}_{x \sim \rho} \left[2\text{D}_{\text{TV}}(\hat{\pi}_{T+1}(\cdot|x), \pi_{T+1}^*(\cdot|x)) \right] \\
&\leq \mathbb{E}_{x \sim \rho, y \sim \pi^*(\cdot|x)} [r^*(x, y)] - \mathbb{E}_{x \sim \rho, y \sim \pi_{T+1}^*(\cdot|x)} [r^*(x, y)] + O \left(\sqrt{\frac{(T+1) \sum_{t=0}^T \kappa_t \cdot \log(T|\Pi|/\delta)}{N\beta^2(1-\alpha)^2}} \right) \\
&\leq \tilde{O} \left(\frac{(1-\alpha)\beta}{T} \right) + O \left(\sqrt{\frac{(T+1) \sum_{t=0}^T \kappa_t \cdot \log(T|\Pi|/\delta)}{N\beta^2(1-\alpha)^2}} \right),
\end{aligned}$$

where the first inequality holds due to the definition of TV-distance with the fact that $r^*(x, y) \in [-1, 1]$, the second inequality holds due to Lemma C.1 and the last inequality holds due to (C.8). Thus, we complete the proof of Theorem B.7. \square

C.3 Proof of Theorem B.11

In this section, we provide the proof of Theorem B.11.

Proof of Theorem B.11. For each iteration $t \in [T]$ and prompt $x \in \mathcal{X}$, according to the definition of π_{t+1}^* in (C.10), we have

$$\begin{aligned}\pi_{t+1}^*(y|x) &= \frac{1}{Z_t^*(x)} \cdot \pi_{\text{ref}}(y|x) \cdot \exp\left(\frac{1}{\beta} \cdot \sum_{i=0}^t \left(\frac{1}{1-\alpha} - \frac{\alpha^{t+1-i}}{1-\alpha}\right) \cdot r^*(x, y)\right) \\ &\propto \pi_{\text{ref}}(y|x) \cdot \exp\left(\frac{1}{\beta} \cdot \sum_{i=0}^t \left(\frac{1}{1-\alpha} - \frac{\alpha^{t+1-i}}{1-\alpha}\right) \cdot r^*(x, y)\right) \\ &= \pi_{\text{ref}}(y|x) \cdot \exp\left(\frac{1}{\beta} \cdot \left(\frac{t+1}{1-\alpha} - \frac{\alpha}{(1-\alpha)^2} + \frac{\alpha^{t+2}}{(1-\alpha)^2}\right) \cdot r^*(x, y)\right).\end{aligned}$$

For simplicity, we set $\gamma = ((t+1)/(1-\alpha) - \alpha/(1-\alpha)^2 + \alpha^{t+1}/(1-\alpha)^2)/\beta$, and for each prompt $x \in \mathcal{X}$, the KL-divergence between $\pi_{t+1}^*(\cdot|x)$ and the optimal policy $\pi^*(x) = \arg \max_{y \in \mathcal{Y}} r^*(x, y)$ can be denoted as:

$$\begin{aligned}\text{KL}(\pi^*(x) \parallel \pi_{t+1}^*(\cdot|x)) &= \log \frac{1}{\pi_{t+1}^*(y_x^*|x)} \\ &= \log \frac{\sum_{y \in \mathcal{Y}} \pi_{\text{ref}}(y|x) \cdot \exp(\gamma \cdot r^*(x, y))}{\pi_{\text{ref}}(y_x^*|x) \cdot \exp(\gamma \cdot r^*(x, y_x^*))} \\ &\leq \frac{\sum_{y \neq y_x^*} \pi_{\text{ref}}(y|x) \cdot \exp(\gamma \cdot r^*(x, y))}{\pi_{\text{ref}}(y_x^*|x) \cdot \exp(\gamma \cdot r^*(x, y_x^*))} \\ &= \frac{\sum_{y \neq y_x^*} \pi_{\text{ref}}(y|x) \cdot \exp(\gamma \cdot (r^*(x, y) - r^*(x, y_x^*)))}{\pi_{\text{ref}}(y_x^*|x)} \\ &\leq \frac{\max_{y \neq y_x^*} \exp(-\gamma \cdot (r^*(x, y_x^*) - r^*(x, y)))}{\pi_{\text{ref}}(y_x^*|x)} \\ &\leq \frac{\exp(-\gamma \Delta)}{\pi_{\text{ref}}(y_x^*|x)} \\ &= \exp\left(-O\left(\frac{t\Delta}{(1-\alpha)} \cdot \frac{1}{\beta}\right)\right),\end{aligned}\tag{C.9}$$

where $y_x^* = \arg \max_{y \in \mathcal{Y}} r^*(x, y)$ denotes the optimal response given the prompt x , the first inequality holds due to the fact that $\log(1+x) \leq x$, the second inequality holds due to $\sum_{y \neq y_x^*} \pi_{\text{ref}}(y|x) \leq 1$ and the last inequality holds due to Assumption B.9. Therefore, the TV distance between optimal policy π^* and $\hat{\pi}_{T+1}$ can be upper bounded as following:

$$\begin{aligned}\mathbb{E}_{x \sim \rho} \left[\text{D}_{\text{TV}}(\hat{\pi}_{T+1}(\cdot|x), \pi^*(\cdot|x)) \right] \\ &\leq \mathbb{E}_{x \sim \rho} \left[\text{D}_{\text{TV}}(\hat{\pi}_{T+1}(\cdot|x), \pi_{T+1}^*(\cdot|x)) \right] + \mathbb{E}_{x \sim \rho} \left[\text{D}_{\text{TV}}(\pi^*(\cdot|x), \pi_{T+1}^*(\cdot|x)) \right] \\ &\leq O\left(\sqrt{\frac{(T+1) \sum_{t=0}^T \kappa_t \cdot \log(T|\Pi|/\delta)}{N\beta^2(1-\alpha)^2}}\right) + \mathbb{E}_{x \sim \rho} \left[\text{D}_{\text{TV}}(\pi^*(\cdot|x), \pi_{T+1}^*(\cdot|x)) \right] \\ &\leq O\left(\sqrt{\frac{(T+1) \sum_{t=0}^T \kappa_t \cdot \log(T|\Pi|/\delta)}{N\beta^2(1-\alpha)^2}}\right) + \mathbb{E}_{x \sim \rho} \left[\sqrt{\frac{1}{2} \cdot \text{D}_{\text{KL}}(\pi^*(\cdot|x) \parallel \pi_{T+1}^*(\cdot|x))} \right] \\ &\leq O\left(\sqrt{\frac{(T+1) \sum_{t=0}^T \kappa_t \cdot \log(T|\Pi|/\delta)}{N\beta^2(1-\alpha)^2}}\right) + \exp\left(-O\left(\frac{t\Delta}{(1-\alpha)} \cdot \frac{1}{\beta}\right)\right),\end{aligned}$$

where the first inequality holds due to $\text{D}_{\text{TV}}(X, Y) \leq \text{D}_{\text{TV}}(X, Z) + \text{D}_{\text{TV}}(Z, Y)$, the second inequality holds due to Lemma C.1, the third inequality holds due to $\text{D}_{\text{TV}}(X, Y) \leq \sqrt{\text{D}_{\text{KL}}(X \parallel Y)}/2$ and the last inequality holds due to (C.9). Thus, we complete the proof of Theorem B.11. \square

C.4 Proof of Theorem B.14

In this section, we provide the proof of Theorem B.14. Similar to the proof of Theorem B.7, we define the auxiliary policy, which is updated according to the dynamics of $\hat{\pi}_{t+1}$ as described in Theorem D.1. However, in environments with general preferences, a latent reward $r^*(x, y)$ may not exist. Instead, we use the win probability of response y against the policy π_t to replace the reparameterized reward $r_t(x, y)$. Specifically, we set:

$$\pi_{t+1}^*(y|x) = \frac{1}{Z_t^*(x)} \cdot \pi_{\text{ref}}(y|x) \cdot \exp\left(\frac{1}{\beta} \cdot \sum_{i=0}^t \left(\frac{1}{1-\alpha} - \frac{\alpha^{t+1-i}}{1-\alpha}\right) \cdot r_t^*(x, y)\right), \quad (\text{C.10})$$

where $r_t^*(x, y) = \mathbb{E}_{y' \sim \pi_t} [\mathbb{P}(y \succ y'|x)]$ and $Z_t^*(x) = \sum_y \pi_{\text{ref}}(y|x) \cdot \exp\left(\sum_{i=0}^t (1/(1-\alpha) - \alpha^{t+1-i}/(1-\alpha)) \cdot r_t^*(x, y)/\beta\right)$ is the partition function for the auxiliary policy. With this notation, the following lemma provide a upper bound for the statistical errors arising from the gap between the reparameterized reward $r_t(x, y)$ and the win probability $r_t^*(x, y)$.

Lemma C.2. Under the mild assumptions of realizability (Assumption E.3) and boundedness (Assumption E.4), the TV-distance between the policy $\hat{\pi}_{T+1}$ and auxiliary policy π_{T+1}^* is upper bounded by:

$$\mathbb{E}_{x \sim \rho} \left[\text{D}_{\text{TV}}(\hat{\pi}_{T+1}(\cdot|x), \pi_{T+1}^*(\cdot|x)) \right] \leq O\left(\sqrt{\frac{(T+1) \sum_{t=0}^T \kappa_t \cdot \log(T|\Pi|/\delta)}{N\beta^2(1-\alpha)^2}}\right),$$

where the coverage coefficient κ_t is defined as:

$$\kappa_t = \max_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \frac{\hat{\pi}_{T+1}(y|x) \pi_{T+1}^*(y|x)}{\pi_t^2(y|x)}.$$

With the help of Lemma C.2, we start the proof of Theorem B.14.

Proof of Theorem B.14. For each iteration $t \in [T]$ and prompt $x \in \mathcal{X}$, according to the definition of π_{t+1}^* in (C.6), we have

$$\begin{aligned} \pi_{t+1}^*(y|x) &= \frac{1}{Z_t^*(x)} \cdot \pi_{\text{ref}}(y|x) \cdot \exp\left(\frac{1}{\beta} \cdot \sum_{i=0}^t \left(\frac{1}{1-\alpha} - \frac{\alpha^{t+1-i}}{1-\alpha}\right) \cdot r_t^*(x, y)\right) \\ &\propto \pi_{\text{ref}}(y|x) \cdot \exp\left(\frac{1}{\beta} \cdot \sum_{i=0}^t \left(\frac{1}{1-\alpha} - \frac{\alpha^{t+1-i}}{1-\alpha}\right) \cdot r_t^*(x, y)\right), \end{aligned}$$

where the win probability $r_t^*(x, y) = \mathbb{E}_{y' \sim \pi_t} [\mathbb{P}(y \succ y'|x)]$. According to the Assumption B.13, for each prompt $x \in \mathcal{X}$ and behavior policy π_t , the unique optimal response $y_x^* \in \mathcal{Y}$ satisfies

$$r_t^*(x, y_x^*) = \mathbb{E}_{y' \sim \pi_t} [\mathbb{P}(y_x^* \succ y'|x)] \geq \mathbb{E}_{y' \sim \pi_t} [\mathbb{P}(y \succ y'|x)] + \Delta = r_t^*(x, y) + \Delta, \forall y \neq y_x^*. \quad (\text{C.11})$$

For simplicity, we set $\gamma_t = (1/(1-\alpha) - \alpha^{t+1-i}/(1-\alpha))/\beta$, and for each prompt $x \in \mathcal{X}$, the KL-divergence between $\pi_{t+1}^*(\cdot|x)$ and the optimal policy $\pi^*(x) = y_x^*$ can be denoted as:

$$\begin{aligned} \text{KL}(\pi^*(x) \|\pi_{t+1}^*(\cdot|x)) &= \log \frac{1}{\pi_{t+1}^*(y_x^*|x)} \\ &= \log \frac{\sum_{y \in \mathcal{Y}} \pi_{\text{ref}}(y|x) \cdot \exp\left(\sum_{i=0}^t \gamma_i \cdot r_i^*(x, y)\right)}{\pi_{\text{ref}}(y_x^*|x) \cdot \exp\left(\sum_{i=0}^t \gamma_i \cdot r_i^*(x, y_x^*)\right)} \\ &\leq \frac{\sum_{y \neq y_x^*} \pi_{\text{ref}}(y|x) \cdot \exp\left(\sum_{i=0}^t \gamma_i \cdot r_i^*(x, y)\right)}{\pi_{\text{ref}}(y_x^*|x) \cdot \exp\left(\sum_{i=0}^t \gamma_i \cdot r_i^*(x, y_x^*)\right)} \\ &= \frac{\sum_{y \neq y_x^*} \pi_{\text{ref}}(y|x) \cdot \exp\left(\sum_{i=0}^t \gamma_i (r_i^*(x, y) - r_i^*(x, y_x^*))\right)}{\pi_{\text{ref}}(y_x^*|x)} \end{aligned}$$

$$\begin{aligned}
&\leq \frac{\max_{y \neq y_x^*} \exp\left(-\sum_{i=0}^t \gamma_i (r_i^*(x, y_x^*) - r_i^*(x, y))\right)}{\pi_{\text{ref}}(y_x^*|x)} \\
&\leq \frac{\exp\left(-\sum_{i=0}^t \gamma_i \Delta\right)}{\pi_{\text{ref}}(y_x^*|x)} \\
&= \exp\left(-O\left(\frac{t\Delta}{(1-\alpha)} \cdot \frac{1}{\beta}\right)\right), \tag{C.12}
\end{aligned}$$

where y_x^* denotes the optimal response given the prompt x , the first inequality holds due to the fact that $\log(1+x) \leq x$, the second inequality holds due to $\sum_{y \neq y_x^*} \pi_{\text{ref}}(y|x) \leq 1$ and the last inequality holds due to (C.11). Therefore, the TV distance between optimal policy π^* and $\hat{\pi}_{T+1}$ can be upper bounded as following:

$$\begin{aligned}
&\mathbb{E}_{x \sim \rho} \left[\mathbf{D}_{\text{TV}}(\hat{\pi}_{T+1}(\cdot|x), \pi^*(\cdot|x)) \right] \\
&\leq \mathbb{E}_{x \sim \rho} \left[\mathbf{D}_{\text{TV}}(\hat{\pi}_{T+1}(\cdot|x), \pi_{T+1}^*(\cdot|x)) \right] + \mathbb{E}_{x \sim \rho} \left[\mathbf{D}_{\text{TV}}(\pi^*(\cdot|x), \pi_{T+1}^*(\cdot|x)) \right] \\
&\leq O\left(\sqrt{\frac{(T+1) \sum_{t=0}^T \kappa_t \cdot \log(T|\Pi|/\delta)}{N\beta^2(1-\alpha)^2}}\right) + \mathbb{E}_{x \sim \rho} \left[\mathbf{D}_{\text{TV}}(\pi^*(\cdot|x), \pi_{T+1}^*(\cdot|x)) \right] \\
&\leq O\left(\sqrt{\frac{(T+1) \sum_{t=0}^T \kappa_t \cdot \log(T|\Pi|/\delta)}{N\beta^2(1-\alpha)^2}}\right) + \mathbb{E}_{x \sim \rho} \left[\sqrt{\frac{1}{2} \cdot \mathbf{D}_{\text{KL}}(\pi^*(\cdot|x) \parallel \pi_{T+1}^*(\cdot|x))} \right] \\
&\leq O\left(\sqrt{\frac{(T+1) \sum_{t=0}^T \kappa_t \cdot \log(T|\Pi|/\delta)}{N\beta^2(1-\alpha)^2}}\right) + \exp\left(-O\left(\frac{t\Delta}{(1-\alpha)} \cdot \frac{1}{\beta}\right)\right),
\end{aligned}$$

where the first inequality holds due to $\mathbf{D}_{\text{TV}}(X, Y) \leq \mathbf{D}_{\text{TV}}(X, Z) + \mathbf{D}_{\text{TV}}(Z, Y)$, the second inequality holds due to Lemma C.2, the third inequality holds due to $\mathbf{D}_{\text{TV}}(X, Y) \leq \sqrt{\mathbf{D}_{\text{KL}}(X \parallel Y)}/2$ and the last inequality holds due to (C.12). Thus, we complete the proof of Theorem B.14. \square

D Proof of Lemmas in Appendix C

D.1 Proof of Lemma C.1

In this subsection, we provide the proof of Lemma C.1 and first propose the following lemmas.

Lemma D.1. Under the mild assumptions of realizability (Assumption B.5) and boundedness (Assumption B.6), for each iteration $t \in [T]$, with probability at least $1 - \delta$, the estimation error can be upper bounded as follows:

$$\mathbb{E}_{x \sim \rho, (y_1, y_2) \sim \pi_t} \left[\left(r^*(x, y_1) - r^*(x, y_2) - r_t(x, y_1) + r_t(x, y_2) \right)^2 \right] \leq O\left(\frac{\log(|\Pi|/\delta)}{N}\right),$$

where reparameterized reward $r_t(x, y) = \beta \log \hat{\pi}_{t+1}(y|x) - \beta \log \pi_t(y|x)$.

Lemma D.2. For any $x \in \mathbb{R}^+$, we have $(1+x) \cdot |\log x| \geq |x-1|$.

With the help of Lemmas D.1 and D.2, we now begin the proof of Lemma C.1.

Proof of Lemma C.1. For each iteration $t \in [T]$ and prompt $x \in \mathcal{X}$, the TV-distance between $\hat{\pi}_{t+1}(\cdot|x)$ and $\pi_{t+1}^*(\cdot|x)$ can be upper bounded as follows:

$$\begin{aligned}
&\mathbf{D}_{\text{TV}}(\hat{\pi}_{t+1}(\cdot|x), \pi_{t+1}^*(\cdot|x)) \\
&= \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x)} \left[\frac{1}{2} \cdot \left| \frac{\hat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} - 1 \right| \right] \\
&\leq \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x)} \left[\frac{1}{2} \cdot \left(1 + \frac{\hat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right) \cdot \left| \log \frac{\hat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right| \right] \\
&= \frac{1}{2} \cdot \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x)} \left[\left| \log \frac{\hat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right| \right] + \frac{1}{2} \cdot \mathbb{E}_{y \sim \hat{\pi}_{t+1}(\cdot|x)} \left[\left| \log \frac{\hat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right| \right]
\end{aligned}$$

$$\begin{aligned}
&\leq \sqrt{\frac{1}{2} \cdot \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x)} \left[\left| \log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right| \right]^2 + \frac{1}{2} \cdot \mathbb{E}_{y \sim \widehat{\pi}_{t+1}(\cdot|x)} \left[\left| \log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right| \right]^2} \\
&\leq \sqrt{\frac{1}{2} \cdot \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right)^2 \right] + \frac{1}{2} \cdot \mathbb{E}_{y \sim \widehat{\pi}_{t+1}(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right)^2 \right]}, \quad (\text{D.1})
\end{aligned}$$

where the first inequality holds due to Lemma D.2, the second inequality holds due to $x + y \leq \sqrt{2(x^2 + y^2)}$ for $x, y > 0$, and the last inequality holds due to $\mathbb{E}[x]^2 \leq \mathbb{E}[x^2]$. In addition, for each iteration $t \in [T]$ and prompt $x \in \mathcal{X}$, we have

$$\begin{aligned}
&\mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right)^2 \right] + \mathbb{E}_{y \sim \widehat{\pi}_{t+1}(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right)^2 \right] \\
&\leq \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right)^2 \right] + \mathbb{E}_{y \sim \widehat{\pi}_{t+1}(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right)^2 \right] \\
&\quad + 2\text{D}_{\text{KL}}(\widehat{\pi}_{t+1}(\cdot|x) \parallel \pi_{t+1}^*(\cdot|x)) \cdot \text{D}_{\text{KL}}(\pi_{t+1}^*(\cdot|x) \parallel \widehat{\pi}_{t+1}(\cdot|x)) \\
&= \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right)^2 \right] + \mathbb{E}_{y \sim \widehat{\pi}_{t+1}(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right)^2 \right] \\
&\quad + 2\mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x)} \left[-\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right] \cdot \mathbb{E}_{y \sim \widehat{\pi}_{t+1}(\cdot|x)} \left[\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right] \\
&= \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x), y' \sim \widehat{\pi}_{t+1}(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} - \log \frac{\widehat{\pi}_{t+1}(y'|x)}{\pi_{t+1}^*(y'|x)} \right)^2 \right], \quad (\text{D.2})
\end{aligned}$$

where the inequality holds due to KL-divergence is non-negative. According to Theorem B.4 and definition of π_{t+1}^* in (C.6), we have

$$\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} = \frac{\log Z_t^*(x)}{\log Z_t(x)} + \sum_{i=0}^t \frac{1}{\beta} \cdot \left(\frac{1}{1-\alpha} - \frac{\alpha^{t+1-i}}{1-\alpha} \right) \cdot (r_i(x, y) - r^*(x, y)). \quad (\text{D.3})$$

Substituting (D.3) into (D.2), we have

$$\begin{aligned}
&\mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right)^2 \right] + \mathbb{E}_{y \sim \widehat{\pi}_{t+1}(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right)^2 \right] \\
&\leq \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x), y' \sim \widehat{\pi}_{t+1}(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} - \log \frac{\widehat{\pi}_{t+1}(y'|x)}{\pi_{t+1}^*(y'|x)} \right)^2 \right] \\
&\leq \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x), y' \sim \widehat{\pi}_{t+1}(\cdot|x)} \left[\left(\sum_{i=0}^t \alpha_{t+1-i} \cdot (r_i(x, y) - r^*(x, y)) - (r_i(x, y') - r^*(x, y')) \right)^2 \right] \\
&\leq \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x), y' \sim \widehat{\pi}_{t+1}(\cdot|x)} \left[(t+1) \cdot \sum_{i=0}^t \alpha_{t+1-i}^2 \cdot (r_i(x, y) - r_i(x, y') - r^*(x, y) + r^*(x, y'))^2 \right], \quad (\text{D.4})
\end{aligned}$$

where $\alpha_i = (1/(1-\alpha) - \alpha^i/(1-\alpha))/\beta$ and the last inequality holds due to Cauchy–Schwarz inequality. According to the definition of coverage coefficient κ_t in (B.1), for policy $\widehat{\pi}_{T+1}$, we have

$$\begin{aligned}
&\mathbb{E}_{y \sim \pi_{T+1}^*(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{T+1}(y|x)}{\pi_{T+1}^*(y|x)} \right)^2 \right] + \mathbb{E}_{y \sim \widehat{\pi}_{T+1}(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{T+1}(y|x)}{\pi_{T+1}^*(y|x)} \right)^2 \right] \\
&\leq (T+1) \cdot \\
&\quad \sum_{t=0}^T \frac{1}{\beta^2(1-\alpha)^2} \cdot \mathbb{E}_{y \sim \pi_{T+1}^*(\cdot|x), y' \sim \widehat{\pi}_{T+1}(\cdot|x)} \left[(r_t(x, y) - r_t(x, y') - r^*(x, y) + r^*(x, y'))^2 \right] \\
&\leq (T+1) \cdot \\
&\quad \sum_{t=0}^T \frac{\kappa_t}{\beta^2(1-\alpha)^2} \cdot \mathbb{E}_{y \sim \pi_t^*(\cdot|x), y' \sim \widehat{\pi}_t(\cdot|x)} \left[(r_t(x, y) - r_t(x, y') - r^*(x, y) + r^*(x, y'))^2 \right]
\end{aligned}$$

$$\leq \frac{1}{\sigma^2(1+R) \cdot (1-\sigma(1+R))^2} \cdot \frac{(T+1) \sum_{t=0}^T \kappa_t \cdot \log(T|\Pi|/\delta)}{N\beta^2(1-\alpha)^2}, \quad (\text{D.5})$$

where the first inequality holds due to (D.4) with the fact that $\alpha_i \leq 1/(1-\alpha)$, the second inequality holds due to the definition of coverage coefficient κ_t in (B.1) and the last inequality holds due to Lemma D.1 with a union bound on the probability across all T iterations. Finally, substituting (D.5) into (D.1), we have

$$\begin{aligned} & \mathbb{E}_{x \sim \rho} \left[\text{D}_{\text{TV}}(\widehat{\pi}_{T+1}(\cdot|x), \pi_{T+1}^*(\cdot|x)) \right] \\ & \leq \mathbb{E}_{x \sim \rho} \left[\sqrt{\frac{1}{2} \cdot \mathbb{E}_{y \sim \pi_{T+1}^*(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{T+1}(y|x)}{\pi_{T+1}^*(y|x)} \right)^2 \right]} + \frac{1}{2} \cdot \mathbb{E}_{y \sim \widehat{\pi}_{T+1}(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{T+1}(y|x)}{\pi_{T+1}^*(y|x)} \right)^2 \right]} \right] \\ & \leq \frac{1}{\sigma(1+R) \cdot (1-\sigma(1+R))} \cdot \sqrt{\frac{(T+1) \sum_{t=0}^T \kappa_t \cdot \log(T|\Pi|/\delta)}{N\beta^2(1-\alpha)^2}}. \end{aligned}$$

Thus, we complete the proof of Lemma C.1. \square

D.2 Proof of Lemma C.2

In this subsection, we provide the proof of Lemma C.2 and first propose the following lemma.

Lemma D.3. Under the mild assumptions of realizability and boundedness (see detailed definitions in Appendix E.3), for each iteration $t \in [T]$, with probability at least $1 - \delta$, the estimation error can be upper bounded as follows:

$$\mathbb{E}_{x \sim \rho, (y_1, y_2) \sim \pi_t} \left[\left(r_t^*(x, y_1) - r_t^*(x, y_2) - r_t(x, y_1) + r_t(x, y_2) \right)^2 \right] \leq O\left(\frac{\log(|\Pi|/\delta)}{N} \right),$$

where the win probability $r_t^*(x, y) = \mathbb{E}_{y' \sim \pi_t} [\mathbb{P}(y \succ y'|x)]$ and reparameterized reward $r_t(x, y) = \beta \log \widehat{\pi}_{t+1}(y|x) - \beta \log \pi_t(y|x)$,

With the help of Lemma C.2, we start the proof of Lemma C.2. The proof is similar to the proof of Lemma C.1, however, it is worth noting that this proof relies on a different auxiliary policy based on the win probability rather than the latent reward:

$$\pi_{t+1}^*(y|x) = \frac{1}{Z_t^*(x)} \cdot \pi_{\text{ref}}(y|x) \cdot \exp\left(\frac{1}{\beta} \cdot \sum_{i=0}^t \left(\frac{1}{1-\alpha} - \frac{\alpha^{t+1-i}}{1-\alpha} \right) \cdot r_t^*(x, y) \right),$$

where the win probability $r_t^*(x, y) = \mathbb{E}_{y' \sim \pi_t} [\mathbb{P}(y \succ y'|x)]$ and $Z_t^*(x) = \sum_y \pi_{\text{ref}}(y|x) \cdot \exp\left(\sum_{i=0}^t (1/(1-\alpha) - \alpha^{t+1-i}/(1-\alpha)) \cdot r_t^*(x, y)/\beta \right)$ is the partition function for the auxiliary policy.

Proof of Lemma C.2. For each iteration $t \in [T]$ and prompt $x \in \mathcal{X}$, the TV-distance between $\widehat{\pi}_{t+1}(\cdot|x)$ and $\pi_{t+1}^*(\cdot|x)$ can be upper bounded as follows:

$$\begin{aligned} & \text{D}_{\text{TV}}(\widehat{\pi}_{t+1}(\cdot|x), \pi_{t+1}^*(\cdot|x)) \\ & = \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x)} \left[\frac{1}{2} \cdot \left| \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} - 1 \right| \right] \\ & \leq \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x)} \left[\frac{1}{2} \cdot \left(1 + \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right) \cdot \left| \log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right| \right] \\ & = \frac{1}{2} \cdot \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x)} \left[\left| \log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right| \right] + \frac{1}{2} \cdot \mathbb{E}_{y \sim \widehat{\pi}_{t+1}(\cdot|x)} \left[\left| \log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right| \right] \\ & \leq \sqrt{\frac{1}{2} \cdot \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x)} \left[\left| \log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right| \right]^2} + \frac{1}{2} \cdot \mathbb{E}_{y \sim \widehat{\pi}_{t+1}(\cdot|x)} \left[\left| \log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right| \right]^2 \\ & \leq \sqrt{\frac{1}{2} \cdot \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right)^2 \right]} + \frac{1}{2} \cdot \mathbb{E}_{y \sim \widehat{\pi}_{t+1}(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right)^2 \right]}, \quad (\text{D.6}) \end{aligned}$$

where the first inequality holds due to Lemma D.2, the second inequality holds due to $x + y \leq \sqrt{2(x^2 + y^2)}$ for $x, y > 0$, and the last inequality holds due to $\mathbb{E}[x]^2 \leq \mathbb{E}[x^2]$. In addition, for each iteration $t \in [T]$ and prompt $x \in \mathcal{X}$, we have

$$\begin{aligned}
& \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right)^2 \right] + \mathbb{E}_{y \sim \widehat{\pi}_{t+1}(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right)^2 \right] \\
& \leq \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right)^2 \right] + \mathbb{E}_{y \sim \widehat{\pi}_{t+1}(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right)^2 \right] \\
& \quad + 2\mathbf{D}_{\text{KL}}(\widehat{\pi}_{t+1}(\cdot|x) \parallel \pi_{t+1}^*(\cdot|x)) \cdot \mathbf{D}_{\text{KL}}(\pi_{t+1}^*(\cdot|x) \parallel \widehat{\pi}_{t+1}(\cdot|x)) \\
& = \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right)^2 \right] + \mathbb{E}_{y \sim \widehat{\pi}_{t+1}(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right)^2 \right] \\
& \quad + 2\mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x)} \left[-\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right] \cdot \mathbb{E}_{y \sim \widehat{\pi}_{t+1}(\cdot|x)} \left[\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right] \\
& = \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x), y' \sim \widehat{\pi}_{t+1}(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} - \log \frac{\widehat{\pi}_{t+1}(y'|x)}{\pi_{t+1}^*(y'|x)} \right)^2 \right], \tag{D.7}
\end{aligned}$$

where the inequality holds due to KL-divergence is non-negative. According to Theorem B.4 and definition of π_{t+1}^* in (C.10), we have

$$\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} = \frac{\log Z_t^*(x)}{\log Z_t(x)} + \sum_{i=0}^t \frac{1}{\beta} \cdot \left(\frac{1}{1-\alpha} - \frac{\alpha^{t+1-i}}{1-\alpha} \right) \cdot (r_i(x, y) - r^*(x, y)). \tag{D.8}$$

Substituting (D.8) into (D.7), we have

$$\begin{aligned}
& \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right)^2 \right] + \mathbb{E}_{y \sim \widehat{\pi}_{t+1}(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} \right)^2 \right] \\
& \leq \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x), y' \sim \widehat{\pi}_{t+1}(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_{t+1}^*(y|x)} - \log \frac{\widehat{\pi}_{t+1}(y'|x)}{\pi_{t+1}^*(y'|x)} \right)^2 \right] \\
& \leq \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x), y' \sim \widehat{\pi}_{t+1}(\cdot|x)} \left[\left(\sum_{i=0}^t \alpha_{t+1-i} \cdot (r_i(x, y) - r_i^*(x, y)) - (r_i(x, y') - r_i^*(x, y')) \right)^2 \right] \\
& \leq \mathbb{E}_{y \sim \pi_{t+1}^*(\cdot|x), y' \sim \widehat{\pi}_{t+1}(\cdot|x)} \left[(t+1) \cdot \sum_{i=0}^t \alpha_{t+1-i}^2 \cdot \left(r_i(x, y) - r_i(x, y') - r_i^*(x, y) + r_i^*(x, y') \right)^2 \right], \tag{D.9}
\end{aligned}$$

where $\alpha_i = (1/(1-\alpha) - \alpha^i/(1-\alpha))/\beta$ and the last inequality holds due to Cauchy–Schwarz inequality. According to the definition of coverage coefficient κ_t in (B.1), for policy $\widehat{\pi}_{T+1}$, we have

$$\begin{aligned}
& \mathbb{E}_{y \sim \pi_{T+1}^*(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{T+1}(y|x)}{\pi_{T+1}^*(y|x)} \right)^2 \right] + \mathbb{E}_{y \sim \widehat{\pi}_{T+1}(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{T+1}(y|x)}{\pi_{T+1}^*(y|x)} \right)^2 \right] \\
& \leq (T+1) \cdot \\
& \quad \sum_{t=0}^T \frac{1}{\beta^2(1-\alpha)^2} \cdot \mathbb{E}_{y \sim \pi_{T+1}^*(\cdot|x), y' \sim \widehat{\pi}_{T+1}(\cdot|x)} \left[\left(r_t(x, y) - r_t(x, y') - r_t^*(x, y) + r_t^*(x, y') \right)^2 \right] \\
& \leq (T+1) \cdot \\
& \quad \sum_{t=0}^T \frac{\kappa_t}{\beta^2(1-\alpha)^2} \cdot \mathbb{E}_{y \sim \pi_t^*(\cdot|x), y' \sim \widehat{\pi}_t(\cdot|x)} \left[\left(r_t(x, y) - r_t(x, y') - r_t^*(x, y) + r_t^*(x, y') \right)^2 \right] \\
& \leq O\left(\frac{(T+1) \sum_{t=0}^T \kappa_t \cdot \log(T|\Pi|/\delta)}{N\beta^2(1-\alpha)^2} \right), \tag{D.10}
\end{aligned}$$

where the first inequality holds due to (D.9) with the fact that $\alpha_i \leq 1/(1-\alpha)$, the second inequality holds due to the definition of coverage coefficient κ_t in (B.1) and the last inequality holds due to

Lemma D.3 with a union bound on the probability across all T iterations. Finally, substituring (D.10) into (D.6), we have

$$\begin{aligned} & \mathbb{E}_{x \sim \rho} \left[\text{D}_{\text{TV}}(\widehat{\pi}_{T+1}(\cdot|x), \pi_{T+1}^*(\cdot|x)) \right] \\ & \leq \mathbb{E}_{x \sim \rho} \left[\sqrt{\frac{1}{2} \cdot \mathbb{E}_{y \sim \pi_{T+1}^*(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{T+1}(y|x)}{\pi_{T+1}^*(y|x)} \right)^2 \right] + \frac{1}{2} \cdot \mathbb{E}_{y \sim \widehat{\pi}_{T+1}(\cdot|x)} \left[\left(\log \frac{\widehat{\pi}_{T+1}(y|x)}{\pi_{T+1}^*(y|x)} \right)^2 \right]} \right] \\ & \leq O \left(\sqrt{\frac{(T+1) \sum_{t=0}^T \kappa_t \cdot \log(T|\Pi|/\delta)}{N\beta^2(1-\alpha)^2}} \right). \end{aligned}$$

Thus, we complete the proof of Lemma C.2. \square

E Proof of Lemmas in Appendix D

E.1 Proof of Lemma D.1

In this section, we adapt the previous results on the estimation error from maximum likelihood estimation to the Bradley-Terry (BT) model (Bradley and Terry, 1952) and provide the proof of Lemma D.1.

We start the analysis with the conditional probability estimation setting. For an instance space \mathcal{X} and a target space \mathcal{Y} , we collect the dataset $\mathcal{D} = (z_i, o_i)_{i=1}^N$, where z_i is sampled from a reference policy $z_i \sim \mu_i$, and o_i is then generated with the latent conditional density $P(o_i|z_i) = f^*(z_i, o_i)$. Under this situation, the maximum likelihood estimator across a function class $\mathcal{F} : \mathcal{Z} \times \mathcal{O} \rightarrow \mathbb{R}$ can be denoted as follows

$$f = \arg \max_{f \in \mathcal{F}} \sum_{i=1}^N \log f(z_i, o_i).$$

For an i.i.d. sampled dataset \mathcal{D} , where $\mu_1 = \mu_2 = \dots = \mu_N$, Geer (2000) (Chapter 7) provides upper bounds for the estimation error between f and the hidden function f^* . Later, Agarwal et al. (2020) extended these results to reinforcement learning, allowing for a martingale sampling process.

Lemma E.1 (Theorem 21, Agarwal et al. 2020). For a finite function class $|\mathcal{F}| < +\infty$, if the latent function $f^* \in \mathcal{F}$, then with probability at least $1 - \delta$, we have

$$\sum_{i=1}^N \mathbb{E}_{z \sim \mu_i} \|f(z, \cdot) - f^*(z, \cdot)\|_{\text{TV}}^2 \leq 2 \log(|\mathcal{F}|/\delta).$$

With the help of lemma E.1, we start to prove Lemma D.1.

Proof of Lemma D.1. Based on the Bradley-Terry (BT) model defined in equation (2.1), for the prompt $x_i \in \mathcal{X}$ and the generated responses y_1, y_2 , we have

$$P(y_1 \succ y_2|x) = \sigma(r^*(x_i, y_{i,1}) - r^*(x_i, y_{i,2})).$$

In this context, we define the instance z_i and target o_i as follows

$$\begin{aligned} z_i &= (x_i, y_{i,1}, y_{i,2}), \\ o_i &= \mathbb{1}(y_{i,1} \succ y_{i,2}) - \mathbb{1}(y_{i,2} \succ y_{i,1}). \end{aligned}$$

For each policy $\pi \in \Pi$, we denote the reparameterized reward $r_\pi(x, y) = \beta \log \pi(y|x) - \beta \log \pi_t(y|x)$. Thus, the density function for an instance $z = (x, y_1, y_2)$ with reward r_π can be represented as

$$f_{r_\pi}(z, o) = \sigma(o \cdot (r_\pi(x, y_1) - r_\pi(x, y_2))),$$

and we denote the density function class \mathcal{F} as $\mathcal{F} = \{f_{r_\pi} | \pi \in \Pi\}$. Based on Assumption B.5 and the definition of the Bradley-Terry (BT) model, we have $f_{r^*} \in \mathcal{F}$, and the latent density function satisfies $P(o|z) = f_{r^*}(z, o)$. Consequently, the maximum likelihood estimator can be expressed as

$$f_{r_t} = \arg \max_{f_{r_\pi} \in \mathcal{F}} \sum_{i=1}^N \log f_{r_\pi}(z_i, o_i),$$

where $r_t(x, y) = \beta \log \widehat{\pi}_{t+1}(y|x) - \beta \log \pi_t(y|x)$. Thus, according to Lemma E.1, the estimation error between r^* and r_t can be upper bounded by

$$\sum_{i=1}^N \mathbb{E}_{z \sim \mu_i} \|f_{r_t}(z, \cdot) - f_{r^*}(z, \cdot)\|_{\text{TV}}^2 \leq 2 \log(|\Pi|/\delta). \quad (\text{E.1})$$

Since the dataset \mathcal{D}_t is collected with reference policy π_t , we have

$$\begin{aligned} & \mathbb{E}_{x \sim \rho, (y_1, y_2) \sim \pi_t} \left[\sigma(r^*(x, y_1) - r^*(x, y_2)) - \sigma(r_t(x, y_1) - r_t(x, y_2)) \right]^2 \\ &= \frac{1}{2} \cdot \mathbb{E}_{x \sim \rho, (y_1, y_2) \sim \pi_t} \left[\sigma(r^*(x, y_1) - r^*(x, y_2)) - \sigma(r_t(x, y_1) - r_t(x, y_2)) \right]^2 \\ & \quad + \frac{1}{2} \cdot \mathbb{E}_{x \sim \rho, (y_1, y_2) \sim \pi_t} \left[\sigma(r^*(x, y_2) - r^*(x, y_1)) - \sigma(r_t(x, y_2) - r_t(x, y_1)) \right]^2 \\ &= \mathbb{E}_{x \sim \rho, (y_1, y_2) \sim \pi_t} \|f_{r_t}(z, \cdot) - f_{r^*}(z, \cdot)\|_{\text{TV}}^2 \\ &\leq \frac{\log(|\Pi|/\delta)}{N}, \end{aligned} \quad (\text{E.2})$$

where the second equation holds due to the definition of TV-distance with $z = (x, y_1, y_2)$ and the inequality holds due to (E.1). According to Assumption B.6, we have $r^*(x, y) \in [-1, 1]$ and $r_t(x, y) \in [-R, R]$, which implies that

$$\begin{aligned} & \mathbb{E}_{x \sim \rho, (y_1, y_2) \sim \pi_t} \left[(r^*(x, y_1) - r^*(x, y_2) - r_t(x, y_1) + r_t(x, y_2))^2 \right] \\ &\leq \frac{1}{\sigma^2(1+R) \cdot (1-\sigma(1+R))^2} \\ & \quad \times \mathbb{E}_{x \sim \rho, (y_1, y_2) \sim \pi_t} \left[\sigma(r^*(x, y_1) - r^*(x, y_2)) - \sigma(r_t(x, y_1) - r_t(x, y_2)) \right]^2 \\ &\leq \frac{1}{\sigma^2(1+R) \cdot (1-\sigma(1+R))^2} \cdot \frac{\log(|\Pi|/\delta)}{N}, \end{aligned}$$

where the first inequality holds due to $|a-b| \cdot \max_{x \in [a,b]} \sigma'(x) \leq |\sigma(a) - \sigma(b)|$ with the fact that $\sigma'(x) = \sigma(x)(1-\sigma(x)) \leq \sigma(1+R)(1-\sigma(1+R))$ and the last inequality holds due to (E.2). Thus, we complete the proof of Lemma D.1. \square

E.2 Proof of Lemma D.2

Proof of Lemma D.2. We prove this inequality based on two situations: $x \geq 1$ or $x < 1$. For the case that $x \geq 1$, we have

$$(1+x) \cdot |\log x| = (1+x) \log x \geq (1+x) \left(1 - \frac{1}{x}\right) \geq x \left(1 - \frac{1}{x}\right) = x - 1,$$

where the first inequality holds due to $\log x \geq 1 - 1/x$ and the second inequality holds due to $x \geq 1$. We now consider the case where $x < 1$. In this case, we have

$$(1+x) \cdot |\log x| = -(1+x) \log x \geq -(1+x)(x-1) \geq -(x-1),$$

where the first inequality holds due to $\log x \leq x-1$ and the second inequality holds due to $x < 1$. Combining the results in these two cases, we complete the proof of Lemma D.2. \square

E.3 Proof of Lemma D.3

In this section, we provide the proof of Lemma D.3, and we summarize our definitions and assumptions as follows. Similar assumptions are used in Rosset et al. (2024) to analyze the statistical error from the reparameterized reward.

Definition E.2 (Feasible Policy Class). For each iteration $t \in [T]$, let $\Pi_t \subseteq \Pi$ represent the feasible policy class, which includes all potential policies π_t generated by Algorithm 1 during iteration t , based on various possible samplings from the data collection process.

Assumption E.3 (Realizability). For each iteration $t \in [T]$ and each potential policy $\pi \in \Pi_t$, we assume that the following updated policy also belongs to the policy class Π :

$$\widehat{\pi}(\cdot|x) = \frac{1}{Z_\pi(x)} \cdot \pi(\cdot|x) \cdot \exp\left(\frac{r_\pi^*(x, \cdot)}{\beta}\right) \in \Pi,$$

where $r_\pi^*(x, y) = \mathbb{E}_{y' \sim \pi}[\mathbb{P}(y \succ y'|x)]$ denotes the win probability against the behavior policy π and $Z_\pi(x) = \sum_{y \in \mathcal{Y}} \pi(y|x) \exp(r_\pi^*(x, y)/\beta)$ is the partition function.

Assumption E.4 (Boundedness). For each iteration $t \in [T]$ and each potential policy $\pi_t \in \Pi_t$, we have

$$\beta \log \frac{\widehat{\pi}_{t+1}(y|x)}{\pi_t(y|x)} \in [-R, R]$$

holds for all $x \in \mathcal{X}$, $u \in \mathcal{Y}$ and any potential collected data set \mathcal{D}_t .

Based on these assumption, we now begin the proof of Lemma D.3

Proof of Lemma D.3. For each policy $\pi \in \Pi_t$ and the corresponding reparameterized reward $r_\pi(x, y) = \beta \log(\pi(y|x)/\pi_t(y|x))$, we have

$$\begin{aligned} & \mathbb{E}_{x \sim \rho, y \sim \pi_t(\cdot|x)} \left[r_\pi(x, y) - r_t^*(x, y) + \log Z_{\pi_t}(x) \right]^2 \\ &= \mathbb{E}_{x \sim \rho, y \sim \pi_t(\cdot|x)} \left[r_\pi(x, y) - \mathbb{E}_{y' \sim \pi_t}[\mathbb{P}(y \succ y'|x)] + \log Z_{\pi_t}(x) \right]^2 \\ &= \mathbb{E}_{x \sim \rho, y, y' \sim \pi_t(\cdot|x)} \left[r_\pi(x, y) - \mathbb{1}(y \succ y'|x) + \log Z_{\pi_t}(x) \right]^2 + C_{\pi_t} \\ &= \mathbb{E}_{x \sim \rho, y, y' \sim \pi_t(\cdot|x)} \left[r_\pi(x, y) - \mathbb{1}(y \succ y'|x) + \log Z_{\pi_t}(x) \right]^2 / 2 \\ &\quad + \mathbb{E}_{x \sim \rho, y, y' \sim \pi_t(\cdot|x)} \left[r_\pi(x, y') - \mathbb{1}(y' \succ y|x) + \log Z_{\pi_t}(x) \right]^2 / 2 + C_{\pi_t} \\ &= \mathbb{E}_{x \sim \rho, y^w, y^l \sim \pi_t} \left[\ell(r_\pi, x, y^w, y^l, \pi_t) \right] + C_{\pi_t}, \end{aligned}$$

where $Z_{\pi_t}(x) = \sum_{y \in \mathcal{Y}} \pi_t(y|x) \exp(r_t^*(x, y)/\beta)$ is the partition function, $C_{\pi_t} = \mathbb{E}_{x \sim \rho, y, y' \sim \pi_t(\cdot|x)} [\mathbb{P}(y \succ y'|x) - \mathbb{1}(y \succ y'|x)]^2$ is the variance of behavior policy π_t and the third equation holds because y, y' collected under the same behavior policy. Conditioned on Assumption E.3, there exists a policy $\pi_{t+1}^* \in \Pi$ such that $r_{\pi_{t+1}^*}(x, y) = \mathbb{E}_{y' \sim \pi_t}[\mathbb{P}(y \succ y'|x)] - \log Z_{\pi_t}(x)$. Under this situation, according to the standard concentration argument based on Bernstein's inequality, the least-squares estimator $\widehat{\pi}_{t+1}$ satisfies

$$\mathbb{E}_{x \sim \rho, y \sim \pi_t(\cdot|x)} \left[r_{\widehat{\pi}_{t+1}}(x, y) - r_{\pi_{t+1}^*}(x, y) \right]^2 \leq O\left(\frac{(R+1)^2 \cdot \log(|\Pi|/\delta)}{N}\right), \quad (\text{E.3})$$

where the inequality holds due to Lemma 6 in Xu and Zeevi (2020) with the Assumption E.4. Therefore, we have

$$\begin{aligned} & \mathbb{E}_{x \sim \rho, (y_1, y_2) \sim \pi_t} \left[\left(r_t^*(x, y_1) - r_t^*(x, y_2) - r_t(x, y_1) + r_t(x, y_2) \right)^2 \right] \\ &= \mathbb{E}_{x \sim \rho, (y_1, y_2) \sim \pi_t} \left[\left(r_{\pi_{t+1}^*}(x, y_1) - r_{\pi_{t+1}^*}(x, y_2) - r_t(x, y_1) + r_t(x, y_2) \right)^2 \right] \\ &\leq 2\mathbb{E}_{x \sim \rho, y_1 \sim \pi_t} \left[r_{\pi_{t+1}^*}(x, y_1) - r_t(x, y_1) \right]^2 + 2\mathbb{E}_{x \sim \rho, y_2 \sim \pi_t} \left[r_{\pi_{t+1}^*}(x, y_2) - r_t(x, y_2) \right]^2 \\ &\leq O\left(\frac{(R+1)^2 \cdot \log(|\Pi|/\delta)}{N}\right), \end{aligned}$$

where the first inequality holds due to $(x+y)^2 \leq 2x^2 + 2y^2$ and the last inequality holds due to (E.3). Thus, we complete the proof of Lemma D.3. \square

F Additional Details in Experiments

F.1 Discussion

Effect of Generation Length. All iterative DPO training baselines, including Snorkel's Mistral-PairRM-DPO and our DPO Iter3, exhibit a trend of increasing response length. This

is inevitable because all response pairs used for training are generated online and ranked by PairRM, where longer sequences are more likely to be chosen as winners. This effect accumulates as the model trains on its own generation. We discuss this effect from the following perspectives: First, both humans and GPT-4 exhibit length bias, and our goal is not to prevent any length growth but to evaluate it from a perspective of fairness. Therefore, we primarily refer to the LC win rate¹, which reflects the predicted win rate when response lengths are at the same level as the baseline. Second, we note that APO Iter2 achieves a significantly higher LC win rate and overall win rate compared to Snorkel’s iterative DPO implementation, while generating much shorter responses (2636 characters compared to 2736). When compared to our DPO Iter3, APO Iter2 achieves a comparable win rate, but also with shorter response lengths. This verifies a faster convergence rate and enhanced performance of our methodology. Finally, this length growth is not inherent to our method but is a result of the online training and PairRM ranking mechanism. It can be mitigated if we query winner/loser pairs and preference choices from GPT-4 or human, with much higher cost.

Effect of Coefficient α . In Table 2, we explore how different choices of α might affect performance. At iteration 1, performance fluctuates when changing α , peaking at $\alpha = 0.3$ with a win rate of 28.23%. Lower and higher values of α yield less stable results, indicating that extreme values may either underutilize or overcompensate the momentum effect. At iteration 2, the model demonstrates performance improvements at both $\alpha = 0.3$ and $\alpha = 0.6$, with the highest win rate of 30.62% observed at $\alpha = 0.6$. Although $\alpha = 0.3$ continues to provide a strong performance, it is surpassed by the higher $\alpha = 0.6$, suggesting that the optimal α might increase with iterations as the model stabilizes. Nevertheless, we continue to use $\alpha = 0.3$ at iteration 2, as our theoretical analysis assumes α as a constant. While an adaptive approach to setting α might offer benefits, $\alpha = 0.3$ has consistently shown to be the most stable and effective choice across different iterations.

Table 2: The effect of different extrapolation strategy and parameters at iteration 1 and 2. Results are reported in terms of length-controlled win rate (%). The row labeled "Start from $\hat{\pi}_0$ " represents results at $t = 1$ (π_1). The row labeled "Start from $\hat{\pi}_1$ " corresponds to results at $t = 2$ (π_2). Additionally, the row labeled "Start from π_1 " shows results when the momentum vector is calculated using $\hat{\pi}_2 - \pi_1$ instead of $\hat{\pi}_2 - \hat{\pi}_1$, which is not covered by our theoretical guarantee.

Starting From	$\alpha = 0.1$	$\alpha = 0.2$	$\alpha = 0.3$	$\alpha = 0.4$	$\alpha = 0.5$	$\alpha = 0.6$
$\hat{\pi}_0$	27.41	25.30	28.23	24.40	25.52	26.57
$\hat{\pi}_1$	28.11	29.15	29.73	28.70	29.33	30.62
π_1	28.23	28.16	27.21	28.44	29.75	28.90

Effect of Extrapolation Strategy. When the momentum vector is calculated using $\hat{\pi}_2 - \pi_1$ instead of $\hat{\pi}_2 - \hat{\pi}_1$, the performance appears less stable, with the lowest LC win rate being 27.21. Despite achieving competitive win rates at certain α values, the results are generally lower than those using the standard Nesterov’s momentum scheme. Intuitively, $\hat{\pi}_2 - \hat{\pi}_1 = (\hat{\pi}_2 - \pi_1) + (\pi_1 - \hat{\pi}_1)$ includes accumulated information, acting as an exponential average of all previous updates. In contrast, if extrapolation is based on π_1 , the direction’s variance is significantly higher, leading to less predictable outcomes. This instability underscores the importance of adhering to the theoretically guaranteed strategy for consistent performance gains.

F.2 Training Pipeline

We now turn to our training pipeline. In Algorithm 1, we begin by setting $\pi_0 = \hat{\pi}_0 = \pi_{\text{ref}}$ as the base model, Mistral-7B-Instruct-v0.2. At each iteration, we sample five pairs of responses under the current policy π_t and rank them using their PairRM score (Jiang et al., 2023a). We designate the top-ranked response as the winner and the bottom-ranked response as the loser. Following the proximal point update described in equation (6), we proceed with a momentum extrapolation. We note that when the policy is a softmax linear function, update (3.6) reduces to a momentum extrapolation in the parameter space. Consequently, we carry out an extrapolation in the parameter space to approximate the corresponding momentum step (3.6) in the probability space.

F.3 Hyperparameters

Our training is conducted on eight A100 GPUs, setting a global batch size of 64. This is done through setting a local batch size of 8 across all GPUs, with a gradient accumulation step of 1. We utilize

Table 3: **MT-Bench Evaluation.** The MT-Bench comprises a total of 8 tasks. We present the average score for all 8 tasks, as well as the average score across 5 dimensions relevant to our training dataset: Writing, Roleplay, Extraction, STEM, and Humanities. Each task includes 2 progressive turns, and we provide the average score for the first turn, the second turn, and the overall average of both turns.

Model	MT-Bench			Average of Five Tasks		
	1st Turn	2nd Turn	Average	1st Turn	2nd Turn	Average
Mistral-7B-Instruct-v0.2	8.08	7.20	7.64	9.42	8.86	9.14
DPO Iter1	7.98	7.29	7.63	9.56	9.06	9.31
DPO Iter2	7.80	7.26	7.52	9.36	8.98	9.17
DPO Iter3	7.61	7.25	7.43	9.23	9.04	9.14
APO Iter1	7.98	7.29	7.63	9.56	9.06	9.31
APO Iter2	7.95	7.26	7.60	9.56	9.23	9.40
APO Iter3	7.72	7.33	7.53	9.69	9.44	9.57

the RMSProp optimizer for each iteration, tuning the learning rate and β parameters. Although we initially considered a range of learning rates from $[5 \times 10^{-7}, 1 \times 10^{-7}, 5 \times 10^{-8}]$, we found that a fixed learning rate of 5×10^{-7} while adjusting β across iterations results in enhanced performance. The final β value is set at 0.01 for the first iteration, 0.1 for the second iteration, and 1.0 for the third iteration. The extrapolation parameter α is set to be a constant 0.3. In addition, the learning rate follows a linear schedule with a warm-up ratio of 0.1. For inference, we employ the `vllm` library (Kwon et al., 2023) for response generation, configured with a temperature of 0.7 and a `top_p` of 0.9. The maximum token length for response generation is set to 2048 tokens. Finally, evaluations across all benchmarks are conducted using eight A6000 GPUs.

F.4 MT-Bench Evaluation

MT-Bench: MT-Bench comprises of 80 multi-turn questions across eight distinct dimensions: Writing, Roleplay, Reasoning, Math, Coding, Extraction, STEM, and Humanities. The evaluation process prompts the GPT-4 judge to assign scores from 1 to 10 to the generated responses. However, this scoring metric has limitations. Particularly, score saturation occurs quickly in certain dimensions, making it difficult to distinguish between models. Additionally, this can disadvantage models that perform exceptionally well in these saturated dimensions, as their distinctiveness is less recognizable.

MT-Bench Evaluation. In addition, we also evaluate our model using the MT-Bench tasks. Among the eight tasks, three—math, coding, and reasoning—are evaluated by a GPT-4 math judge, while the other five—writing, roleplay, extraction, STEM, and humanities—are evaluated by a GPT-4 general judge. Table 3 illustrates the results of iterative DPO and APO across all three iterations. Since the UltraFeedback dataset is not designed for math/code tasks, and our model has never seen correct answers, the ability to solve math questions is not directly related to our training objective. Therefore, we also provide an average of scores for the five tasks focusing on general instruction-following abilities. From Table 3, we see that APO exhibits superior performance on the five general tasks, showing consistent improvement with each iteration, with average scores progressing from 9.31 to 9.40 to 9.57. In contrast, both the base model and the iteratively trained DPO model maintain an average score of around 9.14. Moreover, even when math, coding and reasoning are included, the performance of APO in the second turn surpasses that of other models. However, because it is not specifically trained on math problems, the overall average score is slightly reduced. This decline in average is primarily due to a saturation in the tasks where our method excels (achieving 9.57 out of 10), with limited room for further improvement in these areas. Consequently, the tasks where our model was not trained on become more detrimental when calculating an average of eight.

F.5 Open LLM Leaderboard Evaluation

Open LLM Leaderboard: The Open LLM Leaderboard evaluates models across six different tasks, focusing on a variety of language modeling capabilities. Each task consists of multiple-choice questions (MCQs), with the most probable choice being selected and compared against the correct answer. The final score is calculated based on accuracy. The ability for answering MCQs are not directly related with instruction-following abilities.

Table 4: **Open LLM Leaderboard Evaluation.** Comparison of APO with state-of-the-art iterative training algorithms, results are reported using v0.4.1 of the `lm-evaluation-harness` library. Tasks and parameters follows [Beeching et al. \(2023\)](#).

Models	Arc	TruthfulQA	WinoGrande	GSM8k	HellaSwag	MMLU	Avg.	Avg. 5
Mistral-7B-Instruct-v0.2	63.74	66.81	77.90	42.00	84.77	59.12	65.72	70.47
DPO Iter1	65.36	69.19	77.11	36.92	85.1	58.89	65.43	71.13
DPO Iter2	66.47	69.93	78.14	42.46	85.44	58.29	66.79	71.65
DPO Iter3	66.72	68.07	77.51	41.47	85.74	58.85	66.39	71.37
APO Iter1	65.36	69.19	77.11	36.92	85.1	58.89	65.43	71.13
APO Iter2	66.98	70.98	77.51	42.61	85.32	58.69	67.02	71.90
APO Iter3	65.78	69.75	77.03	37.83	85.43	58.76	65.76	71.35

Open LLM Leaderboard Evaluation: We provide the result of Open LLM Leaderboard. We calculate the average scores across six tasks, as well as the average of five tasks excluding GSM8k. This exclusion is due to our training’s emphasis on general instruction-following ability, with no exposure to correct answers for math problems. Table 4 shows that APO demonstrates robust performance across multiple tasks. Notably, at iteration 2, APO achieves the highest overall average score of 67.02, outperforming the baseline models. When focusing on the average of five key tasks, excluding GSM8k, APO iteration 2 again leads with an impressive score of 71.90. Although APO iteration 3 shows a slight decrease in the overall average score, it maintains a competitive edge. This decline is primarily attributed to the low correlation of multiple-choice question tasks with instruction-following abilities. we note that all trained baselines exhibit similar performance suboptimality on this metric, while the AlpacaEval performance increases.

F.6 Response Examples in Different Iterations

Table 5: Generation example of our fine-tuned model by APO at different iterations.

Prompt	Hi, my sister and her girlfriends want me to play kickball with them. Can you explain how the game is played, so they don't take advantage of me?
Iteration 1	<p>Hello! I'd be happy to help explain the rules of kickball to ensure that you feel confident and prepared for your game with your sister and her girlfriends. Kickball is a fun, social, and inclusive game that combines elements of baseball and soccer. Here's a breakdown of the essential rules:</p> <ol style="list-style-type: none"> 1. Teams and Setup: - Divide into two teams, ideally of equal numbers. - Designate a field with a home base, a first base, a second base, a third base, and a kickball diamond. - Choose a team to be in the field (defense) and one to be at bat (offense). 2. Equipment: - A standard soccer ball (or a large, soft kickball) is used. - No specialized equipment is required for players, except for comfortable clothing and athletic shoes. 3. Objective: - The objective of the game is for the team at bat to score runs by kicking the ball and advancing around the bases, while the fielding team tries to prevent this by getting the runners out. 4. Batting: - The game begins with a pitch (roll) from the pitcher (field player) to the kicker (batter). - The kicker attempts to kick the ball with their foot (no toe-kicks or using hands are allowed). - After a valid kick, the runner(s) may advance around the bases, trying to reach the next base before the ball reaches that base or a defensive player tags them. 5. Fielding: - The fielding team's goal is to get the runners out. - Players can attempt to catch a kicked ball in mid-air to retire the side (three consecutive outs), or they can touch a base with the ball before a runner reaches it to force them out. 6. Scoring: - A run is scored when a player advances around all four bases and returns to home plate. - Each inning is typically limited to a specific number of outs or runs scored (agree on this beforehand). 7. Fair Play: - Encourage sportsmanship and ensure all players understand and respect the rules to maintain a fun and enjoyable experience for everyone involved. 8. Additional rules: - Discuss any specific house rules or modifications to the game before starting, such as allowing multiple kicks per at-bat or implementing a mercy rule if one team is significantly ahead in runs. <p>I hope this information is helpful! If you have any further questions or need clarification on any rules, feel free to ask. Good luck and have a great time playing kickball with your sister and her girlfriends!</p>

Table 6: Generation example of our fine-tuned model by APO at different iterations.

Prompt	Hi, my sister and her girlfriends want me to play kickball with them. Can you explain how the game is played, so they don't take advantage of me?
Iteration 2	<p>Hello! It's great that your sister and her friends have invited you to join them for a game of kickball. Kickball is a fun, social, and inclusive team sport that combines elements of baseball and soccer. Here's a basic explanation of how the game is played to ensure you feel prepared and confident on the field. Keep in mind that variations may exist depending on the specific rules your group chooses to follow, but this outline should provide a solid foundation.</p> <ol style="list-style-type: none"> Teams and Setup: Divide into two even teams, ideally with similar numbers of players (typically 8-12 per team). Designate one team as the home team (they'll be up to bat first), and the other as the visiting team. Set up a diamond-shaped kicking area (the "kickline") approximately 60 feet from a base diamond, similar to baseball. Mark out four bases (first, second, third, and home) in a square shape around the diamond, with home plate located at the end of the kickline. Objective: The objective of kickball is to score runs by hitting the ball with your foot and advancing around the bases before returning to home plate safely. The opposing team aims to get outs by fielding the ball and tagging runners with the ball or touching them with it while they're not on a base. Batting: The batter stands at home plate and kicks the ball when it's pitched towards them by the pitcher from the opposing team. A runner may advance from their base at the start of each pitch, but they must return to that base before the ball is touched by a fielder or thrown to that base. If a runner advances too far before returning to their base, they risk being tagged out in a force play. Fielding: Fielders position themselves around the diamond and outfield to catch fly balls or field grounders. They can also attempt to tag runners with the ball while they're advancing between bases or when they reach a base. Once a fielder catches a fly ball or tags a runner, they record an out. A maximum of three outs per inning is typical. Scoring: Runners score by advancing around the bases and returning safely to home plate. A run is scored when a player reaches home plate after kicking the ball and successfully advances around all four bases without being tagged out or reaching base before a preceding runner is retired (resulting in that runner being forced out). Inning Structure: Each team gets an opportunity to bat in each half-inning. The inning ends when three outs are recorded by the fielding team. Typically, games consist of multiple innings until one team reaches a predetermined number of runs (e.g., five or ten). Fair Play: Ensure that everyone follows the rules and plays in good spirits. Kickball is meant to be an enjoyable experience for all involved, so focus on having fun while staying mindful of safety and fairness. If you feel unsure about any aspect of the game or suspect your sister and her friends might be taking advantage of you, communicate openly with them about your concerns and work together to find solutions that maintain a balanced and enjoyable playing field for everyone.

Table 7: Generation example of our fine-tuned model by APO at different iterations.

Prompt	Hi, my sister and her girlfriends want me to play kickball with them. Can you explain how the game is played, so they don't take advantage of me?
Iteration 3	<p>Hello! It's great that your sister and her friends have invited you to join their kickball game. Kickball is a fun, social, and inclusive sport, well-loved for its simplicity and energy. To ensure you feel confident and prepared for the game, here's a basic explanation of how kickball rules typically work. Keep in mind that variations exist, so always clarify any specific rules with your group before playing.</p> <ol style="list-style-type: none"> Teams and Setup: Divide into two teams, ideally through an evenly distributed random draw or team selection process. Designate a field with clear boundaries using cones or markers. A diamond shape or rectangle works well. Each team defends a designated home base (typically marked by a large cone or object). Equipment: A standard soccer ball is commonly used for kickball games. No specialized equipment is required beyond that, although gloves may be worn for comfort or to improve grip on the ball. Objective: The objective of the game is to score points by kicking the ball and successfully advancing around the bases (first base, second base, third base, and home plate) before returning to home base. A run is scored when a player reaches home base after completing a full circuit around the bases. Starting the Game: A coin toss determines which team bats first. The fielding team positions themselves around the bases and in the outfield, ready to receive the kicked ball. The batting team gathers near home plate with a designated kicker ready to start the inning. Kicking and Running: The kicker takes a stance and kicks the ball forcefully with their foot, aiming for a solid contact to launch the ball as far as possible. Runners on base attempt to advance based on where the ball lands and how the fielding team reacts. Fielding Rules: Fielders must retrieve or touch the ball before a runner advances to the next base. Once a fielder has the ball, they can tag runners with it to record outs. However, runners can advance on a missed catch or an overthrow. If a fielder touches a base before a runner arrives, the runner is out. Scoring: A run is scored when a batter safely kicks the ball and reaches first base, while teammates advance around the bases and return to home plate before the inning ends (three outs). Inning Structure: Each team takes turns at-bat and in the field. Typically, an inning lasts until the fielding team records three outs, at which point the teams switch roles. A game consists of multiple innings, with the team with the most runs at the end declared the winner. Fair Play and Sportsmanship: Encourage a friendly and respectful environment, following the spirit of the game. Ensure all players understand the rules, and focus on having fun rather than taking advantage of one another. If needed, consult your group to establish additional guidelines or modifications to accommodate skill levels and preferences. Additional Tips: <ul style="list-style-type: none"> - Wear comfortable clothing and shoes appropriate for running and kicking. - Stay hydrated and take breaks as needed. - Communicate clearly with teammates and opponents. - Focus on improving your skills, both as a kicker and a runner, to contribute effectively to your team. <p>With this foundational knowledge, you should feel more confident joining your sister and her friends for a kickball game. Enjoy the experience and have fun!</p>

Table 8: Another generation example of our fine-tuned model by APO at different iterations.

Prompt	How did US states get their names?
Iteration 1	<p>The naming of US states reflects a rich and varied history, influenced by explorers, Native American cultures, natural features, and political considerations. Here are some common ways US states got their names:</p> <ol style="list-style-type: none"> Names derived from Native American words: Many states' names come from Native American languages or place names. For example, California comes from the Spanish rendition of the Native American word "Quercus californius," meaning "californio," or "californian," which referred to the native people or the abundant california trees. Florida's name comes from the Spanish word "Pascuan Floridus," meaning "flowery land," a term used by explorer Juan Ponce de León, likely inspired by the abundant natural beauty he encountered. Names inspired by explorers and European settlers: Some states are named after famous explorers, settlers, or colonial founders. For instance, Virginia was named after Queen Elizabeth I of England, who was known as the "Virgin Queen" due to her status as an unmarried monarch. Georgia was named after King George II of Great Britain in honor of his support for the founding of a colony there. Names based on geographical features: Several states are named after prominent geographical features, such as mountains, rivers, or bodies of water. For example, Colorado's name comes from the Spanish word "colorado," meaning "red," a term often used to describe the reddish hue of the area's soil and water. Montana's name means "mountainous" in Latin. Names reflecting historical events or early settlement patterns: Some states' names are connected to significant historical events or early settlement patterns. For instance, Maine was originally called "Norumbega" based on tales of a mythical northern land of abundance, but was later renamed by English explorer John Smith to reflect its northern location relative to Massachusetts. Texas was named by the Spanish as "Tejas," meaning "friends" or "allies," due to the friendly reception given to Spanish explorers by the local Caddo and Tonkawa tribes. Names honoring political or ideological affiliations: A few states' names reflect political or ideological affiliations. For example, Vermont was originally called "New Vermont" or "New England's Plymouth" due to its ties to the English county of Vermont and the Pilgrims of Plymouth Colony. Later, when it sought independence from New York, the name was adopted as a tribute to the revolutionary ideals of the Republic of Vermont. <p>Overall, US states' names offer a fascinating glimpse into the rich history of exploration, settlement, and political development that shaped the American nation.</p>

Table 9: Another generation example of our fine-tuned model by APO at different iterations.

Prompt	How did US states get their names?
Iteration 2	<p>US states have varied histories regarding how they received their names, with some derived from native place names or geographical features, others from explorers or colonial founding figures, and still others from natural elements or symbolic concepts. Here's a brief overview of how some common state names came to be: 1. Alabama: Derived from the Native American Chickasaw and Creek words "alياهو" or "allapaha," meaning "thicketed place" or "place with ticklenut trees." 2. Alaska: Russian explorer Vitus Bering, who first documented the region in 1741, named it "Alaska" from the Russian words "alas k0430044f," meaning "great land" or "large country." 3. Arizona: Named after the Arizona Mining District, which in turn was derived from the Spanish term "Arizonia," meaning "arid" or "dry," likely in reference to the desert climate. 4. Arkansas: Originally called the "Land of Hieroglyphics" by early European explorers due to the abundance of Native American petroglyphs found in the region. Later, it was named Arkansas after the Quapaw tribe's name for a village near present-day Osceola, which translates to "people of the south wind." 5. California: Spanish explorer Juan Rodriguez Cabrillo is credited with giving the name "California" to the region in 1542, possibly inspired by the ancient mythical island of California described by Greek geographer Ptolemy as lying offshore west of the Pacific Ocean, abundant with gold and paradisiacal conditions. 6. Colorado: Derived from the Spanish word "colorado," meaning "red," likely in reference to the red soil and red sandstone formations found in the region. 7. Connecticut: Named by Dutch explorer Adriaen Block in honor of the Mohegan and Montauk Sachems (tribal leaders) Huskera or Quinipet and Sinigtot or Sitenico, whom he met in 1614. The name Connecticut comes from the Mohegan word "quinnipiac," meaning "long tidal river." 8. Delaware: Named after Sir Thomas West, 3rd Baron De La Warr, an English nobleman who served as the first colonial governor in 1630. The name Delaware comes from the Delaware River, which was named after the Lenape Native American tribe, whose name means "upstream people" or "people of the cleared land." 9. Florida: Spanish explorer Juan Ponce de Le00f3n is credited with discovering Florida in 1513 and naming it after "Pascua Florida," meaning "feast of flowers," likely due to the abundant blooming vegetation he encountered during his expedition. 10. Georgia: Named after King George II of Great Britain in honor of his support for establishing a colony there in 1732. 11. Hawaii: The islands' indigenous Polynesian name, "Hawaii Loa," translates to "long home" or "homeland," reflecting the archipelago's unique identity as a distinct and unified cultural entity. The name Hawaii became widely known to Europeans through British explorer James Cook, who visited in 1778 and adopted it for the larger island of Hawaii in his charts. 12. Idaho: Derived from a Shoshone Native American word, "hee-da-ho," meaning "gem of the mountains" or "the sun comes from the waters." This name was popularized during the late 19th century when Idaho was being considered for statehood. 13. Illinois: Derived from a Native American word, "Illiniwek," meaning "tribe of superior men" or "tribe dwelling properly." This term referred to various tribes within the Illini Confederacy, which inhabited much of present-day Illinois prior to European contact. 14. Indiana: Named after the Native American term "Indian00e1," meaning "land of the Indians," reflecting that the region was inhabited by various Native American tribes when European settlers arrived. 15. Iowa: Derived from a Native American word, "ayoua," meaning "prairie" or "beautiful land." This name was popularized during the late 19th century when Iowa was being considered for statehood. These examples represent just a fraction of US states and their unique origin stories. Many other states have names rooted in Native American languages, European explorers, geographical features, or symbolic concepts.</p>

Table 10: Another generation example of our fine-tuned model by APO at different iterations.

Prompt	How did US states get their names?
Iteration 3	<p>US states have varied histories regarding how they earned their names. Here are some common ways and specific examples:</p> <ol style="list-style-type: none"> Indigenous Place Names: Many state names originate from Native American languages spoken by tribes inhabiting the regions prior to European exploration. For instance: <ul style="list-style-type: none"> - Florida: Derived from the Native American word "Pyscaleola" or "Scal scal lla," meaning "flowing water" or "land of flowers," depending on the specific indigenous tribe. European explorer Juan Ponce de Leon may have chosen this name based on reports of native peoples and their designations for the land. - Missouri: Derived from the Missouri River, which in turn took its name from the Miwoksi or Missouri Nation, a Siouan-speaking people inhabiting the area at the time of French explorer Louis Jolliet and Jacques Marquette's passage in 1673. Their name translates roughly to "people of large canoes." European Explorers and Settlers: Many states were named after early European explorers, settlers, or monarchs. For example: <ul style="list-style-type: none"> - Virginia: Named after Queen Elizabeth I of England in honor of her reign and patronage, specifically her title "the Virgin Queen." Sir Walter Raleigh, who received a royal charter to establish colonies in North America, chose this name. - California: Named by Spanish explorer Vasco Núñez de Balboa in 1542, who gave it the Latinized name "California" after the Classical mythological figure Californias, who was believed to inhabit a paradisiacal island on the western edge of the known world. Natural Features: Some states bear names inspired by distinctive geographic or topographical features. For instance: <ul style="list-style-type: none"> - Colorado: Derived from the Spanish "colorado," meaning "red," a reference to the red soil and rock formations found throughout the region. - Montana: Named for the Latin word "montanus," meaning "mountainous." Historical Events: A few states have names connected to significant historical events. For example: <ul style="list-style-type: none"> - Texas: Derived from the Spanish "Tejas," meaning "friends" or "allies." The name referred to the indigenous peoples who initially welcomed Spanish explorers in the late 16th century. Later, Mexico used the name when it ruled the region, and after Texas gained independence in 1836, the name remained. - Wisconsin: Named by French explorer Jean Nicolet in 1634 after the native peoples he encountered, the "Wisconsin" or "Meskousing," an Algonquian-speaking people. The name "Wisconsin" is believed to mean "gathering place" or "place of the red people." Symbolic Meanings: In some cases, state names reflect symbolic or metaphorical concepts. For example: <ul style="list-style-type: none"> - Nevada: Derived from the Spanish "nevado," meaning "snow-covered mountain range," which aptly describes the state's mountainous terrain and its name's association with the silver rush that brought settlers to the area in the late 19th century (the silver being referred to as "white gold" or "silver snow"). <p>These examples illustrate just a few ways US states have acquired their names throughout history. The specific origins of each state name are complex and often intertwined with the rich tapestry of indigenous, European, and American cultural influences that have shaped the United States.</p>

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: We provide theoretical and experimental analysis to justify the main claims.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: Limitations are discussed in conclusion.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory Assumptions and Proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: We provide assumptions for each theorem and proof presented.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: For reproducibility, we provide details on our training pipeline, along with our hyperparameters.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We are not able to reorganize the code at the time of submission.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental Setting/Details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide experimental settings and details in our experimental section.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment Statistical Significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Error bars are not involved in our paper.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96 CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments Compute Resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: We describe the computing resource with experimental settings.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code Of Ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines?>

Answer: [Yes]

Justification: The research conducted in this paper conform with the NeurIPS Code of Ethics.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader Impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: We describe the compute resources in the Experiment Setup section.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper proposes a general methodology that is applicable to any large language model, and poses no particular such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite the original papers that produced the dataset and base models.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset’s creators.

13. **New Assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release new assets

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.