PROVABLY AVOIDING OVER-OPTIMIZATION IN DIRECT PREFERENCE OPTIMIZATION WITHOUT KNOWING THE DATA DISTRIBUTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce PEPO (*Pessimistic Ensemble based Preference Optimization*), a single-step Direct Preference Optimization (DPO)-like algorithm to mitigate the well-known over-optimization issue in preference learning without requiring knowledge of the data-generating distribution π_{data} or learning an explicit reward model. PEPO achieves pessimism via an ensemble of preference-optimized policies and then aggregates them through a worst-case construction that favors the agreement across models. In the tabular setting, PEPO achieves sample complexity guarantees depending only on a single-policy concentrability coefficient, thus avoiding the all-policy concentrability which affects the guarantees of algorithms prone to over-optimization, such as DPO. The theoretical findings are corroborated by convincing practical performance, while retaining the simplicity and the practicality of DPO-style training.

1 INTRODUCTION

Learning from binary preferences is a fundamental technique for the post-training of large language models (Christiano et al., 2017; Ouyang et al., 2022). Among existing approaches, *Direct Preference Optimization* (DPO) (Rafailov et al., 2023) has gained widespread adoption due to its strong empirical performance and simplicity of implementation: indeed, DPO elegantly bypasses the intermediate step of learning a reward function from the preference dataset.

Despite these advantages, DPO is known to suffer from *over-optimization*: as training progresses, improvements in optimizing the training loss do not necessarily translate into better generation quality, and may even degrade performance relative to the initial model (Gao et al., 2023; Park et al., 2024; Xu et al., 2024). Moreover, early stopping using the validation loss as a criterion is not reliable because it is only a loose proxy for what really matters, i.e. the quality of the responses that the model can generate in practice.

Several new DPO variants mitigate over-optimization via *pessimistic regularization*. Notably, approaches such as χ^2-regularized preference optimization (χ^2PO) (Huang et al., 2024) and DPO augmented with supervised fine-tuning (DPO+SFT) (Liu et al., 2024) achieve guarantees that depend only on a *single-policy* concentrability coefficient, which is the theoretical certificate of robustness to over-optimization (Huang et al., 2024). Indeed, these guarantees imply that the learned policy is competitive with any policy sufficiently covered by the data. In stark contrast, non-pessimistic approaches suffer from an *all-policy* concentrability coefficient, meaning that the output policy can be significantly suboptimal even compared to well-covered policies unless the dataset covers the whole policy class.

A notable merit of the guarantees in Huang et al. (2024); Liu et al. (2024) is that they apply to general function classes, even to non-convex ones in the case of χ^2PO. However, these guarantees rely on the assumption that we have access to the data-generating distribution π_{data}, or to a reference/base policy that is well-covered by π_{data}. In practice, this assumption is often violated, for instance, when smaller models are post-trained using preference data generated by larger, proprietary systems whose logits cannot be accessed when fine-tuning the smaller models. A common example is the UltraFeedback dataset (Cui et al., 2023) in which the responses are generated via a pool of LLMs, including proprietary ones such as GPT-4.

This gap motivates the following research question:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Open Question: Can we design a DPO-like algorithm (i.e., which avoids learning an explicit reward) that is provably robust to over-optimization without knowledge of the data-generating distribution π_{data}?

Our work answers this question affirmatively by introducing PEPO (*Pessimistic Ensemble based Preference Optimization*). The central idea behind PEPO is to resort to *pessimism via ensembles*: a technique that is as simple as it is unexplored in the context of LLM alignment, especially in the DPO setting.

Concretely, PEPO trains an ensemble of lightweight DPO-style policies on disjoint subsets of the preference dataset, implemented via low-rank adapters (Hu et al., 2022). Finally, at generation time, we use a certain notion of ensemble agreement to generate responses either via rejection sampling or approximating the (otherwise intractable) response-level normalization constant by its token-level counterpart.

Theoretical contribution Our main theoretical result shows that, in the tabular setting, PEPO achieves sample complexity guarantees depending only on a single-policy concentrability coefficient, despite making no assumptions about π_{data} and not learning a reward model. This contrasts with standard DPO (Rafailov et al., 2023), which inherently incurs an all-policy concentrability dependence, with existing pessimistic DPO variants that require access to the data-generating distribution (Huang et al., 2024; Liu et al., 2024), and with purely theoretical works which are not DPO-like because they require a reward modeling step (Zhu et al., 2023; Zhan et al., 2023; Schlaginhaufen et al., 2025; Li et al., 2023b). Moreover, we characterize the optimal PEPO ensemble size: it is the smallest value that is sufficiently high to guarantee pessimism.

Practical contribution Empirically, we corroborate our analysis, demonstrating that PEPO consistently improves over DPO in post-training the following 7/8B-parameter models: Zephyr-7B[1], Llama-3.1-8B[2], and Mistral-7B[3] and even at 34B scale (Yi-34B[4]), where PEPO still performs convincingly.

Paper Organization The next section clarifies the setting and the notation used in this work. Section 3 presents in greater detail our algorithm, i.e. PEPO. Then, Section 3.4 formally develops the guarantees for PEPO, and, finally Section 4 demonstrates its practical performance.

2 Preliminaries

In *learning from preferences*, we consider a prompt/state space $\mathcal{X}$ and a response/action space $\mathcal{A}$; the learning algorithm receives as input a starting model $\pi_{\text{ref}} : \mathcal{X} \to \Delta_{\mathcal{A}}$ which is usually a model obtained after a *Supervised Fine-Tuning* (SFT) routine, and a preference dataset of the form $\mathcal{D} = \{X_n, A_n^+, A_n^-\}_{n=1}^N$ where $X_n \in \mathcal{X}$, $(A_n^+, A_n^-) \in \mathcal{A} \times \mathcal{A}$ for all $n \in [N]$. In the dataset $\mathcal{D}$, we observe the initial prompts sampled from an initial distribution ν_0, i.e. $X_n \sim \nu_0$ for each $n \in [N]$. Intuitively, $\nu_0(x)$ is the probability with which a human presents prompt x to the language model. Moreover, for each X_n, $\mathcal{D}$ contains two responses A_n^+, A_n^- sampled from an *unknown* distribution π_{data}. Finally, the action A_n^+ has been judged the winner over A_n^- by a Bradley-Terry preference model defined as follows.

Definition 2.1. Bradley-Terry (BT) preference model Given a prompt $x \in \mathcal{X}$, two responses $(a, b) \in \mathcal{A} \times \mathcal{A}$ and access to the *ground truth* reward function $r^\star : \mathcal{X} \times \mathcal{A} \to [0, R_{\max}]$ for some $R_{\max} > 0$, the Bradley-Terry preference model computes the probability of a being a winner over b in response to the prompt x (event denoted as $a \succ b|x$) as $\mathbb{P}^{\text{true}}_{r^\star}(a \succ b|x) = \sigma(\Delta_{r^\star}(x, a, b))$ where $\Delta_{\mathfrak{f}}(x, a, b) = \mathfrak{f}(x, a) - \mathfrak{f}(x, b)$ for any function $\mathfrak{f} : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ and $\sigma(x) = \frac{1}{1+e^{-x}}$ denotes the sigmoid function.

Given the BT model above, the dataset is labeled by sampling $Z \sim \text{Bernoulli}(\mathbb{P}^{\text{true}}_{r^\star}(a \succ b|x))$ for each x, a, b appearing in the dataset. If $Z = 1$, a is declared as the winner; vice versa, if $Z = 0$, b wins the comparison.

[1] https://huggingface.co/alignment-handbook/zephyr-7b-sft-full
[2] https://huggingface.co/allenai/Llama-3.1-Tulu-3-8B-SFT
[3] https://huggingface.co/HuggingFaceH4/mistral-7b-sft-beta
[4] https://huggingface.co/01-ai/Yi-34B

The goal of the learning algorithm is to update the starting model π_{ref} to obtain a post-trained model π_{out} such that it approximately equals

$$\arg\max_{\pi\in\Pi} J_\beta(\pi) := \langle \pi, r^\star \rangle - \beta D_{\text{KL}}(\pi, \pi_{\text{ref}}) \tag{1}$$

where Π is the set of all mappings from prompts to distributions over responses $\Pi = \{\pi : \mathcal{X} \to \Delta_{\mathcal{A}}\}$. The inner product notation is defined for each policy $\pi \in \Pi$ and for any function $r : \mathcal{X} \times \mathcal{A} \to [0, R_{\max}]$ as $\langle \pi, r \rangle := \sum_{x,a\in\mathcal{X}\times\mathcal{A}} \nu_0(x)\pi(a|x)r(x,a)$ and the average (over prompts) D_{KL} is defined for any policy pair $\pi, \pi' \in \Pi \times \Pi$ as $D_{\text{KL}}(\pi, \pi') = \sum_{x,a\in\mathcal{X}\times\mathcal{A}} \nu_0(x)\pi(a|x) \log {}^{\pi(a|x)}/_{\pi'(a|x)}$.

The objective in equation 1 therefore prescribes to find a trade-off between two desiderata: (i) maximizing the *ground truth* reward $r^\star$ used by the Bradley-Terry preference model to label the dataset and (ii) minimizing the shift as measured by the D_{KL} divergence from the initial model π_{ref}. The value of $\beta \in (0, \infty)$ should be larger the more important it is to stay close to π_{ref}, so that π_{out} inherits the useful properties learned by π_{ref} from the data observed before entering the learning from preferences phase.

We now describe two fundamental ways to solve equation 1, which will serve as building blocks for our approach.

Reinforcement Learning from Human Feedback: `RLHF` In `RLHF`, the idea is to optimize equation 1 under two approximations. First, we approximate $r^\star$, which is unknown to the learner, with the maximum likelihood estimator, denoted by $\hat{r}$ and learned on the dataset $\mathcal{D}$ as follows:

$$\hat{r} = \arg\max_{r:\mathcal{X}\times\mathcal{A}\to[0,R_{\max}]} \sum_{n=1}^{N} \log \sigma(\Delta_r(X_n, A_n^+, A_n^-)). \tag{2}$$

Secondly, in equation 1, the prompt distribution ν_0 is unknown; therefore, it must be approximated using, for example, the states observed in the dataset. All in all, `RLHF` methods output the policy maximizing the following empirical objective over $\pi \in \Pi$,

$$\sum_{n=1}^{N} \sum_{a\in\mathcal{A}} \pi(a|X_n) \left(\hat{r}(X_n, a) - \beta \log \frac{\pi(a|X_n)}{\pi_{\text{ref}}(a|X_n)} \right). \tag{3}$$

`RLHF` has the merit of being the conceptual core of the large majority of, if not all, alignment algorithms. Sadly, it has the drawback of being split into two training steps represented by equations equation 2 and equation 3 respectively. This has been elegantly avoided by `DPO` (Rafailov et al., 2023) with a reparameterization trick explained next.

Direct Preference Optimization The core observation is that equation 1, for any reward function r, admits the following closed form solution

$$\pi_r(a|x) = \frac{\pi_{\text{ref}}(a|x) \exp(\frac{r(x,a)}{\beta})}{Z_r(x)},$$

where $Z_r(x) := \sum_b \pi_{\text{ref}}(b|x) \exp(\frac{r(x,b)}{\beta})$. Albeit incomputable due to the hardness of computing $Z_r(x)$, the closed-form expression for π_r allows us to write the reward function as

$$r(x, a) = \beta \log {}^{\pi_r(a|x)}/_{\pi_{\text{ref}}(a|x)} + \beta \log Z_r(x). \tag{4}$$

Then, replacing this expression in equation 2, we can ensure that $\pi_{\hat{r}}$ is in the solution set of the following single-step optimization problem, which is known as the `DPO` objective,

$$\pi_{\hat{r}} \in \arg\max_{\pi\in\Pi} J_{\texttt{DPO}}(\pi; \mathcal{D}) := \sum_{n=1}^{N} \log \sigma(\beta \Delta_{\log \pi/\pi_{\text{ref}}}(X_n, A_n^+, A_n^-)). \tag{5}$$

Importantly, we have used that for each reward r, $\beta\Delta_{\log \pi_r/\pi_{\text{ref}}}(x, a, b) := \beta \log \frac{\pi_r(a|x)}{\pi_{\text{ref}}(a|x)} - \beta \log \frac{\pi_r(b|x)}{\pi_{\text{ref}}(b|x)}$ equals $\Delta_r(x, a, b)$ because the normalization constant $Z_r(x)$ cancels out.

Algorithm 1 PEPO

Require number of ensemble members, i.e. L, exponential tilting $m^\ell : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ for each $\ell \in [L]$.
% Train time.
Create dataset partitions $\cup_{\ell \in [L]} \mathcal{D}^\ell = \mathcal{D}$.
for $\ell \in [L]$ **do**
    Run DPO on each subset

$$\tilde{\pi}^\ell \in \arg\max_{\pi \in \bar{\Pi}} \sum_{X,A,B \in \mathcal{D}^\ell} \log \sigma(\beta \Delta_{\log \pi / \pi_{\text{ref}}}(X, A, B))$$

end for
Define the output distribution

$$\pi_{\text{out}}(a|x) = \frac{\min_{\ell \in [L]} \tilde{\pi}^\ell(a|x) \exp\left(-m^\ell(x,a)\right)}{\sum_{a \in \mathcal{A}} \min_{\ell \in [L]} \tilde{\pi}^\ell(a|x) \exp\left(-m^\ell(x,a)\right)}$$

% Test time.
Require User prompt $x_{\text{test}} \in \mathcal{X}$
Invoke Algorithm 2 to sample $A \sim \pi_{\text{out}}(\cdot|x_{\text{test}})$.

Limitations of RLHF and DPO (i.e., the lack of pessimism): The missing ingredient above is that both methods maximize $\hat{r}$ (directly or indirectly) blindly, i.e., treating it as if it were a perfect estimate of the ground truth $r^\star$. However, in practical situations where N is finite, this is not the case: $\hat{r}$ might not be accurate, and taking into account its uncertainty becomes fundamentally important. As an example, think of a single-state, two-action case in which $\hat{r}(a_1) = \hat{r}(a_2) + \epsilon$ for some small ϵ but the uncertainty[5] of $\hat{r}(a_1)$ is much higher than that of $\hat{r}(a_2)$. Then, choosing a_2 is more reasonable, while RLHF and DPO, which purely maximize $\hat{r}$, would output the highly risky action a_1. In a more realistic example targeted to LLMs, DPO and RLHF might achieve high reward under $\hat{r}$ exploiting hallucinations of $\hat{r}$, i.e., actions that have in fact small reward under the ground truth $r^\star$.

To mitigate these issues, the next section presents our techniques to incorporate pessimism in DPO. In contrast to previous pessimistic DPO versions (Huang et al., 2024; Liu et al., 2024), we *assume no knowledge at all* of π_{data} whereas Huang et al. (2024) assumed that $\pi_{\text{data}} = \pi_{\text{ref}}$ and Liu et al. (2024) leverages access to a policy π_{base} such that $\left\| \pi_{\text{base}} / \pi_{\text{data}} \right\|_\infty$ is small.

3 The Algorithm

Our algorithm, PEPO, applies two important changes on top of DPO which will enable our guarantees featuring only the *single-policy* concentrability coefficient:

- We split the dataset in L subsets of equal size and we maximize L different instances of the standard DPO objective with respect to independent LoRA parameters over each subset.
- Finally, our theory motivates outputting a policy obtained aggregating the ensemble in a pessimistic manner. Unfortunately, computing such a policy is intractable but we present a rejection sampling method to sample exactly from it. We also show how to approximate such routine at token-level for more efficient generation.

We now move to a detailed overview of these two components whose interplay is key for the theoretical guarantees presented in the next section.

3.1 Creating the models ensemble

We build L independent ensemble members (L independent Low Rank Adapters in practice) by training them on disjoint[6] subsets of the preference dataset. Formally, we partition the data $\mathcal{D} = \{X_n, A_n^+, A_n^-\}_{n=1}^N$ in L subsets, i.e., $\mathcal{D} = \cup_{\ell=1}^L \mathcal{D}^\ell$, where we denote the data in the ℓth chunk as

[5] We keep uncertainty vague here for the sake of intuition.

[6] This assumption is needed for our worst-case theoretical analysis; however, in practice, we observe that overlapping datasets can lead to improved performance.

Algorithm 2 `Efficient Response-Level Sampling via Rejection Sampling`

Require Initial Prompt x, Numerator of π_{out}

$$f_{\text{out}}(x,a) = \min_{\ell\in[L]} \tilde{\pi}^{\ell}(a|x)e^{-m^{\ell}(x,a)}, \quad M(x) := \max_{a\in\mathcal{A}} e^{-\min_{\ell\in[L]} m^{\ell}(x,a)}$$

Require: Number of trials $N_{\text{RS}} = \left\lceil \frac{M(x)\log(1/\delta)}{\sum_{a\in\mathcal{A}} f_{\text{out}}(x,a)} \right\rceil$
for $n = 1, \ldots, N_{\text{RS}}$ **do**
  Sample $A^n \sim \pi_{\text{prop}}(\cdot|x)$.
  Compute $\alpha = \frac{f_{\text{out}}(x,A^n)}{M(x)\pi_{\text{prop}}(A^n|x)}$ % Notice that $\alpha \in [0,1]$.
  Sample $U \sim \text{Unif}([0,1])$.
  If $U \leq \alpha$, output A^n otherwise continue.
end for

$\mathcal{D}^{\ell} = \left\{X_{n,\ell}, A^{+}_{n,\ell}, A^{-}_{n,\ell}\right\}$. At this point, we compute $\tilde{\pi}^{\ell}$ for each $\ell \in [L]$ as the solution of the standard `DPO` objective

$$\tilde{\pi}^{\ell} \in \arg\max_{\pi\in\overline{\Pi}} J_{\text{DPO}}(\pi; \mathcal{D}^{\ell}), \tag{6}$$

where the maximization is carried out over the subset of Π defined as $\overline{\Pi} := \left\{\pi \in \Pi : \left|\beta\Delta_{\log\pi/\pi_{\text{ref}}}\right| \leq 2R_{\max}\right\}$ to ensure that the implicit rewards of the ensemble members remain on the same scale as the ground-truth reward.

3.2 Pessimism via Ensemble

We now aggregate the L models in an ensemble in a pessimistic manner. Towards this goal, recall that by the correspondence between policy and reward space given by equation 4, each implicit reward obtained as $r^{\ell}(x,a) := \beta(\log\frac{\tilde{\pi}^{\ell}(a|x)}{\pi_{\text{ref}}(a|x)} - \zeta^{\ell}(x))$ can be considered as an independent estimate of the ground truth reward for an appropriate choice of the *centering factor* $\zeta^{\ell}(x)$. In particular, assuming without loss of generality that $\sum_b \pi_{\text{data}}(b|x)r^{\star}(x,b) = 0$, we can choose the *centering factor* as $\zeta^{\ell}(x) = \frac{1}{N^{-\ell}(x)}\sum_{X,A\in\mathcal{D}\setminus\mathcal{D}^{\ell}} \log\frac{\tilde{\pi}^{\ell}(A|X)}{\pi_{\text{ref}}(A|X)}\mathbb{1}\{X = x\}$ where the sum runs over all the responses A, chosen and rejected, recorded for the prompt x in $\mathcal{D}\setminus\mathcal{D}^{\ell}$ (all of them are sampled from $\pi_{\text{data}}(\cdot|x)$), and $N^{-\ell}(x)$ denotes the number of such responses, with the convention that $0/0 = 0$. [7]

At this point, similarly to Coste et al. (2023), we aim at optimizing a notion of worst-case reward over the ensemble. In our case, we choose the worst-case reward as $r^{-}(x,a) = \min_{\ell\in[L]} r^{\ell}(x,a) - C(x,a)$ where $C(x,a)$ is an additional pessimistic penalty that injects further pessimism, which will turn out to be helpful in the analysis. Then, we define our output policy as the policy learned by the following *virtual* `RLHF` problem:

$$\pi_{\text{out}} = \arg\max_{\pi\in\Pi} \langle\pi, r^{-}\rangle - \beta D_{\text{KL}}(\pi, \pi_{\text{ref}}). \tag{7}$$

We dub the above problem virtual because the program in equation 7 never needs to be actually implemented in practice. Indeed, we will show next how to efficiently sample from π_{out} via rejection sampling, thus bypassing the need for computing π_{out} explicitly.

3.3 Efficient Sampling from π_{out}

The structure of our efficient sampling scheme leverages the closed form solution of equation 7 which is offered by the following lemma.

Lemma 3.1. *Let us choose $m^{\ell}(x,a) = \zeta^{\ell}(x) + C(x,a)/\beta$. Then, the policy π_{out} defined as the solution of the optimization problem in equation 7 satisfies for all $x, a \in \mathcal{X}\times\mathcal{A}$*

$$\pi_{\text{out}}(a|x) = \frac{\min_{\ell\in[L]} \tilde{\pi}^{\ell}(a|x)e^{-m^{\ell}(x,a)}}{\sum_{a\in\mathcal{A}} \min_{\ell\in[L]} \tilde{\pi}^{\ell}(a|x)e^{-m^{\ell}(x,a)}}.$$

[7] Centering leaves all reward *differences* within a model unchanged, $r^{\ell}(x,a) - r^{\ell}(x,b) = \beta\Delta_{\log\tilde{\pi}^{\ell}/\pi_{\text{ref}}}(x,a,b)$, while making the L estimates comparable with one another; moreover, whenever $N^{-\ell}(x) > 0$, $\left|r^{\ell}(x,a)\right| \leq 2R_{\max}$, since $r^{\ell}(x,a)$ is an average of the bounded quantities $\beta\Delta_{\log\tilde{\pi}^{\ell}/\pi_{\text{ref}}}(x,a,\cdot)$.

For the following discussion, let us denote the numerator of π_{out} as $f_{\text{out}}(x,a) := \min_{\ell\in[L]} \tilde{\pi}^\ell(a|x)e^{-m^\ell(x,a)}$, and recall that a stands here for a full response, not simply a token. Therefore, the computation of the denominator, which involves a sum over the response space $\mathcal{A}$, is not tractable. However, rejection sampling (see Algorithm 2) is a very attractive solution in our scenario. Indeed, any model in the ensemble can serve as an ideal proposal distribution since it upper bounds $f_{\text{out}}(x,a)$ everywhere. To see this, notice that $\min_{\ell\in[L]} \tilde{\pi}^\ell(a|x) \leq \tilde{\pi}^{\ell'}(a|x)$ for any ℓ'. These facts ensure that for all $x, a \in \mathcal{X}\times\mathcal{A}$ and for any model $\tilde{\pi}^\ell$ the ratio $f_{\text{out}}(x,a)/\tilde{\pi}^\ell(a|x) \leq M(x)$ for $M(x) := \max_{a\in\mathcal{A}} e^{-\min_{\ell\in[L]} m^\ell(x,a)}$. Formally, the following lemma shows that with high probability Algorithm 2 will output a valid sample from π_{out} in polynomial time.

Lemma 3.2. *Let us consider an arbitrary prompt $x \in \mathcal{X}$ and recall that $f_{\text{out}}(x,a)$ satisfies that $\pi_{\text{out}}(a|x) = f_{\text{out}}(x,a)/\sum_{a\in\mathcal{A}} f_{\text{out}}(x,a)$. Then, Algorithm 2, with access to a proposal distribution π_{prop} such that $f_{\text{out}}(x,a)/(M(x)\pi_{\text{prop}}(a|x)) \leq 1$, outputs a valid sample from $\pi_{\text{out}}(\cdot|x)$ after sampling at most $N_{\text{RS}}(x) = \left\lceil \frac{M(x)\log(1/\delta)}{\sum_{a\in\mathcal{A}} f_{\text{out}}(x,a)} \right\rceil$ times from $\pi_{\text{prop}}(\cdot|x)$ with probability $1-\delta$.*

An interesting remark is that the number of trials $N_{\text{RS}}(x)$ depends on the inverse of the *ensemble agreement* measured via $\sum_{a\in\mathcal{A}} f_{\text{out}}(x,a)$. In words, the rejection sampling routine is expected to take longer to respond to prompts for which the ensemble members do not put a reasonably large probability on a common answer.

Discussion of the hyperparameter In our practical implementation, L is treated as a hyperparameter. Its value is useful to inject the right amount of pessimism. We present an ablation on L in Appendix C, while the next section provides theoretical guidance for the choice of L together with an end-to-end theoretical analysis of PEPO, whose pseudocode is given in Algorithm 1.

3.4 Theoretical Guarantees

We present here the theoretical guarantees for PEPO focusing on the tabular setting where the number of prompts and responses is finite. Our theoretical goal is to quantify the required dataset size N, so that with high probability $J_\beta(\pi^\star) - J_\beta(\pi_{\text{out}}) \leq \varepsilon$ for the soft optimal policy $\pi^\star = \arg\max_{\pi\in\Pi} J_\beta(\pi)$. In words, we want the suboptimality of the policy learned by PEPO w.r.t. the comparator $\pi^\star$ to be at most ε. In particular, as motivated by Huang et al. (2024); Liu et al. (2024), the gold standard for arguing that over-optimization is avoided is to prove a bound on N scaling only with a *single-policy concentrability* coefficient defined as

$$C^\star := \sum_{x\in\mathcal{X}} \nu_0(x) \sum_{a\in\mathcal{A}} \frac{(\pi^\star(a|x))^2}{\pi_{\text{data}}(a|x)}.$$

In stark contrast, methods prone to over-optimization, such as DPO, incur an *all-policy concentrability* coefficient (Song et al., 2024) defined as $C^{\text{all}} := \max_{\pi\in\Pi} \sum_{x\in\mathcal{X}} \nu_0(x) \sum_{a\in\mathcal{A}} \frac{(\pi(a|x))^2}{\pi_{\text{data}}(a|x)}$. We highlight that $C^\star$ and C^{all} have qualitatively different behavior. On the one hand, $C^\star$ is reasonably small as soon as sampling from π_{data} allows to observe with high enough probability all the actions which are often chosen by $\pi^\star$. On the other hand, C^{all} essentially requires observing all possible responses in the dataset with positive probability[8], even those which would never be taken by $\pi^\star$. For our fast rate, we will also need a stronger notion of single policy concentrability coefficient, sometimes known as L_∞ concentrability coefficient, defined as

$$C^\star_\infty = \max_{x,a} \frac{\pi^\star(a|x)}{\pi_{\text{data}}(a|x)}.$$

Clearly, it holds that $C^\star \leq C^\star_\infty$. We are now ready to present our main result: the sample complexity guarantees for PEPO reported in Theorem 3.3, which depend only on $C^\star$ and not on C^{all}.

Theorem 3.3. *Let us consider running PEPO (Algorithm 1) with $L = \left\lceil \frac{7\log(|\mathcal{X}||\mathcal{A}|^2/\delta)}{2} \right\rceil$ many ensemble members and a preference dataset $\mathcal{D}$ of size N and for $C(x,a) = \widetilde{\mathcal{O}}\left(\sqrt{\frac{|\mathcal{A}|^3 e^{2R_{\max}}(\log(1/\delta))^2}{N(x,a)+|\mathcal{A}|}}\right)$ (see Lemma F.1 for the exact expression). Then, with probability at least $1-\delta$, we have that the suboptimality of the PEPO output, π_{out}, against $\pi^\star = \arg\max_{\pi\in\Pi} J_\beta(\pi)$ and for any value of*

[8] If Π contains the uniform policy over responses.

$\beta \in (0, \infty)$ *is upper bounded as follows,*

$$J_\beta(\pi^\star) - J_\beta(\pi_{\text{out}}) \leq \widetilde{\mathcal{O}}\left(\min\left\{ \sqrt{\frac{C^\star e^{4R_{\max}} |\mathcal{A}|^4 |\mathcal{X}| \log \delta^{-1}}{N}}, \frac{C^\star_\infty e^{4R_{\max}} |\mathcal{A}|^4 |\mathcal{X}| \log \delta^{-1}}{\beta N} \right\} \right).$$

In Theorem 3.3, the notation $\widetilde{\mathcal{O}}(\cdot)$ hides logarithmic factors in $|\mathcal{X}|$, $|\mathcal{A}|$, N and polynomial factors in $R_{\max}$. The exponential dependence on $R_{\max}$ is conjectured to be unavoidable in the offline setting (Das et al., 2024)[9]. An important consequence of the bound is that if the dataset size is $\widetilde{\mathcal{O}}\left(\frac{C^\star e^{4R_{\max}} |\mathcal{A}|^4 |\mathcal{X}| \log \delta^{-1}}{\varepsilon^2}\right)$, we can provably attain our learning goal, i.e., $J_\beta(\pi^\star) - J_\beta(\pi_{\text{out}}) \leq \varepsilon$.

Moreover, in the high regularization regime, i.e., for high values of β, and in settings in which $C^\star_\infty$ is comparable to $C^\star$, we have that the sample complexity improves to $\widetilde{\mathcal{O}}\left(\frac{C^\star_\infty e^{4R_{\max}} |\mathcal{A}|^4 |\mathcal{X}| \log \delta^{-1}}{\beta \varepsilon}\right)$.

Critically, our bound does not require prior knowledge of $C^\star$ while this is needed for obtaining comparable rates in Huang et al. (2024); Liu et al. (2024), which, however, work in the function approximation rather than the tabular setting.[10]

The dependence on the action cardinality can be improved to $|\mathcal{A}|^2$ at the cost of introducing some graph-theoretic quantities depending on the observed preferences data. We present this result in an instance-dependent version of Theorem 3.3 (Theorem F.2) presented in Appendix F.

For our analysis, we consider that the dataset partition $\cup_{\ell \in [L]} \mathcal{D}^\ell = \mathcal{D}$ is such that $N(x,a,b)/L - 1 \leq N^\ell(x,a,b) \leq N(x,a,b)/L + 1$ where $N(x,a,b)$ is the number of times x, a, b is observed in $\mathcal{D}$ and $N^\ell(x,a,b)$ is the number of observations in $\mathcal{D}^\ell$. That is, if data points are repeated, they are split evenly across the partition. We now briefly present the proof sketch of Theorem 3.3. The analysis has three main conceptual steps.

Step 1: Establishing pessimism A fundamental quantity of the analysis is the implicitly estimated reward gap, defined as $\widehat{\Delta}(x,a,b) := \min_{\ell \in [L]} r^\ell(x,a) - C(x,a) - \max_{\ell \in [L]} r^\ell(x,b)$, which with high probability is a lower bound on the average reward gap under $r^\star$ when the second action is sampled from π_{data}. The result is formalized by the following lemma.

Lemma 3.4. *Let us denote* $\sum_{b \in \mathcal{A}} \pi_{\text{data}}(b|x) \Delta(x,a,b)$ *as* $\Delta(x,a,\pi_{\text{data}})$ *for* $\Delta \in \{\Delta_{r^\star}, \widehat{\Delta}\}$ *and let us assume that* $L \geq \frac{7 \log(|\mathcal{X}||\mathcal{A}|^2/\delta)}{2}$. *Then, the policies* $\{\tilde{\pi}^\ell\}_{\ell=1}^L$, *solutions of equation 6 together with the penalty* $C(x,a)$ *of Theorem 3.3, induce an estimated gap* $\widehat{\Delta}$ *such that with probability at least* $1 - 4\delta$,

$$\widehat{\Delta}(x,a,\pi_{\text{data}}) \leq \Delta_{r^\star}(x,a,\pi_{\text{data}}) \;\; \forall x, a \in \mathcal{X} \times \mathcal{A}.$$

Step 2: From suboptimality to estimation error The second step of the analysis consists of upper bounding the suboptimality of π_{out} in terms of how accurately $\widehat{\Delta}$ estimates $\Delta_{r^\star}$ on average under π_{data}. The pessimistic event that holds with high probability in light of Lemma 3.4 ensures that in this conversion only $C^\star$ appears and not C^{all}. The formal result reads as follows.

Lemma 3.5. *Let us assume that the policy ensemble* $\{\tilde{\pi}^\ell\}_{\ell=1}^L$ *induces an estimated gap* $\widehat{\Delta}$ *such that* $\widehat{\Delta}(x,a,\pi_{\text{data}}) \leq \Delta_{r^\star}(x,a,\pi_{\text{data}})$ *for all* $x, a \in \mathcal{X} \times \mathcal{A}$. *Then, for any* $\beta > 0$,

$$J_\beta(\pi^\star) - J_\beta(\pi_{\text{out}}) \leq \min\left\{ \sqrt{C^\star \left\langle \pi_{\text{data}}, \text{Err}^2 \right\rangle}, \beta^{-1} C^\star_\infty \left\langle \pi_{\text{data}}, \text{Err}^2 \right\rangle \right\}$$

where π_{out} *is computed from* $\{\tilde{\pi}^\ell\}_{\ell=1}^L$ *as in Lemma 3.1 and,*

$$\text{Err}(x,a) := \widehat{\Delta}(x,a,\pi_{\text{data}}) - \Delta_{r^\star}(x,a,\pi_{\text{data}}).$$

[9] Chen et al. (2025) moved the $e^{R_{\max}}$ dependence to lower-order terms, but in the online setting.

[10] Without knowledge of $C^\star$, Huang et al. (2024); Liu et al. (2024) present rates of order $\sqrt{(C^\star)^2/N}$.

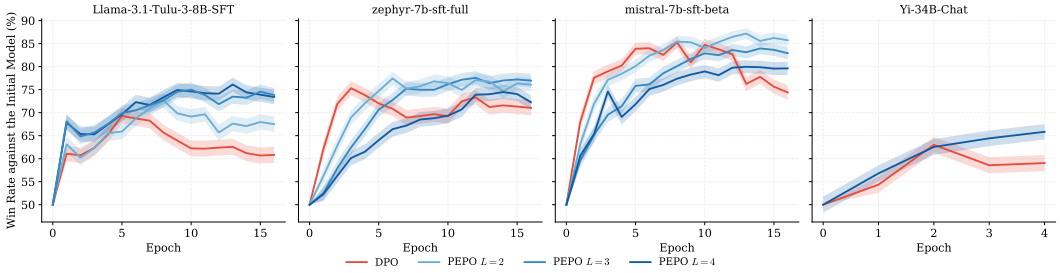


Figure 1: AlpacaEval2 win rate (%) against the initial SFT model across epochs. Shaded regions indicate standard error across evaluation prompts (not across training seeds). Each panel compares DPO with token-level PEPO at different ensemble sizes.

Step 3: Bounding the estimation error Given Lemma 3.5, the last step of the proof is to provide a high probability bound on the estimation error $\langle \pi_{\text{data}}, \text{Err}^2 \rangle$. The following lemma provides precisely this bound and leverages the intuition that equation 6 can be seen as a maximum likelihood problem for estimating $\Delta_{r^\star}$.

Lemma 3.6. *It holds that* $\mathbb{P}\left[\langle \pi_{\text{data}}, \text{Err}^2 \rangle \leq \widetilde{\mathcal{O}}\left(\frac{e^{4R_{\max}}|\mathcal{A}|^4|\mathcal{X}|L^2\log(\delta^{-1})}{N}\right)\right] \geq 1 - 7\delta.$

At this point, the proof of Theorem 3.3 clearly follows by combining Lemmas 3.4,3.5 and 3.6. The detailed analysis is presented in Appendix F.

On the optimal ensemble size Our analysis provides precise guidelines for choosing L. In particular, Lemma 3.6 shows that the generalization error increases with L. Therefore, we would like to choose it as small as possible. However, Lemma 3.4 imposes a lower bound on L for the pessimism to hold. Hence, we conclude that L should be the smallest value which ensures that the estimator $\widehat{\Delta}$ is pessimistic with high enough probability.

4 Experiments

We test whether PEPO mitigates the late-training over-optimization of DPO. Across four SFT backbones (Llama-3.1-Tulu-3-8B-SFT, Zephyr-7B, Mistral-7B, and Yi-34B), we compare token-level PEPO with DPO, χ^2PO, and DPO+SFT on AlpacaEval2 and MT-Bench. Scores are measured against each model's initial SFT checkpoint. Further setup details and results are reported in Appendix C.3.

PEPO avoids over-optimization while DPO does not. In Figure 1, DPO improves initially but degrades at later epochs for all four models. Token-level PEPO instead improves and then stabilizes during training[11]. Because DPO can degrade later in training and validation loss does not reliably identify the best checkpoint (Appendix C.3.6), selecting a DPO checkpoint requires evaluating generated responses during training. PEPO allows us to use the final checkpoint. This agrees with our theory, in which the pessimism obtained from the ensemble protects against over-optimization. For Zephyr-7B and Mistral-7B, increasing L can reduce performance, consistent with the data-efficiency trade-off in Lemma 3.6 because each member sees fewer training examples.

PEPO outperforms pessimistic baselines without access to π_{data}. The methods χ^2PO and DPO+SFT are designed to mitigate over-optimization, but their guarantees rely on exact or approximate knowledge of π_{data}. The responses in UltraFeedback Binarized are generated by a diverse pool of LLMs, including proprietary models such as GPT-4, rather than by the reference model, so π_{data} need not equal π_{ref}. PEPO, by contrast, uses a pessimism mechanism that does not require knowledge of π_{data}. As Table 1 shows, it attains the highest reported score among the available methods at the last recorded checkpoint in every model–benchmark configuration. Table 3 in the appendix reports each method's highest score over the recorded epochs. PEPO attains the highest best-observed score in five of the eight model and benchmark configurations. For PEPO, the largest difference between the best observed score and the final score is 1.2 percentage points on AlpacaEval2 and 1.5 on MT-Bench. The corresponding largest differences are 10.9 and 11.2 for DPO, 7.9 and 9.7 for χ^2PO, and 12.8 and 11.6 for DPO+SFT (see Table 4 for differences across models). Full AlpacaEval2 results across training epochs for all methods are shown in Figure 7. Appendix C.1.2 complements this with a synthetic experiment that controls directly whether $\pi_{\text{ref}} = \pi_{\text{data}}$ or $\pi_{\text{ref}} \neq \pi_{\text{data}}$. PEPO outperforms DPO in both regimes, while χ^2PO matches PEPO only when $\pi_{\text{ref}} = \pi_{\text{data}}$ and fails once $\pi_{\text{ref}} \neq \pi_{\text{data}}$.

[11] Provided L is sufficiently large.

AlpacaEval2	PEPO	χ^2PO	DPO	DPO+SFT
Llama-3-8B	**73.8** $\pm$ **1.6**	65.0 $\pm$ 1.7	60.8 $\pm$ 1.7	64.9 $\pm$ 1.7
Zephyr-7B	**77.0** $\pm$ **1.5**	74.0 $\pm$ 1.6	71.0 $\pm$ 1.6	69.6 $\pm$ 1.6
Mistral-7B	**82.9** $\pm$ **1.3**	81.4 $\pm$ 1.4	74.4 $\pm$ 1.6	63.9 $\pm$ 1.7
Yi-34B	**65.8** $\pm$ **1.7**	—	59.0 $\pm$ 1.7	—

MT-Bench	PEPO	χ^2PO	DPO	DPO+SFT
Llama-3-8B	**69.7**	54.7	54.7	48.4
Zephyr-7B	**66.9**	57.2	60.6	55.9
Mistral-7B	**68.9**	62.0	64.5	53.8
Yi-34B	**62.6**	—	56.0	—

Table 1: Final-checkpoint win rates (%) against the initial SFT model. Each method uses its last recorded checkpoint (epoch 16 for all except Yi-34B for which it is epoch 4). PEPO uses token-level aggregation with $L = 3$, except Yi-34B ($L = 4$). For AlpacaEval2, $\pm$ values are standard errors across evaluation prompts.

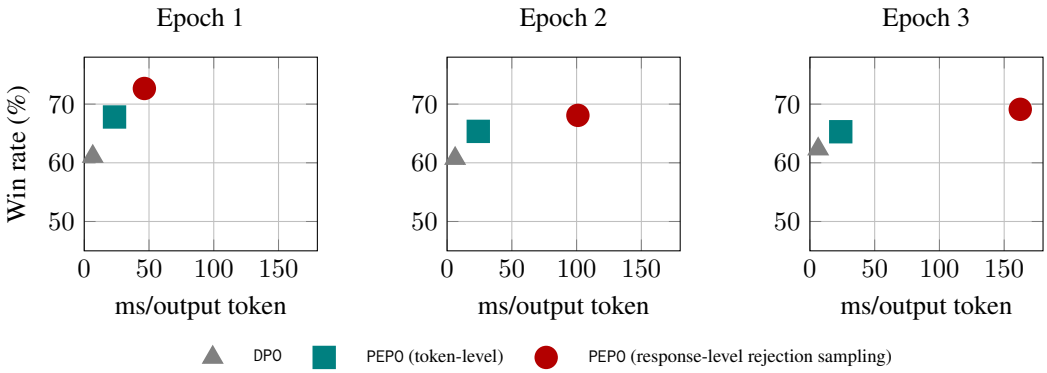


Figure 2: AlpacaEval2 win rate (%) against the initial SFT model versus inference latency for Llama-3-8B across epochs 1–3. Prompts for which the rejection sampler exhausts its trial budget return the best candidate seen.

Rejection sampling vs token-level PEPO In Figure 2, we compare the generation-time trade-offs of DPO and two pessimistic PEPO aggregation rules: response-level rejection sampling (Algorithm 2) and token-level aggregation. Rejection sampling achieves the highest win rate, DPO has the lowest generation latency, and token-level PEPO offers the best overall trade-off. We therefore use the token-level variant in our large-scale experiments. In Appendix D, we show that the token-level implementation of PEPO is the per-token normalization of a response-level policy which is even more pessimistic than π_{out} and hence avoids dependence on C^{all}. The token-level policy inherits a guarantee of the same form up to an additive term, which measures the mismatch between the two normalizations. This term vanishes when they coincide, but not in general.

Disjointness is required by the proof, but its practical necessity remains open. The theory requires independent ensemble members, so our main experiments use disjoint shards. In one Llama-3-8B run, overlapping shards scored 75.9% at epoch 16 versus 73.4% for disjoint shards. Since the overlapping run also changes training-set coverage (68.4% versus 100%) and uses one seed, it does not establish that overlap is generally safe. See Table 5.

5 Conclusions

In this paper, we introduced PEPO, a DPO-like algorithm that provably (in the tabular setting) and practically mitigates over-optimization without knowing the data distribution of the responses recorded in the dataset. This situation is very common in practice when post-training models by leveraging responses generated either by humans or by proprietary models. This contribution opens several exciting directions, the most important of all being to establish the same result beyond the tabular setup. Indeed, while the good experimental results suggest that the tabular setting might be an accurate enough model for practical post-training, there are theoretical results (Huang et al., 2024; Liu et al., 2024) which go beyond it but at the cost of knowing π_{data}. We consider getting the best of both worlds an extremely exciting open problem.

Another interesting direction inspired by Yadkori et al. (2024) is to investigate if the notion of ensemble disagreement can be used to detect the risk of hallucinations and intervene, forcing the model to abstain from answering. In particular, we could force the model to respond "*I do not know*." if Algorithm 2 does not terminate within a maximum number of attempts. There are several other interesting open questions raised by our work. We highlight the most exciting ones in Appendix A.

REPRODUCIBILITY STATEMENT

Source code is provided as supplementary material. Algorithm 1 and Algorithm 2 give the pseudocode of PEPO and of its response-level sampler, and Appendix D describes the token-level implementation used in our large-scale experiments. The assumptions and the statements of our theoretical results are in Section 3.4, and complete proofs are in Appendix F. The setup of the LLM experiments (models, LoRA configuration, optimizer and hyperparameters, seeds, and evaluation protocol with the judge model) is in Appendix C.3, and the synthetic experiments are described in Appendix C. All models and datasets we use are publicly available.

AI USE STATEMENT

We used generative AI tools to review the manuscript and to edit wording in parts of the text. We also used generative AI tools to fix minor mathematical issues in earlier versions of our proofs. We have reviewed all AI-assisted work. We take responsibility for the content of this work produced with the aid of generative AI.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2011.

Alekh Agarwal, Christoph Dann, and Teodor V Marinov. Design considerations in offline preference-based rl. *arXiv preprint arXiv:2502.06861*, 2025.

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

Udbhav Bamba, Minghao Fang, Yifan Yu, Haizhong Zheng, and Fan Lai. Xrpo: Pushing the limits of grpo with targeted exploration and exploitation. *arXiv preprint arXiv:2510.06672*, 2025.

Asaf Cassel, Orin Levy, and Yishay Mansour. Batch ensemble for variance dependent regret in stochastic bandits. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 15678–15685, 2025.

Shicong Cen, Jincheng Mei, Katayoon Goshvadi, Hanjun Dai, Tong Yang, Sherry Yang, Dale Schuurmans, Yuejie Chi, and Bo Dai. Value-incentivized preference optimization: A unified approach to online and offline rlhf. *arXiv preprint arXiv:2405.19320*, 2024.

Mingyu Chen, Yiding Chen, Wen Sun, and Xuezhou Zhang. Avoiding $\mathcal{O}(e^{R_{\max}})$ scaling in rlhf through preference-based exploration. *arXiv preprint arXiv:2502.00666*, 2025.

Richard Y Chen, Szymon Sidor, Pieter Abbeel, and John Schulman. Ucb exploration via q-ensembles. *arXiv preprint arXiv:1706.01502*, 2017.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization. *arXiv preprint arXiv:2310.02743*, 2023.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Wei Zhu, Yuan Ni, Guotong Xie, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback, 2023.

Nirjhar Das, Souradip Chakraborty, Aldo Pacchiano, and Sayak Ray Chowdhury. Active preference optimization for sample efficient rlhf. *arXiv preprint arXiv:2402.10500*, 2024.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruibin Yuan, Qihao Zhu, Wanghai Xu, Wentao Li, Y. K. Li, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

Fei Ding, Baiqiao Wang, Zijian Zeng, and Youwei Wang. Multi-layer grpo: Enhancing reasoning and self-correction in large language models. *arXiv preprint arXiv:2506.04746*, 2025.

Adam Fisch, Jacob Eisenstein, Vicky Zayats, Alekh Agarwal, Ahmad Beirami, Chirag Nagpal, Pete Shaw, and Jonathan Berant. Robust preference optimization through reward model distillation. *arXiv preprint arXiv:2405.19316*, 2024.

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

Alexey Gorbatovski, Boris Shaposhnikov, Alexey Malakhov, Nikita Surnachev, Yaroslav Aksenov, Ian Maksimov, Nikita Balagansky, and Daniil Gavrilov. Learn your reference model for real good alignment. *arXiv preprint arXiv:2404.09656*, 2024.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, Johan Ferret, and Mathieu Blondel. Direct language model alignment from online ai feedback, 2024. URL https://arxiv.org/abs/2402.04792.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Audrey Huang, Wenhao Zhan, Tengyang Xie, Jason D Lee, Wen Sun, Akshay Krishnamurthy, and Dylan J Foster. Correcting the mythos of kl-regularization: Direct alignment without overoptimization via chi-squared preference optimization. *arXiv preprint arXiv:2407.13399*, 2024.

Haque Ishfaq, Qiwen Cui, Viet Nguyen, Alex Ayoub, Zhuoran Yang, Zhaoran Wang, Doina Precup, and Lin Yang. Randomized exploration in reinforcement learning with general value function approximation. In *International Conference on Machine Learning*, pages 4607–4616. PMLR, 2021.

Haque Ishfaq, Qingfeng Lan, Pan Xu, A Rupam Mahmood, Doina Precup, Anima Anandkumar, and Kamyar Azizzadenesheli. Provable and practical: Efficient exploration in reinforcement learning via langevin monte carlo. *arXiv preprint arXiv:2305.18246*, 2023.

Haque Ishfaq, Yixin Tan, Yu Yang, Qingfeng Lan, Jianfeng Lu, A Rupam Mahmood, Doina Precup, and Pan Xu. More efficient randomized exploration for reinforcement learning via approximate sampling. *arXiv preprint arXiv:2406.12241*, 2024.

Haque Ishfaq, Guangyuan Wang, Sami Nur Islam, and Doina Precup. Langevin soft actor-critic: Efficient exploration through uncertainty-driven critic learning. *arXiv preprint arXiv:2501.17827*, 2025.

Kaixuan Ji, Qiwei Di, Heyang Zhao, Qingyue Zhao, and Quanquan Gu. On the optimal sample complexity of offline multi-armed bandits with kl regularization. *arXiv preprint arXiv:2605.02141*, 2026.

Xiang Ji, Sanjeev Kulkarni, Mengdi Wang, and Tengyang Xie. Self-play with adversarial critic: Provable and scalable offline alignment for language models. *arXiv preprint arXiv:2406.04274*, 2024.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 5 2023a.

Zihao Li, Zhuoran Yang, and Mengdi Wang. Reinforcement learning with human feedback: Learning dynamic choices via pessimism. *arXiv preprint arXiv:2305.18438*, 2023b.

Zhihan Liu, Miao Lu, Shenao Zhang, Boyi Liu, Hongyi Guo, Yingxiang Yang, Jose Blanchet, and Zhaoran Wang. Provably mitigating overoptimization in rlhf: Your sft loss is implicitly an adversarial regularizer. *Advances in Neural Information Processing Systems*, 37:138663–138697, 2024.

Zichen Liu et al. Understanding r1-zero-like training: A critical perspective, 2025. URL https://arxiv.org/abs/2503.20783.

Yu Meng, Mengzhou Xia, and Danqi Chen. Simpo: Simple preference optimization with a reference-free reward. *Advances in Neural Information Processing Systems*, 37:124198–124235, 2024.

Antoine Moulin, Gergely Neu, and Luca Viano. Optimistically optimistic exploration for provably efficient infinite-horizon reinforcement and imitation learning. *arXiv preprint arXiv:2502.13900*, 2025.

Rémi Munos, Michal Valko, Daniele Calandriello, Mohammad Gheshlaghi Azar, Mark Rowland, Zhaohan Daniel Guo, Yunhao Tang, Matthieu Geist, Thomas Mesnard, Côme Fiegel, Andrea Michi, Marco Selvi, Sertan Girgin, Nikola Momchev, Olivier Bachem, Daniel J Mankowitz, Doina Precup, and Bilal Piot. Nash learning from human feedback. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2312.00886.

Brendan O'Donoghue. Variational bayesian reinforcement learning with regret bounds. *Advances in Neural Information Processing Systems*, 34:28208–28221, 2021.

Brendan O'Donoghue. Efficient exploration via epistemic-risk-seeking policy optimization. *arXiv preprint arXiv:2302.09339*, 2023.

Ian Osband and Benjamin Van Roy. Near-optimal reinforcement learning in factored mdps. *Advances in Neural Information Processing Systems*, 27, 2014.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. *Advances in neural information processing systems*, 29, 2016a.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pages 2377–2386. PMLR, 2016b.

Ian Osband, John Aslanides, and Albin Cassirer. Randomized prior functions for deep reinforcement learning. *Advances in Neural Information Processing Systems*, 31, 2018.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Approximate thompson sampling via epistemic neural networks. In *Uncertainty in Artificial Intelligence*, pages 1586–1595. PMLR, 2023a.

Ian Osband, Zheng Wen, Seyed Mohammad Asghari, Vikranth Dwaracherla, Morteza Ibrahimi, Xiuyuan Lu, and Benjamin Van Roy. Epistemic neural networks. *Advances in Neural Information Processing Systems*, 36:2795–2823, 2023b.

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

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 4998–5017. Association for Computational Linguistics, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

Aviv Rosenberg, Alon Cohen, Yishay Mansour, and Haim Kaplan. Near-optimal regret bounds for stochastic shortest path. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 8210–8219. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/rosenberg20a.html.

Andreas Schlaginhaufen, Reda Ouhamma, and Maryam Kamgarpour. Efficient preference-based reinforcement learning: Randomized exploration meets experimental design. *arXiv preprint arXiv:2506.09508*, 2025.

Lior Shani, Tom Zahavy, and Shie Mannor. Online apprenticeship learning. *arXiv:2102.06924*, 2021.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Yuda Song, Gokul Swamy, Aarti Singh, J Bagnell, and Wen Sun. The importance of online data: Understanding preference fine-tuning via coverage. *Advances in Neural Information Processing Systems*, 37:12243–12270, 2024.

Jean Tarbouriech, Tor Lattimore, and Brendan O'Donoghue. Probabilistic inference in reinforcement learning done right. *Advances in Neural Information Processing Systems*, 36, 2024.

Jens Tuyls, Dylan J Foster, Akshay Krishnamurthy, and Jordan T Ash. Representation-based exploration for language models: From test-time to post-training. *arXiv preprint arXiv:2510.11686*, 2025.

Luca Viano, Stratis Skoulakis, and Volkan Cevher. Imitation learning in discounted linear MDPs without exploration assumptions. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=DChQpB4AJy.

Stefano Viel, Luca Viano, and Volkan Cevher. Il-soar: Imitation learning with soft optimistic actor critic. *arXiv preprint arXiv:2502.19859*, 2025.

Tengyang Xie, Dylan J Foster, Akshay Krishnamurthy, Corby Rosset, Ahmed Awadallah, and Alexander Rakhlin. Exploratory preference optimization: Harnessing implicit q*-approximation for sample-efficient rlhf. *arXiv preprint arXiv:2405.21046*, 2024.

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

Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. To believe or not to believe your llm. *arXiv preprint arXiv:2406.02543*, 2024.

Yurun Yuan, Fan Chen, Zeyu Jia, Alexander Rakhlin, and Tengyang Xie. Trajectory bellman residual minimization: A simple value-based method for llm reasoning. *arXiv preprint arXiv:2505.15311*, 2025.

Wenhao Zhan, Masatoshi Uehara, Nathan Kallus, Jason D Lee, and Wen Sun. Provable offline preference-based reinforcement learning. *arXiv preprint arXiv:2305.14816*, 2023.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In *International Conference on Machine Learning*, pages 43037–43067. PMLR, 2023.

A Open Directions

We list here the most interesting directions left open by our work.

Pessimism in general function classes with unknown π_{data} As already mentioned in the main text, we think that the most interesting open question is to develop a DPO-like method applicable to general function approximation that achieves sample complexity bounds depending only on $C^\star$. We highlight this open question in Table 2. In other words, Table 2 shows that PEPO, DPO+SFT and χ^2PO all apply beyond the easiest case of tabular setting with known π_{data} but *along different axes*. Achieving those two improvements simultaneously, i.e. designing an approach which applies for general policy classes without knowledge about π_{data}, is in our opinion the most important open question arising from our work. Moreover, we notice that the analysis in Huang et al. (2024); Liu et al. (2024) still requires the policy class to be very expressive since it assumes that it realizes the solution of the objective in equation 1 over the domain of all tabular policies. Bypassing such an assumption is another interesting open direction.

Table 2: Comparison of offline DPO-like algorithms with sample complexity scaling only with $C^\star$.

	Tabular Setting	**General Policy Class**
Known π_{data}	PEPO (Ours) DPO+SFT (Liu et al., 2024) χ^2PO (Huang et al., 2024)	Liu et al. (2024) (Convex only) χ^2PO (Huang et al., 2024)
Unknown π_{data}	PEPO (Ours)	***Open Question***

Optimism via Ensemble In the context of LLM post-training there are applications in which at training time it is possible to generate sentences from the model which is currently learning and to get feedback for these responses in the form of a reward signal or a preference bit. Examples of the first kind are the emerging family of algorithms for reasoning such as GRPO (Shao et al., 2024; DeepSeek-AI et al., 2025), GRPO Done Right (Liu et al., 2025) and TBRM (Yuan et al., 2025), while for online algorithms learning from preference signals we highlight Online DPO (Guo et al., 2024) and Nash Learning from Human Feedback (Munos et al., 2024). While promising at an empirical level, they all miss an exploration mechanism that should be included to ensure that the learning model queries the reward or preference oracle for responses which are uncertain, i.e. for which new information should really be acquired. Such lack of exploration has been also observed in certain experiments (Bamba et al., 2025; Ding et al., 2025). While recently some exploration mechanisms have been proposed for improving GRPO, Best of N and Online DPO (Xie et al., 2024; Tuyls et al., 2025), none of these works leverages the ensemble mechanism which we show to be effective for pessimism. Therefore, it remains open to investigate optimistic ensemble based versions of the above algorithms and see if they can lead to any theoretical and practical advantage over the existing exploration methods.

Outputting a parametric policy Our work successfully avoids over-optimization without knowledge of π_{data} at the cost of outputting a non-parametric policy from which we can sample efficiently but that we cannot compute in closed form. Therefore, our method is ideal for being the last step of an LLM training pipeline (which is expected to be most often the case for DPO-like algorithms). However, PEPO is not suitable as an intermediate step of the training pipeline because it does not output a parametric policy, i.e. a training checkpoint in a more applied jargon. It is therefore interesting for future work to solve the exact same question this work solves but using parametric policies.

B Related Work

Some heuristic techniques have been proposed to mitigate over-optimization: imposing a trust region (Gorbatovski et al., 2024), using length normalization (Meng et al., 2024), or using a loss function with a bounded minimizer, or better curvature (Azar et al., 2024; Agarwal et al., 2025). These methods can sometimes be effective in practice, but unfortunately, none of these approaches can provably prevent over-optimization.

The literature is also rich in works affected by the opposite issue: pessimistic RLHF approaches such as (Zhu et al., 2023; Zhan et al., 2023; Schlaginhaufen et al., 2025; Li et al., 2023b) enjoy strong bounds but are rarely applicable. The common idea is first to learn a reward function and its uncertainty from the preference dataset and, then, to optimize the reward minus the uncertainty. This procedure is provably robust to over-optimization but there are two main downsides: (i) relying on a two-step procedure and (ii) the fact that the uncertainty estimate is usually computed leveraging elliptical bonuses (Abbasi-Yadkori et al., 2011), which is a technique that does not offer guidance towards a practical approach for modern LLMs[12].

Moreover, there are previous works using ensembles in alignment (Coste et al., 2023; Fisch et al., 2024) but they also rely on a two-step recipe. Both works (Coste et al., 2023; Fisch et al., 2024) empirically verified that the technique helps fight over-optimization.

Existing theoretically grounded DPO variants that achieve single policy concentrability guarantees require either that the responses in the dataset are generated according to the SFT checkpoint π_{ref} (Huang et al., 2024) or according to a policy that covers well a known policy, denoted as π_{base} in Liu et al. (2024). Liu et al. (2024) suggested to use the distribution of the chosen responses as π_{base}. Therefore, in this case, imposing their assumption implies knowledge of π_{data}. Other pessimistic analyses of DPO (Cen et al., 2024) still incur an all-policy concentrability term; see the factor $\kappa_{\mathcal{D}}$ in Cen et al. (2024, Theorem 2). Another way to attain single policy concentrability bounds is to allow regeneration according to the policy sequence generated during learning (Ji et al., 2024); however, this process makes the implementation more costly compared to the original DPO algorithm.

Finally, we remark that the idea of using ensembles for uncertainty-aware decision making has been previously used in reinforcement and imitation learning, both at a purely theoretical level (Osband and Van Roy, 2014; Osband et al., 2016b; O'Donoghue, 2021; 2023) and at a practical level (Osband et al., 2016a; Chen et al., 2017; Osband et al., 2018; 2023a;b; Tarbouriech et al., 2024; Ishfaq et al., 2021; 2023; 2024; 2025). Moreover, at a technical level, the most inspiring works for our analysis are Cassel et al. (2025), which studied ensemble-based exploration in bandits, and Viel et al. (2025), which studies Imitation Learning (IL) with access to the environment. In particular, Viel et al. (2025) used an ensemble-based exploration scheme to make previous theoretical insights on the importance of exploration in IL from states or reward features-only (Shani et al., 2021; Viano et al., 2024; Moulin et al., 2025) applicable to scenarios where neural network function approximation is required.

C Additional Experiments

In this section, we report some experiments omitted from the main text and provide additional details about our setup.

C.1 Synthetic preference experiments

C.1.1 Ablation for the ensemble size L in a controlled setting.

We run an ablation for the ensemble size L in a simple and controlled setting. In the environment under consideration, we consider an empty prompt set and 20 possible answers. The preferences are generated via the Bradley-Terry model with rewards which are all Gaussian distributed. In particular, all reward distributions except that of the optimal arm have mean equal to 1 and variance equal to 1. On the other hand, the optimal arm's distribution has mean equal to 2.9 and variance equal to $1/2$. We generate preference datasets of different sizes $\mathcal{D}$ and run PEPO for different ensemble sizes L. We report the results in Figure 3 where we see that for all the values we tried PEPO finds the optimal action for all the 5 considered seeds. However, the strongest performance is obtained for $L = 3$. Increasing L above 3 slows down learning. This is consistent with the theoretical guideline of setting L to the smallest value which is large enough to ensure pessimism.

C.1.2 Known vs unknown π_{data} in a controlled setting.

We consider an empty prompt and 3 possible actions. The preferences dataset is generated via the Bradley-Terry model with Gaussian rewards with means $[1, 1.5, 1]$ and variances $[1.5, 0.5, 1.5]$, i.e., the second action is the optimal one and also has lower variance. The other two actions are identical and suboptimal. In our experiment in Figure 4a, we assume knowledge of π_{data} and we run all

[12] Tuyls et al. (2025) is a notable exception which, however, uses an elliptical-bonus-inspired technique for optimism in online algorithms and not for pessimistic purposes in the offline setting.

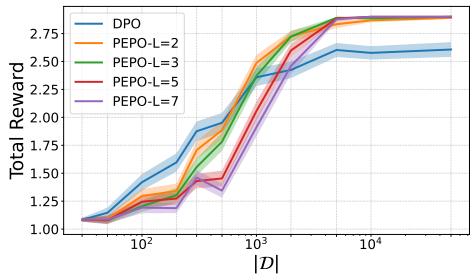
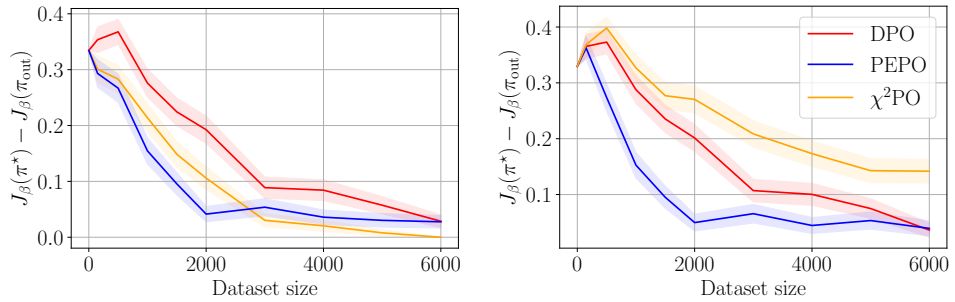


Figure 3: Ablation for L in a bandit setting.



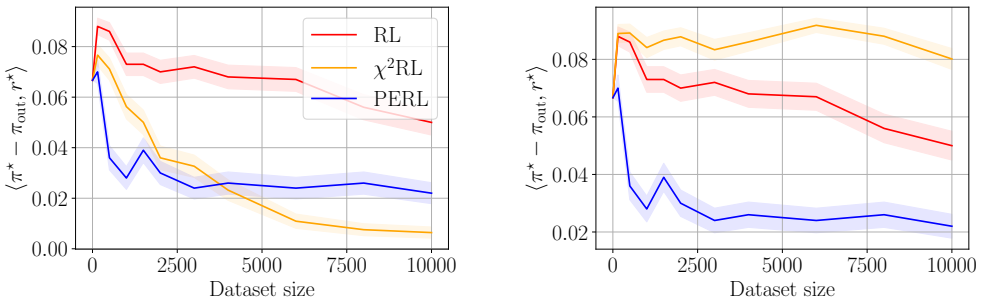
(a) Known π_{data} (b) Unknown π_{data}

Figure 4: Experiment in a 3-armed bandit setting for the situation of known and unknown π_{data}.

the algorithms with $\pi_{\text{ref}} = \pi_{\text{data}}$. Consistently with the theoretical results, both χ^2PO and PEPO outperform DPO in this setting.

However, when considering an unknown π_{data} (see Figure 4b), we cannot set $\pi_{\text{ref}} = \pi_{\text{data}}$. For this experiment, we set π_{ref} to put most of the mass on a suboptimal action while π_{data} puts the largest weight on the optimal one. In this situation, PEPO still works reliably while χ^2PO fails because it interprets the high mean estimate for the optimal arm as affected by high uncertainty, as the optimal action has a small probability under π_{ref}. For Figure 4a, we used $\pi_{\text{data}} = \pi_{\text{ref}} = [0.04, 0.93, 0.03]$. That is, the largest mass is on the central action, which is the optimal one. This makes the optimal action well-covered, i.e. small $C^\star$, therefore χ^2PO and PEPO can work reliably. At the same time, the all-policy concentrability is large because the two suboptimal actions are poorly covered. This explains why DPO is outperformed in Figure 4a. For the unknown case, we consider instead $\pi_{\text{data}} = [0.04, 0.93, 0.03]$ and $\pi_{\text{ref}} = [0.00025, 0.00025, 0.9995]$ and $\beta = 0.001$. In this case, we have two distinct failure modes for DPO and χ^2PO while PEPO still works reliably. DPO fails because the all-policy concentrability is still very large. This implies that the reward of the suboptimal actions cannot be estimated accurately, and, therefore, there is the risk of playing a suboptimal action whose reward is wrongly overestimated. The failure mode of χ^2PO is instead different. The mismatch between π_{data} and π_{ref} leads χ^2PO to perceive the suboptimal action with high probability under π_{ref} as accurately estimated. On the contrary, since the optimal action has low mass under π_{ref}, this action is perceived as quite uncertain. All in all, χ^2PO concludes that the high reward estimate for the optimal action cannot be trusted and plays a suboptimal action. PEPO still succeeds in this setting because under π_{data}, we have that the optimal action is actually estimated more accurately than the suboptimal ones. PEPO can use this fact to successfully identify and play the optimal action if the dataset size is large enough. This is shown in Figure 4b.

To develop intuition and see this mechanism in an even more evident manner, we consider the same task with $\beta = 0$. In this case, we cannot derive DPO-like algorithms but we compute the maximum likelihood estimator $\hat{r}$ for the reward function from the preference dataset and then consider three approaches: (i) RL outputs the action with highest estimated reward, (ii) a method that we name χ^2RL which outputs $\pi_{\text{out}} = \arg\max_{\pi \in \Pi} \langle \pi, \hat{r} \rangle - \gamma D_{\chi^2}(\pi, \pi_{\text{ref}})$ where we set $\gamma = 40$ and (iii) a PEPO-like approach, dubbed PERL that instead of a single reward estimate computes L of them, i.e.



(a) Known π_{data}

(b) Unknown π_{data}

Figure 5: Result in the controlled setting without regularization, i.e. with $\beta = 0$.

$\{\hat{r}_\ell\}_{\ell=1}^L$ and then outputs the action with highest *robust reward* defined as $\min_{\ell \in [L]} \hat{r}_\ell$. We report the results of this experiment in Figure 5.

C.2 Other worst-case aggregation rules for the ensemble.

The rejection sampling rule with the minimum in the numerator as defined in Algorithm 2 can be overly conservative in practice and lead to a very high rejection rate. One possible alternative to reduce the generation time of the model is to consider this alternative numerator, $f_{\text{out}}(x, a) = \max\left\{0, \frac{1}{L}\sum_{\ell=1}^L \tilde{\pi}^\ell(a|x)e^{-m^\ell(x,a)} - \eta \cdot \text{std}\left(\left\{\tilde{\pi}^\ell(a|x)e^{-m^\ell(x,a)}\right\}_{\ell=1}^L\right)\right\}$ where η is a hyperparameter that can be used to tune the level of pessimism. With the same technique used in Viel et al. (2025), we can still establish the sample complexity guarantees for this variant if we set $\eta = \sqrt{L-1}$, because for this choice Samuelson's inequality gives $f_{\text{out}}(x, a) \leq \min_{\ell \in [L]} \tilde{\pi}^\ell(a|x)e^{-m^\ell(x,a)}$ (with std the population standard deviation). For the practical epoch-by-epoch comparison below, we use the recorded run with the original minimum-probability rejection sampler from Algorithm 2. This minimum-aggregation rule is the response-level method compared with token-level PEPO in Figure 2.

C.3 LLM experiments

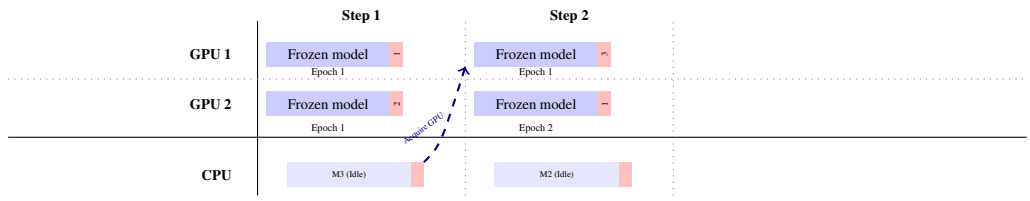
We first describe the LLM setup, then report checkpoint-selection, diagnostic tests, and generation-cost diagnostics.

C.3.1 Models and fine-tuning setup

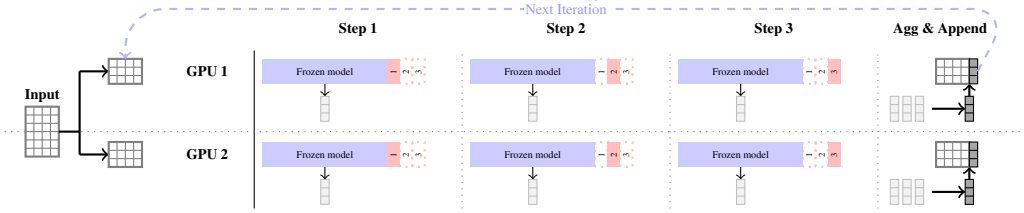
We evaluate PEPO on four instruction-tuned language models spanning a range of sizes. At the 7–8B scale, we consider Llama-3-8B (the Llama-3.1-Tulu-3-8B-SFT checkpoint), which serves as a strong open-source baseline, along with Zephyr-7B-SFT and Mistral-7B-SFT. To demonstrate scalability, we also include Yi-34B-Chat, a 34B parameter model. All models have undergone supervised fine-tuning (SFT) prior to preference optimization.

Rather than full fine-tuning, we employ Low-Rank Adaptation (LoRA) (Hu et al., 2022) for parameter-efficient training. All LoRA adapters use rank $r = 16$ with scaling factor $\alpha = 16$ (the low-rank updates are scaled by α/r), a dropout rate of 5%, and are applied to all linear layers in the model. This adds approximately 0.5–1.5% (depending on the model) trainable parameters while the underlying model stays frozen.

Maintaining an ensemble of L models does *not* require L times more GPU memory compared to a single model. Since all ensemble members share the same frozen model and only differ in their lightweight LoRA adapters, the total memory footprint is approximately $M_{\text{base}} + L \cdot M_{\text{adapter}}$ rather than $L \cdot M_{\text{base}}$. Given that each adapter constitutes only a small fraction of the base model size, the memory overhead of the ensemble is modest even for moderate values of L. The schematic implementation of this process is shown in Figure 6.



(a) Parallel training



(b) Token-level parallel inference with sequential adapter switching

Figure 6: Parallel training (top) and token-level inference with sequential adapter switching (bottom). Each ensemble member produces per-token output that is aggregated at every decoding step.

C.3.2 Dataset and Preprocessing

For all large-scale experiments, we use UltraFeedback Binarized (Cui et al., 2023). This version of the dataset is derived from the original UltraFeedback corpus by "binarizing" multi-model outputs into discrete preference pairs. Specifically, for each prompt, four responses were generated by diverse language models and scored by GPT-4 across multiple dimensions (e.g., helpfulness and honesty). The binarized version identifies the highest-scoring response as the *chosen* sample and selects one of the remaining lower-scoring responses as the *rejected* sample. We use the standard `train_sft` split for training and `test_sft` for validation[13].

During preprocessing, we truncate sequences to a maximum length of 1024 tokens, with prompts limited to 512 tokens to ensure sufficient space for generated responses.

For the main PEPO comparisons, we partition the training data into L disjoint subsets of approximately equal size. This partitioning ensures statistical independence across ensemble members, as required by our theoretical analysis (Lemma 3.4). We note that in practice, data efficiency could potentially be improved by using overlapping subsets. We compare disjoint and overlapping partitions in Table 5.

C.3.3 Training Procedure

We train all models using the AdamW optimizer with `bfloat16` mixed precision for computational efficiency. Training is carried out with a learning rate of 1×10^{-5} and a weight decay of 0.01. We maintain a consistent effective batch size of 64 across all experiments, using gradient accumulation where necessary. To accelerate training, we pre-compute and cache the reference model's log-probabilities for all training examples, avoiding redundant forward passes through the frozen model.

For PEPO, the β parameter controlling KL regularization is set to 0.1, matching the DPO baseline. Note that $L = 1$ corresponds to a single model trained with the DPO loss. We vary the ensemble size $L \in \{2, 3, 4\}$ to study the trade-off between pessimism strength and data efficiency. For the baselines, χ^2PO uses $\beta = 0.1$, $\gamma = 1.0$, and $R_{\max} = 10.0$ with internal gradient clipping, as in Huang et al. (2024). DPO+SFT combines the DPO objective with a supervised fine-tuning regularizer weighted by $\lambda_{\text{SFT}} = 0.005$, as in Liu et al. (2024). In DPO+SFT, we use both chosen and rejected responses from UltraFeedback to estimate π_{data}. All large-scale runs use seed 42.

C.3.4 Ensemble Training and Inference

The PEPO ensemble is trained by assigning each LoRA adapter to its corresponding data partition. Ensemble members can be trained sequentially or in parallel across multiple GPUs without any required communication between the members of the ensemble.

[13]We point out that validation loss is a poor proxy for downstream evaluation performance in Appendix C.3.6.

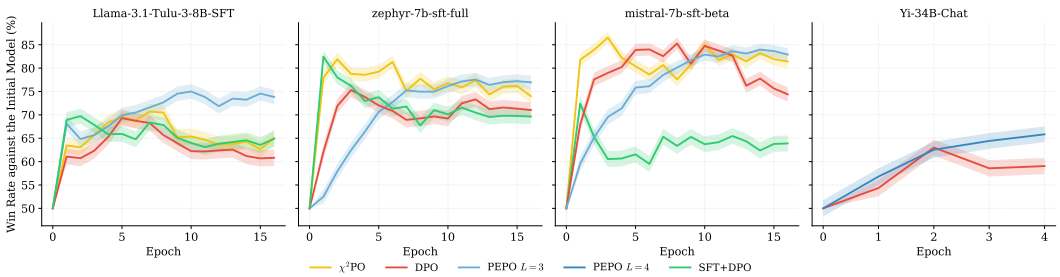


Figure 7: AlpacaEval2 win rate (%) against the initial SFT model across training epochs for all compared methods. Shading shows prompt-level standard error. PEPO uses token-level aggregation with the fixed $L = 3$ comparison for Llama-3-8B, Zephyr-7B, and Mistral-7B, and the available $L = 4$ run for Yi-34B. All runs use seed 42.

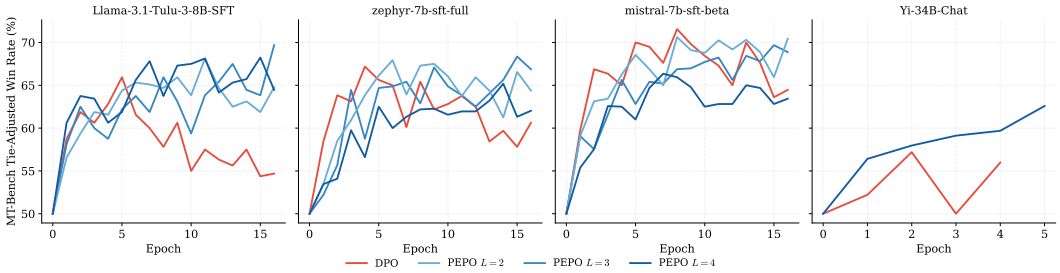


Figure 8: MT-Bench tie-adjusted win rate (%) against the initial SFT model across training epochs. The curves use the same training protocol as Figure 1 and show DPO and token-level PEPO for the different ensemble sizes.

While we introduced a rejection sampling scheme in Section 3 that samples exactly from π_{out}, it can be computationally expensive at inference time. For our large-scale experiments, we instead employ the token-level approximation detailed in Appendix D. This variant applies the pessimistic aggregation at each decoding step, enabling efficient autoregressive generation. We use greedy decoding (temperature $T = 0$) with this token-level aggregation for all reported results unless stated otherwise.

During inference of PEPO, we load multiple adapters together with one shared frozen model. This allows us to efficiently switch the adapters on and off to run the forward pass with different members of the ensemble without reloading the model to and from device each time. We make use of multiple GPUs by distributing a copy of this shared frozen model to each available device.

C.3.5 Evaluation Protocol

We evaluate all models on AlpacaEval2 (Li et al., 2023a), a benchmark that measures pairwise win rate against a reference model across 805 diverse prompts spanning open-ended generation, reasoning, and instruction-following tasks. We report the win rate against the initial (epoch 0) SFT model, which measures the improvement from preference optimization. To ensure reproducibility and manage costs, we use Llama-3-70B-Instruct as the preference judge (run locally), utilizing the `alpaca_eval_llama3_70b_fn` configuration.

All responses use at most 1024 new tokens. We evaluate the recorded checkpoints throughout training to capture learning dynamics and potential over-optimization. The win rates at the last and best epochs are in Table 1 and Table 3.

For MT-Bench, we use the standard 80 two-turn questions and bundled pairwise judge prompts. We run the same local Llama-3-70B-Instruct judge at temperature 0 and report tie-adjusted win rate, assigning half credit to ties. The reference is the initial SFT model. The corresponding training curves are shown in Figure 8.

C.3.6 VALIDATION LOSS DOES NOT TRACK ALPACAEVAL2 WIN RATE

At training time, generation quality is expensive to measure, so one might use validation loss to select a checkpoint. In this setting, however, validation loss is a poor proxy for AlpacaEval2 win rate. In Figure 9, the checkpoint with minimum DPO validation loss differs from the checkpoint with maximum AlpacaEval2 win rate for each of the three plotted models. We therefore use the last recorded checkpoint in the main tables rather than selecting by validation loss or by either evaluation benchmark. This avoids both an unreliable proxy and benchmark-based checkpoint selection.

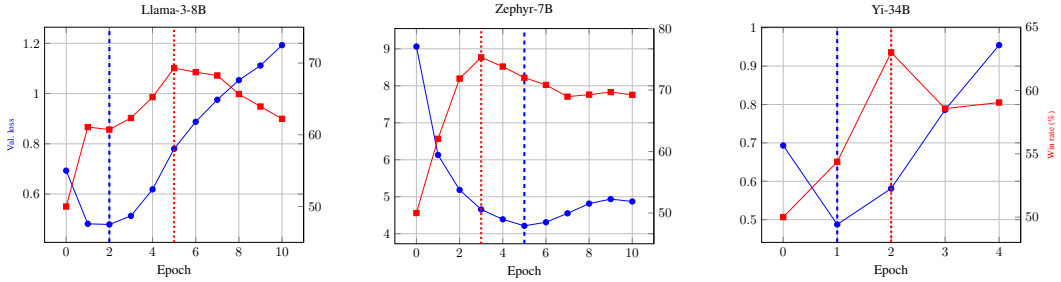


Figure 9: DPO validation loss and AlpacaEval2 win rate for Llama-3-8B, Zephyr-7B, and Yi-34B. Blue dashed lines mark the minimum validation loss. Red dotted lines mark the maximum win rate. The optima differ for all three models, so validation-loss early stopping does not select the best observed AlpacaEval2 checkpoint.

C.3.7 BEST-CHECKPOINT COMPARISON

Table 3 provides a descriptive peak-performance comparison. For each method, model, and benchmark, it reports the highest observed score over the recorded checkpoints. The corresponding epoch is shown as a superscript. These values are descriptive, not held-out model-selection results, and are not used for the final-checkpoint claim in the main text.

AlpacaEval2

	PEPO	χ^2PO	DPO	DPO+SFT
Llama-3-8B	**75.0** $\pm$ **1.5**^{E10}	70.7 $\pm$ 1.6^{E7}	69.3 $\pm$ 1.6^{E5}	69.7 $\pm$ 1.6^{E2}
Zephyr-7B	77.6 $\pm$ 1.4^{E12}	81.9 $\pm$ 1.4^{E2}	75.3 $\pm$ 1.5^{E3}	**82.5** $\pm$ **1.3**^{E1}
Mistral-7B	84.0 $\pm$ 1.3^{E14}	**86.5** $\pm$ **1.2**^{E3}	85.3 $\pm$ 1.3^{E8}	72.4 $\pm$ 1.6^{E1}
Yi-34B	**65.8** $\pm$ **1.7**^{E4}	—	63.0 $\pm$ 1.7^{E2}	—

MT-Bench

	PEPO	χ^2PO	DPO	DPO+SFT
Llama-3-8B	**69.7**^{E16}	64.4^{E4}	65.9^{E5}	60.0^{E1}
Zephyr-7B	**68.4**^{E15}	65.3^{E6}	67.2^{E4}	60.4^{E5}
Mistral-7B	69.7^{E15}	70.6^{E4}	**71.6**^{E8}	63.5^{E1}
Yi-34B	**62.6**^{E5}	—	57.2^{E2}	—

Table 3: Best-checkpoint win rates (%) against the initial SFT model. Superscripts indicate the recorded epoch (for example, ^{E10} denotes epoch 10). PEPO uses token-level aggregation with $L = 3$, except for Yi-34B ($L = 4$). Dashes indicate unavailable runs. For AlpacaEval2, $\pm$ values are standard errors across prompts, not training seeds. All runs use seed 42. These values are descriptive and should not be interpreted as a held-out model-selection result.

AlpacaEval2

	PEPO	χ^2PO	DPO	DPO+SFT
Llama-3-8B	**1.2**	5.7	8.5	4.8
Zephyr-7B	**0.6**	7.9	4.3	12.8
Mistral-7B	**1.1**	5.1	10.9	8.5
Yi-34B	**0.0**	—	4.0	—

MT-Bench

	PEPO	χ^2PO	DPO	DPO+SFT
Llama-3-8B	**0.0**	9.7	11.2	11.6
Zephyr-7B	**1.5**	8.1	6.6	4.5
Mistral-7B	**0.8**	8.6	7.1	9.7
Yi-34B	**0.0**	—	1.2	—

Table 4: Difference between each method's best observed and final win rate (percentage points). Lower values indicate that the final checkpoint remains closer to the method's best observed checkpoint.

C.3.8 TEST OF NECESSITY OF DISJOINT DATA SPLITS

We report a comparison in Table 5 that tests whether disjoint partitions are needed in practice. We use just one run and one configuration, so it is a diagnostic probe rather than formal ablation.

Method	E0	E2	E4	E6	E8	E10	E12	E14	E16
Disjoint splits (L=4)	50.0	69.9	68.4	72.3	73.3	74.6	74.1	74.4	73.4
Random overlapping splits (L=4)	50.0	65.9	71.0	74.3	75.0	75.0	78.8	77.2	75.9

Table 5: Epoch-by-epoch comparison of AlpacaEval2 win rate (%) of two Llama-3-8B PEPO runs. Data splits for $L = 4$ members of the ensemble are *disjoint* (no pairwise overlap, 100% training-set coverage) in one while *overlapping* (≈25% mean pairwise overlap, 68.4% coverage) in the other.

C.3.9 Computational Resources

Experiments were conducted on NVIDIA GH200 120GB and NVIDIA A100-SXM4-80GB GPUs. We allocated one GPU per ensemble member for training (e.g., 4 GPUs for a PEPO model with $L = 4$, trained in parallel) and two GPUs for evaluation and generation (mainly to accommodate the Llama-3-70B-Instruct judge). Since we parallelize the training of the ensemble across devices and partition the dataset such that each member sees only $1/L$ of the data, the wall-clock time required for training is significantly reduced (roughly proportional to $1/L$) compared to training a single model on the full dataset.

For training, we observed that the average wall-clock time per epoch for the Llama-3-8B model dropped from 3.61 hours ($L = 1$) to 0.90 hours ($L = 4$). Similarly, for Yi-34B, training took approximately 4.0 hours per epoch with $L = 4$.

For evaluation on AlpacaEval2, generating responses and computing win rates using the token-level sampling detailed in Appendix D for the 8B model takes approximately 30 minutes on a single NVIDIA GH200 GPU and approximately 1.25 hours on an NVIDIA A100 GPU. In contrast, full rejection sampling using the minimum probability criterion is significantly slower. We observe that the acceptance probability decreases towards 0 as the ensemble is trained for more epochs. This causes the generator to saturate the trial limit (~ 16 attempts per answer), increasing the evaluation time to approximately 80 hours on an NVIDIA A100 GPU.

C.3.10 Details for the results shown in Figure 2

Figure 2 compares AlpacaEval2 win rate and inference latency for DPO, token-level PEPO, and response-level PEPO after 1, 2, and 3 training epochs. The response-level points use the recorded minimum-probability rejection sampler defined in Algorithm 2. This comparison shows the computational cost of exact response-level pessimistic aggregation and motivates the token-level approximation used in the main experiments.

D GUARANTEES FOR THE TOKEN-LEVEL IMPLEMENTATION OF PEPO

As mentioned in the main text, we also introduce a version of PEPO which implements the pessimistic aggregation rule given by Lemma 3.1 at the token level rather than at the response level. This approach is faster at generation time because it does not reject generations, while Algorithm 2 does, and the acceptance probability of the latter shrinks as the ensemble members diverge. We therefore preferred this version for our LLM evaluations.

We show that the token-level aggregation rule induces a pessimistic gap estimate which is even more pessimistic than the response-level one. The response-level policy associated with this estimate, which we denote by $\pi_{\text{out}}^{\text{prod}}$, therefore avoids dependence on C^{all}. The token-level policy $\pi_{\text{out}}^{\text{token}}$ used in our experiments differs from $\pi_{\text{out}}^{\text{prod}}$ only in the normalization, which is carried out token by token rather than over whole responses, and we quantify the price of this approximation exactly: it is a D_{KL} term between the two policies which vanishes when the two normalizations coincide.

For this result, we consider the following token-level MDP. We consider that it has horizon (sequence length) equal to $H \in \mathbb{N}$. As in the main paper we denote as x the initial user prompt and as a the model response. We denote the individual tokens of the response as $\{a_h\}_{h=1}^{H}$ and use square brackets to denote concatenation $a = [a_1, \dots, a_H]$. At each stage $h \in [H]$, we consider the state, $x_h = [x, a_1, \dots, a_{h-1}]$, which is the concatenation of the initial prompt and the tokens the model produced so far. We consider that π_{ref} and π^ℓ are autoregressive models over a token vocabulary $\mathcal{V}$. That is, $\pi_{\text{ref}}(a|x) = \prod_{h=1}^{H} \pi_{\text{ref}}(a_h|x_h)$ and $\pi^\ell(a|x) = \prod_{h=1}^{H} \pi^\ell(a_h|x_h)$, where $x_h := [x, a_1, \dots, a_{h-1}]$ is the prefix built from the prompt x and the first $h-1$ tokens of the response a, so that $x_1 = x$.

A token-level factorization of the pessimism bonus. The bonus $C(x, a)$ is a response-level quantity, but it admits an *exact* per-token factorization. For $0 \le h \le H$, let $\nu(x, a_{1:h})$ be the number of comparisons at prompt x in which at least one of the two recorded responses starts with the prefix $a_{1:h} = [a_1, \dots, a_h]$, so that these counts are non-increasing in h and equal the response-level counter at the final stage, i.e. $\nu(x, a_{1:H}) = N(x, a)$. The penalty $C(x, a)$ of Lemma F.1 depends on the response a only through the count $N(x, a)$, so we can evaluate it at a prefix: we write $C(x, a_{1:h})$ for the quantity obtained from $C(x, a)$ by replacing $N(x, a)$ with $\nu(x, a_{1:h})$, that is

$$C(x, a_{1:h}) := \frac{c_0\, L\, |\mathcal{A}|\, e^{2R_{\max}}\, \iota^2}{\nu(x, a_{1:h}) + |\mathcal{A}|\, L} \;+\; c_1\, e^{R_{\max}}\, \iota^{3/2} \sqrt{\frac{L\, |\mathcal{A}|^3}{\nu(x, a_{1:h}) + |\mathcal{A}|\, L}}, \tag{8}$$

so that $C(x, a_{1:H}) = C(x, a)$ because $\nu(x, a_{1:H}) = N(x, a)$. With this notation, we define the token-level penalties as

$$\begin{aligned} c^{\text{token}}(x_1, a_1) &:= C(x, a_{1:1}), \\ c^{\text{token}}(x_h, a_h) &:= C(x, a_{1:h}) - C(x, a_{1:h-1}), \quad h \ge 2. \end{aligned} \tag{9}$$

Every term is non-negative, because $\nu(x, a_{1:h})$ is non-increasing in h while both $t \mapsto 1/t$ and $t \mapsto 1/\sqrt{t}$ are decreasing, so that $h \mapsto C(x, a_{1:h})$ is non-decreasing; each term is a function of the pair (x_h, a_h) only, since the state x_h contains the prefix $a_{1:h-1}$; and the sum telescopes:

$$\sum_{h=1}^{H} c^{\text{token}}(x_h, a_h) = C(x, a_{1:H}) = C(x, a). \tag{10}$$

At this point, we introduce the token-estimated gap

$$\begin{aligned} \widehat{\Delta}^{\text{token}}(x, a, b) := \beta \sum_{h=1}^{H} \left(\min_{\ell \in [L]} \log \frac{\pi^\ell(a_h|x_h)}{\pi_{\text{ref}}(a_h|x_h)} - \max_{\ell \in [L]} \log \frac{\pi^\ell(b_h|x_h^b)}{\pi_{\text{ref}}(b_h|x_h^b)} \right) \\ - \sum_{h=1}^{H} c^{\text{token}}(x_h, a_h) - \beta \left(\max_{\ell \in [L]} \zeta^\ell(x) - \min_{\ell \in [L]} \zeta^\ell(x) \right), \end{aligned}$$

where $x_h^b := [x, b_1, \dots, b_{h-1}]$ is the prefix built from the response b. We also introduce the per-token normalizers, the accumulated normalization loss and the response-level normalization of the product

of the token-level minima,

$$Z(x_h) := \sum_{a' \in \mathcal{V}} \min_{\ell \in [L]} \pi^\ell(a'|x_h)\, e^{-c^{\text{token}}(x_h, a')/\beta} \;\leq 1, \qquad \Lambda(x,a) := \sum_{h=1}^{H} \log \frac{1}{Z(x_h)} \;\geq 0, \tag{11}$$

$$f^{\text{token}}(x,a) := \prod_{h=1}^{H} \min_{\ell \in [L]} \pi^\ell(a_h|x_h)\, e^{-c^{\text{token}}(x_h, a_h)/\beta}, \qquad \pi^{\text{prod}}_{\text{out}}(a|x) := \frac{f^{\text{token}}(x,a)}{\sum_{a' \in \mathcal{A}} f^{\text{token}}(x,a')}. \tag{12}$$

Notice that $Z(x_h) \leq 1$ because the token-level minimum is dominated by any single member, which is a probability distribution over $\mathcal{V}$, and because $c^{\text{token}} \geq 0$. These quantities will be crucial in proving the following result.

Lemma D.1. *Let* $\pi^{\text{token}}_{\text{out}}(a_h|x_h) = \frac{\min_{\ell \in [L]} \pi^\ell(a_h|x_h)\, e^{-c^{\text{token}}(x_h, a_h)/\beta}}{\sum_{a' \in \mathcal{V}} \min_{\ell \in [L]} \pi^\ell(a'|x_h)\, e^{-c^{\text{token}}(x_h, a')/\beta}}$*. Then, with probability* $1 - 4\delta$*, for any* $\beta > 0$*, it holds that*

$$J_\beta(\pi^\star) - J_\beta(\pi^{\text{prod}}_{\text{out}}) \leq \sqrt{C^\star \left\langle \pi_{\text{data}}, \text{Err}^2_{\text{token}} \right\rangle}, \tag{13}$$

$$J_\beta(\pi^\star) - J_\beta(\pi^{\text{token}}_{\text{out}}) \leq \sqrt{C^\star \left\langle \pi_{\text{data}}, \text{Err}^2_{\text{token}} \right\rangle} + \beta \left[D_{\text{KL}}(\pi^{\text{token}}_{\text{out}}, \pi^{\text{prod}}_{\text{out}}) - D_{\text{KL}}(\pi^\star, \pi^{\text{prod}}_{\text{out}}) \right], \tag{14}$$

where $\text{Err}_{\text{token}}(x,a) := \widehat{\Delta}^{\text{token}}(x,a,\pi_{\text{data}}) - \Delta_{r^\star}(x,a,\pi_{\text{data}})$.

Proof. The proof strategy is to prove that the token-level approach introduces even more pessimism. That is,

$$\widehat{\Delta}^{\text{token}}(x,a,b) \leq \widehat{\Delta}(x,a,b) \quad \forall x,a,b \in \mathcal{X} \times \mathcal{A} \times \mathcal{A}.$$

Towards this end, let us recall that

$$\widehat{\Delta}(x,a,b) = \min_{\ell \in [L]} \left(\beta \log \frac{\pi^\ell(a|x)}{\pi_{\text{ref}}(a|x)} - \beta \zeta^\ell(x) \right) - C(x,a) - \max_{\ell \in [L]} \left(\beta \log \frac{\pi^\ell(b|x)}{\pi_{\text{ref}}(b|x)} - \beta \zeta^\ell(x) \right),$$

and we start by lower bounding the first term as follows, using $\min_\ell(u_\ell + v_\ell) \geq \min_\ell u_\ell + \min_\ell v_\ell$.

$$\begin{aligned}
\min_{\ell \in [L]} \left(\beta \log \frac{\pi^\ell(a|x)}{\pi_{\text{ref}}(a|x)} - \beta \zeta^\ell(x) \right) &\geq \beta \min_{\ell \in [L]} \log \prod_{h=1}^{H} \frac{\pi^\ell(a_h|x_h)}{\pi_{\text{ref}}(a_h|x_h)} - \beta \max_{\ell \in [L]} \zeta^\ell(x) \\
&= \beta \min_{\ell \in [L]} \sum_{h=1}^{H} \log \frac{\pi^\ell(a_h|x_h)}{\pi_{\text{ref}}(a_h|x_h)} - \beta \max_{\ell \in [L]} \zeta^\ell(x) \\
&\geq \sum_{h=1}^{H} \beta \min_{\ell \in [L]} \log \frac{\pi^\ell(a_h|x_h)}{\pi_{\text{ref}}(a_h|x_h)} - \beta \max_{\ell \in [L]} \zeta^\ell(x)
\end{aligned}$$

Then, we can lower bound the second term as follows, using $-\max_\ell(u_\ell - v_\ell) \geq -\max_\ell u_\ell + \min_\ell v_\ell$.

$$\begin{aligned}
-\max_{\ell \in [L]} \left(\beta \log \frac{\pi^\ell(b|x)}{\pi_{\text{ref}}(b|x)} - \beta \zeta^\ell(x) \right) &\geq -\beta \max_{\ell \in [L]} \log \prod_{h=1}^{H} \frac{\pi^\ell(b_h|x^b_h)}{\pi_{\text{ref}}(b_h|x^b_h)} + \beta \min_{\ell \in [L]} \zeta^\ell(x) \\
&= -\beta \max_{\ell \in [L]} \sum_{h=1}^{H} \log \frac{\pi^\ell(b_h|x^b_h)}{\pi_{\text{ref}}(b_h|x^b_h)} + \beta \min_{\ell \in [L]} \zeta^\ell(x) \\
&\geq -\sum_{h=1}^{H} \beta \max_{\ell \in [L]} \log \frac{\pi^\ell(b_h|x^b_h)}{\pi_{\text{ref}}(b_h|x^b_h)} + \beta \min_{\ell \in [L]} \zeta^\ell(x)
\end{aligned}$$

Summing the two lower bounds and subtracting $\sum_h c^{\text{token}}(x_h, a_h) = C(x, a)$, which is the identity equation 10, from both sides, we can conclude that

$$\widehat{\Delta}^{\text{token}}(x, a, b) \le \widehat{\Delta}(x, a, b) \quad \forall x, a, b \in \mathcal{X} \times \mathcal{A} \times \mathcal{A}.$$

Step 2: the response-level policy $\pi_{\text{out}}^{\text{prod}}$. By the telescoping identity equation 10, the token-estimated gap decomposes as $\widehat{\Delta}^{\text{token}}(x, a, b) = \widehat{r}^{\text{token}}(x, a) - u(x, b)$, where

$$\widehat{r}^{\text{token}}(x, a) := \beta \sum_{h=1}^{H} \min_{\ell \in [L]} \log \frac{\pi^\ell(a_h|x_h)}{\pi_{\text{ref}}(a_h|x_h)} - C(x, a),$$

$$u(x, b) := \beta \sum_{h=1}^{H} \max_{\ell \in [L]} \log \frac{\pi^\ell(b_h|x_h^b)}{\pi_{\text{ref}}(b_h|x_h^b)} + \beta \left(\max_{\ell \in [L]} \zeta^\ell(x) - \min_{\ell \in [L]} \zeta^\ell(x) \right).$$

Therefore, $\widehat{\Delta}^{\text{token}}(x, a, \pi_{\text{data}}) = \widehat{r}^{\text{token}}(x, a) - \sum_b \pi_{\text{data}}(b|x) u(x, b)$ differs from $\widehat{r}^{\text{token}}(x, a)$ by a quantity which does not depend on a, and the same first order conditions used in the proof of Lemma 3.1 show that the maximizer over Π of $\left\langle \pi, \widehat{\Delta}^{\text{token}}(\cdot, \cdot, \pi_{\text{data}}) \right\rangle - \beta D_{\text{KL}}(\pi, \pi_{\text{ref}})$ is

$$\pi(a|x) \propto \pi_{\text{ref}}(a|x) \exp\left(\frac{\widehat{r}^{\text{token}}(x, a)}{\beta} \right) = \prod_{h=1}^{H} \min_{\ell \in [L]} \pi^\ell(a_h|x_h)\, e^{-c^{\text{token}}(x_h, a_h)/\beta} = f^{\text{token}}(x, a),$$

where we used that $\pi_{\text{ref}}(a|x) = \prod_h \pi_{\text{ref}}(a_h|x_h)$ and, once more, equation 10. That is, the maximizer is exactly $\pi_{\text{out}}^{\text{prod}}$ defined in equation 12. Moreover, since $\widehat{\Delta}^{\text{token}}(x, a, b) \le \widehat{\Delta}(x, a, b)$ for all x, a, b by **Step 1** and $\widehat{\Delta}(x, a, \pi_{\text{data}}) \le \Delta_{r^\star}(x, a, \pi_{\text{data}})$ for all $x, a \in \mathcal{X} \times \mathcal{A}$ with probability at least $1 - 4\delta$ according to Lemma 3.4, averaging over $b \sim \pi_{\text{data}}(\cdot|x)$ gives $\widehat{\Delta}^{\text{token}}(x, a, \pi_{\text{data}}) \le \Delta_{r^\star}(x, a, \pi_{\text{data}})$ for all x, a on the same event. The proof of Lemma 3.5 uses only these two properties of π_{out}, namely optimality for the reward $\widehat{\Delta}(\cdot, \cdot, \pi_{\text{data}})$ and pessimism of the latter. Hence, invoking it with $\widehat{\Delta}^{\text{token}}$ and $\pi_{\text{out}}^{\text{prod}}$ in place of $\widehat{\Delta}$ and π_{out} proves equation 13.

Step 3: the token-level policy $\pi_{\text{out}}^{\text{token}}$. The policy $\pi_{\text{out}}^{\text{token}}$ is not the maximizer above: it normalizes each factor of f^{token} separately, so that

$$\begin{aligned} \pi_{\text{out}}^{\text{token}}(a|x) &= \prod_{h=1}^{H} \frac{\min_{\ell \in [L]} \pi^\ell(a_h|x_h)\, e^{-c^{\text{token}}(x_h, a_h)/\beta}}{Z(x_h)} \\ &= f^{\text{token}}(x, a)\, e^{\Lambda(x, a)} = \pi_{\text{out}}^{\text{prod}}(a|x)\, F^{\text{token}}(x)\, e^{\Lambda(x, a)}, \end{aligned} \tag{15}$$

where $F^{\text{token}}(x) := \sum_{a \in \mathcal{A}} f^{\text{token}}(x, a)$. Two consequences of equation 15 will be used. First, since $\pi_{\text{out}}^{\text{token}}(\cdot|x)$ is a probability distribution, taking logarithms gives

$$\beta \log \frac{\pi_{\text{out}}^{\text{token}}(a|x)}{\pi_{\text{ref}}(a|x)} = \widehat{r}^{\text{token}}(x, a) + \beta \Lambda(x, a), \tag{16}$$

that is, $\pi_{\text{out}}^{\text{token}}$ is the exact maximizer of the regularized objective for the reward $\widehat{r}^{\text{token}} + \beta\Lambda$, which is more optimistic than $\widehat{r}^{\text{token}}$ because $\Lambda \ge 0$. Second,

$$\Lambda(x, a) = \log \frac{\pi_{\text{out}}^{\text{token}}(a|x)}{\pi_{\text{out}}^{\text{prod}}(a|x)} + \log \frac{1}{F^{\text{token}}(x)}, \tag{17}$$

so that $\Lambda(x, \cdot)$ equals a log-likelihood ratio between the two policies up to a quantity which does not depend on a.

We now repeat the proof of Lemma 3.5 with $\pi_{\text{out}}^{\text{token}}$ in place of π_{out}. The decomposition equation 27 and the bound on its first term are unchanged, because they only use pessimism of $\widehat{\Delta}^{\text{token}}(x, a, \pi_{\text{data}})$ and Cauchy–Schwarz, and give

$$\begin{aligned} \left\langle \pi^\star - \pi_{\text{out}}^{\text{token}}, r^\star \right\rangle \le & \sqrt{C^\star \left\langle \pi_{\text{data}}, \text{Err}_{\text{token}}^2 \right\rangle} \\ & + \sum_x \nu_0(x) \sum_a \left(\pi^\star(a|x) - \pi_{\text{out}}^{\text{token}}(a|x) \right) \widehat{\Delta}^{\text{token}}(x, a, \pi_{\text{data}}). \end{aligned}$$

For the last term, fix $x \in \mathcal{X}$ and abbreviate $\mathrm{KL}_x(\pi, \pi') := \sum_a \pi(a|x) \log {}^{\pi(a|x)}/_{\pi'(a|x)}$. Since $\widehat{\Delta}^{\mathrm{token}}(x, a, \pi_{\mathrm{data}})$ and $\widehat{r}^{\mathrm{token}}(x, a)$ differ by a quantity independent of a, which cancels in the difference of the two distributions, we have by equation 16

$$\begin{aligned}
&\sum_a \left(\pi^\star(a|x) - \pi^{\mathrm{token}}_{\mathrm{out}}(a|x)\right) \widehat{\Delta}^{\mathrm{token}}(x, a, \pi_{\mathrm{data}}) \\
&= \sum_a \left(\pi^\star(a|x) - \pi^{\mathrm{token}}_{\mathrm{out}}(a|x)\right) \left[\beta \log \frac{\pi^{\mathrm{token}}_{\mathrm{out}}(a|x)}{\pi_{\mathrm{ref}}(a|x)} - \beta\Lambda(x, a)\right] \\
&= \beta \left[\mathrm{KL}_x(\pi^\star, \pi_{\mathrm{ref}}) - \mathrm{KL}_x(\pi^\star, \pi^{\mathrm{token}}_{\mathrm{out}}) - \mathrm{KL}_x(\pi^{\mathrm{token}}_{\mathrm{out}}, \pi_{\mathrm{ref}})\right] \\
&\quad - \beta \sum_a \left(\pi^\star(a|x) - \pi^{\mathrm{token}}_{\mathrm{out}}(a|x)\right) \Lambda(x, a),
\end{aligned}$$

where we used $\sum_a \pi(a|x) \log {}^{\pi'(a|x)}/_{\pi''(a|x)} = \mathrm{KL}_x(\pi, \pi'') - \mathrm{KL}_x(\pi, \pi')$, valid for any three policies, with $\pi = \pi^\star$ and with $\pi = \pi^{\mathrm{token}}_{\mathrm{out}}$. We now turn the remaining sum involving Λ into KL terms. Plugging in equation 17 and using that $\log {}^1/_{F^{\mathrm{token}}(x)}$ does not depend on a, so that it is multiplied by $\sum_a \left(\pi^\star(a|x) - \pi^{\mathrm{token}}_{\mathrm{out}}(a|x)\right) = 0$, we obtain

$$\begin{aligned}
&\sum_a \left(\pi^\star(a|x) - \pi^{\mathrm{token}}_{\mathrm{out}}(a|x)\right) \Lambda(x, a) \\
&= \sum_a \left(\pi^\star(a|x) - \pi^{\mathrm{token}}_{\mathrm{out}}(a|x)\right) \left[\log \frac{\pi^{\mathrm{token}}_{\mathrm{out}}(a|x)}{\pi^{\mathrm{prod}}_{\mathrm{out}}(a|x)} + \log {}^1/_{F^{\mathrm{token}}(x)}\right] \\
&= \sum_a \pi^\star(a|x) \log \frac{\pi^{\mathrm{token}}_{\mathrm{out}}(a|x)}{\pi^{\mathrm{prod}}_{\mathrm{out}}(a|x)} - \mathrm{KL}_x(\pi^{\mathrm{token}}_{\mathrm{out}}, \pi^{\mathrm{prod}}_{\mathrm{out}}) \\
&= \mathrm{KL}_x(\pi^\star, \pi^{\mathrm{prod}}_{\mathrm{out}}) - \mathrm{KL}_x(\pi^\star, \pi^{\mathrm{token}}_{\mathrm{out}}) - \mathrm{KL}_x(\pi^{\mathrm{token}}_{\mathrm{out}}, \pi^{\mathrm{prod}}_{\mathrm{out}}),
\end{aligned}$$

where the second step uses that the terms multiplying $\log {}^1/_{F^{\mathrm{token}}(x)}$ sum to zero and the definition of KL_x, while the last one uses the same identity as above with $\pi = \pi^\star$. Substituting this expression in the previous display, the two terms $\mathrm{KL}_x(\pi^\star, \pi^{\mathrm{token}}_{\mathrm{out}})$ cancel out and we are left with

$$\begin{aligned}
&\sum_a \left(\pi^\star(a|x) - \pi^{\mathrm{token}}_{\mathrm{out}}(a|x)\right) \widehat{\Delta}^{\mathrm{token}}(x, a, \pi_{\mathrm{data}}) \\
&= \beta \left[\mathrm{KL}_x(\pi^\star, \pi_{\mathrm{ref}}) - \mathrm{KL}_x(\pi^{\mathrm{token}}_{\mathrm{out}}, \pi_{\mathrm{ref}})\right] \\
&\quad + \beta \left[\mathrm{KL}_x(\pi^{\mathrm{token}}_{\mathrm{out}}, \pi^{\mathrm{prod}}_{\mathrm{out}}) - \mathrm{KL}_x(\pi^\star, \pi^{\mathrm{prod}}_{\mathrm{out}})\right].
\end{aligned}$$

Averaging over $x \sim \nu_0$, the first bracket is exactly $\beta D_{\mathrm{KL}}(\pi^\star, \pi_{\mathrm{ref}}) - \beta D_{\mathrm{KL}}(\pi^{\mathrm{token}}_{\mathrm{out}}, \pi_{\mathrm{ref}})$, which combined with the display above and the definition of J_β gives equation 14. □

E DETAILED COMPARISON WITH χ^2PO AND DPO+SFT IN THE TABULAR SETTING.

Looking at the bound in Huang et al. (2024, Theorem 3.1), we notice that χ^2PO is upper bounded by $\mathcal{O}(e^{2R_{\max}}\sqrt{\frac{C^\star \log(|\Pi|/\delta)}{N}})$. When we are in the tabular case, we need to consider $|\Pi|$ to be a $1/N$-covering set of all Markov policies, i.e. all mappings $\mathcal{X} \to \Delta_{\mathcal{A}}$. The size of this set can be computed to be $N^{|\mathcal{X}||\mathcal{A}|}$. Therefore plugging into the χ^2PO bound we obtain $\mathcal{O}(e^{2R_{\max}}\sqrt{\frac{C^\star |\mathcal{X}||\mathcal{A}| \log(N/\delta)}{N}})$ which also includes polynomial dependence on the number of states and actions as our bound does. The χ^2PO bound has a better action dependence, but it requires knowledge of $C^\star$. When unavailable, the χ^2PO bound becomes

$$\mathcal{O}\left(e^{2R_{\max}}\sqrt{\frac{(C^\star)^2|\mathcal{X}||\mathcal{A}| \log(N/\delta)}{N}}\right),$$

which would lead to a sample complexity of order

$$\widetilde{\mathcal{O}}\left(e^{4R_{\max}}\frac{(C^\star)^2|\mathcal{X}||\mathcal{A}| \log(1/\delta)}{\epsilon^2}\right)$$

which has a worse dependence on $C^\star$ compared to our analysis. Moreover, the χ^2PO bound requires knowledge of π_{data}.

Looking at the bound in Liu et al. (2024, Theorem 5.3), we have that in the tabular setting the covering number for the reward class scales as $N^{|\mathcal{X}||\mathcal{A}|}$. Therefore, in the tabular setting $\iota = \mathcal{O}(\sqrt{|\mathcal{X}||\mathcal{A}| \log(N/\delta)})$. Plugging this into the main bound of Liu et al. (2024, Theorem 5.3), assuming that π_{ref} (or π_{base} in their notation) equals π_{data} and neglecting the lower order terms we obtain an upper bound of order $\frac{(C^\star)^2 e^{2R_{\max}}\sqrt{|\mathcal{X}||\mathcal{A}| \log(N/\delta)}}{\sqrt{N}}$ which leads to a sample complexity of order

$$\widetilde{\mathcal{O}}\left(e^{4R_{\max}}\frac{(C^\star)^4|\mathcal{X}||\mathcal{A}| \log(1/\delta)}{\epsilon^2}\right).$$

Therefore, again the action dependence is better but the dependence on $C^\star$ is worse than the one attained by PEPO, even without knowledge of π_{data}. Moreover, our analysis achieves a fast rate in the large β regime when $C^\star$ is comparable to $C^\star_\infty$. It is unclear if χ^2PO and DPO+SFT can attain the same.

F ANALYSIS OF PEPO

This section presents the analysis of PEPO.

F.1 PROOF OF LEMMA 3.1: CLOSED FORM SOLUTION OF EQUATION (7)

Proof. The first order optimality conditions for Equation (7) give

$$r^-(x,a) = \beta \log\left(\frac{\pi_{\text{out}}(a|x)}{\pi_{\text{ref}}(a|x)}\right) + \beta \log Z(x)$$

where $Z(x)$ needs to be chosen to guarantee normalization of π_{out}. Plugging in the definition of $r^-(x,a)$, we obtain that

$$\beta \min_{\ell\in[L]}\left(\log\frac{\tilde{\pi}^\ell(a|x)}{\pi_{\text{ref}}(a|x)} - m^\ell(x,a)\right) = \beta\log\left(\frac{\pi_{\text{out}}(a|x)}{\pi_{\text{ref}}(a|x)}\right) + \beta\log Z(x)$$
$$\implies \min_{\ell\in[L]}\frac{\tilde{\pi}^\ell(a|x)\exp\left(-m^\ell(x,a)\right)}{\pi_{\text{ref}}(a|x)} = \frac{\pi_{\text{out}}(a|x)Z(x)}{\pi_{\text{ref}}(a|x)}.$$

Then, choosing $Z(x)$ to ensure that $\sum_{a\in\mathcal{A}}\pi_{\text{out}}(a|x) = 1$ for all $x\in\mathcal{X}$, we obtain that

$$\pi_{\text{out}}(a|x) = \frac{\min_{\ell\in[L]}\tilde{\pi}^\ell(a|x)\exp\left(-m^\ell(x,a)\right)}{\sum_{a\in\mathcal{A}}\min_{\ell\in[L]}\tilde{\pi}^\ell(a|x)\exp\left(-m^\ell(x,a)\right)}.$$

□

F.2 PROOF OF LEMMA 3.4: ESTABLISHING PESSIMISM

Lemma 3.4. *Let us denote* $\sum_{b\in\mathcal{A}}\pi_{\text{data}}(b|x)\Delta(x,a,b)$ *as* $\Delta(x,a,\pi_{\text{data}})$ *for* $\Delta\in\{\Delta_{r^\star},\widehat{\Delta}\}$ *and let us assume that* $L \geq \frac{7\log(|\mathcal{X}||\mathcal{A}|^2/\delta)}{2}$. *Then, the policies* $\{\tilde{\pi}^\ell\}_{\ell=1}^L$, *solutions of equation 6 together with the penalty* $C(x,a)$ *of Theorem 3.3, induce an estimated gap* $\widehat{\Delta}$ *such that with probability at least* $1-4\delta$,

$$\widehat{\Delta}(x,a,\pi_{\text{data}}) \leq \Delta_{r^\star}(x,a,\pi_{\text{data}}) \;\; \forall x,a\in\mathcal{X}\times\mathcal{A}.$$

Proof. The proof is rather long and is divided into three steps.

1. Let us recall the convention $0/0 = 0$ and define the *inconsistency slack*

$$\text{slack}^\ell(x,a,b) := \sqrt{\frac{24\,e^{2R_{\max}}\,L\,D_x^\ell}{N(x,a,b)+L}} \quad \text{if } N^\ell(x,a,b) > 0, \quad \text{and 0 otherwise,}$$

where we used D_x^ℓ defined as follows

$$D_x^\ell := \sum_{a,b:\,N^\ell(x,a,b)>0}\left[\underbrace{\max_{|u|\leq 2R_{\max}} F_{x,a,b}^\ell(\sigma(u))}_{\text{best attainable without transitivity}} - \underbrace{F_{x,a,b}^\ell\left(\sigma\left(\beta\Delta_{\log\tilde{\pi}^\ell/\pi_{\text{ref}}}(x,a,b)\right)\right)}_{\text{best attainable by a policy-induced sigmoid}}\right] \geq 0, \tag{18}$$

where for any $x\in\mathcal{X}$ $a,b\in\mathcal{A}$ and $\ell\in[L]$, $F_{x,a,b}^\ell : [0,1]\to\mathbb{R}$ is defined as

$$F_{x,a,b}^\ell(p) = \log(p)N^\ell(x,a\succ b) + \log(1-p)N^\ell(x,b\succ a).$$

Then, we show that for every prompt x and every pair with $N^\ell(x,a,b) > 0$, the fitted margin is the clipped empirical log-odds up to the inconsistency slack,

$$\left|\beta\Delta_{\log\tilde{\pi}^\ell/\pi_{\text{ref}}}(x,a,b) - \left[\sigma^{-1}\left(\frac{N^\ell(x,a\succ b)}{N^\ell(x,a,b)}\right)\right]_{-2R_{\max}}^{2R_{\max}}\right| \leq \text{slack}^\ell(x,a,b),$$

where $[z]_{-M}^{M} := \max\{-M,\min\{M,z\}\}$, with equality of the two margins whenever $D_x^\ell = 0$. For the pairs with $N^\ell(x,a,b) = 0$, we instead set by convention $\left[\sigma^{-1}\left(N^\ell(x,a\succ b)/N^\ell(x,a,b)\right)\right]_{-2R_{\max}}^{2R_{\max}} := 2R_{\max}$, so that the one-sided bound $\beta\Delta_{\log\tilde{\pi}^\ell/\pi_{\text{ref}}}(x,a,b) \leq \left[\sigma^{-1}\left(N^\ell(x,a\succ b)/N^\ell(x,a,b)\right)\right]_{-2R_{\max}}^{2R_{\max}} + \text{slack}^\ell(x,a,b)$ holds for every pair, trivially for the unobserved ones since $\tilde{\pi}^\ell\in\overline{\Pi}$.

2. Let us define the *shrinkage gap*

$$\gamma(x,a,b) := \max_{\ell\in[L]} \left(\left[\sigma^{-1}\left(\frac{N^\ell(x,a\succ b)}{N^\ell(x,a,b)}\right)\right]_{-2R_{\max}}^{2R_{\max}} - \left[\sigma^{-1}\left(\frac{N^\ell(x,a\succ b)}{N^\ell(x,a,b)+2}\right)\right]_{-2R_{\max}}^{2R_{\max}} \right).$$

Notice that $\gamma(x,a,b)\geq 0$ and it accounts for the fact that DPO fits the empirical frequency $N^\ell(x,a\succ b)/N^\ell(x,a,b)$ while the ensemble certifies pessimism only for the shrunk frequency $N^\ell(x,a\succ b)/\big(N^\ell(x,a,b)+2\big)$. Then, with probability at least $1-\delta$, simultaneously for all x,a,b,

$$\min_{\ell\in[L]} \left[\sigma^{-1}\left(\frac{N^\ell(x,a\succ b)}{N^\ell(x,a,b)}\right)\right]_{-2R_{\max}}^{2R_{\max}} \;\leq\; \Delta_{r^\star}(x,a,b)+\gamma(x,a,b);$$

3. on the same event, if the bonus satisfies, for all x,a,

$$C(x,a) \;\geq\; \sum_{b\in\mathcal{A}} \pi_{\mathrm{data}}(b|x)\left[\gamma + \max_{\ell\in[L]} \mathrm{slack}^\ell\right](x,a,b),$$

then the estimated gap satisfies $\widehat{\Delta}(x,a,\pi_{\mathrm{data}})\leq \Delta_{r^\star}(x,a,\pi_{\mathrm{data}})$ for all $x,a\in\mathcal{X}\times\mathcal{A}$. The explicit choice for $C(x,a)$ stated in Theorem 3.3 enjoys this property with probability at least $1-3\delta$ by Lemma F.1, so that the conclusion holds with probability at least $1-4\delta$ overall.

Step 1 Let us denote $\widehat{\mathbb{P}}^\ell(x,a\succ b)$ to be the unconstrained MLE estimator for the winning probability of a over b in response to x. We now show that

$$\widehat{\mathbb{P}}^\ell(x,a\succ b) = \frac{N^\ell(x,a\succ b)}{N^\ell(x,a,b)} \quad \text{if } N^\ell(x,a,b)>0,$$

and arbitrary otherwise.

$$\begin{aligned}
&\sum_{X,A^+,A^-\in\mathcal{D}^\ell} \log\left(\widehat{\mathbb{P}}^\ell(X,A^+\succ A^-)\right)\\
&= \sum_{X,A^+,A^-\in\mathcal{D}^\ell}\sum_{x,a\in\mathcal{X}\times\mathcal{A}}\sum_{b\succ a} \log\left(\widehat{\mathbb{P}}^\ell(x,a\succ b)\right)\mathbb{1}_{\{X,A^+,A^-=x,a,b\}}\\
&\quad + \log\left(\widehat{\mathbb{P}}^\ell(x,b\succ a)\right)\mathbb{1}_{\{X,A^+,A^-=x,b,a\}}\\
&= \sum_{x,a\in\mathcal{X}\times\mathcal{A}}\sum_{b\succ a} \underbrace{\log\left(\widehat{\mathbb{P}}^\ell(x,a\succ b)\right)N^\ell(x,a\succ b) + \log\left(1-\widehat{\mathbb{P}}^\ell(x,a\succ b)\right)N^\ell(x,b\succ a)}_{:=F^\ell_{x,a,b}(\widehat{\mathbb{P}}^\ell(x,a\succ b))},
\end{aligned}$$

where we use the notation $b\succ a$ to indicate all actions b that follow action a for some canonical ordering of the action space. Then, noticing that $F(\cdot)$ is concave, taking the first derivative and equating it to zero, we obtain that the unconstrained maximizer $\widehat{\mathbb{P}}^\ell(x,a\succ b)$ satisfies

$$\frac{N^\ell(x,a\succ b)}{\widehat{\mathbb{P}}^\ell(x,a\succ b;\tilde{\pi}^\ell)} = \frac{N^\ell(x,b\succ a)}{1-\widehat{\mathbb{P}}^\ell(x,a\succ b;\tilde{\pi}^\ell)} \quad \forall\; x,a,b\in\mathcal{X}\times\mathcal{A}\times\mathcal{A},$$

where we defined the counts $N^\ell(x,a,b)=\sum_{X,A^+,A^-\in\mathcal{D}^\ell}\mathbb{1}_{\{X,A^+,A^-=x,a,b\}}+\mathbb{1}_{\{X,A^+,A^-=x,b,a\}}$ and $N^\ell(x,a\succ b)=\sum_{X,A^+,A^-\in\mathcal{D}^\ell}\mathbb{1}_{\{X,A^+,A^-=x,a,b\}}$. Rearranging yields the result.

At this point, we notice that $\tilde{\pi}^\ell$ is the projection of $\widehat{\mathbb{P}}^\ell(x,a\succ b)$ on the space of winning distributions that can be represented as $\sigma(\beta\Delta_{\log\pi/\pi_{\mathrm{ref}}})$ for some $\pi\in\overline{\Pi}$. Towards characterizing this object we introduce,

$$u^\star(x,a,b) = \left[\sigma^{-1}(\widehat{\mathbb{P}}^\ell(x,a\succ b))\right]_{-2R_{\max}}^{2R_{\max}}$$

where $[\cdot]_l^u := \max\{\min\{\cdot,u\},l\}$. The projection of $\widehat{\mathbb{P}}^\ell(x,a\succ b)$ on the space of probabilities induced by $\overline{\Pi}$ however is not $u^\star$ because, given the constraints, the DPO maximum likelihood problem does not decompose over each state-action-action triplet. To account for the difference between $u^\star$

and $\Delta_{\log \tilde{\pi}^\ell/\pi_{\text{ref}}}$ (where $\tilde{\pi}^\ell$ is the DPO solution over the dataset $\mathcal{D}^\ell$) we can use strong concavity of F. First of all, notice that, since $u \mapsto F^\ell_{x,a,b}(\sigma(u))$ is concave with unconstrained maximizer $\sigma^{-1}(\widehat{\mathbb{P}}^\ell(x, a \succ b))$, its maximizer over $[-2R_{\max}, 2R_{\max}]$ is precisely $u^\star(x, a, b)$, so that the first underbraced term of equation 18 equals $F^\ell_{x,a,b}(\sigma(u^\star(x, a, b)))$.

On $[-2R_{\max}, 2R_{\max}]$, the same map has strictly negative second derivative

$$-\frac{\mathrm{d}^2}{\mathrm{d}u^2} F^\ell_{x,a,b}(\sigma(u)) = N^\ell(x, a, b)\, \sigma(u)\, (1 - \sigma(u))$$
$$= \frac{N^\ell(x, a, b)}{\left(e^{u/2} + e^{-u/2}\right)^2} \;\geq\; \frac{N^\ell(x, a, b)}{4} e^{-2R_{\max}} =: \mu^\ell(x, a, b),$$

so $F^\ell_{x,a,b}(\sigma(\cdot))$ is $\mu^\ell(x, a, b)$-strongly concave there. Since $u^\star(x, a, b)$ maximizes it over that interval, and since $\beta\Delta_{\log \tilde{\pi}^\ell/\pi_{\text{ref}}}(x, a, b)$ belongs to the same interval because $\tilde{\pi}^\ell \in \overline{\Pi}$, the first order optimality conditions give

$$\frac{\mathrm{d}}{\mathrm{d}u} F^\ell_{x,a,b}(\sigma(u))\Big|_{u=u^\star(x,a,b)} \left(\beta\Delta_{\log \tilde{\pi}^\ell/\pi_{\text{ref}}}(x, a, b) - u^\star(x, a, b)\right) \;\leq\; 0,$$

hence, abbreviating $\widehat{u}^\ell(x, a, b) := \beta\Delta_{\log \tilde{\pi}^\ell/\pi_{\text{ref}}}(x, a, b)$,

$$\begin{aligned} F^\ell_{x,a,b}\left(\sigma\left(\widehat{u}^\ell(x, a, b)\right)\right) \;&\leq\; F^\ell_{x,a,b}\left(\sigma\left(u^\star(x, a, b)\right)\right) \\ &\quad + \frac{\mathrm{d}}{\mathrm{d}u} F^\ell_{x,a,b}(\sigma(u))\Big|_{u=u^\star(x,a,b)} \left(\widehat{u}^\ell(x, a, b) - u^\star(x, a, b)\right) \\ &\quad - \frac{\mu^\ell(x, a, b)}{2} \left(\widehat{u}^\ell(x, a, b) - u^\star(x, a, b)\right)^2 \\ &\leq\; F^\ell_{x,a,b}\left(\sigma\left(u^\star(x, a, b)\right)\right) - \frac{\mu^\ell(x, a, b)}{2} \left(\widehat{u}^\ell(x, a, b) - u^\star(x, a, b)\right)^2. \end{aligned}$$

Summing over the pairs observed at x in the fold ℓ and recalling the definition equation 18 of D^ℓ_x,

$$\sum_{a,b:\, N^\ell(x,a,b)>0} \frac{\mu^\ell(x, a, b)}{2} \left(\widehat{u}^\ell(x, a, b) - u^\star(x, a, b)\right)^2 \;\leq\; D^\ell_x,$$

so that, pair by pair,

$$\left|\widehat{u}^\ell(x, a, b) - u^\star(x, a, b)\right| \;\leq\; \sqrt{\frac{2D^\ell_x}{\mu^\ell(x, a, b)}} \;=\; \sqrt{\frac{8e^{2R_{\max}} D^\ell_x}{N^\ell(x, a, b)}}.$$

Finally, the partition property stated in Section 3.4 gives $N^\ell(x, a, b) \geq \max\{1, {}^{N(x,a,b)}/_{L} - 1\} \geq {}^{(N(x,a,b)+L)}/_{(3L)}$ for every pair observed in the fold ℓ, so that the right hand side above is at most $\text{slack}^\ell(x, a, b)$. This proves the first claim, since $u^\star(x, a, b) = \left[\sigma^{-1}\left({}^{N^\ell(x,a\succ b)}/_{N^\ell(x,a,b)}\right)\right]_{-2R_{\max}}^{2R_{\max}}$ by Step 1 and the definition of $u^\star$.

In particular, $D^\ell_x = 0$ forces $\widehat{u}^\ell(x, a, b) = u^\star(x, a, b)$ for every pair observed in the fold ℓ, i.e. DPO reproduces the clipped empirical frequencies exactly. This is the case, for instance, whenever the pairs compared at x are vertex-disjoint: the profile $\{u^\star(x, a, b)\}_{a,b}$ is then realized inside $\overline{\Pi}$ by the policy with $\beta \log {}^{\pi(c|x)}/_{\pi_{\text{ref}}(c|x)}$ equal to $\pm {}^{u^\star(x,a,b)}/_{2}$ on the two responses of each compared pair and to 0 on every other response, whose margins are bounded by $2R_{\max}$ within a compared pair, by $2R_{\max}$ across two compared pairs and by $R_{\max}$ between a compared and an uncompared response.

Step 2 In order to prove the second step in our proof roadmap, we invoke Cassel et al. (2025, Lemma 3), which gives that for any fixed state $x \in \mathcal{X}$, actions $a, b \in \mathcal{A} \times \mathcal{A}$ and any possible value of $N^\ell(x, a, b)$, it holds that

$$\mathbb{P}\left[\min_{\ell \in [L]} \frac{N^\ell(x, a \succ b)}{N^\ell(x, a, b) + 2} > \mathbb{P}^{\text{true}}_{r^\star}(a \succ b|x)\right] \leq e^{-2L/7}.$$

Therefore, choosing $L \geq \frac{7\log(|\mathcal{X}||\mathcal{A}|^2/\delta)}{2}$, we obtain that

$$\mathbb{P}\left[\exists x,a,b \in \mathcal{X}\times\mathcal{A}\times\mathcal{A} \;\; \min_{\ell\in[L]} \frac{N^\ell(x, a\succ b)}{N^\ell(x,a,b)+2} > \mathbb{P}^{\text{true}}_{r^\star}(a\succ b|x)\right] \leq \delta.$$

Then, passing to the complementary event and using that $\left[\sigma^{-1}(\cdot)\right]^{2R_{\max}}_{-2R_{\max}}$ is a non-decreasing function, together with $\left[\sigma^{-1}\left(\mathbb{P}^{\text{true}}_{r^\star}(a\succ b|x)\right)\right]^{2R_{\max}}_{-2R_{\max}} = \Delta_{r^\star}(x,a,b)$, which holds because $|\Delta_{r^\star}(x,a,b)| \leq R_{\max} \leq 2R_{\max}$, we obtain that

$$\mathbb{P}\left[\min_{\ell\in[L]}\left[\sigma^{-1}\left(\frac{N^\ell(x,a\succ b)}{N^\ell(x,a,b)+2}\right)\right]^{2R_{\max}}_{-2R_{\max}} \leq \Delta_{r^\star}(x,a,b) \quad \forall x,a,b\in\mathcal{X}\times\mathcal{A}\times\mathcal{A}\right] \geq 1-\delta. \tag{19}$$

The quantity controlled in equation 19 is built from the *shrunk* frequencies $N^\ell(x,a\succ b)/(N^\ell(x,a,b)+2)$, whereas **Step 1** characterizes the trained margins in terms of the *empirical* frequencies $N^\ell(x,a\succ b)/N^\ell(x,a,b)$, which are larger. The difference between the two is exactly what the shrinkage gap $\gamma(x,a,b)$ measures: by its definition, for every $\ell\in[L]$,

$$\left[\sigma^{-1}\left(\frac{N^\ell(x,a\succ b)}{N^\ell(x,a,b)}\right)\right]^{2R_{\max}}_{-2R_{\max}} \leq \left[\sigma^{-1}\left(\frac{N^\ell(x,a\succ b)}{N^\ell(x,a,b)+2}\right)\right]^{2R_{\max}}_{-2R_{\max}} + \gamma(x,a,b).$$

Taking the minimum over $\ell\in[L]$ on both sides, since $\gamma(x,a,b)$ does not depend on ℓ, and combining with equation 19, we conclude that with probability at least $1-\delta$

$$\min_{\ell\in[L]}\left[\sigma^{-1}\left(\frac{N^\ell(x,a\succ b)}{N^\ell(x,a,b)}\right)\right]^{2R_{\max}}_{-2R_{\max}} \leq \Delta_{r^\star}(x,a,b)+\gamma(x,a,b) \qquad \forall x,a,b\in\mathcal{X}\times\mathcal{A}\times\mathcal{A},$$

which is the second step of our proof roadmap. **Step 3**

$$\begin{aligned}
\widehat{\Delta}(x,a,b) &= \min_{\ell\in[L]}\left(\beta\log\frac{\tilde{\pi}^\ell(a|x)}{\pi_{\text{ref}}(a|x)} - \beta\zeta^\ell(x)\right) - \max_{\ell'\in[L]}\left(\beta\log\frac{\tilde{\pi}^{\ell'}(b|x)}{\pi_{\text{ref}}(b|x)} - \beta\zeta^{\ell'}(x)\right) - C(x,a)\\
&\leq \min_{\ell\in[L]}\beta\left(\log\frac{\tilde{\pi}^\ell(a|x)}{\pi_{\text{ref}}(a|x)} - \log\frac{\tilde{\pi}^\ell(b|x)}{\pi_{\text{ref}}(b|x)}\right) - C(x,a)\\
&\leq \min_{\ell\in[L]}\left[\sigma^{-1}\left(\frac{N^\ell(x,a\succ b)}{N^\ell(x,a,b)}\right)\right]^{2R_{\max}}_{-2R_{\max}} + \max_{\ell\in[L]}\text{slack}^\ell(x,a,b) - C(x,a)\\
&\leq \Delta_{r^\star}(x,a,b) + \gamma(x,a,b) + \max_{\ell\in[L]}\text{slack}^\ell(x,a,b) - C(x,a),
\end{aligned}$$

where the second inequality uses **Step 1** in its one-sided form, which holds for every pair thanks to the convention adopted for the unobserved pairs, and the last one holds with probability at least $1-\delta$ by virtue of **Step 2**. Then, by monotonicity of expectation we have that for any $x,a\in\mathcal{X}\times\mathcal{A}$,

$$\begin{aligned}
\widehat{\Delta}(x,a,\pi_{\text{data}}) &\leq \Delta_{r^\star}(x,a,\pi_{\text{data}}) + \sum_{b\in\mathcal{A}}\pi_{\text{data}}(b|x)\left[\gamma + \max_{\ell\in[L]}\text{slack}^\ell\right](x,a,b) - C(x,a)\\
&\leq \Delta_{r^\star}(x,a,\pi_{\text{data}}).
\end{aligned}$$

The last step follows from the assumption on the penalty $C(x,a)$ and it concludes the proof. □

F.3 Properties of the penalty

The proof of Lemma 3.4 left the penalty $C(x,a)$ implicit: it only requires that $C(x,a)$ dominates the π_{data}-average of the shrinkage gaps $\gamma(x,a,b)$ and of the inconsistency slacks $\text{slack}^\ell(x,a,b)$, both defined in Lemma 3.4. We now verify that the explicit choice used in Theorem 3.3 has this property, and that it is small enough not to dominate the final rate. Throughout this section we abbreviate the logarithmic factor

$$\iota := \log\left(\frac{2NL\,|\mathcal{X}|\,|\mathcal{A}|^2}{\delta}\right),$$

which depends on the size of the problem and on δ only, and not on x, a, b or ℓ.

Lemma F.1. *Let the penalty be*

$$C(x,a) := \frac{c_0\, L\, |\mathcal{A}|\, e^{2R_{\max}}\, \iota^2}{N(x,a) + |\mathcal{A}|\, L} \;+\; c_1\, e^{R_{\max}}\, \iota^{3/2} \sqrt{\frac{L\, |\mathcal{A}|^3}{N(x,a) + |\mathcal{A}|\, L}}$$

for two large enough absolute constants c_0 and c_1. Then the following three statements hold.

1. *With probability at least $1-\delta$, for every prompt $x \in \mathcal{X}$ and every $\ell \in [L]$,*

$$D_x^\ell \;\le\; \frac{|\mathcal{A}|^2}{2}\, \iota.$$

2. *With probability at least $1-3\delta$, for every $x \in \mathcal{X}$ and every $a \in \mathcal{A}$,*

$$C(x,a) \;\ge\; \sum_{b \in \mathcal{A}} \pi_{\mathrm{data}}(b|x) \left[\gamma(x,a,b) + \max_{\ell \in [L]} \mathrm{slack}^\ell(x,a,b) \right].$$

3. *With probability at least $1-4\delta$,*

$$\sum_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}} \nu_0(x) \pi_{\mathrm{data}}(a|x)\, C(x,a)^2 \;\le\; \widetilde{\mathcal{O}}\left(\frac{e^{4R_{\max}}\, |\mathcal{A}|^4\, |\mathcal{X}|\; L\, \log(1/\delta)}{N} \right).$$

Proof. The proof is divided into five steps. The first one bounds the inconsistency D_x^ℓ, the second and the third one relate the counts appearing in $\gamma(x,a,b)$ and in $\mathrm{slack}^\ell(x,a,b)$ to the quantity $N(x,a)$ appearing in $C(x,a)$, the fourth one bounds $\gamma(x,a,b)$ and proves the domination property and the fifth one the bound on the squared penalty.

Step 1: the inconsistency D_x^ℓ is at most $|\mathcal{A}|^2\iota/2$. The ground truth reward $r^\star$ induces the margins $\Delta_{r^\star}(x,a,b)$, which satisfy $|\Delta_{r^\star}(x,a,b)| \le R_{\max} \le 2R_{\max}$ for every $x,a,b \in \mathcal{X} \times \mathcal{A} \times \mathcal{A}$. Therefore the policy inducing them belongs to $\overline{\Pi}$ for *every* comparison graph, and since $\tilde{\pi}^\ell$ maximizes $\sum_{a,b} F^\ell_{x,a,b}(\sigma(\cdot))$ over $\overline{\Pi}$,

$$\sum_{a,b:\, N^\ell(x,a,b)>0} F^\ell_{x,a,b}\left(\sigma\left(\beta \Delta_{\log \tilde{\pi}^\ell/\pi_{\mathrm{ref}}}(x,a,b)\right)\right) \;\ge\; \sum_{a,b:\, N^\ell(x,a,b)>0} F^\ell_{x,a,b}\left(\mathbb{P}^{\mathrm{true}}_{r^\star}(a \succ b|x)\right),$$

where we used that $\sigma(\Delta_{r^\star}(x,a,b)) = \mathbb{P}^{\mathrm{true}}_{r^\star}(a \succ b|x)$. Moreover, the maximum over $|u| \le 2R_{\max}$ in equation 18 is at most the maximum over all $u \in \mathbb{R}$, which by **Step 1** of the proof of Lemma 3.4 is attained at $\widehat{\mathbb{P}}^\ell(x, a \succ b)$. Plugging both facts into equation 18,

$$D_x^\ell \;\le\; \sum_{a,b:\, N^\ell(x,a,b)>0} \left(N^\ell(x, a \succ b) \log \frac{\widehat{\mathbb{P}}^\ell(x, a \succ b)}{\mathbb{P}^{\mathrm{true}}_{r^\star}(a \succ b|x)} + N^\ell(x, b \succ a) \log \frac{1 - \widehat{\mathbb{P}}^\ell(x, a \succ b)}{1 - \mathbb{P}^{\mathrm{true}}_{r^\star}(a \succ b|x)} \right).$$

Conditionally on the counts, $N^\ell(x, a \succ b)$ is a Binomial random variable with $N^\ell(x,a,b)$ trials and success probability $\mathbb{P}^{\mathrm{true}}_{r^\star}(a \succ b|x)$, and the summand above is exactly $N^\ell(x,a,b)$ times the Kullback–Leibler divergence between the empirical frequency $\widehat{\mathbb{P}}^\ell(x, a \succ b)$ and $\mathbb{P}^{\mathrm{true}}_{r^\star}(a \succ b|x)$. The method of types (see Lemma F.6) gives, for every $t \ge 0$, that this summand exceeds t with probability at most $(N^\ell(x,a,b)+1)e^{-t}$. Choosing $t = \iota$ and taking a union bound over the at most $L\,|\mathcal{X}|\,|\mathcal{A}|^2$ quadruples (ℓ, x, a, b), we obtain that with probability at least $1-\delta$ every summand is at most ι, simultaneously for all ℓ and x. Since the number of unordered pairs a, b with $N^\ell(x,a,b) > 0$ is at most $\binom{|\mathcal{A}|}{2} \le |\mathcal{A}|^2/2$, the first claim follows.

Step 2: the penalty is large enough to induce pessimism The proof of the second statement is split into three steps.

Step 2.1: the pair counts $N(x,a,b)$ are not much smaller than their expectation. Each element of $\mathcal{D}$ records a comparison of the unordered pair a, b at prompt x with probability at least $\nu_0(x)\pi_{\mathrm{data}}(a|x)\pi_{\mathrm{data}}(b|x)$. Applying Lemma F.4 to the indicators of this event, which are non-negative and bounded by 1, and taking a union bound over the at most $|\mathcal{X}|\,|\mathcal{A}|^2$ triples x, a, b, we obtain that with probability at least $1-\delta$

$$N\, \nu_0(x)\pi_{\mathrm{data}}(a|x)\pi_{\mathrm{data}}(b|x) \;\le\; 2N(x,a,b) + 4\iota \qquad \forall\, x,a,b \in \mathcal{X} \times \mathcal{A} \times \mathcal{A}. \tag{20}$$

We claim that this implies

$$\frac{1}{N(x,a,b)+L} \;\le\; \frac{9\,\iota}{N\,\nu_0(x)\pi_{\text{data}}(a|x)\pi_{\text{data}}(b|x)+L} \qquad \forall\, x,a,b \in \mathcal{X}\times\mathcal{A}\times\mathcal{A}. \tag{21}$$

Indeed, if $N\nu_0(x)\pi_{\text{data}}(a|x)\pi_{\text{data}}(b|x) \ge 8\iota$, then equation 20 gives $2N(x,a,b) \ge {N\nu_0(x)\pi_{\text{data}}(a|x)\pi_{\text{data}}(b|x)}/{2}$, hence $N(x,a,b)+L \ge {(N\nu_0(x)\pi_{\text{data}}(a|x)\pi_{\text{data}}(b|x)+L)}/{4}$; and if $N\nu_0(x)\pi_{\text{data}}(a|x)\pi_{\text{data}}(b|x) < 8\iota$, then $N\nu_0(x)\pi_{\text{data}}(a|x)\pi_{\text{data}}(b|x)+L < 8\iota+L \le 9\iota L$, while $N(x,a,b)+L \ge L$.

Step 2.2: the sum over the opponents. Each element of $\mathcal{D}$ records a comparison involving the response a at prompt x with probability at most $2\nu_0(x)\pi_{\text{data}}(a|x)$. Applying Lemma F.5 to the indicators of this event, which are non-negative and bounded by 1, and taking a union bound over the $|\mathcal{X}|\,|\mathcal{A}|$ pairs x,a, we obtain that with probability at least $1-\delta$

$$N(x,a) \;\le\; 4N\,\nu_0(x)\pi_{\text{data}}(a|x) + 4\iota \qquad \forall\, x,a \in \mathcal{X}\times\mathcal{A},$$

and, by the same case distinction as in **Step 2.1**,

$$\frac{1}{N\,\nu_0(x)\pi_{\text{data}}(a|x)+|\mathcal{A}|\,L} \;\le\; \frac{9\,\iota}{N(x,a)+|\mathcal{A}|\,L} \qquad \forall\, x,a \in \mathcal{X}\times\mathcal{A}. \tag{22}$$

Now considering that the intersection of the events of **Step 2.1** and **Step 2.2** holds, we average under $\pi_{\text{data}}(\cdot|x)$ the action b on both sides of equation 21,

$$\sum_{b\in\mathcal{A}} \frac{\pi_{\text{data}}(b|x)}{N(x,a,b)+L} \;\le\; 9\,\iota \sum_{b\in\mathcal{A}} \frac{\pi_{\text{data}}(b|x)}{N\,\nu_0(x)\pi_{\text{data}}(a|x)\,\pi_{\text{data}}(b|x)+L}. \tag{23}$$

For every fixed x,a, the function $u \mapsto {u}/{(N\nu_0(x)\pi_{\text{data}}(a|x)u+L)}$ is concave on $[0,\infty)$. Since the $|\mathcal{A}|$ numbers $\{\pi_{\text{data}}(b|x)\}_{b\in\mathcal{A}}$ are non-negative and sum to one, Jensen's inequality applied to this function gives

$$\begin{aligned}
\sum_{b\in\mathcal{A}} \frac{\pi_{\text{data}}(b|x)}{N\,\nu_0(x)\pi_{\text{data}}(a|x)\,\pi_{\text{data}}(b|x)+L} \;&\le\; |\mathcal{A}| \cdot \frac{{1}/{|\mathcal{A}|}\sum_{b\in\mathcal{A}}\pi_{\text{data}}(b|x)}{{N\nu_0(x)\pi_{\text{data}}(a|x)}/{|\mathcal{A}|}\sum_{b\in\mathcal{A}}\pi_{\text{data}}(b|x)+L} \\
&= |\mathcal{A}| \cdot \frac{{1}/{|\mathcal{A}|}}{{N\nu_0(x)\pi_{\text{data}}(a|x)}/{|\mathcal{A}|}+L} \\
&= \frac{|\mathcal{A}|}{N\,\nu_0(x)\pi_{\text{data}}(a|x)+|\mathcal{A}|\,L} \qquad (24) \\
&\le \frac{9\,\iota\;|\mathcal{A}|}{N(x,a)+|\mathcal{A}|\,L}. \qquad (25)
\end{aligned}$$

where the last step follows from equation 22. Recalling equation 23, we conclude that, on the same event,

$$\sum_{b\in\mathcal{A}} \frac{\pi_{\text{data}}(b|x)}{N(x,a,b)+L} \;\le\; \frac{81\,\iota^2\;|\mathcal{A}|}{N(x,a)+|\mathcal{A}|\,L} \qquad \forall\, x,a \in \mathcal{X}\times\mathcal{A}. \tag{26}$$

Step 2.3: putting everything together. We work on the intersection of the events of **Step 1**, **Step 2.1** and **Step 2.2**, which has probability at least $1-3\delta$, and we bound the two terms of the requirement separately. For the shrinkage gaps, we first bound $\gamma(x,a,b)$, as defined in **Step 2** of the proof of Lemma 3.4, by a purely count-based quantity. Fix $\ell \in [L]$ and abbreviate the corresponding summand of the maximum defining $\gamma(x,a,b)$ as

$$\gamma^\ell(x,a,b) := \left[\sigma^{-1}\left(\frac{N^\ell(x,a\succ b)}{N^\ell(x,a,b)}\right)\right]_{-2R_{\max}}^{2R_{\max}} - \left[\sigma^{-1}\left(\frac{N^\ell(x,a\succ b)}{N^\ell(x,a,b)+2}\right)\right]_{-2R_{\max}}^{2R_{\max}}.$$

If $N^\ell(x,a,b)=0$, then by the convention introduced in the proof of Lemma 3.4 the first projection equals $2R_{\max}$, while the second one equals $-2R_{\max}$ because its argument is ${0}/{2}=0$, so that $\gamma^\ell(x,a,b)=4R_{\max}$. In this case the partition property gives $N(x,a,b)\le L$, hence $\gamma^\ell(x,a,b) = 4R_{\max} \le 2\left(1+e^{2R_{\max}}\right) \le {6\left(1+e^{2R_{\max}}\right)L}/{(N(x,a,b)+L)}$, where we used $e^{2R_{\max}} \ge 1+2R_{\max}$. If $N^\ell(x,a,b) \ge 1$, we distinguish two cases. When $\sigma^{-1}\left({N^\ell(x,a\succ b)}/{N^\ell(x,a,b)}\right) < 2R_{\max}$, that is

$N^\ell(x,a\succ b)/N^\ell(x,b\succ a) < e^{2R_{\max}}$ and therefore $N^\ell(x,b\succ a) > N^\ell(x,a,b)/(1+e^{2R_{\max}})$, we use that the projection on $[-2R_{\max}, 2R_{\max}]$ is 1-Lipschitz and that $\log(1+z)\le z$ to get

$$\begin{aligned}\gamma^\ell(x,a,b) &\le \sigma^{-1}\left(\frac{N^\ell(x,a\succ b)}{N^\ell(x,a,b)}\right) - \sigma^{-1}\left(\frac{N^\ell(x,a\succ b)}{N^\ell(x,a,b)+2}\right)\\ &= \log\left(\frac{1}{N^\ell(x,a,b)/N^\ell(x,a\succ b)-1}\right) - \log\left(\frac{1}{(N^\ell(x,a,b)+2)/N^\ell(x,a\succ b)-1}\right)\\ &= \log\left(\frac{(N^\ell(x,a,b)+2)/N^\ell(x,a\succ b)-1}{N^\ell(x,a,b)/N^\ell(x,a\succ b)-1}\right)\\ &= \log\left(1+\frac{2}{N^\ell(x,b\succ a)}\right) \le \frac{2\left(1+e^{2R_{\max}}\right)}{N^\ell(x,a,b)}.\end{aligned}$$

When instead $\sigma^{-1}\left(N^\ell(x,a\succ b)/N^\ell(x,a,b)\right) \ge 2R_{\max}$, that is $N^\ell(x,a\succ b)\ge e^{2R_{\max}}N^\ell(x,b\succ a)$ and therefore $N^\ell(x,b\succ a)\le N^\ell(x,a,b)/(1+e^{2R_{\max}})$, the first projection equals $2R_{\max}$. Abbreviating $n := N^\ell(x,a,b)$ and $n_- := N^\ell(x,b\succ a)$, and using that the projection of z is at least $\min\{z, 2R_{\max}\}$, we get $\gamma^\ell(x,a,b)\le \max\left\{0, \log\left(e^{2R_{\max}}(n_-+2)/(n-n_-)\right)\right\}$. The function $n_-\mapsto \log\left(e^{2R_{\max}}(n_-+2)/(n-n_-)\right)$ is increasing, so evaluating it at the largest admissible value $n_- = n/(1+e^{2R_{\max}})$ gives $\gamma^\ell(x,a,b)\le \log\left(1+2(1+e^{2R_{\max}})/n\right) \le 2(1+e^{2R_{\max}})/N^\ell(x,a,b)$. Taking the maximum over $\ell\in[L]$, using the partition property in the form $N^\ell(x,a,b)\ge (N(x,a,b)+L)/(3L)$ for the folds that observe the pair and the bound above for the folds that do not, we obtain

$$\gamma(x,a,b) \le \frac{6\left(1+e^{2R_{\max}}\right)L}{N(x,a,b)+L} \qquad \forall\, x,a,b\in\mathcal{X}\times\mathcal{A}\times\mathcal{A}.$$

Combining this with equation 26,

$$\sum_{b\in\mathcal{A}}\pi_{\mathrm{data}}(b|x)\gamma(x,a,b) \le 6\left(1+e^{2R_{\max}}\right)L\sum_{b\in\mathcal{A}}\frac{\pi_{\mathrm{data}}(b|x)}{N(x,a,b)+L} \le \frac{972\,e^{2R_{\max}}L\,|\mathcal{A}|\,\iota^2}{N(x,a)+|\mathcal{A}|\,L},$$

where we also used $1+e^{2R_{\max}}\le 2e^{2R_{\max}}$. This is at most the first term of $C(x,a)$ as soon as $c_0\ge 972$. For the slacks, recalling their definition from the statement of Lemma 3.4 and applying the Cauchy–Schwarz inequality with respect to $\pi_{\mathrm{data}}(\cdot|x)$,

$$\sum_{b\in\mathcal{A}}\pi_{\mathrm{data}}(b|x)\max_{\ell\in[L]}\mathrm{slack}^\ell(x,a,b) \le \sqrt{24\,e^{2R_{\max}}L\max_{\ell\in[L]}D_x^\ell}\sqrt{\sum_{b\in\mathcal{A}}\frac{\pi_{\mathrm{data}}(b|x)}{N(x,a,b)+L}}.$$

Bounding the first square root with **Step 2.1** and the second one with equation 26,

$$\begin{aligned}\sum_{b\in\mathcal{A}}\pi_{\mathrm{data}}(b|x)\max_{\ell\in[L]}\mathrm{slack}^\ell(x,a,b) &\le 9\,\iota\sqrt{\frac{24\,e^{2R_{\max}}L\,|\mathcal{A}|^3\,\iota}{2\,(N(x,a)+|\mathcal{A}|\,L)}}\\ &\le 32\,e^{R_{\max}}\,\iota^{3/2}\sqrt{\frac{L\,|\mathcal{A}|^3}{N(x,a)+|\mathcal{A}|\,L}},\end{aligned}$$

which is at most the second term of $C(x,a)$ as soon as $c_1\ge 32$. Summing the two displays proves the second claim.

Step 3: the penalty is small enough. Since $(u+v)^2\le 2u^2+2v^2$, and since $(N(x,a)+|\mathcal{A}|\,L)^2 \ge |\mathcal{A}|\,L\,(N(x,a)+|\mathcal{A}|\,L) \ge |\mathcal{A}|\,L\,(N(x,a)+L)$ as well as $N(x,a)+|\mathcal{A}|\,L\ge N(x,a)+L$, we have

$$C(x,a)^2 \le \frac{2c_0^2\,L\,|\mathcal{A}|\,e^{4R_{\max}}\,\iota^4 + 2c_1^2\,e^{2R_{\max}}\,\iota^3 L\,|\mathcal{A}|^3}{N(x,a)+L}.$$

It therefore suffices to bound $\sum_{x,a}\nu_0(x)\pi_{\mathrm{data}}(a|x)\,(N(x,a)+L)^{-1}$. To this end we apply Lemma F.4 to the random variables

$$Z_n := \frac{1}{\mathfrak{f}(X_n,A_n^+)+L} + \frac{1}{\mathfrak{f}(X_n,A_n^-)+L} \in \left[0,\frac{2}{L}\right], \qquad n\in[N],$$

whose conditional expectation is $2\sum_{x,a}\nu_0(x)\pi_{\text{data}}(a|x)\left(\mathfrak{f}(x,a)+L\right)^{-1}$, for a fixed function $\mathfrak{f}:\mathcal{X}\times\mathcal{A}\to\{0,1,\dots,N\}$. Taking a union bound over the at most $(N+1)^{|\mathcal{X}||\mathcal{A}|}$ such functions and then choosing $\mathfrak{f}=N(\cdot,\cdot)$, we obtain that with probability at least $1-\delta$

$$2N\sum_{x,a}\frac{\nu_0(x)\pi_{\text{data}}(a|x)}{N(x,a)+L}\;\le\;2\sum_{x,a}\frac{N(x,a)}{N(x,a)+L}+\frac{4}{L}\log\left(\frac{(N+1)^{|\mathcal{X}||\mathcal{A}|}}{\delta}\right),$$

whose right hand side is at most $2\,|\mathcal{X}|\,|\mathcal{A}|+{}^{4(|\mathcal{X}||\mathcal{A}|\log(N+1)+\log(1/\delta))}/_{L}$, so that

$$\sum_{x,a}\frac{\nu_0(x)\pi_{\text{data}}(a|x)}{N(x,a)+L}\;\le\;\widetilde{\mathcal{O}}\left(\frac{|\mathcal{X}|\,|\mathcal{A}|}{N}\right).$$

Combining the last two displays with the bound on $C(x,a)^2$ gives the third claim, on the intersection with the event of **Step 2.3**. □

F.4 Proof of Lemma 3.5: from suboptimality to estimation error.

At this point, we can present the upper bound on the suboptimality, i.e. $J_\beta(\pi^\star)-J_\beta(\pi_{\text{out}})$, which depends only on the single policy concentrability and the generalization error of the estimated gap $\widehat{\Delta}$. Such upper bound holds under the pessimistic event defined as follows:

$$\mathcal{E}_{\text{pessimism}}=\left\{\sum_b\pi_{\text{data}}(b|x)\widehat{\Delta}(x,a,b)\le\sum_b\pi_{\text{data}}(b|x)\Delta_{r^\star}(x,a,b)\quad\forall x,a\in\mathcal{X}\times\mathcal{A}\right\}.$$

The result is proven formally in the following after restating the lemma for convenience.

Lemma 3.5. *Let us assume that the policy ensemble $\{\tilde{\pi}^\ell\}_{\ell=1}^L$ induces an estimated gap $\widehat{\Delta}$ such that $\widehat{\Delta}(x,a,\pi_{\text{data}})\le\Delta_{r^\star}(x,a,\pi_{\text{data}})$ for all $x,a\in\mathcal{X}\times\mathcal{A}$. Then, for any $\beta>0$,*

$$J_\beta(\pi^\star)-J_\beta(\pi_{\text{out}})\le\min\left\{\sqrt{C^\star\left\langle\pi_{\text{data}},\text{Err}^2\right\rangle},\beta^{-1}C^\star_\infty\left\langle\pi_{\text{data}},\text{Err}^2\right\rangle\right\}$$

where π_{out} is computed from $\{\tilde{\pi}^\ell\}_{\ell=1}^L$ as in Lemma 3.1 and,

$$\text{Err}(x,a):=\widehat{\Delta}(x,a,\pi_{\text{data}})-\Delta_{r^\star}(x,a,\pi_{\text{data}}).$$

Proof. The proof follows from these calculations.

$$\begin{aligned}
\langle\pi^\star-\pi_{\text{out}},r^\star\rangle&=\sum_x\nu_0(x)\sum_a(\pi^\star(a|x)-\pi_{\text{out}}(a|x))r^\star(x,a)\\
&=\sum_x\nu_0(x)\sum_{a,b}(\pi^\star(a|x)-\pi_{\text{out}}(a|x))\pi_{\text{data}}(b|x)(r^\star(x,a)-r^\star(x,b))\\
&=\sum_x\nu_0(x)\sum_{a,b}(\pi^\star(a|x)-\pi_{\text{out}}(a|x))\pi_{\text{data}}(b|x)(\Delta_{r^\star}(x,a,b))\\
&=\sum_x\nu_0(x)\sum_{a,b}(\pi^\star(a|x)-\pi_{\text{out}}(a|x))\pi_{\text{data}}(b|x)(\Delta_{r^\star}(x,a,b)-\widehat{\Delta}(x,a,b))\\
&\quad+\sum_x\nu_0(x)\sum_{a,b}(\pi^\star(a|x)-\pi_{\text{out}}(a|x))\pi_{\text{data}}(b|x)\widehat{\Delta}(x,a,b)
\end{aligned}\tag{27}$$

Looking at the first term, we obtain that

$$\begin{aligned}
&\sum_x \nu_0(x) \sum_{a,b} (\pi^\star(a|x) - \pi_{\mathrm{out}}(a|x)) \pi_{\mathrm{data}}(b|x) (\Delta_{r^\star}(x,a,b) - \widehat{\Delta}(x,a,b)) \\
&\qquad = \sum_x \nu_0(x) \sum_{a,b} \pi^\star(a|x) \pi_{\mathrm{data}}(b|x) (\Delta_{r^\star}(x,a,b) - \widehat{\Delta}(x,a,b)) \\
&\qquad\quad + \sum_x \nu_0(x) \sum_{a,b} \pi_{\mathrm{out}}(a|x) \pi_{\mathrm{data}}(b|x) (\widehat{\Delta}(x,a,b) - \Delta_{r^\star}(x,a,b)) \\
&\qquad \le \sum_x \nu_0(x) \sum_{a,b} \pi^\star(a|x) \pi_{\mathrm{data}}(b|x) (\Delta_{r^\star}(x,a,b) - \widehat{\Delta}(x,a,b)) \quad \text{(Using } \mathcal{E}_{\mathrm{pessimism}}\text{)} \\
&\qquad = \sum_x \nu_0(x) \sum_{a,b} \frac{\pi^\star(a|x)}{\sqrt{\pi_{\mathrm{data}}(a|x)}} \sqrt{\pi_{\mathrm{data}}(a|x)} \pi_{\mathrm{data}}(b|x) (\Delta_{r^\star}(x,a,b) - \widehat{\Delta}(x,a,b)) \\
&\qquad \le \sqrt{\sum_x \nu_0(x) \sum_a \frac{(\pi^\star(a|x))^2}{\pi_{\mathrm{data}}(a|x)}} \\
&\qquad\quad \cdot \sqrt{\sum_x \nu_0(x) \sum_a \pi_{\mathrm{data}}(a|x) (\sum_b \pi_{\mathrm{data}}(b|x) (\Delta_{r^\star}(x,a,b) - \widehat{\Delta}(x,a,b)))^2}
\end{aligned}$$

Finally, we upper bound Equation (27) by the fact that

$$\pi_{\mathrm{out}} = \operatorname*{arg\,max}_{\pi \in \Pi} \sum_x \nu_0(x) \sum_a \pi(a|x) r^-(x,a) - \beta D_{\mathrm{KL}}(\pi, \pi_{\mathrm{ref}})$$

where $D_{\mathrm{KL}}(\pi, \pi') = \sum_x \nu_0(x) \sum_a \pi(a|x) \log \frac{\pi(a|x)}{\pi'(a|x)}$ and

$$\begin{aligned}
r^-(x,a) &:= \widehat{\Delta}(x,a,b) + \beta \max_{\ell \in [L]} \left(\log \frac{\tilde{\pi}^\ell(b|x)}{\pi_{\mathrm{ref}}(b|x)} - \zeta^\ell(x) \right) \\
&= \beta \min_{\ell \in [L]} \left(\log \frac{\tilde{\pi}^\ell(a|x)}{\pi_{\mathrm{ref}}(a|x)} - m^\ell(x,a) \right).
\end{aligned}$$

Indeed, notice that

$$\begin{aligned}
&\sum_x \nu_0(x) \sum_{a,b} (\pi^\star(a|x) - \pi_{\mathrm{out}}(a|x)) \pi_{\mathrm{data}}(b|x) \widehat{\Delta}(x,a,b) = \\
&\sum_x \nu_0(x) \sum_{a,b} (\pi^\star(a|x) - \pi_{\mathrm{out}}(a|x)) \pi_{\mathrm{data}}(b|x) r^-(x,a) \\
&\quad \underbrace{- \sum_x \nu_0(x) \sum_{a,b} (\pi^\star(a|x) - \pi_{\mathrm{out}}(a|x)) \pi_{\mathrm{data}}(b|x) \beta \max_{\ell \in [L]} \left(\log \frac{\tilde{\pi}^\ell(b|x)}{\pi_{\mathrm{ref}}(b|x)} - \zeta^\ell(x) \right)}_{=0} \\
&= \sum_x \nu_0(x) \sum_a (\pi^\star(a|x) - \pi_{\mathrm{out}}(a|x)) r^-(x,a) \\
&\le \beta D_{\mathrm{KL}}(\pi^\star, \pi_{\mathrm{ref}}) - \beta D_{\mathrm{KL}}(\pi_{\mathrm{out}}, \pi_{\mathrm{ref}})
\end{aligned}$$

where the last inequality follows from the optimality of π_{out}. Then, putting all together we get

$$\begin{aligned}
\langle \pi^\star - \pi_{\mathrm{out}}, r^\star \rangle + \beta D_{\mathrm{KL}}(\pi_{\mathrm{out}}, \pi_{\mathrm{ref}}) - \beta D_{\mathrm{KL}}(\pi^\star, \pi_{\mathrm{ref}}) &\le \sqrt{\sum_x \nu_0(x) \sum_a \frac{(\pi^\star(a|x))^2}{\pi_{\mathrm{data}}(a|x)}} \\
&\quad \cdot \sqrt{\sum_x \nu_0(x) \sum_a \pi_{\mathrm{data}}(a|x) (\sum_b \pi_{\mathrm{data}}(b|x) (\Delta_{r^\star}(x,a,b) - \widehat{\Delta}(x,a,b)))^2},
\end{aligned}$$

which implies the statement.

Conversion with fast rates Now, we aim to prove the fast rate, i.e. $1/N$ type, result. This is enabled by a different suboptimality to estimation error conversion. To this end, let us notice that

- π_{out} is optimal for the RLHF problem (see equation 1) with reward function $\widehat{\Delta}(x, a, \pi_{\text{data}})$.
- $\pi^\star$ is optimal for the RLHF problem with reward function $\Delta_{r^\star}(x, a, \pi_{\text{data}})$.
- Thanks to Lemma 3.4, we have that with high probability $\Delta_{r^\star}(x, a, \pi_{\text{data}}) \geq \widehat{\Delta}(x, a, \pi_{\text{data}})$ for any $x, a \in \mathcal{X} \times \mathcal{A}$.

Under the conditions above, we can invoke Ji et al. (2026, Lemma C.2) to guarantee that

$$
\begin{aligned}
J_\beta(\pi^\star) - J_\beta(\pi_{\text{out}}) &\leq \beta^{-1} \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} \nu_0(x)\pi^\star(a|x) \left(\sum_{b\in\mathcal{A}} \pi_{\text{data}}(b|x)(\Delta_{r^\star}(x,a,b) - \widehat{\Delta}(x,a,b)) \right)^2 \\
&= \beta^{-1} \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} \nu_0(x) \frac{\pi^\star(a|x)}{\pi_{\text{data}}(a|x)} \pi_{\text{data}}(a|x) \left(\sum_{b\in\mathcal{A}} \pi_{\text{data}}(b|x)(\Delta_{r^\star}(x,a,b) - \widehat{\Delta}(x,a,b)) \right)^2 \\
&\leq \beta^{-1} C^\star_\infty \sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} \nu_0(x)\pi_{\text{data}}(a|x) \left(\sum_{b\in\mathcal{A}} \pi_{\text{data}}(b|x)(\Delta_{r^\star}(x,a,b) - \widehat{\Delta}(x,a,b)) \right)^2 \\
&= \beta^{-1} C^\star_\infty \left\langle \pi_{\text{data}}, \text{Err}^2 \right\rangle .
\end{aligned}
$$

□

F.5 Proof of Lemma 3.6: bounding the estimation error

The goal of this subsection is to provide a high probability bound on the following generalization error

$$
\left\langle \pi_{\text{data}}, \text{Err}^2 \right\rangle := \sum_{x\in\mathcal{X}} \nu_0(x) \sum_a \pi_{\text{data}}(a|x) (\sum_b \pi_{\text{data}}(b|x)(\Delta_{r^\star}(x,a,b) - \widehat{\Delta}(x,a,b)))^2.
$$

The formal results are presented in Lemma 3.6.

Lemma 3.6. *It holds that* $\mathbb{P}\left[\left\langle \pi_{\text{data}}, \text{Err}^2 \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{e^{4R_{\max}} |\mathcal{A}|^4 |\mathcal{X}| L^2 \log(\delta^{-1})}{N} \right) \right] \geq 1 - 7\delta.$

Proof. Let us notice that denoting for a fixed $x, a, b \in \mathcal{X} \times \mathcal{A}^2$,

$$
\bar{\ell} = \arg\max_{\ell\in[L]} \beta \log \frac{\tilde{\pi}^\ell(b|x)}{\pi_{\text{ref}}(b|x)} - \beta\zeta^\ell(x)
$$

and

$$
\underline{\ell} = \arg\min_{\ell\in[L]} \beta \log \frac{\tilde{\pi}^\ell(a|x)}{\pi_{\text{ref}}(a|x)} - \beta\zeta^\ell(x).
$$

Then, we have that

$$
\begin{aligned}
\left| \widehat{\Delta}(x,a,b) - \Delta_{r^\star}(x,a,b) \right| &\leq \left| \beta \log \frac{\tilde{\pi}^{\underline{\ell}}(a|x)}{\pi_{\text{ref}}(a|x)} - \beta\zeta^{\underline{\ell}}(x) - r^\star(x,a) \right| \\
&\quad + \left| \beta \log \frac{\tilde{\pi}^{\bar{\ell}}(b|x)}{\pi_{\text{ref}}(b|x)} - \beta\zeta^{\bar{\ell}}(x) - r^\star(x,b) \right| + C(x,a) \\
&\leq \sum_{\mathfrak{a}\in\{a,b\}} \max_{\ell\in[L]} \left| \beta \log \frac{\tilde{\pi}^\ell(\mathfrak{a}|x)}{\pi_{\text{ref}}(\mathfrak{a}|x)} - \beta\zeta^\ell(x) - r^\star(x,\mathfrak{a}) \right| + C(x,a),
\end{aligned}
$$

where the penalty $C(x, a)$ appears because it enters the definition of $\widehat{\Delta}(x, a, b)$; we carry it along and bound its contribution through Lemma F.1 at the end.

Then, using the fact that $r^\star(x, \pi_{\text{data}}) = 0$ without loss of generality and that by the Azuma-Hoeffding inequality, we have that for any action $a \in \mathcal{A}$ and state $x \in \mathcal{X}$

$$\left|\log(\tilde{\pi}^\ell(a|x)/\pi_{\text{ref}}(a|x)) - \zeta^\ell(x) - \underbrace{\left(\log(\tilde{\pi}^\ell(a|x)/\pi_{\text{ref}}(a|x)) - \sum_{\mathfrak{a}\in\mathcal{A}} \pi_{\text{data}}(\mathfrak{a}|x) \log \frac{\tilde{\pi}^\ell(\mathfrak{a}|x)}{\pi_{\text{ref}}(\mathfrak{a}|x)}\right)}_{:=g(x,a)}\right|$$
$$= \left|\frac{N^{-\ell}(x)(\log(\tilde{\pi}^\ell(a|x)/\pi_{\text{ref}}(a|x)) - \zeta^\ell(x) - g(x,a))}{N^{-\ell}(x)+2} - \frac{2(g(x,a) - \log(\tilde{\pi}^\ell(a|x)/\pi_{\text{ref}}(a|x)) + \zeta^\ell(x))}{N^{-\ell}(x)+2}\right|$$
$$\leq \mathcal{O}\left(\sqrt{\frac{R_{\max}^2/\beta^2 \log(N|\mathcal{X}||\mathcal{A}|/\delta)}{N^{-\ell}(x)+2}} + \frac{R_{\max}/\beta}{N^{-\ell}(x)+2}\right),$$

where the last inequality follows with probability $1-\delta$ by the Azuma-Hoeffding inequality, which can be applied because

$$\begin{aligned}
&N^{-\ell}(x)(\log(\tilde{\pi}^\ell(a|x)/\pi_{\text{ref}}(a|x)) - \zeta^\ell(x) - g(x,a)) \\
&\quad= \sum_{X,A\in\mathcal{D}\setminus\mathcal{D}^\ell} (\log(\tilde{\pi}^\ell(a|x)/\pi_{\text{ref}}(a|x)) - g(x,a)) - \log\frac{\tilde{\pi}^\ell(A|X)}{\pi_{\text{ref}}(A|X)} \mathbb{1}_{\{X=x\}} \\
&\quad= \sum_{X,A\in\mathcal{D}\setminus\mathcal{D}^\ell} \left(\sum_{a\in\mathcal{A}} \pi_{\text{data}}(a|x) \log\frac{\tilde{\pi}^\ell(a|x)}{\pi_{\text{ref}}(a|x)} - \log\frac{\tilde{\pi}^\ell(A|X)}{\pi_{\text{ref}}(A|X)}\right) \mathbb{1}_{\{X=x\}},
\end{aligned}$$

is the sum of a martingale difference sequence which is almost surely bounded by $2R_{\max}/\beta$ because of the constraints imposed on $\tilde{\pi}^\ell$ in Equation (6). For the same reason, we also have that

$$\left|g(x,a) - \log(\tilde{\pi}^\ell(a|x)/\pi_{\text{ref}}(a|x)) + \zeta^\ell(x)\right| \leq \frac{2R_{\max}}{\beta},$$

which allows to bound the second term as $\frac{2R_{\max}/\beta}{N^{-\ell}(x)+2}$. Then, noticing that $g(x,a) = \Delta_{\log\tilde{\pi}^\ell/\pi_{\text{ref}}}(x, a, \pi_{\text{data}})$, we obtain that

$$\begin{aligned}
&\left|\log(\tilde{\pi}^\ell(a|x)/\pi_{\text{ref}}(a|x)) - \zeta^\ell(x) - \Delta_{\log\tilde{\pi}^\ell/\pi_{\text{ref}}}(x,a,\pi_{\text{data}})\right| \\
&\leq \mathcal{O}\left(\sqrt{\frac{R_{\max}^2/\beta^2 \log(N|\mathcal{X}||\mathcal{A}|/\delta)}{N^{-\ell}(x)+2}} + \frac{R_{\max}/\beta}{N^{-\ell}(x)+2}\right).
\end{aligned}$$

Therefore, using this fact,

$$\begin{aligned}
&\left|\widehat{\Delta}(x,a,b) - \Delta_{r^\star}(x,a,b)\right| \\
&\leq \sum_{\mathfrak{a}\in\{a,b\}} \max_{\ell\in[L]} \Big(\left|\beta\Delta_{\log\tilde{\pi}^\ell/\pi_{\text{ref}}}(x,\mathfrak{a},\pi_{\text{data}}) - \Delta_{r^\star}(x,\mathfrak{a},\pi_{\text{data}})\right| \\
&\quad + \beta\left|\log(\tilde{\pi}^\ell(\mathfrak{a}|x)/\pi_{\text{ref}}(\mathfrak{a}|x)) - \zeta^\ell(x) - \Delta_{\log\tilde{\pi}^\ell/\pi_{\text{ref}}}(x,\mathfrak{a},\pi_{\text{data}})\right| \Big) + C(x,a) \\
&\leq \sum_{\mathfrak{a}\in\{a,b\}} \max_{\ell\in[L]} \left|\beta\Delta_{\log\tilde{\pi}^\ell/\pi_{\text{ref}}}(x,\mathfrak{a},\pi_{\text{data}}) - \Delta_{r^\star}(x,\mathfrak{a},\pi_{\text{data}})\right| \\
&+ \mathcal{O}\left(\sqrt{\frac{R_{\max}^2 \log(N|\mathcal{X}||\mathcal{A}|/\delta)}{N^{-\ell}(x)+2}} + \frac{R_{\max}}{N^{-\ell}(x)+2}\right) + C(x,a).
\end{aligned}$$

To bound the first term, we use **Step 1** of the proof of Lemma 3.4. Abbreviating $\widehat{p}^\ell(x,\mathfrak{a},a') := {}^{N^\ell(x,\mathfrak{a}\succ a')}/_{N^\ell(x,\mathfrak{a},a')}$, it gives $\left|\beta\Delta_{\log\tilde{\pi}^\ell/\pi_{\text{ref}}}(x,\mathfrak{a},a') - \left[\sigma^{-1}(\widehat{p}^\ell(x,\mathfrak{a},a'))\right]_{-2R_{\max}}^{2R_{\max}}\right| \leq \text{slack}^\ell(x,\mathfrak{a},a')$ for every pair with $N^\ell(x,\mathfrak{a},a') > 0$. Moreover, $\Delta_{r^\star}(x,\mathfrak{a},a') = \left[\sigma^{-1}(\mathbb{P}_{r^\star}^{\text{true}}(\mathfrak{a}\succ a'|x))\right]_{-2R_{\max}}^{2R_{\max}}$, and the map $p \mapsto \left[\sigma^{-1}(p)\right]_{-2R_{\max}}^{2R_{\max}}$ is $4e^{2R_{\max}}$-Lipschitz on

$[0,1]$, because $1/(p(1-p)) \leq \left(e^{R_{\max}} + e^{-R_{\max}}\right)^2 \leq 4e^{2R_{\max}}$ for $p \in \sigma([-2R_{\max}, 2R_{\max}])$. For the pairs with $N^\ell(x, \mathfrak{a}, a') = 0$, we simply use $\left|\beta\Delta_{\log \tilde{\pi}^\ell/\pi_{\text{ref}}}(x, \mathfrak{a}, a') - \Delta_{r^\star}(x, \mathfrak{a}, a')\right| \leq 3R_{\max}$ together with the fact that, by the partition property, $N(x, \mathfrak{a}, a') \leq L$ in this case, so that $3R_{\max} \leq 3\sqrt{2}R_{\max}\sqrt{L/(N(x,\mathfrak{a},a')+L)}$. Therefore,

$$
\begin{aligned}
&\sum_{\mathfrak{a}\in\{a,b\}} \max_{\ell\in[L]} \left|\beta\Delta_{\log \tilde{\pi}^\ell/\pi_{\text{ref}}}(x, \mathfrak{a}, \pi_{\text{data}}) - \Delta_{r^\star}(x, \mathfrak{a}, \pi_{\text{data}})\right| \\
&\leq \sum_{\mathfrak{a}\in\{a,b\}} \sum_{a'\in\mathcal{A}} \pi_{\text{data}}(a'|x) \max_{\ell\in[L]} \left|\beta\Delta_{\log \tilde{\pi}^\ell/\pi_{\text{ref}}}(x, \mathfrak{a}, a') - \Delta_{r^\star}(x, \mathfrak{a}, a')\right| \\
&\leq \sum_{\mathfrak{a}\in\{a,b\}} \sum_{a'\in\mathcal{A}} \pi_{\text{data}}(a'|x) \left(4e^{2R_{\max}} \max_{\ell:\, N^\ell(x,\mathfrak{a},a')>0} \left|\widehat{p}^\ell(x, \mathfrak{a}, a') - \mathbb{P}^{\text{true}}_{r^\star}(\mathfrak{a} \succ a'|x)\right|\right. \\
&\qquad\qquad \left. + 3\sqrt{2}R_{\max}\sqrt{\frac{L}{N(x,\mathfrak{a},a')+L}} + \max_{\ell\in[L]} \text{slack}^\ell(x, \mathfrak{a}, a')\right) \\
&\leq \widetilde{\mathcal{O}}\left(\sum_{\mathfrak{a}\in\{a,b\}} e^{2R_{\max}} \sum_{a'\in\mathcal{A}} \pi_{\text{data}}(a'|x)\sqrt{\frac{L\log(|\mathcal{A}|^2|\mathcal{X}|NL/\delta)}{N(x,\mathfrak{a},a')+L}}\right) + \sum_{\mathfrak{a}\in\{a,b\}} C(x, \mathfrak{a}),
\end{aligned}
$$

where the last inequality holds with probability $1-\delta$ thanks to the Azuma-Hoeffding inequality applied to the empirical frequencies $\widehat{p}^\ell(x, \mathfrak{a}, a')$ and a union bound over $[N] \times \mathcal{X} \times \mathcal{A}^2 \times [L]$, that is over the possible values of the count $N^\ell(x, \mathfrak{a}, a')$ and over the indices $\ell, x, \mathfrak{a}, a'$, together with $N^\ell(x, \mathfrak{a}, a') \geq (N(x,\mathfrak{a},a')+L)/(3L)$ for the folds that observe the pair. For the slacks, we used that, on the events of Lemma F.1, **Step 2.3** of its proof gives $\sum_{a'\in\mathcal{A}} \pi_{\text{data}}(a'|x) \max_{\ell\in[L]} \text{slack}^\ell(x, \mathfrak{a}, a') \leq C(x, \mathfrak{a})$. Hence, putting all together, we obtain that with probability $1-2\delta$,

$$
\begin{aligned}
\left|\widehat{\Delta}(x, a, b) - \Delta_{r^\star}(x, a, b)\right| &\leq \widetilde{\mathcal{O}}\left(\sum_{\mathfrak{a}\in\{a,b\}} e^{2R_{\max}} \sum_{a'\in\mathcal{A}} \pi_{\text{data}}(a'|x)\sqrt{\frac{L\log(|\mathcal{A}|^2|\mathcal{X}|NL/\delta)}{N(x,\mathfrak{a},a')+L}}\right) \\
&\quad + \widetilde{\mathcal{O}}\left(\sqrt{\frac{R^2_{\max}\log(N|\mathcal{X}||\mathcal{A}|/\delta)}{N(x)+1}}\right) + 2C(x, a) + C(x, b),
\end{aligned}
$$

where we have also used that $N^{-\ell}(x) \geq \frac{L-1}{L}N(x) - 1 \geq \frac{N(x)}{2} - 1$ for $L \geq 2$. Then, towards controlling

$$
\begin{aligned}
\left\langle \pi_{\text{data}}, \text{Err}^2 \right\rangle &\leq \sum_{x\in\mathcal{X}} \sum_{a,b\in\mathcal{A}\times\mathcal{A}} \nu_0(x)\pi_{\text{data}}(a|x)\pi_{\text{data}}(b|x) \left|\widehat{\Delta}(x, a, b) - \Delta_{r^\star}(x, a, b)\right|^2 \\
&\leq \widetilde{\mathcal{O}}\left(e^{4R_{\max}} \sum_{x\in\mathcal{X}} \sum_{a,b\in\mathcal{A}\times\mathcal{A}} \nu_0(x)\pi_{\text{data}}(a|x)\pi_{\text{data}}(b|x)\frac{L\log(|\mathcal{A}|^2|\mathcal{X}|NL/\delta)}{N(x,a,b)+L}\right) \\
&\quad + 15\sum_{x\in\mathcal{X}} \sum_{a\in\mathcal{A}} \nu_0(x)\pi_{\text{data}}(a|x)C(x, a)^2,
\end{aligned}
$$

where we also used that $N(x) \geq N(x, a, b)$, that $(u+v+w)^2 \leq 3u^2 + 3v^2 + 3w^2$ to separate the penalties, Jensen's inequality for the average over a', and that squaring the per point bound squares its Lipschitz factor $e^{2R_{\max}}$. Therefore, the terms which decay as $1/N(x)$ are lower order terms absorbed into the $\widetilde{\mathcal{O}}$ notation. Now, using Rosenberg et al. (2020, Lemma D.4), we have that for any sequence of positive i.i.d. random variables $Z_1, \ldots, Z_n \in [0, B]$ with mean μ it holds that

$$
N\mu \leq 2\sum_{i=1}^{N} Z_i + 4B\log(N\delta^{-1}).
$$

In our case, the bound does not apply directly because $N(\cdot,\cdot,\cdot) : \mathcal{X} \times \mathcal{A} \times \mathcal{A} \to [0, N]$ is a random function. However, for any $\mathfrak{f} : \mathcal{X} \times \mathcal{A} \times \mathcal{A} \to [0, N]$, we can apply Rosenberg et al. (2020,

Lemma D.4) for $Z_i = (\mathfrak{f}(X_i, A_i, A_i'))$, where we recall that $\mathcal{D} = \{X_i, A_i, A_i'\}_{i=1}^N$, and with failure probability $\frac{\delta}{N|\mathcal{X}||\mathcal{A}|^2}$, we obtain that

$$N \sum_{x\in\mathcal{X}} \sum_{a,b\in\mathcal{A}\times\mathcal{A}} \frac{\nu_0(x)\pi_{\mathrm{data}}(a|x)\pi_{\mathrm{data}}(b|x)}{\mathfrak{f}(x,a,b)+L} \leq 2\sum_{i=1}^N \frac{1}{\mathfrak{f}(X_i,A_i,A_i')+L} + 4\,|\mathcal{X}|\,|\mathcal{A}|^2\,L\log(N^2\delta^{-1}).$$

Therefore, choosing specifically $\mathfrak{f}(\cdot,\cdot,\cdot) = N(\cdot,\cdot,\cdot)$ we obtain

$$\begin{aligned}
N \sum_{x\in\mathcal{X}} \sum_{a,b\in\mathcal{A}\times\mathcal{A}} \frac{\nu_0(x)\pi_{\mathrm{data}}(a|x)\pi_{\mathrm{data}}(b|x)}{N(x,a,b)+L} &\leq 2\sum_{i=1}^N \frac{1}{N(X_i,A_i,A_i')+L} + 4\,|\mathcal{X}|\,|\mathcal{A}|^2\,L\log(N^2\delta^{-1}) \\
&= 2 \sum_{x,a,b\in\mathcal{X}\times\mathcal{A}\times\mathcal{A}} \frac{\sum_{i=1}^N \mathbb{1}_{\{X_i,A_i,A_i'=x,a,b\}}}{N(x,a,b)+L} + 4\,|\mathcal{X}|\,|\mathcal{A}|^2\,L\log(N^2\delta^{-1}) \\
&\leq 2 \sum_{x,a,b\in\mathcal{X}\times\mathcal{A}\times\mathcal{A}} \frac{N(x,a,b)}{N(x,a,b)+L} + 4\,|\mathcal{X}|\,|\mathcal{A}|^2\,L\log(N^2\delta^{-1}) \\
&\leq 2\,|\mathcal{X}|\,|\mathcal{A}|^2 + 4\,|\mathcal{X}|\,|\mathcal{A}|^2\,L\log(N^2\delta^{-1}) \\
&= \mathcal{O}\left(|\mathcal{X}|\,|\mathcal{A}|^2\,L\log(N^2\delta^{-1})\right).
\end{aligned}$$

The second term is bounded by Lemma F.1, which gives $\sum_{x,a}\nu_0(x)\pi_{\mathrm{data}}(a|x)C(x,a)^2 \leq \widetilde{\mathcal{O}}\left(e^{4R_{\max}}|\mathcal{A}|^4|\mathcal{X}|L\log(1/\delta)/N\right)$, a quantity decreasing in N of the same order as the first one. Therefore, we can conclude that with probability $1-7\delta$, by a union bound over the two Azuma events, the event above and the events of Lemma F.1, it holds that

$$\left\langle \pi_{\mathrm{data}}, \mathrm{Err}^2 \right\rangle \leq \widetilde{\mathcal{O}}\left(\frac{e^{4R_{\max}}\,|\mathcal{X}|\,|\mathcal{A}|^4 L^2\log(N^2\delta^{-1})}{N}\right).$$

This concludes the proof. □

Finally, we conclude the section by stating an instance-dependent version of Theorem 3.3 which improves the action space dependence greatly in the case of sparse empirical preference graphs, as happens in the case of UltraFeedback — the training set used in our experiments.

Theorem F.2. *Let us consider running* PEPO *(Algorithm 1) with* $L = \left\lceil \frac{7\log(|\mathcal{X}||\mathcal{A}|^2/\delta)}{2} \right\rceil$ *many ensemble members and a preference dataset* $\mathcal{D}$ *of size* N *and for* $C(x,a) := \frac{c_0 L|\mathcal{A}|e^{2R_{\max}}\iota^2}{N(x,a)+|\mathcal{A}|L} + c_1 e^{R_{\max}}\iota\sqrt{\frac{L|\mathcal{A}|D_{\max}}{N(x,a)+|\mathcal{A}|L}}$ *with large enough absolute constants* c_0, c_1*, where* $D_{\max} := \max_{x\in\mathcal{X},\ell\in[L]} D_x^\ell$*. Then, with probability at least* $1-\delta$*, we have that the suboptimality of the* PEPO *output,* π_{out}*, against* $\pi^\star = \arg\max_{\pi\in\Pi} J_\beta(\pi)$ *and for any value of* $\beta \in (0,\infty)$ *is upper bounded as follows,*

$$\begin{aligned}
&J_\beta(\pi^\star) - J_\beta(\pi_{\mathrm{out}}) \\
&\leq \widetilde{\mathcal{O}}\Bigg(\min\Bigg\{\sqrt{\frac{C^\star e^{4R_{\max}}\,|\mathcal{A}|^2\,(1+D_{\max})\,|\mathcal{X}|\log\delta^{-1}}{N}}, \\
&\qquad\qquad \frac{C_\infty^\star e^{4R_{\max}}\,|\mathcal{A}|^2\,(1+D_{\max})\,|\mathcal{X}|\log\delta^{-1}}{\beta N}\Bigg\}\Bigg).
\end{aligned}$$

The proof follows by changing the bonus definition to make it depend on the preference-dataset-dependent quantity $\max_{x\in\mathcal{X},\ell\in[L]} D_x^\ell$ and avoiding bounding it by its worst-case, i.e. $|\mathcal{A}|^2/2$ as done in **Step 1** of Lemma F.1. The additive 1 accounts for the terms that do not depend on $D_{\max}$, namely the first term of the penalty and the statistical error of the empirical frequencies in the proof of Lemma 3.6, which are of order $\widetilde{\mathcal{O}}\left(e^{4R_{\max}}|\mathcal{A}|^2|\mathcal{X}|L^2/N\right)$.

F.6 Technical Results

In this appendix, we provide the proofs of the auxiliary results we used throughout our analysis. We start by presenting some known results from the literature that we used in our analysis.

Lemma F.3. ***Azuma-Hoeffding inequality*** *Let $\{X_k\}_{k=1}^K$ be a martingale difference sequence adapted to the filtration $\mathcal{F}_k$ which is almost surely bounded by B, i.e. $|X_k| \leq B$ for all $k \in [K]$. Then, we have that*

$$\mathbb{P}\left[\left|\sum_{k=1}^{K} X_k\right| \geq \sqrt{2KB^2 \log(2/\delta)}\right] \leq \delta.$$

Lemma F.4. ***Multiplicative Freedman inequality. (Variant of Rosenberg et al. (2020, Lemma D.4))*** *Let $\{X_j\}_{j\geq 1}$ be a sequence of non negative random variables adapted to the filtration $\mathcal{F}_j$ and almost surely bounded by B, i.e. $0 \leq X_j \leq B$ for all $j \geq 1$. Then, we have that with probability at least $1-\delta$, simultaneously for all $k \geq 1$,*

$$\sum_{j=1}^{k} \mathbb{E}\left[X_j | \mathcal{F}_{j-1}\right] \leq 2\sum_{j=1}^{k} X_j + 2B\log(1/\delta).$$

Proof. Without loss of generality we take $B = 1$, the general case following by applying the result to X_j/B. Fix $\lambda \in \mathbb{R}$ and define $M_0 := 1$ and, for $k \geq 1$,

$$M_k := \exp\left(\lambda \sum_{j=1}^{k} X_j - \left(e^\lambda - 1\right)\sum_{j=1}^{k} \mathbb{E}\left[X_j | \mathcal{F}_{j-1}\right]\right).$$

Since $X_j \in [0,1]$ and $u \mapsto e^{\lambda u}$ is convex, bounding it on $[0,1]$ by the chord through its endpoints gives $e^{\lambda X_j} \leq 1 + \left(e^\lambda - 1\right) X_j$, hence, using $1 + u \leq e^u$,

$$\mathbb{E}\left[e^{\lambda X_j} | \mathcal{F}_{j-1}\right] \leq 1 + \left(e^\lambda - 1\right)\mathbb{E}\left[X_j | \mathcal{F}_{j-1}\right] \leq \exp\left(\left(e^\lambda - 1\right)\mathbb{E}\left[X_j | \mathcal{F}_{j-1}\right]\right).$$

Therefore $\mathbb{E}\left[M_k | \mathcal{F}_{k-1}\right] \leq M_{k-1}$, that is $\{M_k\}_{k\geq 0}$ is a non negative supermartingale started at $M_0 = 1$. For such a process, Ville's inequality gives $\mathbb{P}\left[\exists k \geq 0 : M_k \geq 1/\delta\right] \leq \delta$, which we prove for completeness: letting $\tau := \inf\{k \geq 0 : M_k \geq 1/\delta\}$, the optional stopping theorem applied to the bounded stopping time $\tau \wedge k$ gives $\mathbb{E}\left[M_{\tau\wedge k}\right] \leq M_0 = 1$, while $M_{\tau\wedge k} \geq 1/\delta \mathbb{1}_{\{\tau \leq k\}}$ because M is non negative, so that $\mathbb{P}\left[\tau \leq k\right] \leq \delta$ for every k and the claim follows letting $k \to \infty$. Consequently, with probability at least $1-\delta$ we have $M_k < 1/\delta$ for every $k \geq 0$, that is, taking logarithms,

$$\lambda \sum_{j=1}^{k} X_j \leq \left(e^\lambda - 1\right)\sum_{j=1}^{k} \mathbb{E}\left[X_j | \mathcal{F}_{j-1}\right] + \log(1/\delta) \qquad \forall k \geq 1. \tag{28}$$

The statement follows choosing $\lambda = -\log 2$, for which $e^\lambda - 1 = -1/2$, and using that $2\log 2 \leq 2$. □

Lemma F.5. ***Reversed form of Lemma F.4*** *Let $\{X_j\}_{j\geq 1}$ be a sequence of non negative random variables adapted to the filtration $\mathcal{F}_j$ and almost surely bounded by B, i.e. $0 \leq X_j \leq B$ for all $j \geq 1$. Then, we have that with probability at least $1-\delta$, simultaneously for all $k \geq 1$,*

$$\sum_{j=1}^{k} X_j \leq 2\sum_{j=1}^{k} \mathbb{E}\left[X_j | \mathcal{F}_{j-1}\right] + 2B\log(1/\delta).$$

Proof. The proof is the one of Lemma F.4 up to the choice of λ in equation 28. Taking there $\lambda = \log 2$, for which $e^\lambda - 1 = 1$, and dividing by λ, the statement follows from $1/\log 2 \leq 2$. □

Lemma F.6. ***Method of types*** *Let $\widehat{\mu}$ be the empirical frequency of n independent* Bernoulli(μ) *samples. Then, for every $t \geq 0$,*

$$\mathbb{P}\left[n\left(\widehat{\mu}\log\frac{\widehat{\mu}}{\mu} + (1-\widehat{\mu})\log\frac{1-\widehat{\mu}}{1-\mu}\right) \geq t\right] \leq (n+1)e^{-t}.$$

Proof. The empirical frequency takes the $n+1$ values k/n for $k \in \{0, \dots, n\}$ and, for each of them, $\mathbb{P}[\widehat{\mu} = k/n] = \binom{n}{k}\mu^k(1-\mu)^{n-k} \leq \exp\left(-n\left(\frac{k}{n}\log\frac{k/n}{\mu} + (1-\frac{k}{n})\log\frac{1-k/n}{1-\mu}\right)\right)$. Summing this bound over the values of $\widehat{\mu}$ for which the event holds gives the claim. □

Lemma F.7. ***Bernoulli anti-concentration (Cassel et al., 2025, Lemma 3)*** *Let $L \geq 1$. Then, consider $\{N^\ell\}_{\ell=1}^L$ for any $N^\ell \geq 0$. Consider now samples $X_{n,\ell} \sim \text{Bernoulli}(\mu)$ for each $\ell \in [L]$ and $n \in [N^\ell]$ and the estimators*

$$\hat{\mu}^\ell = \frac{\sum_{n=1}^{N^\ell} X_{n,\ell}}{N^\ell + 2}.$$

Then, it holds that

$$\mathbb{P}\left[\exists \ell \in [L] \quad \hat{\mu}^\ell \leq \mu\right] \geq 1 - e^{-2L/7}.$$

Lemma F.8. ***Suboptimality to squared reward estimation error. (Ji et al., 2026, Lemma C.2)*** *Let $r : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ be any reward function such that $r \leq r^\star$. Moreover, let $\pi^\star$ be the optimal policy for the regularized RL problem with reward $r^\star$ and π_r the one when the reward is r. That is, $\pi^\star(a|x) \propto \pi_{\text{ref}}(a|x)e^{r^\star(x,a)/\beta}$ and $\pi_r(a|x) \propto \pi_{\text{ref}}(a|x)e^{r(x,a)/\beta}$. Then, it holds that*

$$J_\beta(\pi^\star) - J_\beta(\pi_r) \leq \beta^{-1}\sum_{x\in\mathcal{X}}\nu_0(x)\sum_{a\in\mathcal{A}}\pi^\star(a|x)(r(x,a) - r^\star(x,a))^2$$

Proof. We have that we can rewrite the suboptimality as follows

$$\begin{aligned} J_\beta(\pi^\star) - J_\beta(\pi_r) &= \langle \pi^\star, r^\star - \beta\log(\pi^\star/\pi_{\text{ref}})\rangle - \langle \pi_r, r^\star - \beta\log(\pi_r/\pi_{\text{ref}})\rangle \\ &= \beta\sum_{x\in\mathcal{X}}\nu_0(x)\log Z_\star(x) - \langle \pi_r, r^\star - r\rangle - \beta\sum_{x\in\mathcal{X}}\nu_0(x)\log Z_r(x), \end{aligned}$$

where we have defined $Z_r(x) = \sum_{a\in\mathcal{A}}\pi_{\text{ref}}(a|x)e^{r(x,a)/\beta}$ and the shorthand $Z_\star := Z_{r^\star}$. Now defining

$$H(r,x) = \log Z_r(x) - \beta^{-1}\sum_{a\in\mathcal{A}}\pi_r(a|x)(r(x,a) - r^\star(x,a)),$$

we can notice that

$$\begin{aligned} J_\beta(\pi^\star) - J_\beta(\pi_r) &= \beta\sum_{x\in\mathcal{X}}\nu_0(x)(H(r^\star,x) - H(r,x)) \\ &= \beta\sum_{x\in\mathcal{X}}\nu_0(x)\sum_{a\in\mathcal{A}}(r^\star(x,a) - r(x,a))\frac{\partial H(r,x)}{\partial r(x,a)}\bigg|_{r=r_\gamma} \end{aligned}$$

for $r_\gamma(x,a) = \gamma r(x,a) + (1-\gamma)r^\star(x,a)$ for some unknown $\gamma \in [0,1]$. Now, we compute $\nabla H(r_\gamma, x, a)$.

$$\begin{aligned} \frac{\partial H(r,x)}{\partial r(x,a)}\bigg|_{r=r_\gamma} &= \beta^{-1}\frac{\pi_{\text{ref}}(a|x)e^{r_\gamma(x,a)/\beta}}{Z_{r_\gamma}(x)} - \beta^{-1}(r_\gamma(x,a) - r^\star(x,a))\frac{\partial\pi_r(a|x)}{\partial r(x,a)} \\ &\quad - \beta^{-1}\pi_{r_\gamma}(a|x) \\ &\quad + \beta^{-1}\sum_{a'\neq a}(r_\gamma(x,a') - r^\star(x,a'))\frac{\pi_{\text{ref}}(a'|x)e^{r_\gamma(x,a')/\beta}}{Z^2_{r_\gamma}(x)}\frac{\partial Z_r(x)}{\partial r(x,a)}\bigg|_{r=r_\gamma} \\ &= -\beta^{-2}(r_\gamma(x,a) - r^\star(x,a))\left(\pi_{r_\gamma}(a|x) - \pi^2_{r_\gamma}(a|x)\right) \\ &\quad + \beta^{-2}\sum_{a'\neq a}(r_\gamma(x,a') - r^\star(x,a'))\pi_{r_\gamma}(a'|x)\pi_{r_\gamma}(a|x), \end{aligned}$$

where we have used that $\frac{\partial Z_r(x)}{\partial r(x,a')} = \frac{\pi_{\text{ref}}(a'|x)e^{r(x,a')/\beta}}{\beta}$ for any $a' \in \mathcal{A}$, $\frac{\partial\pi_r(a|x)}{\partial r(x,a)} = \beta^{-1}(\pi_r(a|x) - \pi^2_r(a|x))$ for any $a \in \mathcal{A}$ and $\frac{\partial\pi_r(a|x)}{\partial r(x,a')} = -\beta^{-1}\pi_r(a|x)\pi_r(a'|x)$ for any $a \in \mathcal{A}$ and $a' \neq a$. Then, we

can rearrange the above expression as

$$\begin{aligned}\left.\frac{\partial H(r,x)}{\partial r(x,a)}\right|_{r=r_\gamma} &= -\beta^{-2}(r_\gamma(x,a)-r^\star(x,a))\pi_{r_\gamma}(a|x)\\ &\quad+\beta^{-2}\sum_{a'\in\mathcal{A}}(r_\gamma(x,a')-r^\star(x,a'))\pi_{r_\gamma}(a'|x)\pi_{r_\gamma}(a|x)\\ &\leq \beta^{-2}(r^\star(x,a)-r_\gamma(x,a))\pi_{r_\gamma}(a|x)\end{aligned}$$

where the last step uses the fact that $r_\gamma - r^\star \leq 0$ because $r \leq r^\star$. Then, we obtain that

$$J_\beta(\pi^\star) - J_\beta(\pi_r) \leq \beta^{-1}\underbrace{\sum_{x\in\mathcal{X}}\nu_0(x)\sum_{a\in\mathcal{A}}\pi_{r_\gamma}(a|x)(r^\star(x,a)-r(x,a))^2}_{f_\gamma},$$

for some $\gamma \in [0,1]$. We conclude the proof by maximizing the above expression with respect to γ. To this end, notice that

$$\begin{aligned}\frac{\partial \pi_{r_\gamma}(a|x)}{\partial\gamma} &= \sum_{a'\in\mathcal{A}}\frac{\partial\pi_{r_\gamma}(a|x)}{\partial r_\gamma(x,a')}\frac{\partial r_\gamma(x,a')}{\partial\gamma}\\ &= -\sum_{a'\in\mathcal{A}}\frac{\partial\pi_{r_\gamma}(a|x)}{\partial r_\gamma(x,a')}(r^\star(x,a')-r(x,a'))\\ &= -\beta^{-1}(\pi_{r_\gamma}(a|x)-\pi^2_{r_\gamma}(a|x))(r^\star(x,a)-r(x,a))\\ &\quad+\beta^{-1}\sum_{a'\neq a}\pi_{r_\gamma}(a|x)\pi_{r_\gamma}(a'|x)(r^\star(x,a')-r(x,a'))\\ &= -\beta^{-1}\pi_{r_\gamma}(a|x)(r^\star(x,a)-r(x,a))\\ &\quad+\beta^{-1}\sum_{a'\in\mathcal{A}}\pi_{r_\gamma}(a|x)\pi_{r_\gamma}(a'|x)(r^\star(x,a')-r(x,a'))\end{aligned}$$

where the first line uses $\partial r_\gamma(x,a')/\partial\gamma = r(x,a') - r^\star(x,a') = -(r^\star(x,a') - r(x,a'))$, since $r_\gamma = \gamma r + (1-\gamma)r^\star$. Therefore,

$$\begin{aligned}\frac{\partial f_\gamma}{\partial\gamma} &= -\beta^{-1}\sum_{x\in\mathcal{X}}\nu_0(x)\sum_{a\in\mathcal{A}}\pi_{r_\gamma}(a|x)(r^\star(x,a)-r(x,a))^3\\ &\quad+\beta^{-1}\sum_{x\in\mathcal{X}}\nu_0(x)\left(\sum_{a\in\mathcal{A}}\pi_{r_\gamma}(a|x)(r^\star(x,a)-r(x,a))^2\right)\left(\sum_{a'\in\mathcal{A}}\pi_{r_\gamma}(a'|x)(r^\star(x,a')-r(x,a'))\right).\end{aligned}$$

Now, we prove that $\frac{\partial f_\gamma}{\partial\gamma} \leq 0$, as a consequence of the following more general statement, $\mathbb{E}[X^3] \geq \mathbb{E}[X^2]\,\mathbb{E}[X]$ for any random variable $X \geq 0$. This is Chebyshev's association inequality: X^2 and X are both non-decreasing functions of X, hence positively correlated, that is

$$\mathbb{E}[X^3] - \mathbb{E}[X^2]\,\mathbb{E}[X] = \mathrm{Cov}(X^2, X) = \tfrac{1}{2}\,\mathbb{E}\left[\left(X^2-\widetilde{X}^2\right)\left(X-\widetilde{X}\right)\right] \geq 0,$$

where $\widetilde{X}$ is an independent copy of X and the inequality holds because $X^2 - \widetilde{X}^2$ and $X - \widetilde{X}$ share the same sign for $X, \widetilde{X} \geq 0$. Applying this with $X = r^\star(x,A) - r(x,A) \geq 0$ for $A \sim \pi_{r_\gamma}(\cdot|x)$, which is non-negative because $r \leq r^\star$, the negative term of the display above dominates the positive one for every x, so that $\frac{\partial f_\gamma}{\partial\gamma} \leq 0$. Therefore, f_γ is decreasing in γ, hence it is maximized for $\gamma = 0$. All in all, noticing that $\pi_{r_0} = \pi^\star$, we reach the conclusion of the proof, i.e.

$$J_\beta(\pi^\star) - J_\beta(\pi_r) \leq \beta^{-1}\sum_{x\in\mathcal{X}}\nu_0(x)\sum_{a\in\mathcal{A}}\pi^\star(a|x)(r^\star(x,a)-r(x,a))^2.$$

□

F.6.1 Rejection Sampling Routine: Proof of Lemma 3.2

Proof. Let us denote

$$f_{\text{out}}(x,a) = \min_{\ell\in[L]} \pi^\ell(a|x)\exp(-m^\ell(x,a))$$

At each round, Algorithm 2 accepts the sample with probability

$$\sum_{a\in\mathcal{A}} \pi_{\text{prop}}(a|x)\frac{f_{\text{out}}(x,a)}{M(x)\pi_{\text{prop}}(a|x)} = \sum_{a\in\mathcal{A}} f_{\text{out}}(x,a)M^{-1}(x) = Z$$

where we defined $Z = M^{-1}(x)\sum_{a\in\mathcal{A}}\min_{\ell\in[L]}\pi^\ell(a|x)\exp(-m^\ell(x,a))$. We omitted the dependence on the prompt x for simplicity. Therefore, the random variable X representing the number of trials until the first successful acceptance follows a Geometric distribution: $X \sim \text{Geometric}(Z)$

We want to find $N_{\text{RS}}(x)$ such that the probability of failing to get a sample after $N_{\text{RS}}(x)$ trials is at most δ:

$$P(X > N_{\text{RS}}(x)) \leq \delta$$

The event $X > N_{\text{RS}}(x)$ occurs if and only if all $N_{\text{RS}}(x)$ independent trials result in a rejection. The probability of rejection in a single trial is $(1-Z)$. Therefore:

$$P(X > N_{\text{RS}}(x)) = (1-Z)^{N_{\text{RS}}(x)}$$

Therefore, our goal can be rewritten as

$$(1-Z)^{N_{\text{RS}}(x)} \leq \delta$$

Taking the natural logarithm of both sides (and noting that $\ln(1-Z)$ is negative, which flips the inequality):

$$N_{\text{RS}}(x)\ln(1-Z) \leq \ln(\delta)$$

which implies

$$N_{\text{RS}}(x) \geq \frac{\ln(\delta)}{\ln(1-Z)}$$

To simplify this into a more interpretable bound, we use the fact that for $Z\in(0,1)$, $\ln(1-Z)\leq -Z$. This implies that:

$$\frac{1}{\ln(1-Z)} \geq -\frac{1}{Z}$$

Therefore $\frac{\ln(\delta)}{\ln(1-Z)} \leq \frac{\ln(1/\delta)}{Z}$, so any $N_{\text{RS}}(x) \geq \frac{\ln(1/\delta)}{Z} = \frac{M(x)\ln(1/\delta)}{\sum_{a\in\mathcal{A}} f_{\text{out}}(x,a)}$ satisfies the requirement, which gives the choice in the statement. □

By virtue of Lemma 3.2, Algorithm 2 fails to output a valid sample with probability at most δ. The maximum number of trials $N_{\text{RS}}(x)$ is related by the above lemma to the desired confidence level. If the routine does not generate valid samples within $N_{\text{RS}}(x)$ trials, then we can output *I don't know* as an answer to the prompt x.

We think that this is a very desirable property for LLMs. It is important that these models can estimate their uncertainties even when they do not have access to a verifier. This is a characteristic that is not granted by the χ^2 regularization in Huang et al. (2024), but only by our ensemble mechanism when applied to DPO. Surprisingly, the same feature does not appear if we consider reward-ensemble-based pessimism as done in Coste et al. (2023).