The Role of Preference Data and Unembeddings in the Convergence Rate of DPO

Gayathri Neela Chandran[*][†]
Indian Institute of Science
cgayathri@alumni.iitm.ac.in

Sai Soumya Nalli[*]
Microsoft Research India
t-snalli@microsoft.com

Sruthi Gorantla
Amazon AGI
gorantlas@iisc.ac.in

Amit Deshpande
Microsoft Research India
amitdesh@microsoft.com

Anand Louis
Indian Institute of Science
anandl@iisc.ac.in

Abstract

Human or AI feedback in the form of preference data over response-pairs plays a crucial role in finetuning Large Language Models (LLMs) using Reinforcement Learning from Human Feedback (RLHF), Direct Preference Optimization (DPO) and their variants. For these methods to be effective, the representations or unembeddings of responses must be expressive enough to align with the preference data. In this paper, we study the convergence of gradient descent for DPO using finite samples in the realizable setting, for example, preferences generated by the Bradley-Terry model of linear reward functions on query & response representations. Unlike previous theoretical analysis with stronger assumptions about the underlying unembeddings, our analysis works with a parameterization that is better representative of LLM implementations and doesn't assume independence of logits.

We derive a linear convergence rate bound for gradient descent on the DPO objective. Our bound crucially depends on the condition number of the matrix of query embeddings, the algebraic connectivity and the maximum degree of the comparison graph over responses. Our bound can guide the selection of preference feedback in order to optimize the cost of data acquisition as well as the cost of training. We show that in addition to the DPO converging to the optimum of the loss function, the learned reward differences also converge to the ground truth. These results, shown for pairwise preference data, can be extended to listwise preference data as well as discrete choice data and are validated through a set of experiments over both synthetic and real world datasets. To ensure the sufficiency of the available data, we study both the identifiability of the ground truth and the generalizability of the aligned model. Additionally, linear convergence results for DPO under tabular parameterization of the policy are also obtained.

1 Introduction

Alignment of large language models (LLMs) involves ensuring these models behave in ways that are consistent with human values and expectations. The pipeline of techniques like reinforcement learning from human/ AI feedback (RLHF/ RLAIF) Ziegler et al. [2020] involves training a reward model for query-response pairs using human/AI feedback in the form of preference data. This reward model then guides the optimization of the parameters of a supervised fine tuned model while

[*] Equal contribution.

[†] Work done while at IISc.

2nd Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2025).

minimizing the deviation from the reference model. Direct preference optimization (DPO) Rafailov et al. [2023] and other direct alignment methods (such as Gheshlaghi Azar et al. [2024], Tang et al. [2024]) fine-tune the LLM directly under the assumption that the language model implicitly expresses the true reward differences, eliminating the need for a separate reward model.

With the increasing popularity of DPO and it's variants both due to the empirical performance and the simpler pipeline, there has also been a growing interest Gheshlaghi Azar et al. [2024], in it's theoretical questions regarding efficiency, robustness and generalization. Several theoretical works have explored preference learning in the context of DPO and RLHF. Through modeling RLHF as a KL-regularized contextual bandit problem or Markov decision process, many worksXiong et al. [2024]Zhu et al. [2023] derive generalization and sample complexity results.

Convergence Analysis of DPO Convergence of DPO has been studied for different parameterised models of DPO. We highlight some key directions: Shi et al. [2025] analyze the convergence rates of DPO with varying sampling strategies under *tabular softmax parametrization* when the graph of comparisons over responses are connected. Nika et al. [2024] studies theoretical comparison between RLHF and DPO and obtains convergence results for policy classes which are log linear in the joint feature mapping of the query and the response. Yuan et al. [2025] studies gradient entanglement between preferred and dispreferred responses in DPO that can lead to their probabilities not being able to move in opposite directions during finetuning. Feng et al. [2025] proposes a new sampling strategy PILAF that explicitly aligns the preference learning for maximizing the oracle reward. They theoretically demonstrate how their sampler favours directions that are more sensitive to the objective function. Calandriello et al. [2024] establishes the equivalence of Nash-MD with Identity Preference Optimization, a variant of DPO, though it falls short of showing convergence. Im and Li [2024] theoretically analyses distinguishability and its role in governing the rate of change of the unembedding matrix.

Our work also draws inspiration from the convergence analysis of Bradley Terry log likelihood in Vojnovic et al. [2020] that learns rewards over a set of items. We non-trivially extend these results to the DPO setting where the implicit rewards depend on the query embeddings.

1.1 Our Contributions

Our parameterization seeks to model the neural network architecture of transformer models used in DPO. As is the case with the recent works that seek to theoretically analyze DPO Im and Li [2024], Razin et al. [2025] we focus on a simpler implementation of DPO to aid our rigorous analysis. We choose the weights in the final layer, the unembedding matrix that generates the output logits from the input embeddings generated by a neural network, as our parameters Jiang et al. [2024], Im and Li [2024] Park et al. [2025] Han et al. [2024]. Explicitly tuning only the output-projection (unembedding) weights while freezing the rest of the model can be interpreted as a special case of head-only fine-tuning, LoRA, and other partial fine-tuning strategies. The main contributions of our work are as follows:

1. **Convergence rates for DPO** We analyze the loss function of DPO to show *linear convergence of gradient descent*. We also unearth *the crucial role of prompt embeddings and the structure of comparison graphs over responses* in the collected data on the rate of convergence. In Lemmas 3.1 and 3.2, we establish critical properties such as strong convexity over a restricted domain, alongside PL-inequality and smoothness of the loss function. Theorem 3.3 presents the number of iterations required to ensure a loss function under ϵ as $O\left(\kappa(\mathbf{X})^2 \frac{\lambda_{mn}(\mathbf{L_M}^{(\text{all})})}{\lambda_{m+1}(\mathbf{L_M}^{(\text{all})})} \log\left(\frac{1}{\epsilon}\right)\right)$ where $\mathbf{X}$ is the prompt embeddings matrix and $\mathbf{L_M}^{(\text{all})}$ is the Laplacian of the graph formed by the union of comparison data over all query instances. This is described further in Section 3. While this is shown for datasets where the graph of comparisons between responses for each prompt is connected, Appendix B.4 shows the same result for linearly independent query embeddings without requiring connectivity for comparison graphs.
2. **Strategic Collection of Preference Data** Given the high costs associated with running DPO and collecting annotation data, our analysis suggest efficient strategies for data collection. Our results suggest that for a given budget of prompt data, *prompts and responses for*

comparison must be chosen such that they form a well-conditioned embedding matrix instead of choosing a random set of prompts. Further, over a fixed set of queries our results support the strategy of picking comparison data in order to minimize the ratio $\frac{\lambda_{\max}(\mathbf{L_M}^{(\text{all})})}{\lambda^{+}_{\min}(\mathbf{L_M}^{(\text{all})})}$. These findings are consistently supported by our results on synthetic data 6, 7 as well as real world data 4, 8. The above results are theoretically proved for both pairwise and listwise comparison data.

3. **Generalizability** Section 4 considers the realizable setting where the implicit rewards can be expressed by our linear model. Given oracle access, Theorem 4.1 explores the convergence of the implicit reward differences of the model with the underlying Bradley-Terry rewards, first, for the training data and second, whether it can predict the right probabilities for a new input. While generalizability requires that the training prompts span the embedding space, connectivity over all the comparison graphs for the inputs is not a necessity. Moreover, the theorem provides an equivalent *easily checkable condition on the training data* to directly conclude the sufficiency of data for generalizability. An alternate analysis of the convergence by considering implicit reward differences and the ground truth inspired by Shi et al. [2025] is shown in lemma 4.2.

Comparison with Related Work Our choice of parameters differs from most of existing works Shi et al. [2025], Nika et al. [2024] in terms of its faithfulness to real-world LLM architectures. It captures how the logits of different responses for a prompt are tied together through that query's hidden state representation unlike other works where these logits are independent. This also does not allow for the direct adoption of the Bradley-Terry convergence analysis Vojnovic et al. [2020] and makes this result a non-trivial extension.

While Shi et al. [2025] shows linear convergence for Vanilla DPO (under tabular parameterization) and a sampling strategy that achieves a quadratic convergence in the tabular setting, it fails to extend to the setting of parameterization of unembeddings. Nika et al. [2024] obtains convergence results for policy classes which are log linear in the joint feature mapping of the query and the response. However, these joint embeddings aren't representative of the independent unembeddings used in transformer models.

Closer to our setting, Razin et al. [2025] study likelihood displacement, and Im and Li [2024] provide bounds on unembedding updates under distributional assumptions. In contrast, our analysis examines properties of the loss function and establishes linear convergence rate of Gradient Descent. These works involve parameterizations that are significantly simpler than any practical implementation of DPO and the current body of research lies far from characterizing the complexity of real-world DPO. Hence, our paper tries to address the question:

Can we give provable convergence guarantees for DPO using a parametrization that is closer to practical implementations of DPO?

To the best of our knowledge, this is the first work to establish linear convergence guarantees for Direct Preference Optimization (DPO) under the unembedding parameterization, a setting that directly reflects modern transformer architectures. Additionally, we prove linear convergence of the tabular parameterization of the probabilities as a theoretical result in Lemma 3.4, contrasting with the softmax parameterization. Our work with both of these parameterizations also suggests that skeleton of the Vojnovic et al. [2020] analysis can be extended to different implementations of DPO.

2 Preliminaries

Notation We denote by $\mathbf{M}_i$ and $\mathbf{M}^j$ the i^{th} row vector and the j^{th} column vector of a matrix $\mathbf{M}$. It's eigen values or singular values are denoted by $\lambda_1(\mathbf{M}), \lambda_2(\mathbf{M}) \cdots \lambda_n(\mathbf{M})$ and $\sigma_1(\mathbf{M}), \cdots \sigma_n(\mathbf{M})$ in increasing order.

Direct Preference Optimization [Rafailov et al. [2023]] Given a human/ AI preference dataset constructed from the set of prompts $\mathcal{X} = \{x_v\}_{v=1}^{m}$ and the set of responses $\mathcal{Y} = \{y_i\}_{i=1}^{n}$, we have $d_{v,i,j}$ denote the number of times y_i was preferred over y_j for the prompt x_v in our dataset. Therefore, we get the dataset $\mathcal{D} = \{d_{v,i,j} \in \mathbb{R} : \forall v \in [m],\ i,j \in [n]\}_{i=1}^{N}$. Model alignment is done through

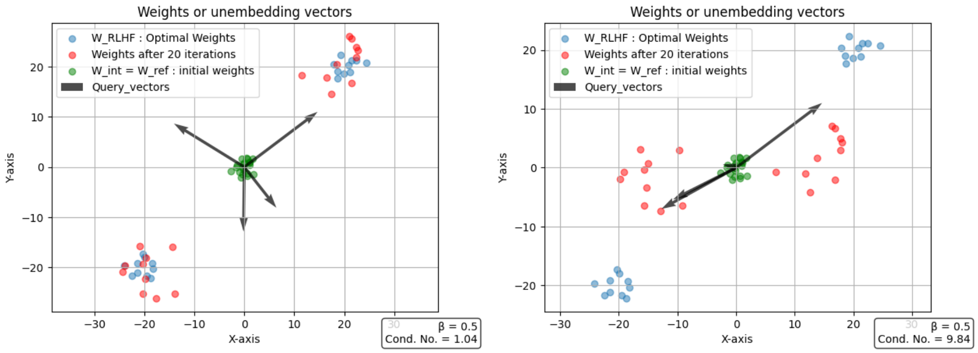

Figure 1: Scatter plots of the umembedding vectors at $t = 20$ iterations for a well-conditioned and ill-conditioned input/query embeddings for the same β, ground truth reward differences and initial weights

optimizing for θ that minimizes

$$\mathcal{L}_{\text{DPO}}(\theta) = -\sum_{v}^{m}\sum_{i}^{n}\sum_{j}^{n} d_{v,i,j}\log\sigma\left(\beta\log\frac{\pi_\theta(y_i \mid x_v)}{\pi_{\text{ref}}(y_i \mid x_v)} - \beta\log\frac{\pi_\theta(y_j \mid x_v)}{\pi_{\text{ref}}(y_j \mid x_v)}\right) \tag{1}$$

where π_{ref} is the reference policy and $\beta \in \mathbb{R}^+$ is a regularizing constant.

Let us denote by π_{DPO} the policy that minimizes the above expression We shall abuse the use of π_{DPO} throughout to talk about the probability distribution over responses as well as the pairwise preference distribution over pairs in different contexts, which shall be specified.

Parameters of LM Current language models typically have two parts: a neural network that produces a *hidden embedding* $\mathbf{x}_v \in \mathbb{R}^d$ for every prompt x_v and *a token unembedding* matrix or the output embedding matrix $\mathbf{W} \in \mathbb{R}^{n\times d}$ whose every row $\mathbf{W}_i$ converts the hidden embedding $\mathbf{x}$ into the logit for the the token $y_i \in \mathcal{Y}$. Razin et al. [2025]

$$\pi_\theta(y_i|x) = \frac{e^{\mathbf{W}_i\mathbf{x}}}{\sum_{j=1}^{n} e^{\mathbf{W}_j\mathbf{x}}}, \text{ where } \theta = \mathbf{W} \tag{2}$$

However note that since DPO is typically a finetuning step we *optimize over only the unembedding matrix* $\mathbf{W}$. Suppose that the reference policy has a parameter matrix $\mathbf{W}^{\text{ref}}$, we use π_W to represent the policy parameterized by $\mathbf{W}$ and π_{ref} to represent the reference policy. For this parametrization we can rewrite the DPO loss as follows,

$$\mathcal{L}_{\text{DPO}}(\mathbf{W}) = -\sum_{v}^{m}\sum_{i}^{n}\sum_{j}^{n} d_{v,i,j}\log\sigma\left(\beta(\mathbf{W}_i - \mathbf{W}_i^{(\text{ref})})\mathbf{x}_v - \beta(\mathbf{W}_j - \mathbf{W}_j^{(\text{ref})})\mathbf{x}_v\right) \tag{3}$$

3 Convergence Rate for Direct Preference Optimization

In this section, we characterize the rate of convergence of gradient descent on the DPO loss function. We demonstrate that it satisfies properties like the Polyak-Lojasiewicz (PL) inequality and smoothness that are integral to understanding convergence rate. This allows us to identify some necessary conditions for convergence as well as how data can be chosen to ensure faster convergence. These results can be extended to the DPO with data in listwise comparison, which is modeled using the Plackett-Luce model, which is an extension of the Bradley-Terry model to handle listwise preferences. The proofs to the results in this section are provided in Appendix B and C. It also includes an additional result with relaxed constraints.

3.1 Polyak-Lojasiewicz (PL) Inequality and Smoothness of the DPO Loss

It is known that for any μ-smooth function satisfying γ-PL inequality, the gradient descent algorithm with a suitable choice of the step size has a linear convergence rate. Similar to Vojnovic et al. [2020] in order to establish this, we observe that the quadratic form of the Hessian of the DPO loss, $\mathbf{v}^T \nabla^2 \mathcal{L}_{\text{DPO}}(\mathbf{W}, \mathbf{X})\mathbf{v}$, in the space of unembedding vectors $\mathbf{v} \in \mathbb{R}^{nd}$, can be expressed as the quadratic form of the Laplacian, $\mathbf{r}^T \mathbf{L_M}^{(\text{all})}\mathbf{r}$, of the graph representing comparisons over the responses for all queries.

The graph is constructed by adding edges of weight $m_{v,i,j} = d_{v,i,j} + d_{v,j,i}$ between any two query-response pairs (x^v, y_i) and (x^v, y_j) whenever they are compared $m_{v,i,j}$ times for query x^v. The full graph is the union of component graphs G_v for all queries $x^v \in X$. The vectors involved in the quadratic form of the Laplacian lie in the implicit reward space and are obtained via a linear transformation of the unembedding vectors $\mathbf{v}$. This is used to show $\gamma-$ strong convexity (a stronger condition than $\gamma-$ PL inequality) in a restricted subspace and $\mu-$smoothness.

Lemma 3.1. *For any upper bound on the reward amplitude $\rho > 0$,*

- *$\mathcal{L}_{DPO}(\mathbf{W}; \mathbf{X})$ is γ'-strongly convex on $\mathbf{W} \in \mathcal{W}_\rho \cap \mathcal{W}_0 \cap \mathcal{W}_X$ and $\mathbf{X} \in [-\theta, \theta]^{d \times m}$ when the graph of comparisons G_v is connected for all responses $x^v \in X$*
- *μ-smooth on $\mathbf{W} \in \mathbb{R}^{n \times d}$ for $\mathbf{X} \in \mathbb{R}^{d \times m}$*

where $\gamma' = c_\rho \beta^2 (\sigma^+_{\min}(\mathbf{X}))^2 \lambda_{m+1}(\mathbf{L}_M^{(all)})$, $\mu = \frac{1}{4}\beta^2 d^2 \sigma^2_{\max}(\mathbf{X})\lambda_{nm}(\mathbf{L}_M^{(all)})$ and subspaces of $\mathbb{R}^{n \times d}$ given by

$$\mathcal{W}_0 := \Big\{\mathbf{W} : (\mathbf{W} - \mathbf{W}^{\text{ref}})^T \mathbf{1}_n = \mathbf{0}_d\Big\}, \mathcal{W}_\rho := \Big\{\mathbf{W} : \beta \left\|(\mathbf{W} - \mathbf{W}^{\text{ref}})\mathbf{x}^v\right\|_\infty \leq \rho \;\; \forall \mathbf{x}^v \in \mathbb{R}^d\Big\},$$
$$\mathcal{W}_X := \Big\{\mathbf{W} : \mathbf{W}_i x = 0 \;\; \forall i \in [n], \; \forall x \in \text{null}(\mathbf{X})\Big\},$$

Further, $\beta \in \mathbb{R}^+$ is the KL regularization coefficient and $c_\rho = 1/(e^\rho + e^{-\rho})^2$

By restricting $\mathbf{W}$ so that the implicit reward vector is orthogonal to the null space of $\mathbf{L_M}^{(\text{all})}$, the quadratic form of the Laplacian remains positive, giving us a non-zero lower bound on $\text{vec}(\mathbf{W})^T \nabla^2 \mathcal{L}_{\text{DPO}}(\mathbf{W}'; \mathbf{X})\text{vec}(\mathbf{W})$ resulting in *strong convexity* in this subspace. When the graphs G_v over the queries are connected, $\chi_0 = \{\mathbf{r} \in \mathbb{R}^{nm} \mid \mathbf{r} \cdot e_i = 0 \forall i\}$ where $e_i \in \mathbb{R}^{nm}$ denotes the indicator vector of the i-th block, is the subspace orthogonal to the null space of $\mathbf{L_M}^{(\text{all})}$. This guides our choice of restricted subspace.

Since strong convexity is satisfied over $\mathbf{W} \in \mathcal{W}_\rho \cap \mathcal{W}_0 \cap \mathcal{W}_X$ and $\mathbf{X} \in \mathcal{X}_\theta$, the DPO loss function also satisfies PL inequality over the same subspace. This PL inequality can be extended to a larger domain for $\mathbf{W}$ using the inherent symmetry in the DPO loss function.

We exploit the property that

1. $\mathcal{L}_{\text{DPO}}(\Pi(\mathbf{W}; \mathbf{W}_0, \mathbf{c}), \mathbf{X}) = \mathcal{L}_{\text{DPO}}(\mathbf{W}, \mathbf{X})$
2. For any i, k, $\nabla_{W_{i,k}} \mathcal{L}_{\text{DPO}}(\Pi(\mathbf{W}; \mathbf{W}_0, \mathbf{c}), \mathbf{X}) = \nabla_{W_{i,k}} \mathcal{L}_{\text{DPO}}(\mathbf{W}, \mathbf{X})$.

where $\Pi(\mathbf{W}; \mathbf{W}_0, \mathbf{c}) = \mathbf{W} + \mathbf{1}_n \mathbf{c}^T + \mathbf{W}_0$

This helps transform any $\mathbf{W}' \in \mathcal{W}_\rho$ to a $\mathbf{W} \in \mathcal{W}_{2\rho} \cap \mathcal{W}_0 \cap \mathcal{W}_X$ for a careful choice of $\mathbf{c}$ and $\mathbf{W}_0$ such that the loss value and gradient remain. Using this property, PL inequality can be shown to hold over $\mathcal{W}_\rho$ as well.

Theorem 3.2. *For any upper bound on the reward amplitude $\rho > 0$, $\mathbf{W} \in \mathcal{W}_\rho$ and $\mathbf{X} \in [-\theta, \theta]^{d \times m}$*

$\mathcal{L}_{DPO}(\mathbf{W}; \mathbf{X})$ satisfies γ''-PL inequality with $\gamma'' = c_{2\rho}\beta^2 \sigma^2_{min}(\mathbf{X})\lambda_{m+1}(\mathbf{L}_M^{(all)})$ when the graph of comparisons G_v for each query x^v is connected.

Here

$$\mathcal{W}_\rho = \left\{\mathbf{W} : \beta \left\|(\mathbf{W} - \mathbf{W}^{ref})\mathbf{x}^v\right\|_\infty \leq \rho \;\; \forall \mathbf{x}^v \in [-\theta, \theta]^d\right\},$$

$\beta \in \mathbb{R}^+$, and $c_\rho = 1/(e^{2\rho} + e^{-2\rho})^2$

Compared to Vojnovic et al. [2020], the challenge to arrive at PL inequality lies in how the rewards for each query-response pair are not independent and share common parameters. This requires identifying a different subspace in which the strong convexity result holds. The extension to a larger subspace also presents its own challenges in the DPO setting which is carefully addressed in our result.

3.2 Analysis of Convergence of DPO

Using the properties of PL inequality and smoothness from Lemma 3.1 and Theorem 3.2 , we obtain a characterization of the number of iterations needed for convergence to an error ϵ.

Theorem 3.3. *Gradient decent over the DPO loss function $\mathcal{L}_{DPO}(\mathbf{W};\mathbf{X})$ in the space where $\mathbf{W} \in \mathcal{W}_\rho$ and $\mathbf{X} \in [-\theta,\theta]^{d\times m}$ requires $T = O\left(\frac{\mu''}{\gamma''}\log(\frac{1}{\epsilon})\right) = O\left(\kappa(\mathbf{X})^2 \frac{\lambda_{mn}({\mathbf{L}_\mathbf{M}}^{(all)})}{\lambda_{m+1}({\mathbf{L}_\mathbf{M}}^{(all)})}\frac{d^2}{c_\rho}\log(\frac{1}{\epsilon})\right)$ iterations for convergence for the error threshold ϵ such that $\mathcal{L}_{DPO}(\mathbf{W}^{(T)};\mathbf{X}) - \mathcal{L}_{DPO}(\mathbf{W}^\star;\mathbf{X}) \leq \epsilon$ when the graph of comparisons G_v for each query x_v is connected.*

Thus, given a query budget k, a good subset of queries is the one whose embedding matrix $\mathbf{X}$ attains a small condition number $\kappa(\mathbf{X})$, since the convergence rate scales inversely with $\kappa(\mathbf{X})$.

3.3 Convergence rates for tabular parameterization

In this subsection we briefly present our convergence rate results for DPO loss under the tabular parameterization of the policy.

$$\mathcal{L}_{\text{DPO}}(\pi) = -\sum_v\sum_i\sum_j d_{v,i,j}\log\sigma\left(\beta\log\frac{\pi(y_i\mid x)}{\pi_{\text{ref}}(y_i\mid x)} - \beta\log\frac{\pi(y_j\mid x)}{\pi_{\text{ref}}(y_j\mid x)}\right) \tag{4}$$

where every $\pi(y\mid x)\forall y,x$ is trainable.

Here, we shall directly optimize over the π's which would be equivalent to fitting in a logarithmically parametrized reward model, where it is not straight-forward to determine whether properties such as PL inequality and smoothness are preserved.

Lemma 3.4. *Gradient descent over the tabular parameters for the 4 takes time $T = \mathcal{O}\left(\frac{\mu_\alpha}{\gamma_\alpha}\log\frac{1}{\epsilon}\right)$ where $\mu_{\alpha_{GD}}$ and γ_α are decreasing and increasing functions on $\alpha = c_{\pi_{ref}}L_0^{\frac{1}{\beta}}$, $\alpha_{GD} = k_{\beta,m}\alpha$ and L_0 is the initial likelihood and m is the number of samples*

We also note that the *tabular softmax parameterization* considered in Shi et al. [2025] the parameters of the policy appear as logits which turns the DPO objective directly into the Bradley Terry negative log-likelihood for a tabular reward model. And hence the analysis of Vojnovic et al. [2020] directly proves the linear rate of convergence. This is particularly noteworthy because this highlights *the potential of the analysis of the Bradley-Terry loss Vojnovic et al. [2020] to act as a skeleton for multiple parameterizations* (linear in unembeddings, tabular and tabular softmax so far) of the DPO policy.

4 Convergence of the Unembeddings

In this section, we work in the realizability setting where the data follows a Bradley Terry distribution guided by a ground truth reward $r^*(x,y)$. We also assume that the policy that maximizes the RLHF objective can be expressed by the LLM, motivated by its necessity for DPO to theoretically discard the reward model and directly optimize for the model parameters. Note that π_{DPO} and π_{RLHF} represent the probability distributions $\mathcal{Y}\times\mathcal{X}\to[0,1]$ over the responses for every input such that they maximize the DPO objective and the RLHF Objective (equations 1 and 4) respectively

$$\pi_{\text{RLHF}} = \max_{\pi\in\Pi}\mathop{\mathbb{E}}_{x\sim\rho(\mathcal{X}),y\sim\pi(\cdot|x)} r^*(x,y) - \beta\text{KL}(\pi\parallel\pi_{ref})$$

Ziebart [2010] shows that the solution for the above can be expressed as

$$\pi_{\text{RLHF}}(y \mid x) = \frac{1}{Z(x)}\pi_{\text{ref}}(y \mid x)\exp(\frac{1}{\beta}r^*(x,y)), \forall x \in \mathcal{X}, y \in \mathcal{Y}$$

and if π_{RLHF} can be realizable by our linearly parametrized LLM through parameters $\mathbf{W}^{\text{RLHF}}$, then

$$\pi_{\text{RLHF}}(y_i \mid x) = \frac{1}{Z(x)}\exp\left(\mathbf{x}(\mathbf{W}_i^{\text{ref}})^T + \frac{1}{\beta}r^*\right) = \frac{\exp(\mathbf{x}\left(\mathbf{W}_i^{\text{RLHF}}\right)^T)}{\sum\limits_{j=1}^{|Y|}\exp(\mathbf{x}\left(\mathbf{W}_j^{\text{RLHF}}\right)^T)}$$

Hence this implies that $r^*(x, y_i)$ has to be expressed as $\mathbf{x}\beta\left(\mathbf{W}_i^{\text{RLHF}} - \mathbf{W}_i^{\text{ref}}\right)^T$ shifted by a global constant. This form of data also can be interpreted as Reinforcement Learning through AI Feedback where the data is generated by another LLM with the same query embeddings. We abuse the notation of π_{RLHF} and π_{DPO} to denote pairwise probability distribution over pairs of responses given a query.

Given oracle access to reward differences between any pair $(y_i, y_j) \in Y \times Y$ for any reward $x \in X$, if DPO is performed, two natural questions are (1) Does the resulting preference distribution converge to $\pi_{\text{RLHF}}(y_i \succ y_j \mid x)$ over the pairs queried? And (2) Do the resulting output unembeddings $\mathbf{W}^{\text{DPO}}$ converge to $\mathbf{W}^{\text{RLHF}}$ apart from a global shift? The two questions talk about output behavior on the training data and generalizability over any other input following the ground truth distribution.

Theorem 4.1. *Suppose the ground truth reward $r^*(x, y_i)$ is of the form $\mathbf{x}\beta\left(\mathbf{W}_i^{RLHF} - \mathbf{W}_i^{ref}\right)^T$, $Q = \{(y_i^\alpha, y_j^\alpha, x^\alpha)\}_{\alpha=1}^k$ are the queried triplets then*

1. *$\pi_{DPO}(y_i \succ y_j \mid x) = \pi_{RLHF}(y_i \succ y_j \mid x)$ for any $\{y,y_j, x\} \in Q$*
2. *$\text{rank}(A_Q) = (n-1)d$ iff $\pi_{DPO}(y_i \succ y_j \mid x) = \pi_{RLHF}(y_i \succ y_j \mid x)$ for any $x \in \mathbb{R}^d$ and $y_i, y_j \in \mathcal{Y}$*

where A_Q is the query coefficient matrix generated as follows: any $(y_i^\alpha, y_j^\alpha, x^\alpha) \in Q$ generates a row vector (in $\mathbb{R}^{nd}$) as $(0, 0, .., x^T, 0, .., 0, -x^T, ...0)$ where x^T is placed at the i^{th} $d-$dimensional block and $-x^T$ at the j^{th}.

The first part states that DPO exactly recovers the ground-truth reward differences on all observed triplets. This happens simply because the DPO loss is equivalent to MLE in the Bradley-Terry model of the implicit rewards which is maximized when the reward differences exactly match the ground truth.

The second provides an exact rank condition on the query coefficient matrix A_Q that characterizes generalization to unseen triplets. Note that $A_Q\vec{\mathbf{W}}$ generates the column vector for the implicit rewards of the training data. This, combined with the fact that unembeddings differences have to be retrieved exactly to generalize to any triplet outside Q leads to the rank equivalence. Note that this is important because it *provides an easily checkable condition to directly determine whether the training data leads to generalizability.*

In particular, achieving $\text{rank}(A_Q) = (n-1)d$ requires at least d linearly independent query embeddings. However, it is *not necessary that all comparison graphs be connected for generalizability* since multiple queries with single comparisons can still allow a rank $(n-1)d$ A_Q.

Convergence of reward differences. In addition to noting that the learned implicit reward differences for DPO at the optimum coincides with the ground truth rewards, we also establish rate of convergence of the learned reward differences to the ground truth by directly trying to extend the result in Shi et al. [2025] by adapting it to the simple parameterized setting.

Lemma 4.2. *We define the error in learned reward difference for the prompt x_v and response y_i and y_j as $\delta_v^t(y_i, y_j; \mathbf{W}^t) = r_{v,i}^\star - r_{v,j}^\star - \left(h(\mathbf{W}_i^t, \mathbf{x}_v) - h(\mathbf{W}_j^t, \mathbf{x}_v)\right)$. In the T^{th} iteration of gradient descent for $\mathcal{L}_{DPO}(\mathbf{W}; \mathbf{X})$ when the graphs G_v are connected for all prompts x_v, the error in learned rewards converges linearly as $|\delta_v^T(y_r, y_l; \mathbf{W}^T)| \leq |\max_{u,i,j}\delta_u^1(y_i, y_j; \mathbf{W}^1)|\zeta^{T-1}$ when $d_{i,j}^v = p_{i,j}^\star(y_i \succ y_j|x_v)$ where the contraction factor $\zeta = \max\left(1 + \eta\beta^2 n\sigma_{\max}^2(\mathbf{X})\left(1 - 8\sigma'_{\min}\right), \eta\beta^2 n\sigma_{\max}^2(\mathbf{X})(2 - 4\sigma'_{\min}) - 1\right)$ and $\sigma'_{\min} = \sigma'(\rho)$*

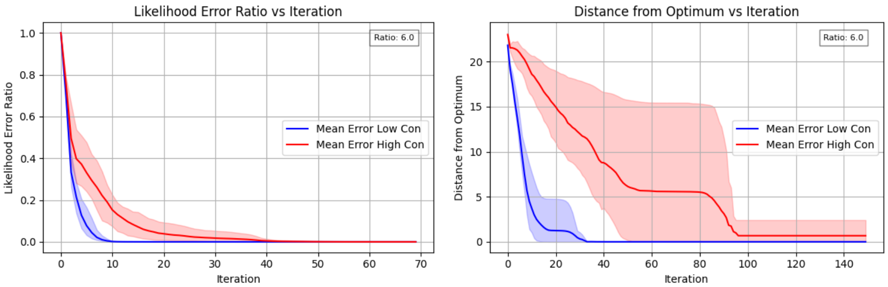

Figure 2: Convergence of $(\mathcal{L}^t_{\text{DPO}} - \mathcal{L}^\star_{\text{DPO}})/(\mathcal{L}^0_{\text{DPO}} - \mathcal{L}^\star_{\text{DPO}})$ and $\|\mathbf{W}' - \mathbf{W}^*\|_\infty$ when they are normalized when compared for different condition numbers over multiple iterations.

Similar to Theorem 3.3, we see linear convergence using the approach from Shi et al. [2025] as well. We obtain the result by using careful algebra to express $\delta^t_v(y_i, y_j; \mathbf{W}^t)$ in terms of $\delta^{t-1}_v(y_i, y_j; \mathbf{W}^{t-1})$ in such a way that we are able to obtain a contraction factor. It uses properties of the sigmoid function and the mean value theorem.

The proofs of generalizability and the convergence of learned reward differences results in this section can be found in the Appendices D and E respectively.

5 Experiments

5.1 Experiments with Synthetic Data

We conduct synthetic experiments that confirm that a well-conditioned $\mathbf{X}$ improves the convergence rate when working with a fixed budget for queries. Additionally, when multiple responses per query are compared higher algebraic connectivity and lower maximum eigenvalue lead to faster convergence. We conduct two sets of synthetic experiments to demonstrate this. In the first set of experiments, the embeddings of queries, $\mathbf{X}$, and the gold standard weights, $\mathbf{W}^\star$, which generate the ground truth rewards are randomly generated. In the second set of experiments, the unembeddings for $\mathbf{W}^\star$ are generated from a mixture of two 2D Gaussian distributions allowing us to observe the separation rate against the condition number of the subset of inputs chosen. The results of the experiment with Gaussian unembeddings can be found in Fig. 2. The details of the experiments and the results of the experiments with randomly generated unembeddings can be found in the Appendix F.

5.2 Experiments with Real World Datasets

Datasets and models We conduct experiments on the dataset Safe-RLHF Dai et al. [2024] using the unified reward model [3] Dai et al. [2023]. We use an instruction tuned GPT2 [4] as our reference model to run DPO on and also generate responses for the prompts of the dataset. We also confirm results on Stanford Human Preferences (SHP)Ethayarajh et al. [2022], see Appendix F. All experiments were run on an NVIDIA A80 GPU.

Generating comparison data We pick $\sim 10\%$ (7300) of the prompts from original training dataset and generate 5 responses for each of these, which are then scored by the reward model.

The test data for our experiment is generated similarly by using 130 prompts from the original test data to (1)ensure a 1:4 ratio with the training data since a subset of 512 from total prompts will be picked later and (2) so that responses for both the train and test data come from the same reference distribution.

Picking the subsets of queries These prompts are then embedded (using the reference model), from which well-conditioned and bad-conditioned subsets of 512 prompts are collected (through random sampling). The subsets are selected based on the condition number $\frac{\sigma_1}{\sigma_k}$, where σ_i denotes

[3] `https://huggingface.co/PKU-Alignment/beaver-7b-unified-reward`

[4] `https://huggingface.co/RaushanTurganbay/GPT2_instruct_tuned`

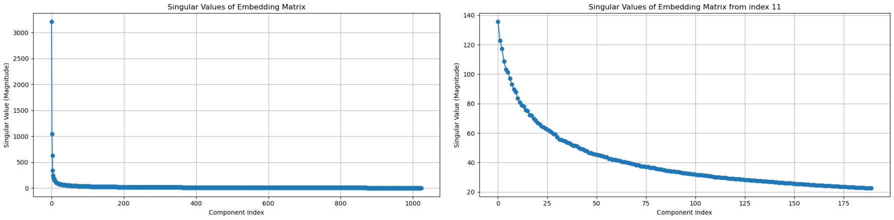

Figure 3: (a)Singular values of the embedding matrix (b)Singular values from λ_{11}

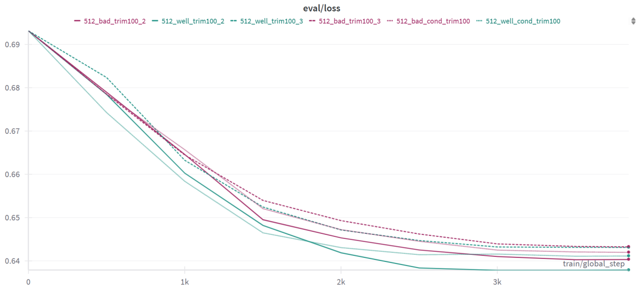

Figure 4: Evaluation loss for different picks of subsets based on condition number(green and red for well and bad-cond. subsets resp.)

the top singular values in decreasing order and $k = 100$. This is motivated by the observation that only a few directions in the 1024-dimensional embedding space appear to be meaningful (fig 3). We repeated the experiment 3 times for robustness and evaluated it on the same held out dataset. .

We then run DPO on the reference policy (unfreezing only the final unembedding layer.) with the training and evaluation loss set to the DPO loss with probabilities.

Results. We observe the following trend (fig 4) in evaluation loss when subsets are picked according to the condition number of the projected matrix (picked for the first 100 principle directions). We chose this as a comparison metric rather than training loss as our metric since the subsets would have different losses. The well (and bad)-conditioned matrices have the condition number in 113-115 (and 160-163). We summarize the average evaluation loss across training steps and the difference in the average steps required to reach various loss thresholds in the table below:

Eval Steps	Well-Cond Loss	Bad-Cond Loss
0	0.6931 ± 0.0000	0.6931 ± 0.0000
1000	0.6606 ± 0.0017	0.6646 ± 0.0002
2000	0.6440 ± 0.0020	0.6473 ± 0.0013
3000	0.6409 ± 0.0020	0.6425 ± 0.0009
3840	0.6407 ± 0.0018	0.6419 ± 0.0010

(a) Average evaluation loss (mean ± std).

Threshold	Well-cond	Bad-cond
0.690	106	110
0.675	595	640
0.660	1028	1190
0.650	1461	1708
0.645	1909	2409
0.652	2415	3471
0.641	2953	> 3840

(b) Steps to reach thresholds.

Table 1: Comparison of well-conditioned vs bad-conditioned subsets.

While the first table demonstrates that the loss remains consistently lower for the well-conditioned subset, the second table is central to our analysis, as it explicitly illustrates the faster convergence rate of the well-conditioned subset.

The code can be found at `https://anonymous.4open.science/r/Convergence-of-DPO-08E1`

6 Conclusion

Our work provides a rigorous theoretical foundation for understanding and improving Direct Preference Optimization (DPO) through a parameterization that closely mirrors the architecture of modern transformer-based language models. We derive linear convergence guarantees for gradient descent under the assumption of connected comparison graphs and analyze how the structure and conditioning of prompt embeddings, along with the graph of response comparisons, fundamentally influence the convergence rate. Our results lead to strategies for data collection, emphasizing the importance of selecting prompts and comparisons that yield a well-conditioned embedding matrix and low spectral ratio in the Laplacian of the comparison graph. We further explore the generalizability of the implicit reward differences learned by the model, establishing conditions under which these learned preferences extend to new queries. Compared to previous works that rely on decoupled parameterizations, our approach captures the interconnected nature of logits through shared prompt embeddings. Our findings are consistently supported by both synthetic and real world data experiments.

Limitations and future work. Our model, while close to the actual architecture of the LLM allows for optimization only over the final layer, which might not explain the rates in different fine-tuning implementations. Our analysis also considers distributions over the next immediate token rather than the autoregressive generation which would change the hidden statement embedding after every predicted token. Understanding the system of parameterizations that have a linear convergence rate for DPO using the Bradley-Terry convergence analysis would be a direct potential extension of this work. Exploring the role of non-uniform samplers as in Shi et al. [2025] and the setting of Online DPO in our model would be interesting further problems to work on.

References

Daniele Calandriello, Zhaohan Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko, Tianqi Liu, Rishabh Joshi, Zeyu Zheng, and Bilal Piot. Human alignment of large language models through online preference optimisation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe rlhf: Safe reinforcement learning from human feedback, 2023. URL https://arxiv.org/abs/2310.12773.

Josef Dai, Xuehai Pan, Ruiyang Sun, Jiaming Ji, Xinbo Xu, Mickel Liu, Yizhou Wang, and Yaodong Yang. Safe RLHF: safe reinforcement learning from human feedback. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=TyFrPOKYXw.

Kawin Ethayarajh, Yejin Choi, and Swabha Swayamdipta. Understanding dataset difficulty with $\mathcal{V}$-usable information. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 5988–6008. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/ethayarajh22a.html.

Yunzhen Feng, Ariel Kwiatkowski, Kunhao Zheng, Julia Kempe, and Yaqi Duan. Pilaf: Optimal human preference sampling for reward modeling. *arXiv preprint arXiv:2502.04270*, 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 Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li, editors, *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 4447–4455. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/gheshlaghi-azar24a.html.

Chi Han, Jialiang Xu, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek Abdelzaher, and Heng Ji. Word embeddings are steers for language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16410–16430, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.864. URL https://aclanthology.org/2024.acl-long.864/.

Shawn Im and Yixuan Li. Understanding the learning dynamics of alignment with human feedback. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and

Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20983–21006. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/im24a.html.

Yibo Jiang, Goutham Rajendran, Pradeep Kumar Ravikumar, Bryon Aragam, and Victor Veitch. On the origins of linear representations in large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 21879–21911. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/jiang24d.html.

Andi Nika, Debmalya Mandal, Parameswaran Kamalaruban, Georgios Tzannetos, Goran Radanovic, and Adish Singla. Reward model learning vs. direct policy optimization: A comparative analysis of learning from human preferences. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 38145–38186. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/nika24a.html.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models, 2025. URL https://arxiv.org/abs/2406.01506.

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. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 53728–53741. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a85b405ed65c6477a4fe8302b5e06ce7-Paper-Conference.pdf.

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Unintentional unalignment: Likelihood displacement in direct preference optimization, 2025. URL https://arxiv.org/abs/2410.08847.

Ruizhe Shi, Runlong Zhou, and Simon S. Du. The crucial role of samplers in online direct preference optimization, 2025. URL https://arxiv.org/abs/2409.19605.

Yunhao Tang, Zhaohan Daniel Guo, Zeyu Zheng, Daniele Calandriello, Remi Munos, Mark Rowland, Pierre Harvey Richemond, Michal Valko, Bernardo Avila Pires, and Bilal Piot. Generalized preference optimization: A unified approach to offline alignment. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 47725–47742. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/tang24b.html.

Milan Vojnovic, Se-Young Yun, and Kaifang Zhou. Convergence rates of gradient descent and mm algorithms for bradley-terry models. In Silvia Chiappa and Roberto Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1254–1264. PMLR, 26–28 Aug 2020. URL https://proceedings.mlr.press/v108/vojnovic20a.html.

Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint, 2024. URL https://arxiv.org/abs/2312.11456.

Hui Yuan, Yifan Zeng, Yue Wu, Huazheng Wang, Mengdi Wang, and Liu Leqi. A common pitfall of margin-based language model alignment: Gradient entanglement, 2025. URL https://arxiv.org/abs/2410.13828.

Banghua Zhu, Michael Jordan, and Jiantao Jiao. Principled reinforcement learning with human feedback from pairwise or k-wise comparisons. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 43037–43067. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/zhu23f.html.

Brian D. Ziebart. Modeling purposeful adaptive behavior with the principle of maximum causal entropy. 2010. URL https://api.semanticscholar.org/CorpusID:11919065.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL https://arxiv.org/abs/1909.08593.

A Preliminaries for Proofs

A.1 Notes on Notation

We use the following shorthand notation across the proofs for simplicity of expression:

- $h(\mathbf{W}_i, \mathbf{x}^v)$ to refer to the implicit reward difference $\beta(\mathbf{W}_i - \mathbf{W}_i^{(\text{ref})})\mathbf{x}^v$. We also use it freely for $h(\mathbf{W}, \mathbf{X}) = \beta(\mathbf{W} - \mathbf{W}^{(\text{ref})})\mathbf{X}$ or $h(\mathbf{W}, \mathbf{x}^v) = \beta(\mathbf{W} - \mathbf{W}^{(\text{ref})})\mathbf{x}^v$.
- The learned/implicit rewards for prompt x^v and response y_i, $r_{v,i} = \beta(\mathbf{W}_i - \mathbf{W}_i^{(\text{ref})})\mathbf{x}^v$ unless a different $\mathbf{W}$ has been specified. Further, $\mathbf{r}_v = h(\mathbf{W}, \mathbf{x}^v)$ and $\mathbf{r} = h(\mathbf{W}, \mathbf{X})$.
- $\alpha_{\mathbf{W},v,i} = \exp(r_{\mathbf{W},v,i}) = \dfrac{e^{\beta \mathbf{W}_i \mathbf{x}^v}}{e^{\beta(\mathbf{W}_i^{(\text{ref})})\mathbf{x}^v}}$
- $\Delta_v(y_i, y_j; \mathbf{W}) = \sigma(r^\star_{v,i} - r^\star_{v,j}) - \sigma\left(h(\mathbf{W}_i, \mathbf{x}^v) - h(\mathbf{W}_j, \mathbf{x}^v)\right)$

Some basic concepts from convex optimization analysis that we use in the paper are defined below:

Strong Convexity A function $f : \mathbb{R}^n \to \mathbb{R}$ is said to be μ-**strongly convex** on X if it satisfies the following subgradient inequality: for all $x, y \in X$,

$$f(x) \geq f(y) + \nabla f(y)^\top (x - y) + \frac{\mu}{2}\|x - y\|^2.$$

Equivalently, the function f is μ-strongly convex on X if and only if $f(x) - \frac{\mu}{2}\|x\|^2$ is convex on X.

Smoothness A function f is said to be L-**smooth** on X if its gradient ∇f is L-Lipschitz on X, i.e.,

$$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|, \quad \forall x, y \in X.$$

For any L-smooth function f on X, the following property holds :

$$f(x) \leq f(y) + \nabla f(y)^\top (x - y) + \frac{L}{2}\|x - y\|^2, \quad \forall x, y \in X.$$

Polyak–Łojasiewicz (PL) Inequality A function f is said to satisfy the **Polyak–Łojasiewicz (PL) inequality** on X if there exists $\mu > 0$ such that

$$f(x) - f(x^*) \leq \frac{1}{2\mu}\|\nabla f(x)\|^2, \quad \forall x \in X,$$

where x^* is a minimizer of f. When the PL inequality holds on X for a specific value of μ, we say that the μ-PL inequality holds on X.

If f is μ-strongly convex on X, then f satisfies the μ-PL inequality on X.

A.2 Convergence Rates for the Bradley-Terry Model

Given a pairwise preference dataset $\mathcal{D}$ over the response set Y of size n (note that this would be for a fixed response in RLHF) and a set of scores $\mathbf{r} = (r_1, r_2 \cdots, r_n)$ each r_i corresponding to a $y \in Y$, the Bradley-Terry negative log-likelihood loss would be:

$$\mathcal{L}_{BT}(\mathbf{r}) = -\sum_{i=1}^{|Y|}\sum_{j \neq i} d_{ij} \log\left(\frac{e^{r_i}}{e^{r_i} + e^{r_j}}\right) = -\sum_{i=1}^{|Y|}\sum_{j \neq i} d_{ij}\left(r_i - \log(e^{r_i} + e^{r_j})\right) \tag{5}$$

Vojnovic et al. [2020] shows that gradient descent on using MLE for fitting in the Bradley Terry model is linear with the rate crucially determined by the algebraic connectivity of the matrix $\mathbf{M}$ of item pair co-occurrences in observed comparison data. Note that we denote by $\mathbf{r}$ the row vector of scores.

Lemma A.1. *For any $\omega \geq 0$ the negative log-likelihood function for the Bradley-Terry model of paired comparisons is γ-strongly convex on $\mathcal{W}_{\omega,0} = \mathcal{W}_\omega \cap \{\mathbf{r} \in \mathbb{R}^{1\times n} : \mathbf{r}\mathbf{1}^T = 0\}$ when $\mathcal{W}_\omega = \{\mathbf{r} \in \mathbb{R}^{1\times n} : \|\mathbf{r}\|_\infty \leq \omega\}$ and $\mu-$smooth over $\mathbb{R}^n$ where*

$$\gamma = c_\omega \lambda_2(\mathbf{L}_\mathbf{M}) \text{ and } \mu = \frac{1}{4}\lambda_n(\mathbf{L}_\mathbf{M})$$

where $c_\omega = \frac{1}{e^{-\omega}+e^{\omega}}$, $\mathbf{L}_\mathbf{M}$ is the Laplacian matrix of $\mathbf{M}$. The function also satisfes γ-PL inequality over $\mathcal{W}_\omega$

By the Gershgorin circle theorem, $\lambda_n(\mathbf{L}_\mathbf{M}) \leq 2d(\mathbf{M})$ where $d(\mathbf{M})$ is the number of paired comparisons per item in $\mathbf{M}$ and $\lambda_2(\mathbf{L}_\mathbf{M})$ is known as the algebraic connectivity of M. This implies a $T = \mathcal{O}\left(\frac{d(\mathbf{M})}{a(\mathbf{M})}\log\frac{1}{\epsilon}\right)$ convergence time bound.

Note that for multiple inputs, the lemma extends easily since all the variables involved are independent from each other in the following loss.

$$\mathcal{L}_{BT}(\mathbf{r}) = -\sum_x \sum_{i=1}^{|Y|} \sum_{j\neq i} d_{x,i,j} \log\left(\frac{e^{r_{x,i}}}{e^{r_{x,i}} + e^{r_{x,j}}}\right) = -\sum_x \sum_{i=1}^{|Y|} \sum_{j\neq i} d_{x,i,j}\left(r_{x,i} - \log(e^{r_{x,i}} + e^{r_{x,j}})\right) \tag{6}$$

Lemma A.2. *For any $\omega \geq 0$ the negative log-likelihood function for the Bradley-Terry model of paired comparisons over multiple inputs 6 is γ-strongly convex on $\mathcal{W}_{\omega,0} = \mathcal{W}_\omega \cap \{\mathbf{r} \in \mathbb{R}^{m\times n} : \mathbf{r}\mathbf{1}^T = 0\}$ when $\mathcal{W}_\omega = \{\mathbf{r} \in \mathbb{R}^{m\times n} : \|\mathbf{r}\|_\infty \leq \omega\}$ and $\mu-$smooth over $\mathbb{R}^{m\times n}$ where*

$$\gamma = c_\omega \min_x \lambda_2(\mathbf{L}_{\mathbf{M}_x}) \text{ and } \mu = \frac{1}{4}\max_x \lambda_n(\mathbf{L}_{\mathbf{M}_x})$$

where $c_\omega = \frac{1}{e^{-\omega}+e^{\omega}}$, $\mathbf{L}_\mathbf{M}$ is the Laplacian matrix of $\mathbf{M}$. The function also satisfes γ-PL inequality over $\mathcal{W}_\omega$

B Linear Convergence of DPO

B.1 Proof of Lemma 3.1

The second derivatives of the DPO loss function can be written as,

$$\frac{\partial^2 \mathcal{L}_{\text{DPO}}(\mathbf{W};\mathbf{X})}{\partial W_{j,k} W_{i,k}} = \begin{cases} \beta^2 \sum_{v=1}^m x_{v,k} x_{v,l} \sum_{j\neq i} m_{v,i,j} \frac{\alpha_{v,i}\alpha_{v,j}}{(\alpha_{v,i}+\alpha_{v,j})^2}, & i = j \\ -\beta^2 \sum_{v=1}^m x_{v,k} x_{v,l} m_{v,i,j} \frac{\alpha_{v,i}\alpha_{v,j}}{(\alpha_{v,i}+\alpha_{v,j})^2}, & i \neq j \end{cases}$$

where we use $\alpha_{\mathbf{W},v,i} = \exp(\beta(\mathbf{W}_i - \mathbf{W}_i^{(\text{ref})})\mathbf{x}^v)$ for simplicity of expression. We now write the quadratic form of the Hessian of the loss function in terms of the vector $\text{vec}(\mathbf{w}) \in \mathbb{R}^{nd}$ with

$\mathbf{w} \in \mathbb{R}^{n \times d}$. The quadratic form is written as follows,

$$\text{vec}(\mathbf{w})^\top \nabla^2 \mathcal{L}_{\text{DPO}}(\mathbf{W}; \mathbf{X}) \text{vec}(\mathbf{w}) \tag{7}$$

$$= \sum_{v=1}^{m} \sum_{i=1}^{n} \sum_{k=1}^{d} \sum_{l=1}^{d} \left[\frac{\partial^2 \mathcal{L}_{\text{DPO}}(\mathbf{W}; \mathbf{X})}{\partial W_{i,k} W_{i,l}} w_{i,k} w_{i,l} + \sum_{\substack{j=1 \\ j \neq i}}^{n} \frac{\partial^2 \mathcal{L}_{\text{DPO}}(\mathbf{W}; \mathbf{X})}{\partial W_{j,l} W_{i,k}} w_{j,l} w_{i,k} \right] \tag{8}$$

$$= \sum_{v=1}^{m} \sum_{i=1}^{n} \sum_{k=1}^{d} \sum_{l=1}^{d} \sum_{\substack{j=1 \\ j \neq i}}^{n} \beta^2 x_{v,k} x_{v,l} m_{v,i,j} \frac{\alpha_{\mathbf{W},v,i} \alpha_{\mathbf{W},v,j}}{(\alpha_{\mathbf{W},v,i} + \alpha_{\mathbf{W},v,j})^2} \left(w_{i,k} w_{i,l} - w_{i,k} w_{j,l} \right) \tag{9}$$

$$= \sum_{v=1}^{m} \sum_{i=1}^{n} \sum_{j>i}^{n} \beta^2 m_{v,i,j} \frac{\alpha_{\mathbf{W},v,i} \alpha_{\mathbf{W},v,j}}{(\alpha_{\mathbf{W},v,i} + \alpha_{\mathbf{W},v,j})^2} \sum_{k=1}^{d} \sum_{l=1}^{d} x_{v,k} x_{v,l} (w_{i,k} - w_{j,k})(w_{i,l} - w_{j,l}) \tag{10}$$

$$= \sum_{v=1}^{m} \sum_{i=1}^{n} \sum_{j>i}^{n} \beta^2 m_{v,i,j} \frac{\alpha_{\mathbf{W},v,i} \alpha_{\mathbf{W},v,j}}{(\alpha_{\mathbf{W},v,i} + \alpha_{\mathbf{W},v,j})^2} \left((\mathbf{w}_i - \mathbf{w}_j) \mathbf{x}^v \right)^2 \tag{11}$$

Say, $\mathbf{w} = \mathbf{W}' - \mathbf{W}^{(\text{ref})}$ and $r'_{v,i} = \mathbf{w}_i \mathbf{x}^v$. For the values of implicit reward $r_{\mathbf{W},v,i} = \beta(\mathbf{W}_i - \mathbf{W}^{(\text{ref})})\mathbf{x}^v$ in $[-\rho, \rho]$, we obtain the lower bound on $\frac{\alpha_{\mathbf{W},v,i} \alpha_{\mathbf{W},v,j}}{(\alpha_{\mathbf{W},v,i} + \alpha_{\mathbf{W},v,j})^2}$ as $c_\rho = \frac{1}{(e^\rho + e^{-\rho})^2}$. This is because it takes the form $z(1-z)$ where $z \in \left\{ \frac{e^{-\rho}}{e^\rho + e^{-\rho}}, \frac{e^\rho}{e^\rho + e^{-\rho}} \right\}$ and $z(1-z)$ obtains its minimum at the boundaries. We then obtain,

$$\begin{aligned}
\text{vec}(\mathbf{w})^\top \nabla^2 \mathcal{L}_{\text{DPO}}(\mathbf{W}; \mathbf{X}) \text{vec}(\mathbf{w}) \geq & c_\rho \frac{\beta^2}{2} \sum_{v=1}^{m} \sum_{i=1}^{n} \sum_{j>i}^{n} m_{v,i,j} (r^2_{\mathbf{w},v,i} + r^2_{\mathbf{w},v,j}) - c_\rho \beta^2 \sum_{i=1}^{n} \sum_{\substack{j=1 \\ j \neq i}}^{n} m_{v,i,j} r_{\mathbf{w},v,i} r_{\mathbf{w},v,j} \\
= & c_\rho \beta^2 \sum_{v=1}^{m} \sum_{i=1}^{n} D_{v,i} r^2_{\mathbf{w},v,i} - c_\rho \beta^2 \sum_{v=1}^{m} \sum_{i=1}^{n} \sum_{j=1}^{n} m_{v,i,j} r_{\mathbf{w},v,i} r_{\mathbf{w},v,j} \\
= & c_\rho \text{vec}(\mathbf{r})^\top \mathbf{L_M}^{(\text{all})} \text{vec}(\mathbf{r})
\end{aligned}$$

where, $D_{v,i} = \sum_{\substack{j=1 \\ j \neq i}}^{n} m_{v,i,j}$. $\mathbf{L_M}^{(\text{all})}$ is the Laplacian of the graph of comparisons G_{all} in which the prompts-response pairs $(x^v, y_i) \in \mathcal{X} \times \mathcal{Y}$ form the nodes. For two responses y_i and y_j that are compared for the input prompt x^v in the data, an edge of weight $m_{v,i,j}$ exists between (x^v, y_i) and (x^v, y_j) in G_{all}. Note that this forms a graph with mn nodes with at least m connected components as no edge or path lies between any two nodes of different prompts. In this graph, $D_{v,i}$ represents the degree of the node (x^v, y_i). Matrix $\mathbf{D} \in \mathbb{R}^{mn}$ formed by $D_{v,i}$ along its diagonals forms the degree matrix and matrix $\mathbf{M} \in \mathbb{R}^{mn}$ with elements $m_{v,i,j}$ forms the adjacency matrix of graph G_{all}.

Let each of the m subgraphs of G_{all} for the prompt $x^v \in \mathcal{X}$ be denoted by G_v. The $(m+1)^{th}$ smallest eigenvalue of the Laplacian of G_{all}, represented by $\lambda_{m+1}(\mathbf{L_M}^{(\text{all})})$, is the smallest non-zero eigen-value when G_v is connected for each $x^v \in \mathcal{X}$. In fact, the Laplacian $\mathbf{L_M}^{(\text{all})}$ is formed by placing the Laplacians of each of the G_v's along the diagonal and its eigenvalues are given by the combination of eigenvalues of the Laplacians of each G_v.

Now, in the space $\mathbf{W}' \in \mathcal{W}_0$, $\beta(\mathbf{W}' - \mathbf{W}^{(\text{ref})})^\top \mathbf{1}_n = 0 \implies \mathbf{r}\mathbf{1}_m = 0$. That is, $\mathcal{W}_0$ is orthogonal to the null space of $\mathbf{L_M}^{(\text{all})}$. Therefore, for $\mathbf{W}' \in \mathcal{W}_0$,

$$\text{vec}(\mathbf{w})^\top \nabla^2 \mathcal{L}_{\text{DPO}}(\mathbf{W}; \mathbf{X}) \text{vec}(\mathbf{w}) \geq c_\rho \lambda_{m+1}(\mathbf{L_M}^{(\text{all})}) \parallel \beta \mathbf{x}^v (\mathbf{W}' - \mathbf{W}^{(\text{ref})})^\top \parallel_F^2$$

Now, when $\mathbf{W} \in \mathcal{W}_X$, every row in $\mathbf{W}$ is orthogonal to the null space of $\mathbf{X}$ as well. Therefore, $\parallel (\mathbf{W}' - \mathbf{W})\mathbf{x}^v \parallel_F^2 \geq \left(\sigma^+_{\min}(\mathbf{X}) \right) \parallel \mathbf{W}' - \mathbf{W}^{(\text{ref})} \parallel^2$. We now obtain,

$$\text{vec}(\mathbf{w})^\top \nabla^2 \mathcal{L}_{\text{DPO}}(\mathbf{W};\mathbf{X})\text{vec}(\mathbf{w}) \geq c_\rho \beta^2 \lambda_{m+1}(\mathbf{L_M}^{(\text{all})}) \left(\sigma_{\min}^{+}(\mathbf{W})\right)^2 \parallel \mathbf{w} \parallel_F^2$$

Therefore, we observe γ'-strong convexity in the space $\mathbf{W} \in \mathcal{W}_0 \cap \mathcal{W}_\rho \cap \mathcal{W}_X$ when G_v is connected for all $x^v \in \mathcal{X}$ with $\gamma' = c_\rho \beta^2 \lambda_{m+1}(\mathbf{L_M}^{(\text{all})}) \left(\sigma_{\min}^{+}(\mathbf{X})\right)^2 \parallel \mathbf{x} \parallel^2$

Now, we observe that the maximum of $z(1-z)$ is observed at $1/4$. Therefore, using the result in eq. 11, we obtain,

$$\begin{aligned}\text{vec}(\mathbf{w})^\top \nabla^2 \mathcal{L}_{\text{DPO}}(\mathbf{W};\mathbf{X})\text{vec}(\mathbf{w}) \leq & \frac{1}{4}\sum_{v=1}^{m}\sum_{i=1}^{n}\sum_{j>i}^{n} \beta^2 m_{v,i,j} \left((\mathbf{w}_i - \mathbf{w}_j)\mathbf{x}^v\right)^2 \\ = & \frac{1}{4}\text{vec}(\mathbf{r})^\top \mathbf{L_M}^{(\text{all})}\text{vec}(\mathbf{r}) \\ \leq & \frac{\beta^2}{4}\lambda_{mn}(\mathbf{L_M}^{(\text{all})})\sigma_{\max}^2(\mathbf{X}) \parallel \mathbf{w} \parallel_F^2\end{aligned}$$

where $\lambda_{mn}(\mathbf{L_M}^{(\text{all})})$ is the largest eigenvalue of $\mathbf{L_M}^{(\text{all})}$. Therefore, it is μ-smooth for all $\mathbf{X} \in \mathbb{R}^{d\times m}$ and $\mathbf{W} \in \mathbb{R}^{n\times d}$ with $\mu = \frac{\beta^2}{4}\lambda_{mn}(\mathbf{L_M}^{(\text{all})})\sigma_{\max}^2(\mathbf{X})$.

B.2 Proof of Theorem 3.2

The gradient of the DPO loss function is obtained as,

$$\frac{\partial \mathcal{L}_{\text{DPO}}(\mathbf{W};\mathbf{X})}{\partial W_{i,k}} = -\beta \sum_{v}^{m} x_{v,k} \sum_{j\neq i} \frac{d_{v,j,i}\alpha_{v,i} - dv,i,j\alpha_{v,j}}{\alpha_{v,i}+\alpha_{v,j}}$$

We define a function $\Pi(\mathbf{W};\mathbf{W}_0,\mathbf{c}) = \mathbf{W} + \mathbf{1}_n\mathbf{c}^\top + \mathbf{W}_0$ for some column vector $\mathbf{c} \in \mathbb{R}^d$ and $\mathbf{W}_0 \in \mathcal{W}'_X$ such that $\mathbf{W}_0 + \mathbf{W}^{(\text{ref})} \in \mathcal{W}_{2\rho} \cap \mathcal{W}_0$ where $\mathcal{W}'_X = \{\mathbf{W} \in \mathbb{R}^{n\times d} : (\mathbf{W}_i - \mathbf{W}_i^{(\text{ref})})\mathbf{x} = 0 \; \forall i \in [n] \; \forall \mathbf{x} \notin \text{null}(\mathbf{X})\}$. That is, to each i^{th} column in $\mathbf{W}$, we are adding an all-ones vector $\mathbf{1}_n$ scaled by c_i, and, to each row in $\mathbf{W}$, we are adding a row vector from the null space of $\mathbf{X}$, shifted by rows in $\mathbf{W}^{(\text{ref})}$. Therefore, we make the following claim:

Claim B.1. *Any matrix $\mathbf{W}' \in \mathcal{W}_\rho$ can be represented as $\mathbf{W}' = \mathbf{W} + \mathbf{1}_n\mathbf{c}^\top + \mathbf{W}_0$ for some choice of $\mathbf{W} \in \mathcal{W}_X \cap \mathcal{W}_{2\rho} \cap \mathcal{W}_0$, $\mathbf{c} \in \mathbb{R}^d$, and $\mathbf{W}_0 \in \mathcal{W}'_X$ such that $\mathbf{W}_0 + \mathbf{W}^{(ref)} \in \mathcal{W}_{2\rho} \cap \mathcal{W}_0$.*

Proof. We observe that for any $\mathbf{W}' \in \mathcal{W}_\rho$, we can define $\mathbf{c}$ such that $c_i = 1/n \sum_j (W'_{j,i} - W_{j,i}^{(\text{ref})})$, $\tilde{\mathbf{W}} = \mathbf{W}' - \mathbf{W}^{(\text{ref})} - \mathbf{1}_n\mathbf{c}^\top$, and $\mathbf{W}_0 = \tilde{\mathbf{W}}(I - P_{\text{row}(\mathbf{X})})$ where $P_{\text{row}(\mathbf{X})} = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top)^\dagger\mathbf{X}$ is the matrix for projection on the row space of $\mathbf{X}$. We also define $P_{\text{null}(\mathbf{X})} = I - P_{\text{row}(\mathbf{X})}$. We now note that $(\mathbf{W}_0 + \mathbf{W}^{(\text{ref})} - \mathbf{W}^{(\text{ref})})^\top \mathbf{1}_n = (I - P_{\text{row}(\mathbf{X})})^\top \tilde{\mathbf{W}}^\top \mathbf{1}_n = 0$ and,

$$\begin{aligned}&\parallel \beta(\mathbf{W}_0 + \mathbf{W}^{(\text{ref})} - \mathbf{W}^{(\text{ref})})\mathbf{x} \parallel_\infty \\ &= \parallel \beta\tilde{\mathbf{W}}P_{\text{row}(\mathbf{X})}\mathbf{x} \parallel_\infty \\ &\leq \parallel \beta\tilde{\mathbf{W}}^\top\mathbf{x} \parallel_\infty \\ &\leq \parallel \beta(\mathbf{W}' - \mathbf{W}^{(\text{ref})})\mathbf{x} \parallel_\infty + \parallel \beta\mathbf{c}\mathbf{1}_n\mathbf{x} \parallel_\infty \\ &\leq 2\rho\end{aligned}$$

We obtain $\parallel \beta\mathbf{c}\mathbf{1}_n\mathbf{x} \parallel_\infty \leq \rho$, using the fact that $\beta\mathbf{c}\mathbf{1}_n\mathbf{x} = [\beta\langle\mathbf{x},\mathbf{c}\rangle, \beta\langle\mathbf{x},\mathbf{c}\rangle, \ldots \beta\langle\mathbf{x},\mathbf{c}\rangle] \in \mathbb{R}^{1\times n}$ and $\langle\mathbf{x},\mathbf{c}\rangle = \langle\mathbf{x}, \frac{1}{n}\sum_j^n(\mathbf{W}'_j - \mathbf{W}_j^{(\text{ref})})\rangle = \frac{1}{n}\sum_j^n\langle\mathbf{x}, \mathbf{W}'_j - \mathbf{W}_j^{(\text{ref})}\rangle \leq \max_j\langle\mathbf{x}, \mathbf{W}'_j - \mathbf{W}_j^{(\text{ref})}\rangle \leq \rho/\beta$, since $\mathbf{W}' \in \mathcal{W}_\rho$.

Hence, we have, $\mathbf{W}_0 \in \mathcal{W}'_X$ and $\mathbf{W}_0 + \mathbf{W}^{(\text{ref})} \in \mathcal{W}_{2\rho} \cap \mathcal{W}_0$. Further, we have $\mathbf{W} = \mathbf{W}' - \mathbf{1}_n\mathbf{c}^\top - \mathbf{W}_0$. Therefore, $(\mathbf{W} - \mathbf{W}^{(\text{ref})})^\top\mathbf{1}_n = \tilde{\mathbf{W}}^\top\mathbf{1}_n - \mathbf{W}_0^\top\mathbf{1}_n = 0$. Further, $\mathbf{W} - \mathbf{W}^{(\text{ref})} = \tilde{\mathbf{W}} - \tilde{\mathbf{W}}(I - P_{\text{row}(\mathbf{X})}) = \tilde{\mathbf{W}}P_{\text{null}(\mathbf{X})}$. This means that, $\parallel (\mathbf{W} - \mathbf{W}^{(\text{ref})})\mathbf{x} \parallel_\infty = \parallel \tilde{\mathbf{W}}P_{\text{row}(\mathbf{X})}\mathbf{x} \parallel_\infty \leq \parallel \tilde{\mathbf{W}}\mathbf{x} \parallel_\infty \leq 2\rho$. Therefore, we can say that $\mathbf{W} \in \mathcal{W}_X \cap \mathcal{W}_0 \cap \mathcal{W}_{2\rho}$.

Hence, for any $\mathbf{W}' \in \mathcal{W}_\rho$, we can identify a $\mathbf{W} \in \mathcal{W}_X \cap \mathcal{W}_0 \cap \mathcal{W}_{2\rho}$ such that $\Pi(\mathbf{W}; \mathbf{W}_0, \mathbf{c}) = \mathbf{W} + \mathbf{1}_n\mathbf{c}^\top + \mathbf{W}_0 = \mathbf{W}'$ for some $\mathbf{c} \in \mathbb{R}^d$ and $\mathbf{W}_0 \in \mathcal{W}'_X$ such that $\mathbf{W}_0 + \mathbf{W}^{\text{(ref)}} \in \mathcal{W}_{2\rho} \cap \mathcal{W}_0$. □

We note that, for some $\mathbf{W}_0 \in \mathcal{W}'_X$ such that $\mathbf{W}_0 + \mathbf{W}^{\text{(ref)}} \in \mathcal{W}_{2\rho} \cap \mathcal{W}_0$,

$$\begin{aligned}
&\tilde{r}(\Pi(\mathbf{W}; \mathbf{W}_0, \mathbf{c}); \mathbf{x}) = \beta\mathbf{x}(\Pi(\mathbf{W}; \mathbf{W}_0, \mathbf{c}) - \mathbf{W}^{ref})^\top = \tilde{r}(\mathbf{W}; \mathbf{x}) + \beta\mathbf{x}\mathbf{c} \\
\implies &\mathcal{L}_{\text{DPO}}(\Pi(\mathbf{W}; \mathbf{W}_0, \mathbf{c}); \mathbf{X}) \\
&= -\sum_v^m \sum_i^n \sum_j^n d_{v,i,j}(\tilde{r}(\Pi(\mathbf{W}; \mathbf{W}_0, \mathbf{c})_i; \mathbf{x}_v) - \log(e^{\tilde{r}(\Pi(\mathbf{W}; \mathbf{W}_0, \mathbf{c})_i; \mathbf{x}_v)} + e^{\tilde{r}(\Pi(\mathbf{W}; \mathbf{W}_0, \mathbf{c})_j; \mathbf{x}_v)})) \\
&= -\sum_v^m \sum_i^n \sum_j^n d_{v,i,j} \log\left(\frac{e^{\tilde{r}(\mathbf{W}_i; \mathbf{x}_v) + \beta\mathbf{x}\mathbf{c}}}{e^{\tilde{r}(\mathbf{W}_i; \mathbf{x}_v) + \beta\mathbf{x}\mathbf{c}} + e^{\tilde{r}(\mathbf{W}_j; \mathbf{x}_v) + \beta\mathbf{x}\mathbf{c}}}\right) \\
&= -\sum_v^m \sum_i^n \sum_j^n d_{v,i,j} \log\left(\frac{e^{\tilde{r}(\mathbf{W}_i; \mathbf{x}_v)}}{e^{\tilde{r}(\mathbf{W}_i; \mathbf{x}_v)} + e^{\tilde{r}(\mathbf{W}_j; \mathbf{x}_v)}}\right) \\
&= \mathcal{L}_{\text{DPO}}(\mathbf{W}; \mathbf{X})
\end{aligned}$$

Similarly,

$$\begin{aligned}
\nabla_{W_{i,k}} \mathcal{L}_{\text{DPO}}(\Pi(\mathbf{W}; \mathbf{W}_0, \mathbf{c}); \mathbf{X}) &= -\sum_{\mathbf{x}_v} x_{v,k} \sum_{j \neq i} \frac{d^v_{j,i} e^{\tilde{r}(\Pi(\mathbf{W}; \mathbf{W}_0, \mathbf{c})_i, \mathbf{x}^\mathbf{v})} - d^v_{i,j} e^{\tilde{r}(\Pi(\mathbf{W}; \mathbf{W}_0, \mathbf{c})_j, \mathbf{x}^\mathbf{v})}}{e^{\tilde{r}(\Pi(\mathbf{W}; \mathbf{W}_0, \mathbf{c})_i, \mathbf{x}^\mathbf{v})} + e^{\tilde{r}(\Pi(\mathbf{W}; \mathbf{W}_0, \mathbf{c})_j, \mathbf{x}^\mathbf{v})}} \\
&= -\sum_{\mathbf{x}_v} x_{v,k} \sum_{j \neq i} \frac{d^v_{j,i} e^{\tilde{r}(\mathbf{W}_i, \mathbf{x}^\mathbf{v})} - d^v_{i,j} e^{\tilde{r}(\mathbf{W}_j, \mathbf{x}^\mathbf{v})}}{e^{\tilde{r}(\mathbf{W}_i, \mathbf{x}^\mathbf{v})} + e^{\tilde{r}(\mathbf{W}_j, \mathbf{x}^\mathbf{v})}} \\
&= \nabla_{W_{i,k}} \mathcal{L}_{\text{DPO}}(\mathbf{W}; \mathbf{X})
\end{aligned}$$

Therefore, we have $\mathcal{L}_{\text{DPO}}(\Pi(\mathbf{W}; \mathbf{W}_0, \mathbf{c}); \mathbf{X}) = \mathcal{L}_{\text{DPO}}(\mathbf{W}; \mathbf{X})$ and $\nabla \mathcal{L}_{\text{DPO}}(\Pi(\mathbf{W}; \mathbf{W}_0, \mathbf{c}); \mathbf{X}) = \nabla \mathcal{L}_{\text{DPO}}(\mathbf{W}; \mathbf{X})$ when $\mathbf{W}_0 \in \mathcal{W}'_X$ such that $\mathbf{W}_0 + \mathbf{W}^{\text{(ref)}} \in \mathcal{W}_{2\rho} \cap \mathcal{W}_0$.

Suppose that $\mathbf{W}^\star$ and $\mathbf{W}^\star_{2\rho,0,X}$ are the minima of $\mathcal{L}_{\text{DPO}}$ over $\mathcal{W}_\rho$ and $\mathcal{W}_{2\rho} \cap \mathcal{W}_0 \cap \mathcal{W}_X$ respectively.

Say we have some $\mathbf{W}' \in \mathcal{W}_\rho$. From the claim above we know that there exists a $\mathbf{W}, \mathbf{W}_1 \in \mathcal{W}_{2\rho} \cap \mathcal{W}_0 \cap \mathcal{W}_X$ such that $\mathbf{W}' = \Pi(\mathbf{W}; \mathbf{W}_{01}, \mathbf{c}_1)$, $\mathbf{W}^\star = \Pi(\mathbf{W}_1; \mathbf{W}_{02}, \mathbf{c}_2)$ for some $\mathbf{c}_1, \mathbf{c}_2 \in \mathbb{R}^d$ and $\mathbf{W}_{01} + \mathbf{W}^{\text{(ref)}}, \mathbf{W}_{02} + \mathbf{W}^{\text{(ref)}} \in \mathcal{W}_{2\rho} \cap \mathcal{W}_0$. We now have

$$\begin{aligned}
\mathcal{L}_{\text{DPO}}(\mathbf{W}'; \mathbf{X}) - \mathcal{L}_{\text{DPO}}(\mathbf{W}^\star; \mathbf{X}) &= \mathcal{L}_{\text{DPO}}(\Pi(\mathbf{W}; \mathbf{W}_{01}, \mathbf{c}_1); \mathbf{X}) - \mathcal{L}_{\text{DPO}}(\Pi(\mathbf{W}_1; \mathbf{W}_{02}, \mathbf{c}_2); \mathbf{X}) \\
&= \mathcal{L}_{\text{DPO}}(\mathbf{W}; \mathbf{X}) - \mathcal{L}_{\text{DPO}}(\mathbf{W}_1; \mathbf{X}) \\
&\leq \mathcal{L}_{\text{DPO}}(\mathbf{W}; \mathbf{X}) - \mathcal{L}_{\text{DPO}}(\mathbf{W}^\star_{2\rho,0,X}; \mathbf{X}) \qquad (12)
\end{aligned}$$

The strong convexity condition from Lemma 3.1 implies that γ''-PL inequality must also be satisfied for $\mathbf{W} \in \mathcal{W}_{2\rho} \cap \mathcal{W}_0 \cap \mathcal{W}_X$ where $\gamma'' = c_{2\rho}\beta^2\sigma^2_{min}(\mathbf{X})\lambda_{m+1}(\mathbf{L}^{(all)}_M)$.

$$\mathcal{L}_{\text{DPO}}(\mathbf{W}; \mathbf{X}) - \mathcal{L}_{\text{DPO}}(\mathbf{W}^\star_{2\rho,0,X}; \mathbf{X}) \leq \frac{1}{2\gamma''} \parallel \nabla \mathcal{L}_{\text{DPO}}(\mathbf{W}; \mathbf{X}) \parallel^2$$

We now have,

$$\begin{aligned}
\implies \mathcal{L}_{\text{DPO}}(\mathbf{W}'; \mathbf{X}) - \mathcal{L}_{\text{DPO}}(\mathbf{W}^\star; \mathbf{X}) &\leq \frac{1}{2\gamma''} \parallel \nabla \mathcal{L}_{\text{DPO}}(\mathbf{W}; \mathbf{X}) \parallel^2 \\
&= \frac{1}{2\gamma''} \parallel \nabla \mathcal{L}_{\text{DPO}}(\Pi(\mathbf{W}; \mathbf{W}_0, \mathbf{c}); \mathbf{X}) \parallel^2 \\
&= \frac{1}{2\gamma''} \parallel \nabla \mathcal{L}_{\text{DPO}}(\mathbf{W}'; \mathbf{X}) \parallel^2 \qquad (13)
\end{aligned}$$

Therefore, $\mathcal{L}_{\text{DPO}}(\mathbf{W}; \mathbf{X})$ satisfies γ''-PL inequality on $\mathbf{W}' \in \mathcal{W}_\rho$ with $\gamma'' = c_{2\rho}\beta^2\sigma^2_{min}(\mathbf{X})\lambda_{m+1}(\mathbf{L}^{(all)}_M)$.

B.3 Proof of Theorem 3.3

Lemma B.2. *Say, function $f(\mathbf{x})$ is μ-smooth on $\mathcal{X}_\mu$ and satisfies the γ-PL inequality on $\mathcal{X}_\gamma$. Further, for the total number of iterations T, we have an error threshold ϵ such that, $f(\mathbf{x}^{(T)}) - f(\mathbf{x}^\star) \leq \epsilon$. When $\mathbf{x}^\star$ is the minimizer of f and $\mathbf{x}^{(t+1)}$ is the next step after $\mathbf{x}^{(t)}$ obtained using the gradient descent algorithm with step size $\eta = 1/\mu$, then, for $\mathbf{x}^{(t)} \in \mathcal{X}_\gamma$ and $\mathbf{x}^{(t+1)} \in \mathcal{X}_\mu$, we have,*

$$f(\mathbf{x}^{(t+1)}) - f(\mathbf{x}^\star) \leq \left(1 - \frac{\gamma}{\mu}\right)(f(\mathbf{x}^{(t)}) - f(\mathbf{x}^\star))$$
$$\textit{and, } T = O\left(\frac{\gamma}{\mu}\log\left(\frac{1}{\epsilon}\right)\right)$$

Proof. This is a well-known result as described in Vojnovic et al. [2020] . At each time step t, we have $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - \eta\nabla f(\mathbf{x}^{(t)})$.

$$\begin{aligned} f(\mathbf{x}^{(t+1)}) - f(\mathbf{x}^\star) =& f(\mathbf{x}^{(t)} - \eta\nabla f(\mathbf{x}^{(t)})) - f(\mathbf{x}^\star) \\ \leq& f(\mathbf{x}^{(t)}) - \eta\|\nabla f(\mathbf{x}^{(t)})\|^2 + \frac{\mu}{2}\eta^2\|\nabla f(\mathbf{x}^{(t)})\|^2 - f(\mathbf{x}^\star) \\ =& f(\mathbf{x}^{(t)}) - f(\mathbf{x}^\star) - \left(\eta - \frac{\mu}{2}\eta^2\right)\|\nabla f(\mathbf{x}^{(t)})\|^2 \\ \leq& f(\mathbf{x}^{(t)}) - f(\mathbf{x}^\star) - \left(2\gamma\eta - \gamma\mu\eta^2\right)(f(\mathbf{x}^{(t)}) - f(\mathbf{x}^\star)) \\ =& \left(1 - 2\gamma\eta + \gamma\mu\eta^2\right)(f(\mathbf{x}^{(t)}) - f(\mathbf{x}^\star)) \end{aligned}$$

The first and second inequalities come from the properties of μ-smoothness and γ-PL inequality respectively. Taking $\eta = 1/\mu$ minimizes the above bound, as follows,

$$f(\mathbf{x}^{(t+1)}) - f(\mathbf{x}^\star) \leq \left(1 - \frac{\gamma}{\mu}\right)(f(\mathbf{x}^{(t)}) - f(\mathbf{x}^\star))$$

Now, for T total iterations starting at $\mathbf{x}^{(0)}$, this implies,

$$f(\mathbf{x}^{(T)}) - f(\mathbf{x}^\star) \leq \left(1 - \frac{\gamma}{\mu}\right)^T(f(\mathbf{x}^{(0)}) - f(\mathbf{x}^\star))$$

But, we want, $f(\mathbf{x}^{(T)}) - f(\mathbf{x}^\star) \leq \epsilon$. Setting the upper bound ϵ on the RHS, we obtain,

$$T \geq \log\left(\frac{f(\mathbf{x}^{(0)}) - f(\mathbf{x}^\star)}{\epsilon}\right)\frac{1}{\log\left(\frac{1}{1-\frac{\gamma}{\mu}}\right)}$$

For small values of γ/μ, we have $-\log(1 - \gamma/\mu) \approx \gamma/\mu$. Therefore, we get,

$$T \geq \frac{\mu}{\gamma}\log\left(\frac{f(\mathbf{x}^{(0)}) - f(\mathbf{x}^\star)}{\epsilon}\right)$$

□

Now, using the properties of γ''-PL inequality and μ-smoothness as described in Theorem 3.2 and 3.1, and applying it to Lemma $B.2$, we obtain the required number of iterations $T = O(\frac{\mu}{\gamma''}\log\left(\frac{1}{\epsilon}\right))$ for gradient descent on $\mathbf{W} \in \mathcal{W}_\rho$ when each graph of comparisons G_v for the response x_v is connected.

B.4 Relaxed convexity requirements for linear convergence

Lemma B.3. *Given linearly independent queries $\mathbf{X}$ whose comparisons are present in the dataset, we show that*

- *γ_ρ-strong convexity is satisfied in the domain $\mathcal{W}_\rho \cap \left(\bigcap_{\mathbf{x}\in X} \mathcal{W}_{\mathbf{x}}\right) \cap \mathcal{W}_{span(\mathbf{X})}$*
- *γ_ρ-PL inequality is satisfied in the domain $\mathcal{W}_\rho$*

where $\mathcal{W}_\rho = \{\mathbf{W} \in \mathbb{R}^{n\times d} : |\beta((\mathbf{W}_i - \mathbf{W}_j) - (\mathbf{W}_i^{ref} - \mathbf{W}_j^{ref}))\mathbf{x}| \leq \rho \forall (x, y_i, y_j) \in Q$,
$\mathcal{W}_{\mathbf{x}} = \{\mathbf{W} \in \mathbb{R}^{n\times d} : (\sum_{i\in C} \mathbf{x}\mathbf{W}_i) = 0$ for all connected components C in a graph $G_{\mathbf{x}}\}$,
$\mathcal{W}_{span(\mathbf{X})} = \left\{\mathbf{W} \in \mathbb{R}^{n\times d} : \mathbf{W}_i \in span(\mathbf{X}) \forall i\right\}$ Further, $\beta \in \mathbb{R}^+$ is the KL regularization coefficient,
$\gamma_\rho = c_\rho \lambda^+_{\min}(\mathbf{L_M}^{(all)})\sigma^+_{\min}(\mathbf{X})^2$ and $c_\rho = 1/(e^\rho + e^{-\rho})^2$

Proof. Note that for any $\mathbf{W} \in \mathcal{W}_\rho$ and $\mathbf{r}$ such that $\mathbf{r}_{v,i} = \beta \mathbf{w}_i \mathbf{x}^v$, the expression as obtained in the proof lemma B.1 is as follows:

$$\text{vec}(\mathbf{w})^\top \mathcal{L}_{\text{DPO}}(\mathbf{W};\mathbf{X})\text{vec}(\mathbf{w}) \geq c_\rho \text{vec}(\mathbf{r}_\mathbf{w})^\top \mathbf{L_M}^{(\text{all})}\text{vec}(\mathbf{r}_\mathbf{w})$$

We make the following 2 observations:

1. Suppose that $\mathbf{w} \in \left(\bigcap_{\mathbf{x}\in X} \mathcal{W}_\mathbf{x}\right)$, then the corresponding $\mathbf{r}$ has the property: for any (maximal) strongly connected component $C \in G_\mathbf{x}$, $\sum_{i\in V(C)} (\mathbf{r}_\mathbf{w})_{\mathbf{x},i} = 0$.
2. Given the Laplacian matrix of connected components (without any directed edges between them) it's null space is spanned by the orthogonal vectors $\{v_{\mathbf{x},C}, \forall C$ is a component of $G_\mathbf{x}, \forall \mathbf{x} \in X\}$ where $v_{\mathbf{x},C}$ is a vector of size nd such that all the indices indexed by the vertices of C are set to 1 and the remaining to 0.

Thus the above observations show that the vec$(\mathbf{r}_\mathbf{w})$ is orthogonal to null space (and hence orthogonal to all the 0-valued eigen vectors) of $\mathbf{L_M}^{(\text{all})}$. Through Eigen decomposition, we can conclude that

$$c_\rho \text{vec}(\mathbf{r}_\mathbf{w})^\top \mathbf{L_M}^{(\text{all})}\text{vec}(\mathbf{r}_\mathbf{w}) \geq c_\rho \lambda^+_{\min}(\mathbf{L_M}^{(\text{all})})^2 \text{vec}(\mathbf{r}_\mathbf{w})^\top \text{vec}(\mathbf{r}_\mathbf{w})$$

where $\lambda^+_{\min}(\mathbf{L_M}^{(\text{all})})$ is the minimum non zero eigen value of $\mathbf{L_M}^{(\text{all})}$.
Similar to the lemma 3.1 if $\mathbf{w} \in \mathcal{W}_{\text{span}(\mathbf{X})}$ as well then we know again by singular value decomposition (on all column vectors of $\mathbf{w}$) that $\text{vec}(\mathbf{r}_\mathbf{w})^\top \text{vec}(\mathbf{r}_\mathbf{w}) \geq \left(\sigma^+_{\min}(\mathbf{X})\right)^2 \text{vec}(\mathbf{w})^\top \text{vec}(\mathbf{w})$
This in turn implies that for any $z \in \left(\bigcap_{\mathbf{x}\in X} \mathcal{W}_\mathbf{x}\right) \cap \mathcal{W}_{\text{span}(\mathbf{X})}$ and $\mathbf{W} \in \mathcal{W}_\rho$,

$$\text{vec}(z)^\top \mathcal{L}_{\text{DPO}}(\mathbf{W};\mathbf{X})\text{vec}(z) \geq c_\rho \lambda^+_{\min}(\mathbf{L_M}^{(\text{all})})\sigma^+_{\min}(\mathbf{X})\|\text{vec}(z)\|_2^2$$

This shows that $\mathcal{L}_{\text{DPO}}(\mathbf{W};\mathbf{X})$ is $c_\rho \lambda^+_{\min}(\mathbf{L_M}^{(\text{all})})\sigma^+_{\min}(\mathbf{X})$-strongly convex over $\mathcal{W}_\rho \cap \left(\bigcap_{\mathbf{x}\in X} \mathcal{W}_\mathbf{x}\right) \cap \mathcal{W}_{\text{span}(\mathbf{X})}$.

Proposition B.4. *Note that for any $\mathbf{w} \in \mathcal{W}_\rho$, there exists an $\mathbf{w}' \in \mathcal{W}_\rho \cap \left(\bigcap_{\mathbf{x}\in X} \mathcal{W}_\mathbf{x}\right) \cap \mathcal{W}_{span(\mathbf{X})}$ such that*

1. *$\mathcal{L}_{DPO}(\mathbf{w};\mathbf{X}) = \mathcal{L}_{DPO}(\mathbf{w}';\mathbf{X})$*
2. *$\nabla\mathcal{L}_{DPO}(\mathbf{w};\mathbf{X}) = \nabla\mathcal{L}_{DPO}(\mathbf{w}';\mathbf{X})$*

Proof. $\mathbf{w}'$ can be constructed from $\mathbf{w}$ as follows:

a) For every component C of $G_\mathbf{x}$, we subtract a vector $\mathbf{x}_C \in \text{null}(\{x : x \text{ column of } \mathbf{X}\} \setminus \mathbf{x})$ from all $\mathbf{w}_i, i \in C$ as follows:

$$\sum_{i\in V(C)} (\mathbf{w}_i - \mathbf{x}_C)^T \mathbf{x} = 0$$

Note that there exists an $\mathbf{x}_C \in \text{null}(\{x : x \text{ row of } \mathbf{X}\} \setminus \mathbf{x})$ such that $\mathbf{x} \cdot \mathbf{x}_C \neq 0$ since we know that $\mathbf{x} \notin \text{span}(\{x : x \text{ column of } \mathbf{X}\} \setminus \mathbf{x})$ due to linear independence. If there was no such $\mathbf{x}_C$, then we would know that $\text{null}(\{x : x \text{ column of } \mathbf{X}\} \setminus \mathbf{x})$ is orthogonal to $\mathbf{x}$ thereby placing $\mathbf{x}$ in the row space of $\{x : x \text{ column of } \mathbf{X}\} \setminus \mathbf{x}$ which is a contradiction.

b) Project $\mathbf{w}^a$ obtained from the above on to the $\text{span}(\mathbf{X})$, in other words, multiply the projection matrix P that projects onto the row space of vectors X with $\mathbf{w}^a$

We now observe that the relevant reward differences remain preserved through the above transformations. Firstly, the removal of $\mathbf{x}_C$ doesn't affect any of the reward differences of the other queries, as it is perpendicular to all of them. Furthermore, any comparisons over C for $\mathbf{x}$ do not have any change in their reward difference since the same value is removed from both of them. Finally, projecting the $\mathbf{w}$ over $\text{span}(\mathbf{X})$ similarly does not change any of the reward differences. We note that both $\mathcal{L}_{\text{DPO}}$ and its gradient are invariant under transformations that preserve the reward differences of the observed comparisons.

To establish the lemma, we now only need to show that $\mathbf{w}' \in \mathcal{W}_\rho$ since tha bove transformations esnure that $\mathbf{w}' \in \left(\bigcap_{\mathbf{x}\in X} \mathcal{W}_{\mathbf{x}}\right) \cap \mathcal{W}_{\text{span}(\mathbf{X})}$ already. We note since the above transformation does not change the reward differences $\mathbf{w}'$ is still in $\mathcal{W}_\rho$.

□

The above lemma extends the $\gamma-$strong convexity in $\mathcal{W}_\rho \cap \left(\bigcap_{\mathbf{x}\in X} \mathcal{W}_{\mathbf{x}}\right) \cap \mathcal{W}_{\text{span}(\mathbf{X})}$ to the $\gamma-$PL inequality in $\mathcal{W}_\rho$ □

B.5 Listwise Preference data - Plackett-Luce model

Given a list-wise preference dataset $\mathcal{D}$ over the response set Y of size n (note that this would be for a fixed response in RLHF) and a set of scores $\mathbf{r} = (r_1, r_2 \cdots, r_n)$ each r_i corresponding to a $y \in Y$, the Bradley-Terry negative log-likelihood loss would be:

$$\mathcal{L}_{BT}(\mathbf{r}) = -\sum_{l\in\mathcal{D}}\sum_{i=1}^{|l|} \log\left(\frac{e^{r_{l_i}}}{\sum_{k=i}^{|l|} e^{r_{l_k}}}\right) \tag{14}$$

Similar to Vojnovic et al. [2020], all the proofs in the Bradley-Terry setting follow with a slight change in variables. The lemma 7.5 in Vojnovic et al. [2020] gives the exact ratio of constants with respect to Bradley Terry.

C Tabular Parameterization

Strong connectivity Note that if the comparison graph in A.2 is not strongly connected then there are strongly connected components that are either incomparable or there are edges only in one direction from some component to another. The first case clearly talks about the redundancy of considering a reward comparison between the responses of the two different components, which can be set to any arbitrary value.

In the other case, if there are only , then there does not exist a global minimum over $\mathbf{r}\mathbf{1}^T = 0$. This can be observed through contradiction; say, $\mathbf{r}^{\text{opt}}$ is the global minimum and C_1 is the strongly connected component that acts as a sink. Consider the following $\mathbf{r}'$ with $k > 0$

$$\mathbf{r}'_i = \begin{cases} \mathbf{r}_i^{\text{opt}} + k & \text{if } i \in C_1 \\ \mathbf{r}_i^{\text{opt}} & \text{ow} \end{cases}$$

Note that $\mathbf{r}'$, when normalized (subtract the average to make the components sum upto 1) gives lesser negative log likelihood. While this is not to say that there does not exist an infimum for the function, the reward differences (between responses in the different components) need to be made arbitrarily large in order to get close to this infimum. Hence, it would be more appropriate to talk only about reward differences within a strongly connected component.

C.1 Gradient Descent for Tabular Parameterization

Note that the tabular parameterization basically has the following loss function for $x \in X, y_i, y_j \in Y$:

$$\mathcal{L}_{\text{DPO}}(\pi) = -\sum_x \sum_i \sum_j d_{v,i,j} \log \sigma \left(\beta \log \frac{\pi(y_i \mid x)}{\pi_{\text{ref}}(y_i \mid x)} - \beta \log \frac{\pi(y_j \mid x)}{\pi_{\text{ref}}(y_j \mid x)} \right) \tag{15}$$

which is essentially a paramaterization of $r_{x,i} = \beta \log \frac{\pi(y_i|x)}{\pi_{\text{ref}}(y_i|x)}$ in the 6

Understanding the behavior of the loss function under the tabular parameterization is crucial to understand the theoretical limitations of DPO in an ideal setting where the probabilities are independent and directly optimized. Nika et al. [2024] analyzes this and concludes that the function is not smooth as there would be no lower bound on the minimum component of the matrix during gradient descent. However, we note the necessity of strong connectivity C in this setting which leads to a convergence of Gradient Descent.

Note that gradient descent *would include normalization at every step* since taking a step along $-\nabla\mathcal{L}_{\text{DPO}}(\pi)$ would possibly give a point outside $\Pi = \{\pi : \pi_i > 0, \sum_i \pi_i = 1\}$ and hence we shall observe smoothness and PL-inequality over $\Pi_{\alpha,K} = \{\pi : \pi_i \geq \alpha, \pi_i \leq K\}$. Note that we could instead have the step along the plane which would give a slightly different version of gradient descent. Here we show the step size needed and then prove the convergence rate of GD with the following update at iteration t:

$$\pi'_{new} = \pi_{t-1} - \eta_t \nabla\mathcal{L}_{DPO}(\pi_{t-1})$$
$$\pi_t = \frac{\pi'_{new}}{\sum_i (\pi'_{new})_i}$$

C.2 Smoothness, PL-inequality and convergence

We first note that the above parameterization is such that any $\pi \in \Pi_\alpha = \{\pi \in \mathbb{R}^{m\times n}, \pi_{x,i} \geq \alpha\}$ has a corresponding $\mathbf{r} \in \mathbb{R}^{m\times n}$ such that $\mathcal{L}_{\text{DPO}}(\pi) = \mathcal{L}_{\text{BT}}(\mathbf{r})$ for any $\alpha > 0$ and vice versa. The same is the case for any $\pi \in \Pi_{0,K} = \{\pi \in \mathbb{R}^{m\times n}, \pi_{x,i} \geq 0, \pi_{x,i} \leq K\}$ where $K > 0$. We also note the following relation between their gradients.

$$\begin{aligned}
\frac{\partial}{\partial \pi_{x,i}} \mathcal{L}_{\text{DPO}}(\pi) =& \frac{\beta}{\pi_{x,i}} \sum_{j \neq i} \left(d_{x,i,j} - (d_{x,i,j} + d_{x,j,i}) \frac{\left(\frac{\pi_{x,i}}{\pi_{\text{ref}} x,i}\right)^\beta}{\left(\frac{\pi_{x,i}}{\pi_{\text{ref}} x,i}\right)^\beta + \left(\frac{\pi_{x,j}}{\pi_{\text{ref}} x,j}\right)^\beta} \right) \\
=& \frac{\beta}{\pi_{x,i}} \frac{\partial}{\partial \mathbf{r}_{x,i}} \mathcal{L}_{\text{BT}}(\mathbf{r})
\end{aligned} \tag{16}$$

where $\mathbf{r}_{x,i} = \beta \log \left(\frac{\pi_{x,i}}{\pi_{\text{ref}} x,i} \right)$ is the implicit reward difference. Note that while we prove the following for multiple inputs, it can be proven for a single input and easily extended since the parameters for $\mathbf{r}_1, \mathbf{r}_2$ corresponding to different inputs are independent of each other.

Lemma C.1. *For any* $1 \geq \alpha > 0$*, we have* γ_α *PL- inequality over* $\Pi'_{\alpha,\frac{1}{\alpha}} = \{\pi : \alpha \leq \frac{\pi_{x,i}}{\pi_{ref_{x,i}}} \leq \frac{1}{\alpha} \forall x, i\}$ *and* μ_α*-smoothness over* $\Pi'_\alpha = \{\pi : \alpha \leq \frac{\pi_{x,i}}{\pi_{ref_{x,i}}}\}$ *with*

$$\gamma_\alpha = \frac{\alpha^2\beta^2}{\max(\pi_{ref_{x,i}}^2)(\alpha^{-\beta} + \alpha^\beta)} \min_x \lambda_2(\mathbf{L}_{M_x}) \quad \textit{and} \quad \mu_\alpha = \frac{\beta d_{\max}}{\alpha^2 \min(\pi_{ref_{x,i}}^2)} \sqrt{\left(\frac{\beta^2}{4} + m^2\right)}$$

Proof. 1. We know that $\mathcal{L}_{BT}(\mathbf{r})$ satisfies $\gamma = c_\omega \min_x \lambda_2(\mathbf{L}_{\mathbf{M}_x})$ PL inequality in $\mathcal{W}_\omega = \{\mathbf{r} \in \mathbb{R}^{m\times n} : \|\mathbf{r}\|_\infty \leq \omega\}$ for any $\omega > 0$

$$\mathcal{L}_{\text{BT}}(\mathbf{r}) - \mathcal{L}^*_{\text{BT}} \leq \frac{1}{2\gamma}\|\nabla\mathcal{L}_{\text{BT}}(\mathbf{r})\|_2^2$$

Consider the parameterization $\mathbf{r}_{x,i} = \beta \log \frac{\pi_{x,i}}{\pi_{\text{ref}x,i}}$. Since we know that every $\mathbf{r} \in \mathcal{W}_\omega$ has corresponding $\pi \in \Pi' = \{\pi : e^{\frac{-\omega}{\beta}} \leq \frac{\pi_{x,i}}{\pi_{\text{ref}x,i}} \leq e^{\frac{\omega}{\beta}}\}$ and that the gradients are related through equation 16 we have the following:

$$\begin{aligned}
\mathcal{L}_{\text{DPO}}(\pi) - \mathcal{L}^*_{\text{DPO}} = \mathcal{L}_{\text{BT}}(\mathbf{r}) - \mathcal{L}^*_{\text{BT}} &\leq \frac{1}{2\gamma}\|\nabla\mathcal{L}_{\text{BT}}(\mathbf{r})\|_2^2 \\
&\leq \frac{\max(\pi_{x,i}^2)}{2\gamma\beta^2}\|\nabla\mathcal{L}_{\text{DPO}}(\pi)\|_2^2 \\
&\leq \frac{\max(\pi_{\text{ref}x,i}^2)e^{\frac{2\omega}{\beta}}}{2\gamma\beta^2}\|\nabla\mathcal{L}_{\text{DPO}}(\pi)\|_2^2
\end{aligned}$$

where $\mathcal{L}^*_{\text{BT}}$ and $\mathcal{L}^*_{\text{DPO}}$ are the minimum values in the given domains. Hence we have the $\gamma_{\text{DPO}} = c'_\omega \min_x \lambda_2(\mathbf{L}_{M_x})$ where $c'_\omega = \frac{\beta^2}{\max(\pi_{\text{ref}x,i}^2)e^{\frac{2\omega}{\beta}}(e^{-\omega} + e^{\omega})}$ in $\Pi' = \{\pi : e^{\frac{-\omega}{\beta}} \leq \frac{\pi_i}{\pi_{\text{ref}i}} \leq e^{\frac{\omega}{\beta}}\}$ for any $\omega > 0$. Supposing that $e^{\frac{-\omega}{\beta}} = \alpha$, for any $\alpha > 0$ we have γ_α PL-inequality over $\Pi'_{\alpha,\frac{1}{\alpha}}$ where γ_α and $\Pi'_{\alpha,\frac{1}{\alpha}}$ are as defined in the lemma.

2. The function $\mathcal{L}_{\text{DPO}}$ is said to be L-smooth in a domain say $\Pi'_{\alpha,K} = \{\pi : \alpha \leq \frac{\pi_{x,i}}{\pi_{\text{ref}x,i}} \leq K \forall (x,i)\}$ where $K > \alpha > 0$ if for any $\pi_1, \pi_2 \in \Pi'_{\alpha,K}$, $\|\nabla\mathcal{L}_{\text{DPO}}(\pi_1) - \nabla\mathcal{L}_{\text{DPO}}(\pi_2)\|_2 \leq L\|\pi_1 - \pi_2\|_2$

   Let $\tilde{r_1}, \tilde{r_2}$ be $\beta \log \frac{\pi_{1x,i}}{\pi_{\text{ref}x,i}}, \beta \log \frac{\pi_{2x,i}}{\pi_{\text{ref}x,i}}$ respectively. We observe the following:

$$\begin{aligned}
\frac{\partial}{\partial \pi_{x,i}}\mathcal{L}_{\text{DPO}}(\pi_1) - \frac{\partial}{\partial \pi_{x,i}}\mathcal{L}_{\text{DPO}}(\pi_2) &\leq \frac{\beta}{\pi_{1x,i}}\frac{\partial}{\partial \mathbf{r}_{x,i}}\mathcal{L}_{\text{BT}}(\tilde{r_1}) - \frac{\beta}{\pi_{2x,i}}\frac{\partial}{\partial \mathbf{r}_{x,i}}\mathcal{L}_{\text{BT}}(\tilde{r_2}) \\
&= \frac{\beta}{\pi_{1x,i}}\left(\frac{\partial}{\partial \mathbf{r}_{x,i}}\mathcal{L}_{\text{BT}}(\tilde{r_1}) - \frac{\partial}{\partial \mathbf{r}_{x,i}}\mathcal{L}_{\text{BT}}(\tilde{r_2})\right) + \beta\left(\frac{1}{\pi_{1x,i}} - \frac{1}{\pi_{2x,i}}\right)\frac{\partial}{\partial \mathbf{r}_{x,i}}\mathcal{L}_{\text{BT}}(\tilde{r_2})
\end{aligned}$$

   Suppose that $|\frac{\partial}{\partial \mathbf{r}_{x,i}}\mathcal{L}_{\text{BT}}(\tilde{r_2})|$ is upper bounded by G, we have

$$\begin{aligned}
\|\nabla\mathcal{L}_{\text{DPO}}(\pi_1) - \nabla\mathcal{L}_{\text{DPO}}(\pi_2)\|_2^2 &= \sum_x\sum_i \|\frac{\partial}{\partial \pi_{x,i}}\mathcal{L}_{\text{DPO}}(\pi_1) - \frac{\partial}{\partial \pi_{x,i}}\mathcal{L}_{\text{DPO}}(\pi_2)\|_2^2 \\
&\leq \frac{\beta^2}{\alpha^2\pi_{\text{ref}_{\min}}^2}\|\nabla\mathcal{L}_{\text{BT}}(\tilde{r_1}) - \mathcal{L}_{\text{BT}}(\tilde{r_2})\|_2^2 + \frac{\beta^2}{\alpha^4\pi_{\text{ref}_{\min}}^4}G^2\|\pi_1 - \pi_2\|_2^2 \\
&\leq \frac{\beta^2\mu^2}{\alpha^2\pi_{\text{ref}_{\min}}^2}\|\tilde{r_1} - \tilde{r_2}\|_2^2 + \frac{\beta^2}{\alpha^4\pi_{\text{ref}_{\min}}^4}G^2\|\pi_1 - \pi_2\|_2^2 \quad (17)
\end{aligned}$$

   where $\mu = \frac{d_{\max}}{2}$ in the last inequality which follows due to smoothness of $\mathcal{L}_{BT}$ over the entire space. We further examine the terms in the above inequality as follows:

   (a) We use the Lipchitz continuity of the logarithmic function when it has a lower bound in order to get an upper bound in terms of $\|\pi_1 - \pi_2\|_2$ for $\|\tilde{r_1} - \tilde{r_2}\|_2$.

$$\|\tilde{r_1} - \tilde{r_2}\|_2 \leq \beta^2 \sum_x\sum_i \|\log(\pi_{1x,i}) - \log(\pi_{2x,i})\|_2^2$$

From mean value theorem we know that $\frac{\log(\pi_{1x,i})-\log(\pi_{2x,i})}{\pi_{1x,i}-\pi_{2x,i}} = \frac{1}{c}$ where $\min(\pi_{1x,i}, \pi_{2x,i}) \leq c \leq \max(\pi_{1x,i}, \pi_{2x,i})$. This leads to the following:

$$\|\tilde{r_1} - \tilde{r_2}\|_2 \leq \frac{\beta^2}{\alpha^2 \pi_{\text{ref}}{}^2_{\text{min}}} \sum_x \sum_i \|\pi_{1x,i} - \pi_{2x,i}\|_2^2 = \frac{\beta^2}{\alpha^2 \pi_{\text{ref}}{}^2_{\text{min}}} \|\pi_1 - \pi_2\|_2^2$$

(b) Since we know that $\frac{\partial}{\partial \mathbf{r}_{x,i}} \mathcal{L}_{\text{BT}}(\tilde{r}_2) = \sum_{j \neq i} \left(d_{x,i,j} - (d_{x,i,j} + d_{x,j,i}) \frac{\left(\frac{\pi_{x,i}}{\pi_{\text{ref}}{}_{x,i}}\right)^\beta}{\left(\frac{\pi_{x,i}}{\pi_{\text{ref}}{}_{x,i}}\right)^\beta + \left(\frac{\pi_{x,j}}{\pi_{\text{ref}}{}_{x,j}}\right)^\beta} \right)$, we upper bound it as follows

$$|\frac{\partial}{\partial \mathbf{r}_{x,i}} \mathcal{L}_{\text{BT}}(\tilde{r}_2)| < \sum_{j \neq i} d_{\text{max}} = (m-1) d_{\text{max}}$$

where d_{max} is the maximum degree of the comparison graph
Substituting these upper bounds in the equation 17 we now have

$$\|\nabla \mathcal{L}_{\text{DPO}}(\pi_1) - \nabla \mathcal{L}_{\text{DPO}}(\pi_2)\|_2^2 \leq \left(\frac{\beta d_{\text{max}}}{\alpha^2 \pi_{\text{ref}}{}^2_{\text{min}}} \right)^2 \left(\frac{\beta^2}{4} + m^2 \right) \|\pi_1 - \pi_2\|_2^2$$

The smoothness constant over $\Pi'_{\alpha,K} = \frac{\beta d_{\text{max}}}{\alpha^2 \pi_{\text{ref}}{}^2_{\text{min}}} \sqrt{\left(\frac{\beta^2}{4} + m^2\right)}$. Note that the independence on K implies the same smoothness over Π'_α

□

C.3 Lower bound and Convergence

We show a lower bound on the infinity norm of $\pi_{\text{DPO}}(\cdot \mid x)$ as a corollary of the following more general lemma. The lemma establishes a lower bound on the infinity norm on any distribution that has a lower negative loglikelihood than a realizable value.

Lemma C.2. *Consider any probability distribution over the responses for input x, $\pi(\cdot \mid x)$ whose likelihood given the data over responses for x is L' (it's corresponding negative log likelihood $-\log L'$) and the comparison graph over x is strongly connected. For any distribution π' if $\mathcal{L}_{DPO,x}(\pi') \leq -\log L'$ then $\pi' \in \{\pi \in \Pi : \alpha \leq \left(\frac{\pi_i}{\pi_{ref_i}}\right) \leq A \quad \forall i, x\}$ where $\alpha = \frac{(L')^{\frac{1}{\beta}}}{n \pi_{ref_{\max}}}$ and $A = \frac{1}{n \pi_{ref_{\min}} (L')^{\frac{1}{\beta}}}$*

Proof. The contrapositive of this would be to show that α such that $\forall \pi' \in \{\pi : \left(\frac{\pi_i}{\pi_{\text{ref}i}}\right)_{\min} \leq \alpha, \sum \pi_i = 1\} \cup \{\pi : \left(\frac{\pi_i}{\pi_{\text{ref}i}}\right)_{\max} \geq A, \sum \pi_i = 1\}$,

$$L(\pi) = \Pi_{i,j} \left(\frac{\left(\frac{\pi_i}{\pi_i^{(ref)}}\right)^\beta}{\left(\frac{\pi_i}{\pi_i^{(ref)}}\right)^\beta + \left(\frac{\pi_j}{\pi_j^{(ref)}}\right)^\beta} \right)^{d_{ij}} \leq L'$$

(1) Given any π, suppose for the sake of contradiction, let π_i be the component with $\min \left(\frac{\pi_i}{\pi_{\text{ref}i}}\right)$ which is less than α, we know that there exists a k, such that $\pi_k \geq \frac{1}{n}$. Due to strong connectivity, there exists a path $p = x_i....x_k$ where $d_{v_i v_{i+1}} \geq 1$ and we have:

$$
\begin{aligned}
L(\pi) \leq& \Pi_{(i,j)\in e(p)} \left(\frac{\left(\frac{\pi_i}{\pi_i^{(ref)}}\right)^\beta}{\left(\frac{\pi_i}{\pi_i^{(ref)}}\right)^\beta + \left(\frac{\pi_j}{\pi_j^{(ref)}}\right)^\beta} \right)^{d_{ij}} \leq \Pi_{(i,j)\in e(p)} \left(\frac{\left(\frac{\pi_i}{\pi_i^{(ref)}}\right)^\beta}{\left(\frac{\pi_i}{\pi_i^{(ref)}}\right)^\beta + \left(\frac{\pi_j}{\pi_j^{(ref)}}\right)^\beta} \right) \\
\leq& \Pi_{(i,j)\in e(p)} \frac{\left(\frac{\pi_i}{\pi_i^{(ref)}}\right)^\beta}{\left(\frac{\pi_j}{\pi_j^{(ref)}}\right)^\beta} = \frac{\left(\frac{\pi_i}{\pi_i^{(ref)}}\right)^\beta}{\left(\frac{\pi_{i_1}}{\pi_{i_1}^{(ref)}}\right)^\beta} \frac{\left(\frac{\pi_{i_1}}{\pi_{i_1}^{(ref)}}\right)^\beta}{\left(\frac{\pi_{i_2}}{\pi_{i_2}^{(ref)}}\right)^\beta} \cdots \frac{\left(\frac{\pi_{i_{a-1}}}{\pi_{i_{a-1}}^{(ref)}}\right)^\beta}{\left(\frac{\pi_k}{\pi_k^{(ref)}}\right)^\beta} = \frac{\left(\frac{\pi_i}{\pi_{\text{ref}i}}\right)^\beta}{\left(\frac{\pi_k}{\pi_{\text{ref}k}}\right)^\beta} \\
\leq& \left(\frac{\pi_i}{\pi_{\text{ref}i}} \frac{\pi_{\text{ref}\max}}{\frac{1}{n}} \right)^\beta
\end{aligned}
\tag{18}
$$

Hence if $\min(\frac{\pi_i}{\pi_{\text{ref}_i}}) \leq \frac{(L')^{\frac{1}{\beta}}}{n\pi_{\text{ref}\max}}$ we know that the required $L(\pi) \leq L'$

(2) Similarly given any π, for the sake of contradiction, let π_i be the component with $\max\left(\frac{\pi_i}{\pi_{\text{ref}_i}}\right)$ which is greater than A. We know that there exists a k, such that $\pi_k \leq \frac{1}{n}$ and due to strong connectivity, there exists a path $p = x_k....x_i$ where $d_{v_i v_{i+1}} \geq 1$ and we have:

$$
\begin{aligned}
L(\pi) \leq& \Pi_{(i,j)\in e(p)} \left(\frac{\left(\frac{\pi_i}{\pi_i^{(ref)}}\right)^\beta}{\left(\frac{\pi_i}{\pi_i^{(ref)}}\right)^\beta + \left(\frac{\pi_j}{\pi_j^{(ref)}}\right)^\beta} \right)^{d_{ij}} \leq \Pi_{(i,j)\in e(p)} \left(\frac{\left(\frac{\pi_i}{\pi_i^{(ref)}}\right)^\beta}{\left(\frac{\pi_i}{\pi_i^{(ref)}}\right)^\beta + \left(\frac{\pi_j}{\pi_j^{(ref)}}\right)^\beta} \right) \\
\leq& \Pi_{(i,j)\in e(p)} \frac{\left(\frac{\pi_i}{\pi_i^{(ref)}}\right)^\beta}{\left(\frac{\pi_j}{\pi_j^{(ref)}}\right)^\beta} = \frac{\left(\frac{\pi_k}{\pi_k^{(ref)}}\right)^\beta}{\left(\frac{\pi_{i_1}}{\pi_{i_1}^{(ref)}}\right)^\beta} \frac{\left(\frac{\pi_{i_1}}{\pi_{i_1}^{(ref)}}\right)^\beta}{\left(\frac{\pi_{i_2}}{\pi_{i_2}^{(ref)}}\right)^\beta} \cdots \frac{\left(\frac{\pi_{i_{a-1}}}{\pi_{i_{a-1}}^{(ref)}}\right)^\beta}{\left(\frac{\pi_i}{\pi_i^{(ref)}}\right)^\beta} = \frac{\left(\frac{\pi_k}{\pi_{\text{ref}k}}\right)^\beta}{\left(\frac{\pi_i}{\pi_{\text{ref}i}}\right)^\beta} \\
\leq& \left(\frac{\frac{1}{n}}{\pi_{\text{ref}\min}} \frac{1}{\frac{\pi_i}{\pi_{\text{ref}_i}}} \right)^\beta
\end{aligned}
\tag{19}
$$

Hence if $\max(\frac{\pi_i}{\pi_{\text{ref}_i}}) \geq \frac{1}{n\pi_{\text{ref}\min}(L')^{\frac{1}{\beta}}}$ we know that the required $L(\pi) \leq L'$ □

Corollary C.3. *Say L' is a realizable likelihood for some probability distribution, if there is a distribution π' such that $\mathcal{L}_{DPO,x}(\pi') \leq -\log L'$ then $\pi' \in \{\pi \in \Pi : \alpha \leq \left(\frac{\pi_i}{\pi_{ref_i}}\right) \leq \frac{1}{\alpha} \quad \forall i, x\}$ where $\alpha = \min\left(\frac{(L')^{\frac{1}{\beta}}}{n\pi_{ref\max}}, n\pi_{ref\min}(L')^{\frac{1}{\beta}}\right)$*

This result shall be used to get the bounds for the points that gradient descent reaches, leaving us with appropriate constants for smoothness and PL-inequality.

C.4 Proof of Gradient Descent

We now prove the convergence rate of GD with the following update at iteration t:

$$
\begin{aligned}
\pi'_t &= \pi_{t-1} - \eta_t \nabla \mathcal{L}_{DPO}(\pi_{t-1}) \\
\pi_t &= \frac{\pi'_t}{\sum_i (\pi'_t)_i}
\end{aligned}
$$

Theorem C.4. *Let π_0 be the initial distribution with likelihood L_0. We show the following:*

1. *Suppose that after $t-1$ iterations, the distribution π_{t-1} has $\mathcal{L_{DPO}} \leq -\log L_0$. Then, under a gradient descent update with the step size $\eta_{GD} = \frac{1}{\mu_{\alpha_{GD}}}$, the following holds:*
   - *a) The intermediate $\pi'_t \in \Pi'_{\alpha_{GD}} = \{\pi : \alpha_{GD} \leq \frac{\pi_i}{\pi_{\mathit{ref}_i}} \quad \forall i \forall x\}$*
   - *b) The normalized distribution π_t has $\mathcal{L}_{DPO}(\pi_t) \leq -\log L_0$*
2. *Gradient descent converges with the error threshold ϵ in the steps $T = \mathcal{O}\left(\frac{\mu_{\alpha_{GD}}}{\gamma_\alpha} \log\left(\frac{1}{\epsilon}\right)\right)$.*

Here the parameters are defined as:

$$\alpha = \min\left(\frac{(L_0)^{\frac{1}{\beta}}}{n\pi_{\mathit{ref}_{\max}}}, n\pi_{\mathit{ref}_{\min}}(L_0)^{\frac{1}{\beta}}\right), \quad \alpha_{GD} = \alpha\left(\frac{1}{1+\sqrt{\left(\frac{\beta^2}{4m^2}+1\right)}}\right)$$

$$\mu_{\alpha_{GD}} = \frac{\beta d_{\max}}{\alpha_{GD}^2 \min(\pi_{\mathit{ref}_{x,i}}^2)}\sqrt{\left(\frac{\beta^2}{4}+m^2\right)}, \quad \gamma_\alpha = \frac{\alpha^2\beta^2}{\max(\pi_{\mathit{ref}_{x,i}}^2)(\alpha^{-\beta}+\alpha^\beta)} \min_x \lambda_2(\mathbf{L}_{M_x})$$

Proof. Note that the corollary C.3 we know that $\pi_{t-1} \in \Pi_{\alpha,\frac{1}{\alpha}} = \{\pi : \alpha \leq \frac{\pi_i}{\pi_{\text{ref}_i}} \leq \frac{1}{\alpha} \quad \forall i \forall x\}$ and from the update step $\pi'_t = \pi_{t-1} - \eta_t \nabla \mathcal{L}_{DPO}(\pi_{t-1})$ we know that

$$\frac{(\pi'_t)_i}{\pi_{\text{ref}_i}} = \frac{(\pi_{t-1})_i}{\pi_{\text{ref}_i}} - \frac{1}{\pi_{\text{ref}_i}\mu_{\alpha_{GD}}}\nabla_i \mathcal{L}_{DPO}(\pi_{t-1})$$

From our examination of the gradient in lemma C.1 we know that $\dfrac{\partial}{\partial \pi_{x,i}}\mathcal{L}_{\text{DPO}}(\pi_{t-1}) < \frac{\beta}{\pi_i} m d_{\max}$ and hence we have

$$\begin{aligned}
\frac{(\pi'_t)_i}{\pi_{\text{ref}_i}} &\geq \min\left(\frac{(\pi_{t-1})_i}{\pi_{\text{ref}_i}}\right) - \frac{\beta}{\pi_{\text{ref}_i}\mu_{\alpha_{GD}}}\frac{1}{\pi_{t-1_i}} m d_{\max} \\
&\geq \alpha - \frac{\beta}{\pi_{\text{ref}_i}\pi_{t-1_i}} \frac{\alpha_{GD}^2 \min(\pi_{\text{ref}_{x,i}}^2)}{\beta d_{\max}\sqrt{\left(\frac{\beta^2}{4}+m^2\right)}} m d_{\max} \\
&\geq \alpha - \frac{\alpha_{GD}^2 \min(\pi_{\text{ref}_{x,i}})}{\pi_{t-1_i}\sqrt{\left(\frac{\beta^2}{4m^2}+1\right)}} \geq \alpha - \frac{\alpha_{GD}\alpha}{\sqrt{\left(\frac{\beta^2}{4m^2}+1\right)}}\frac{\pi_{\text{ref}_i}}{\pi_{t-1_i}} \\
&\geq \alpha - \frac{\alpha_{GD}}{\sqrt{\left(\frac{\beta^2}{4}+m^2\right)}} = \alpha_{GD}
\end{aligned}$$

Hence we can conclude that $\pi'_t \in \Pi'_{\alpha_{GD}}$ and since $\alpha_{GD} \leq \alpha, \pi_{t-1} \in \Pi'_{\alpha_{GD}}$. Since we know that $\mathcal{L}_{\text{DPO}}$ is $\mu_{\alpha_{GD}}$-smooth in this domain we know the following:

$$\begin{aligned}
\mathcal{L}_{\text{DPO}}(\pi_t) = \mathcal{L}_{\text{DPO}}(\pi'_t) &\leq \mathcal{L}_{\text{DPO}}(\pi_{t-1}) + \nabla\mathcal{L}_{\text{DPO}}(\pi_{t-1}) \cdot (\pi'_t - \pi_{t-1}) + \frac{\mu_{\alpha_{GD}}}{2}\|\pi'_t - \pi_{t-1}\|_2^2 \\
&= \mathcal{L}_{\text{DPO}}(\pi_{t-1}) - \frac{1}{\mu_{\alpha_{GD}}}\|\nabla\mathcal{L}_{\text{DPO}}(\pi_{t-1})\|_2^2 + \frac{\mu_{\alpha_{GD}}}{2}\frac{1}{\mu_{\alpha_{GD}}^2}\|\nabla\mathcal{L}_{\text{DPO}}(\pi_{t-1})\|_2^2 \\
&= \mathcal{L}_{\text{DPO}}(\pi_{t-1}) - \frac{1}{2\mu_{\alpha_{GD}}}\|\nabla\mathcal{L}_{\text{DPO}}(\pi_{t-1})\|_2^2
\end{aligned}$$

Since we have γ_α inequality over the space $\Pi_{\alpha,\frac{1}{\alpha}}$ in which both π_t and the optimal distribution (with loss $\mathcal{L}^*_{\text{DPO}}$ we have

$$\begin{aligned}\mathcal{L}_{\text{DPO}}(\pi_t) - \mathcal{L}^*_{\text{DPO}} &\leq \mathcal{L}_{\text{DPO}}(\pi_{t-1}) - \mathcal{L}^*_{\text{DPO}} - \frac{1}{2\mu_{\alpha_{GD}}}\|\nabla\mathcal{L}_{\text{DPO}}(\pi_{t-1})\|_2^2 \\ &\leq (\mathcal{L}_{\text{DPO}}(\pi_{t-1}) - \mathcal{L}^*_{\text{DPO}}) - \frac{\gamma_\alpha}{\mu_{\alpha_{GD}}}(\mathcal{L}_{\text{DPO}}(\pi_{t-1}) - \mathcal{L}^*_{\text{DPO}}) \\ &= \left(1 - \frac{\gamma_\alpha}{\mu_{\alpha_{GD}}}\right)(\mathcal{L}_{\text{DPO}}(\pi_{t-1}) - \mathcal{L}^*_{\text{DPO}})\end{aligned}$$

Through induction, we can conclude from the above theorem that GD never steps outside $\Pi_{\alpha_{GD}}$ and the normalized distributions(along with the optimal distribution) always lie inside $\Pi_{\alpha,\frac{1}{\alpha}}$. This proves the convergence of gradient descent with the error threshold ϵ in the steps $T = \mathcal{O}\left(\frac{\mu_{\alpha_{GD}}}{\gamma_\alpha}\log\left(\frac{1}{\epsilon}\right)\right)$.
□

D Generalizability

1. Note that this is equivalent to showing that
$$(\mathbf{W}_i - \mathbf{W}_j)\mathbf{x} = (\mathbf{W}_i^{\text{RLHF}} - \mathbf{W}_j^{\text{RLHF}})\mathbf{x}$$
for all $(y_i^\alpha, y_j^\alpha, x^\alpha) \in Q$, where $\mathbf{W}$ is any parameter matrix that minimizes $\mathcal{L}_{\text{DPO}}$.
Consider the maximum of the likelihood:
$$L(\mathbf{W}) = \prod_x \prod_{\{i,j\}} L(\pi)_{x,\{i,j\}},$$
where
$$L(\pi)_{x,\{i,j\}} = \left(\frac{e^{\beta(\mathbf{W}_i - \mathbf{W}_i^{\text{ref}})\mathbf{x}}}{e^{\beta(\mathbf{W}_i - \mathbf{W}_i^{\text{ref}})\mathbf{x}} + e^{\beta(\mathbf{W}_j - \mathbf{W}_j^{\text{ref}})\mathbf{x}}}\right)^{d_{x,i,j}} \left(\frac{e^{\beta(\mathbf{W}_j - \mathbf{W}_j^{\text{ref}})\mathbf{x}}}{e^{\beta(\mathbf{W}_i - \mathbf{W}_i^{\text{ref}})\mathbf{x}} + e^{\beta(\mathbf{W}_j - \mathbf{W}_j^{\text{ref}})\mathbf{x}}}\right)^{d_{x,j,i}}.$$
Let
$$f = \frac{e^{\beta(\mathbf{W}_i - \mathbf{W}_i^{\text{ref}})\mathbf{x}}}{e^{\beta(\mathbf{W}_i - \mathbf{W}_i^{\text{ref}})\mathbf{x}} + e^{\beta(\mathbf{W}_j - \mathbf{W}_j^{\text{ref}})\mathbf{x}}}, \quad a = d_{x,i,j}, \quad b = d_{x,j,i}.$$
Then we have
$$\begin{aligned}L(\pi)_{x,\{i,j\}} &= a^a b^b \left(\frac{f}{a}\right)^a \left(\frac{1-f}{b}\right)^b \\ &\leq a^a b^b \left(\frac{a\left(\frac{f}{a}\right) + b\left(\frac{1-f}{b}\right)}{a+b}\right)^{a+b} = \frac{a^a b^b}{(a+b)^{a+b}}.\end{aligned}$$
The inequality follows from the AM–GM inequality, implying that $L(\mathbf{W})_{x,\{i,j\}}$ is maximized when
$$\frac{e^{\beta(\mathbf{W}_i - \mathbf{W}_i^{\text{ref}})\mathbf{x}}}{e^{\beta(\mathbf{W}_j - \mathbf{W}_j^{\text{ref}})\mathbf{x}}} = \frac{d_{x,i,j}}{d_{x,j,i}}.$$
Because $d_{x,i,j}$ and $d_{x,j,i}$ are obtained from oracle comparisons, this ratio equals the exponential of the true reward difference. Since $\mathbf{W} = \mathbf{W}^{\text{RLHF}}$ satisfies this ratio for every triplet $\{x, \{i,j\}\}$, the maximum can occur for all triplets simultaneously. Therefore, any $\mathbf{W}$ that maximizes the likelihood must satisfy for all $(y_i, y_j, x) \in Q$:
$$(\mathbf{W}_i - \mathbf{W}_j)\mathbf{x} = (\mathbf{W}_i^{\text{RLHF}} - \mathbf{W}_j^{\text{RLHF}})\mathbf{x}$$
which implies that $\pi_{DPO}(y_i \succ y_j \mid x) = \pi_{RLHF}(y_i \succ y_j \mid x)$ for all $(y_i, y_j, x) \in Q$
2. The preceding result implies that the set of maximum-likelihood estimators (MLEs) $\mathbf{W}$ is precisely the set of matrices satisfying
$$A_Q\,\text{vec}(\mathbf{W}) = A_Q\,\text{vec}(\mathbf{W}^{\text{RLHF}}),$$
where $A_Q \in \mathbb{R}^{|Q| \times nd}$ is the query coefficient matrix defined previously.

(a) Condition for generalization. To ensure generalization to any unseen input x and any pair of responses, the retrieved unembedding matrices must belong to the affine subspace

$$\mathcal{S} = \left\{ \mathbf{W} \in \mathbb{R}^{n \times d} \ : \ \mathbf{W} = \mathbf{W}^{\mathrm{RLHF}} + c \otimes \mathbf{1}^n, \ c \in \mathbb{R}^d \right\}.$$

Equivalently, this requires that the kernel of A_Q be contained within the direction subspace of $\mathcal{S}$:

$$\ker(A_Q) \subseteq \{\, c \otimes \mathbf{1}^n : c \in \mathbb{R}^d \,\}.$$

Since the right-hand side is trivially contained in $\ker(A_Q)$, equality must hold:

$$\ker(A_Q) = \{\, c \otimes \mathbf{1}^n : c \in \mathbb{R}^d \,\}.$$

Consequently, the rank of A_Q must satisfy

$$\mathrm{rank}(A_Q) = nd - d = (n-1)d.$$

(b) Necessity of the rank condition. Assume for contradiction that $\mathrm{rank}(A_Q) \neq (n-1)d$. Then $\mathrm{rank}(A_Q) < (n-1)d$, implying that

$$\dim(\ker(A_Q)) > d.$$

Because $\{c \otimes \mathbf{1}^n : c \in \mathbb{R}^d\} \subseteq \ker(A_Q)$, there exists a non-trivial

$$\Delta \notin \{c \otimes \mathbf{1}^n : c \in \mathbb{R}^d\}$$

such that another feasible parameter matrix

$$\mathbf{W} = \mathbf{W}^{\mathrm{RLHF}} + \Delta$$

also satisfies $A_Q \operatorname{vec}(\mathbf{W}) = A_Q \operatorname{vec}(\mathbf{W}^{\mathrm{RLHF}})$.

Let $\Delta = [\, \Delta_1, \ldots, \Delta_n \,]$ with columns $\Delta_i \in \mathbb{R}^d$. Since $\Delta \notin \{c \otimes \mathbf{1}^n\}$, there exist indices $i \neq j$ such that $\Delta_i \neq \Delta_j$. Choose a coordinate k where they differ and denote by $e_k \in \mathbb{R}^d$ the k-th standard basis vector. Then

$$e_k^\top(\mathbf{W}_i - \mathbf{W}_j) \neq e_k^\top(\mathbf{W}_i^{\mathrm{RLHF}} - \mathbf{W}_j^{\mathrm{RLHF}}),$$

showing that the reward difference between responses i and j for feature dimension k deviates from the RLHF solution. Thus, $\mathbf{W}$ fails to preserve the pairwise structure required for unseen inputs x and cannot generalize. Hence, the necessary and sufficient condition for generalization is

$$\mathrm{rank}(A_Q) = (n-1)d.$$

E Convergence of Learned Reward Differences

E.1 Proof of Lemma 4.2

We extend the analysis of convergence with uniform sampling in Shi et al. [2025] which was carried out for a tabular parameterization of rewards to our setting where we parameterize the embeddings. The case of uniform sampling in Shi et al. [2025] corresponds to setting $d_{v,i,j} = p^\star(y_i \succ y_j | x_v) = \sigma(r^\star_{v,i} - r^\star_{v,j}) \ \forall x_v \in \mathcal{X}, \ y_i, y_j \in \mathcal{Y}$.

We then have,

$$\begin{aligned}
\nabla_{W_{i,k}} \mathcal{L}_{\mathrm{DPO}}(\mathbf{W}; \mathbf{X}) =& \beta \sum_v^m x_{v,k} \sum_{j \neq i}^n \frac{d_{v,j,i}\alpha_{v,i} - d_{v,i,j}\alpha_{v,j}}{\alpha_{v,i} + \alpha_{v,j}} \\
=& \beta \sum_v^m x_{v,k} \sum_{j \neq i}^n \left(d_{v,j,i} \frac{\alpha_{v,i}/\alpha_{v,j}}{1 + \alpha_{v,i}/\alpha_{v,j}} - d_{v,i,j} \frac{\alpha_{v,j}/\alpha_{v,i}}{1 + \alpha_{v,j}/\alpha_{v,i}} \right) \\
=& \beta \sum_v^m x_{v,k} \sum_{j \neq i}^n \sigma(r^\star_{v,j} - r^\star_{v,i}) \sigma(h(\mathbf{W}_i; \mathbf{x}_v) - h(\mathbf{W}_j; \mathbf{x}_v)) \\
& - \beta \sum_v^m x_{v,k} \sum_{j \neq i}^n \sigma(r^\star_{v,i} - r^\star_{v,j}) \sigma(h(\mathbf{W}_j; \mathbf{x}_v) - h(\mathbf{W}_i; \mathbf{x}_v))
\end{aligned}$$

We note that $\sigma(-z)\sigma(u) - \sigma(z)\sigma(-u) = (1-\sigma(z))\sigma(u) - \sigma(z)(1-\sigma(u)) = \sigma(u) - \sigma(z)$. If we take, $z = r^\star_{v,i} - r^\star_{v,j}$ and $u = h(\mathbf{W}_i;\mathbf{x}_v) - h(\mathbf{W}_j;\mathbf{x}_v)$, we obtain, $\sigma(r^\star_{v,j} - r^\star_{v,i})\sigma(h(\mathbf{W}_i;\mathbf{x}_v) - h(\mathbf{W}_j;\mathbf{x}_v)) - \sigma(r^\star_{v,i} - r^\star_{v,j})\sigma(h(\mathbf{W}_j;\mathbf{x}_v) - h(\mathbf{W}_i;\mathbf{x}_v)) = \sigma(h(\mathbf{W}_i;\mathbf{x}_v) - h(\mathbf{W}_j;\mathbf{x}_v)) - \sigma(r^\star_{v,i} - r^\star_{v,j})$

$$\begin{aligned}
\nabla_{W_{i,k}}\mathcal{L}_{\text{DPO}}(\mathbf{W};\mathbf{X}) =& -\beta\sum_u^m x_{u,k}\sum_{j\neq i}^n \big(\sigma(r^\star_{u,j} - r^\star_{u,i}) - \sigma(h(\mathbf{W}_i;\mathbf{x}_u) - h(\mathbf{W}_j;\mathbf{x}_u))\big)\\
=& -\beta\sum_u^m x_{u,k}\sum_{j\neq i}^n \Delta_u(y_i,y_j;\mathbf{W})
\end{aligned}$$

Further, for any pair of responses (y_i, y_j),

$$\begin{aligned}
W_{i,k}^{(t+1)} =& W_{i,k}^{(t)} - \eta\beta\sum_u^m\sum_{j\neq i}^n \Delta_u(y_i,y_j;\mathbf{W}^{(t)})\\
\implies W_{i,k}^{(t+1)} - W_{j,k}^{(t+1)} =& W_{i,k}^{(t)} - W_{j,k}^{(t)} - \eta\beta\sum_u^m\sum_s^n x_{u,k}\Big(\Delta_u(y_i,y_s;\mathbf{w}^{(t)}) - \Delta_u(y_j,y_s;\mathbf{w}^{(t)})\Big)\\
\implies \beta(\mathbf{W}_i^{(t+1)} - \mathbf{W}_j^{(t+1)})\mathbf{x}_v =& \beta(\mathbf{W}_i^{(t)} - \mathbf{W}_j^{(t)})\mathbf{x}_v\\
& - \eta\beta^2\sum_k^d x_{v,k}\sum_u^m\sum_s^n x_{u,k}\Big(\Delta_u(y_i,y_s;\mathbf{w}^{(t)}) - \Delta_u(y_j,y_s;\mathbf{w}^{(t)})\Big)
\end{aligned}$$

Recall that. $\Delta_v(y_i,y_j;\mathbf{W}) = \sigma(r^\star_{v,i} - r^\star_{v,j}) - \sigma\left(h(\mathbf{W}_i,\mathbf{x}^v) - h(\mathbf{W}_j,\mathbf{x}^v)\right)$. We use $\delta_v(y_i,y_j;\mathbf{W}^t) = \delta_v^t(i,j)$ and $\Delta_v(y_i,y_j;\mathbf{W}^t) = \Delta_v^t(i,j)$ as short hand notation. Further, $\boldsymbol{\delta}^t_{i,j} = [\delta_v^t(i,j)]$ is a vector for different inputs x_v.

We also note that the lowest value of $\sigma'(z) = \sigma(z)(1-\sigma(z))$ is $\sigma'_{\text{min}} = \sigma(\rho)\sigma(-\rho) = c_\rho$ and its highest value is $1/4$ when $z \in [-\rho,\rho]$. Now, using the mean value theorem, we observe,

$$\begin{aligned}
&\frac{\sigma(x)-\sigma(y)}{x-y} = \sigma'(c) \text{ for some } c \in [-\rho,\rho] \text{ when } x,y \in [-\rho,\rho]\\
\implies & c_\rho \leq \frac{\sigma(x)-\sigma(y)}{x-y} \leq \frac{1}{4} \text{ when } x,y \in [-\rho,\rho]
\end{aligned}$$

Now, at each time step t, we order the actions such that $\delta_v^{(t)}(i,j) \geq 0$ if $i > j, \forall y_i, y_j \in \mathcal{Y},\ x_u \in \mathcal{X}$. Now, without loss of generalization, we assume $l < r$. We obtain an upper bound on $\delta_v^{t+1}(r,l)$ as follows,

$$\begin{aligned}
\delta_v^{(t+1)}(r,l) =& \delta_v^{(t)}(r,l) - \eta\beta^2 \sum_k^d x_{v,k} \sum_u^m x_{u,k} \sum_i^n \left(\Delta_u^{(t)}(r,i) - \Delta_u^{(t)}(l,i) \right) \\
\leq& \delta_v^{(t)}(r,l) - \eta\beta^2 \sum_k^d x_{v,k} \sum_u^m x_{u,k} \sum_{i=1}^{l-1} \left(c_\rho(r,i) - \frac{1}{4}\delta_u^{(t)}(l,i) \right) \\
&- \eta\beta^2 \sum_k^d x_{v,k} \sum_u^m x_{u,k} \sum_{i=l}^{r} \left(c_\rho \delta_u^{(t)}(r,i) - c_\rho \delta_u^{(t)}(l,i) \right) \\
&- \eta\beta^2 \sum_k^d x_{v,k} \sum_u^m x_{u,k} \sum_{i=r+1}^{n} \left(\frac{1}{4}\delta_u^{(t)}(r,i) - c_\rho \delta_u^{(t)}(l,i) \right) \\
=& \delta_v^{(t)}(r,l) - \eta\beta^2 \left[c_\rho(l-1) \sum_k^d x_{v,k} \sum_u^m x_{u,k}\delta_u^{(t)}(r,l) - \left(\frac{1}{4} - c_\rho \right) \sum_k^d x_{v,k} \sum_u^m x_{u,k} \sum_{i=1}^{l-1} \delta_u^{(t)}(l,i) \right] \\
&- \eta\beta^2 c_\rho (r-l+1) \sum_k^d x_{v,k} \sum_u^m x_{u,k}\delta_u^{(t)}(r,l) \\
&- \eta\beta^2 \left[c_\rho(n-r) \sum_k^d x_{v,k} \sum_u^m x_{u,k}\delta_u^{(t)}(r,l) - \left(\frac{1}{4} - c_\rho \right) \sum_k^d x_{v,k} \sum_u^m x_{u,k} \sum_{i=r+1}^{n} \delta_u^{(t)}(i,r) \right] \\
=& \delta_v^{(t)}(r,l) - \eta\beta^2 n c_\rho \sum_k^d x_{v,k} \sum_u^m x_{u,k}\delta_u^{(t)}(r,l) \\
&+ \eta\beta^2 \left(\frac{1}{4} - c_\rho \right) \sum_k^d x_{v,k} \sum_u^m x_{u,k} \left(\sum_{i=1}^{l-1} \delta_u^{(t)}(l,i) + \sum_{i=r+1}^{n} \delta_u^{(t)}(i,r) \right)
\end{aligned}$$

Note that the vector formed by $\sum_k^d x_{v,k} \sum_u^m x_{u,k}\delta_v^{(t)}(r,l)$ over v gives us $\mathbf{X}^T\mathbf{X}\boldsymbol{\delta}_{r,l}^{(t)}$.

$$\begin{aligned}
\implies \boldsymbol{\delta}_{r,l}^{(t+1)} =& \boldsymbol{\delta}_{r,l}^{(t)} - \eta\beta^2 n c_\rho \mathbf{X}^T\mathbf{X}\boldsymbol{\delta}_{r,l}^{(t)} + \eta\beta^2 \left(\frac{1}{4} - c_\rho \right) \mathbf{X}^T\mathbf{X} \left(\sum_{i=1}^{l+1} \boldsymbol{\delta}_{l,i}^{(t)} + \sum_{i=r+1}^{n} \boldsymbol{\delta}_{i,r}^{(t)} \right) \\
\leq& \boldsymbol{\delta}_{r,l}^{(t)} - \eta\beta^2 n c_\rho \sigma_{\min}^2(\mathbf{X})\boldsymbol{\delta}_{r,l}^{(t)} + \eta\beta^2 \left(\frac{1}{4} - c_\rho \right) \sigma_{\max}^2(\mathbf{X}) \left(\sum_{i=1}^{l+1} \boldsymbol{\delta}_{l,i}^{(t)} + \sum_{i=r+1}^{n} \boldsymbol{\delta}_{i,r}^{(t)} \right) \\
\implies \delta_v^{(t+1)}(r,l) \leq& \max_{u,i,j} \delta_u^{(t)}(i,j) \left(1 - \eta\beta^2 n c_\rho \sigma_{\min}^2(\mathbf{X}) + \frac{1}{4}\eta\beta^2 n \sigma_{\max}^2(\mathbf{X}) - \eta\beta^2 n c_\rho \sigma_{\max}^2(\mathbf{X}) \right) \\
\leq& \max_{u,i,j} \delta_u^{(t)}(i,j) \left(1 + \eta\beta^2 n \sigma_{\max}^2(\mathbf{X}) \left(\frac{1}{4} - 2c_\rho \right) \right)
\end{aligned}$$

Similarly, for a lower limit on $\delta_v^{(t)}(r,l)$, we observe,

$$
\begin{aligned}
& -\delta_v^{(t+1)}(r,l) \\
&= -\delta_v^{(t)}(r,l) + \eta\beta^2 \sum_k^d x_{v,k} \sum_u^m x_{u,k} \sum_i^n \left(\Delta_u^{(t)}(r,i) - \Delta_u^{(t)}(l,i)\right) \\
&\leq -\delta_v^{(t)}(r,l) + \eta\beta^2 \sum_k^d x_{v,k} \sum_u^m x_{u,k} \sum_{i=1}^{l-1} \left(\frac{1}{4}\delta_u^{(t)}(r,i) - c_\rho \delta_u^{(t)}(l,i)\right) \\
&\quad + \eta\beta^2 \sum_k^d x_{v,k} \sum_u^m x_{u,k} \left[\sum_{i=l}^{r} \left(\frac{1}{4}\delta_u^{(t)}(r,i) - \frac{1}{4}\delta_u^{(t)}(l,i)\right) + \sum_{i=r+1}^{n} \left(c_\rho \delta_u^{(t)}(r,i) - \frac{1}{4}\delta_u^{(t)}(l,i)\right)\right] \\
&= -\delta_v^{(t)}(r,l) + \eta\beta^2 \left[\frac{1}{4}(l-1) \sum_k^d x_{v,k} \sum_u^m x_{u,k}\delta_u^{(t)}(r,l) + \left(\frac{1}{4} - c_\rho\right) \sum_k^d x_{v,k} \sum_u^m x_{u,k} \sum_{i=1}^{l-1} \delta_u^{(t)}(l,i)\right] \\
&\quad + \eta\beta^2 \frac{1}{4}(r-l+1) \sum_k^d x_{v,k} \sum_u^m x_{u,k}\delta_u^{(t)}(r,l) \\
&\quad + \eta\beta^2 \left[\frac{1}{4}(n-r) \sum_k^d x_{v,k} \sum_u^m x_{u,k}\delta_u^{(t)}(r,l) + \left(\frac{1}{4} - c_\rho\right) \sum_k^d x_{v,k} \sum_u^m x_{u,k} \sum_{i=r+1}^{n} \delta_u^{(t)}(i,r)\right] \\
&= -\delta_v^{(t)}(r,l) + \frac{1}{4}\eta\beta^2 n \sum_k^d x_{v,k} \sum_u^m x_{u,k}\delta_u^{(t)}(r,l) \\
&\quad + \eta\beta^2 \left(\frac{1}{4} - c_\rho\right) \sum_k^d x_{v,k} \sum_u^m x_{u,k} \left(\sum_{i=1}^{l-1} \delta_u^{(t)}(l,i) + \sum_{i=r+1}^{n} \delta_u^{(t)}(i,r)\right)
\end{aligned}
$$

$$
\begin{aligned}
\implies -\boldsymbol{\delta}_{r,l}^{t+1} &= \frac{1}{4}\eta\beta^2 n \mathbf{X}^T\mathbf{X}\boldsymbol{\delta}_{r,l}^{(t)} - \boldsymbol{\delta}_{r,l}^{(t)} + \eta\beta^2 \left(\frac{1}{4} - c_\rho\right) \mathbf{X}^T\mathbf{X} \left(\sum_{i=1}^{l-1} \boldsymbol{\delta}_{l,i}^{t} + \sum_{i=r+1}^{n} \boldsymbol{\delta}_{i,r}^{t}\right) \\
&\leq \left(\frac{1}{4}\eta\beta^2 n \sigma_{\max}^2(\mathbf{X}) - 1\right) \boldsymbol{\delta}_{r,l}^{(t)} + \eta\beta^2 \left(\frac{1}{4} - c_\rho\right) \sigma_{\max}^2(\mathbf{X}) \left(\sum_{i=1}^{l-1} \boldsymbol{\delta}_{l,i}^{t} + \sum_{i=r+1}^{n} \boldsymbol{\delta}_{i,r}^{t}\right)
\end{aligned}
$$

$$
\implies -\delta_v^{t+1}(r,l) \leq \left(\max_{u,i,j} \delta_u^t(i,j)\right) \left(\eta\beta^2 n \sigma_{\max}^2(\mathbf{X})(\frac{1}{2} - c_\rho) - 1\right)
$$

Say, $\max\left(1 + \eta\beta^2 n \sigma_{\max}^2(\mathbf{X})\left(\frac{1}{2} - 2c_\rho\right), \eta\beta^2 n \sigma_{\max}^2(\mathbf{X})(\frac{1}{2} - c_\rho) - 1\right) = \zeta$ and $\max_{u,i,j} |\delta_u^t(i,j)| = \delta_{\max}^t$. We must choose η such that $0 < \zeta < 1$. Therefore, we obtain:

$$
|\delta_v^{t+1}(r,l)| \leq \zeta |\delta_{\max}^t|
$$

Therefore, we obtain,

$$
|\delta_v^{t+1}(r,l)| \leq |\delta_{\max}^1| \zeta^t
$$

E.2 Analysis of Quadratic Convergence of Learned Reward Differences

We follow an exercise similar to the proof of quadratic convergence with policy-guided mixed sampler and see that quadratic convergence does not hold for our choice of parameterization.

In the policy-guided sampling, the DPO loss function becomes,

$$\mathcal{L}_{\text{DPO}}(\theta = \mathbf{W}; \mathbf{X}) = -\sum_{v}^{m}\sum_{i}^{n}\sum_{j}^{n} d_{v,i,j} \log\sigma\left(\beta \log \frac{\pi_\theta(y_i \mid x_v)}{\pi_{\text{ref}}(y_i \mid x_v)} - \beta \log \frac{\pi_\theta(y_j \mid x_v)}{\pi_{\text{ref}}(y_j \mid x_v)}\right)$$

$$\begin{aligned}\text{where, } d_{v,i,j} =& p^\star(y_i \succ y_j | x_v) \operatorname{sg}\left(1 + \frac{1}{2}\left(\left(\frac{\pi_\theta(y_i \mid x_v)}{\pi_{\text{ref}}(y_i \mid x_v)} \frac{\pi_{\text{ref}}(y_j \mid x_v)}{\pi_\theta(y_j \mid x_v)}\right)^\beta + \left(\frac{\pi_\theta(y_j \mid x_v)}{\pi_{\text{ref}}(y_j \mid x_v)} \frac{\pi_{\text{ref}}(y_i \mid x_v)}{\pi_\theta(y_i \mid x_v)}\right)^\beta\right)\right) \\ =& p^\star(y_i \succ y_j | x_v) \operatorname{sg}\left(1 + \frac{1}{2}\left(\frac{\alpha_{v,i}}{\alpha_{v,j}} + \frac{\alpha_{v,i}}{\alpha_{v,j}}\right)\right)\end{aligned}$$

where $\operatorname{sg}(\cdot)$ is a stopping-gradient operator and $\pi_\theta(y_i|x_v) = \frac{\exp(\mathbf{W}_i \mathbf{x}_v)}{\sum_j^n \exp(\mathbf{W}_j \mathbf{x}_v)}$. We can simplify this as,

$$\begin{aligned}&\mathcal{L}_{\text{DPO}}(\mathbf{W}; \mathbf{X}) \\ =& -\sum_{v}^{m}\sum_{i}^{n}\sum_{j>i}^{n} \operatorname{sg}\left(1 + \frac{1}{2}\left(\frac{\alpha_{v,i}}{\alpha_{v,j}} + \frac{\alpha_{v,i}}{\alpha_{v,j}}\right)\right) \left[p^\star(y_i \succ y_j|x_v) \log\sigma\left(\log\frac{\alpha_{v,i}}{\alpha_{v,j}}\right) + p^\star(y_j \succ y_i|x_v) \log\sigma\left(\log\frac{\alpha_{v,j}}{\alpha_{v,i}}\right)\right] \\ =& -\sum_{v}^{m}\sum_{i}^{n}\sum_{j}^{n} \frac{1}{2} \operatorname{sg}\left(1 + \frac{\alpha_{v,i}}{\alpha_{v,j}}\right) \left[p^\star(y_i \succ y_j|x_v) \log\sigma\left(\log\frac{\alpha_{v,i}}{\alpha_{v,j}}\right) + p^\star(y_j \succ y_i|x_v) \log\sigma\left(\log\frac{\alpha_{v,j}}{\alpha_{v,i}}\right)\right]\end{aligned}$$

Now, we observe the gradient is,

$$\begin{aligned}&\nabla_{W_{i,k}} \mathcal{L}_{\text{DPO}}(\mathbf{W}; \mathbf{X}) \\ =& -\sum_{v}^{m} x_{v,k} \sum_{j}^{n} \frac{\beta}{2} \operatorname{sg}\left(2 + \left(\frac{\alpha_{v,i}}{\alpha_{v,j}} + \frac{\alpha_{v,i}}{\alpha_{v,j}}\right)\right) \Delta_v(y_i, y_j; \mathbf{W})\end{aligned}$$

We observe that,

$$\begin{aligned}2 + \frac{\alpha_{v,i}}{\alpha_{v,j}} + \frac{\alpha_{v,i}}{\alpha_{v,j}} =& 2 + \exp(h(\mathbf{W}_i, \mathbf{x}_v) - h(\mathbf{W}_j, \mathbf{x}_v)) + \exp(h(\mathbf{W}_j, \mathbf{x}_v) - h(\mathbf{W}_i, \mathbf{x}_v)) \\ =& (1 + \exp(h(\mathbf{W}_i, \mathbf{x}_v) - h(\mathbf{W}_j, \mathbf{x}_v)))(1 + \exp(h(\mathbf{W}_j, \mathbf{x}_v) - h(\mathbf{W}_i, \mathbf{x}_v))) \\ =& \frac{1}{\sigma'(h(\mathbf{W}_i, \mathbf{x}_v) - h(\mathbf{W}_j, \mathbf{x}_v)))}\end{aligned}$$

Therefore,

$$\nabla_{W_{i,k}} \mathcal{L}_{\text{DPO}}(\mathbf{W}; \mathbf{X}) = -\frac{\beta}{2} \sum_{v}^{m} x_{v,k} \sum_{j}^{n} \frac{\Delta_v(y_i, y_j; \mathbf{W})}{\sigma'(h(\mathbf{W}_i, \mathbf{x}_v) - h(\mathbf{W}_j, \mathbf{x}_v))}$$

Through Taylor expansion, we have that,

$$\Delta_v(y_i, y_j; \mathbf{W}) = \sigma'(h(\mathbf{W}_i, \mathbf{x}_v) - h(\mathbf{W}_j; \mathbf{x}_v))\delta_v(y_i, y_j; \mathbf{W}) + \frac{1}{2}\sigma''(\xi_v(y_i, y_j; \mathbf{W}))\delta(y_i, y_j; \mathbf{W})^2$$

where $\xi_v(y_i, y_j; \mathbf{W})$ lies between $h(\mathbf{W}_i, \mathbf{x}_v) - h(\mathbf{W}_j; \mathbf{x}_v)$ and $r^\star_{v,i} - r^\star_{v,j}$.

At any times step t during gradient descent, for any pair of responses y_i, y_j and the prompt x_v,

$$\begin{aligned}&\delta_v^{(t+1)}(i, j) \\ =& \delta_v^{(t)}(i, j) - \frac{\eta\beta^2}{2} \sum_{k}^{d} x_{v,k} \sum_{u}^{m} x_{u,k} \sum_{l}^{n} \left(\frac{\Delta_u(y_i, y_l; \mathbf{W}^{(t)})}{\sigma'(h(\mathbf{W}_i^{(t)}, \mathbf{x}_u) - h(\mathbf{W}_l^{(t)}, \mathbf{x}_u))} - \frac{\Delta_u(y_j, y_l; \mathbf{W}^{(t)})}{\sigma'(h(\mathbf{W}_j^{(t)}, \mathbf{x}_u) - h(\mathbf{W}_l^{(t)}, \mathbf{x}_u))}\right) \\ =& \delta_v^{(t)}(i, j) - \frac{\eta\beta^2}{2} \sum_{k}^{d} x_{v,k} \sum_{u}^{m} x_{u,k} \sum_{l}^{n} \left(\delta_u^{(t)}(i, l) - \delta_u^{(t)}(j, l)\right) \\ &- \frac{\eta\beta^2}{4} \sum_{k}^{d} x_{v,k} \sum_{u}^{m} x_{u,k} \sum_{l}^{n} \left[\frac{\sigma''(\xi_u(y_i, y_l; \mathbf{W}))(\delta_u^{(t)}(i, l))^2}{\sigma'(h(\mathbf{W}_i^{(t)}, \mathbf{x}_u) - h(\mathbf{W}_l^{(t)}, \mathbf{x}_u))} - \frac{\sigma''(\xi_u(y_j, y_l; \mathbf{W}))(\delta_u^{(t)}(j, l))^2}{\sigma'(h(\mathbf{W}_j^{(t)}, \mathbf{x}_u) - h(\mathbf{W}_l^{(t)}, \mathbf{x}_u))}\right]\end{aligned}$$

Rewriting this in terms of the vector $\boldsymbol{\delta}_{i,j}^{(t)}$,

$$\boldsymbol{\delta}_{i,j}^{(t+1)} = (\mathbb{I} - \frac{\eta\beta^2 n}{2}\mathbf{X}^T\mathbf{X})\boldsymbol{\delta}_{i,j}^{(t)} - \frac{\eta\beta^2}{4}\sum_{l}^{n}\mathbf{X}^T\mathbf{X}(\boldsymbol{\Phi}_{i,l}^{(t)} - \boldsymbol{\Phi}_{j,l}^{(t)}) \tag{20}$$

where the vector $\boldsymbol{\Phi}_{i,j}^{(t)} \in \mathbb{R}^m$ is a vector of size m where each element corresponding to the prompt x_u is of the form,

$$\Phi_u^{(t)}(i,j) = \frac{\sigma''(\xi_u(y_i, y_j; \mathbf{W}))(\delta_u^{(t)}(i,j))^2}{\sigma'(h(\mathbf{W}_i^{(t)}, \mathbf{x}_u) - h(\mathbf{W}_j^{(t)}, \mathbf{x}_u))}$$

We note that for quadratic convergence, we require the terms barring the $(\delta_u(i,j)^{(t)})^2$ to go to zero. However, it is not possible to choose a single value of step size η such that the $(\mathbb{I} - \frac{\eta\beta^2 n}{2}\mathbf{X}^T\mathbf{X})\boldsymbol{\delta}_{i,j}^{(t)}$ becomes zero for all x_v. Here, we cannot have $\mathbb{I} = \frac{\eta\beta^2 n}{2}\mathbf{X}^T\mathbf{X}$ unless $\mathbf{X}^T\mathbf{X}$ is a scaled version of the identity matrix, hence it is possible to imagine many examples where the linear term cannot be gotten rid of. Similarly, restricting $\boldsymbol{\delta}_{i,j}^{(t)}$ to the null space of $\mathbb{I} - \frac{\eta\beta^2 n}{2}\mathbf{X}^T\mathbf{X}$ will fail in many examples as well. To have a non-trivial null space of $\mathbb{I} - \frac{\eta\beta^2 n}{2}\mathbf{X}^T\mathbf{X}$, we remove the space corresponding to specific eigenvalues of $\mathbf{X}^T\mathbf{X}$ by setting $\eta = \frac{2}{\eta\beta^2 n\lambda_i}$. Then the component of $\delta_{i,j}^{(t)}$ in the subspace spanned by the eigenvectors of λ_i becomes zero. But restricting $\delta_{i,j}^{(t)}$ to this space can be too restrictive as it's most likely to be a low-dimensional space.

Since, we have a tight characterization of $\boldsymbol{\delta}_{i,j}^{(t+1)}$ in eq. 20, it suggests that quadratic convergence cannot be observed with a parameterization of the embeddings defined in 2. This suggests that the assumptions in Shi et al. [2025] were simplistic and the quadratic convergence doesn't seem to hold under more complex realistic assumptions. This speaks to the merit of our choice of parameterization that suggests that quadratic convergence will be ruled out.

F Supplementary Results

F.1 Synthetic Experiments with Random Embeddings

Experimental setup. In this setup, the embeddings of the queries $\mathbf{X}$ and the gold standard weights, $\mathbf{W}^\star$, are randomly generated. The ground truth rewards are calculated as $\mathbf{r}^\star = \beta(\mathbf{W}^\star - \mathbf{W}^{(\text{ref})})\mathbf{X}$. The coefficients in the loss function, $d_{i,j}^v$, are set to the preference probabilities calculated using the Bradley Terry Model, $\frac{e^{r^\star_{v,i}}}{e^{r^\star_{v,i}} + e^{r^\star_{v,j}}}$. A subset $\mathcal{X}_{\text{train}}$ of size defined by a budget k is chosen for performing gradient descent on the DPO loss function. This experiment is carried out for four different condition numbers of $\mathbf{X}$ and three different ratios of the largest and smallest eigenvalues of the Laplacian. The learned parameters are also tested by being used to estimate the rewards for a test set of queries.

Results. As predicted by our theoretical results, the rate of convergence of gradient descent is slower for higher condition numbers of $\mathbf{X}$ and higher finite condition number $\lambda_{\max}/\lambda_{\text{min}}^{+}$ of the Laplacian of comparisons. We observe that the parameters also converge to the gold standard parameters and the implicit reward differences converge to the ground truth reward differences.

F.2 Synthetic Experiments with Gaussian Embeddings

Experimental setup. In this setup, the embeddings of the queries $\mathbf{X}$ are randomly generated whereas the gold standard weights, $\mathbf{W}^{(\text{ref})}$ are obtained from two Gaussian clusters. The initial weights are set to match with the original clusters when centers around $(0,0)$ The ground truth rewards are calculated as $\mathbf{r}^\star = \beta(\mathbf{W}^\star - \mathbf{W}^{(\text{ref})})\mathbf{X}$. The convergence of the weights to the optimal weights is observed for different subset of queries with varying condition number.

Results. The final weights converge to the optimal weights with rates inversely proportional to the condition number of the input submatrix picked (fig 1), which supports our theoretical results.

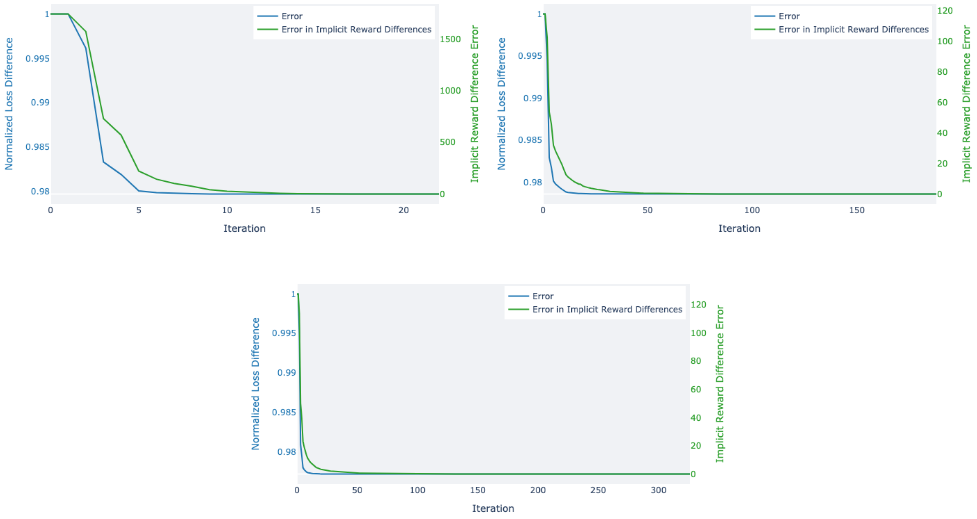

Figure 5: Convergence of loss function and error in $\mathbf{W}$ for different finite condition numbers of the Laplacian of preference data: (from left to right) $\kappa = 1.00$, $\kappa = 106.49$, and $\kappa = 364.09$.

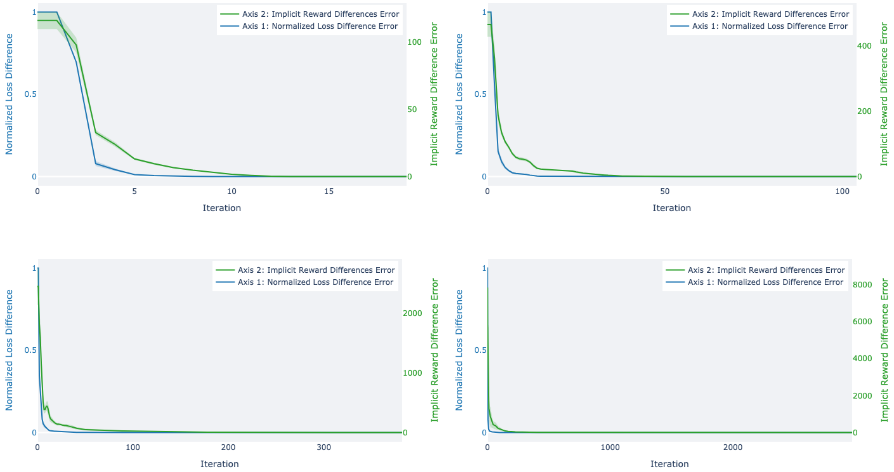

Figure 6: Convergence of $(\mathcal{L}^t_{\text{DPO}} - \mathcal{L}^\star_{\text{DPO}})/(\mathcal{L}^0_{\text{DPO}} - \mathcal{L}^\star_{\text{DPO}})$ and $\sum_v^m \sum_i^n \sum_j^n \delta_v^t(y_i, y_j)$ for different condition numbers of the query embedding matrix: (from left to right) $\kappa = 3.16$, $\kappa = 17.78$, $\kappa = 100.00$, and $\kappa = 316.23$.

Note that we also perform multiple iterations over different generated matrices with a fixed condition number ratio in order to show a general trend in the dependence as shown in fig 7.

F.3 Real World dataset experiments - Stanford Human Preferences

Experimental setup. We now add results for a set of initial experiments have been done using the SHP (Stanford Human Preferences) Ethayarajh et al. [2022]dataset with a similar pipeline as 5.2. From the prompts that appear at least 10 times, 5 responses each are generated from the reference model and all the comparison scores are given using the reward model [5] Dai et al. [2023]. These prompts are then embedded, from which well and bad conditioned subset of 512 prompts are collected (through random sampling). The subsets are selected based on the condition number $\frac{\sigma_1}{\sigma_k}$, where σ_i

[5] `https://huggingface.co/PKU-Alignment/beaver-7b-unified-reward`

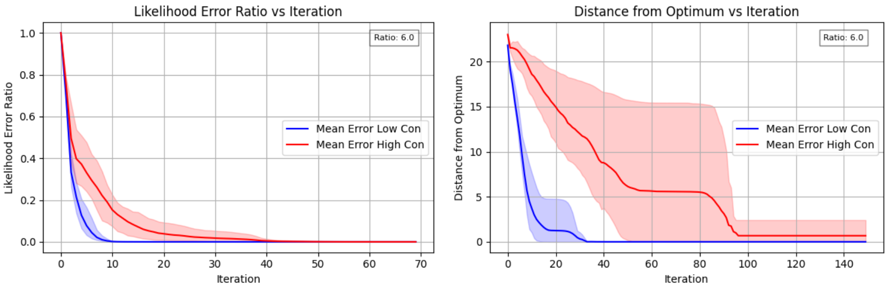

Figure 7: Convergence of $(\mathcal{L}^t_{\text{DPO}} - \mathcal{L}^\star_{\text{DPO}})/(\mathcal{L}^0_{\text{DPO}} - \mathcal{L}^\star_{\text{DPO}})$ and $\|\mathbf{W}' - \mathbf{W}^*\|_\infty$ when they are normalized when compared for different condition numbers over multiple iterations.

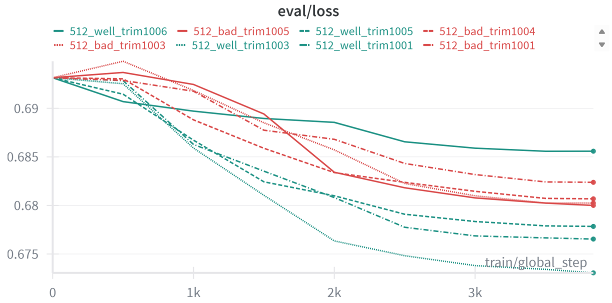

Figure 8: Evaluation loss for different picks of subsets based on condition number

denotes the top singular values in decreasing order and $k = 100$. This is motivated by the observation that only a few directions in the 1024-dimensional embedding space appear to be meaningful. We ran the experiment 4 times and evaluated it on the same held out dataset.

Results. The following trend (fig 8) is observed in the evaluation loss.

The ratio $\frac{\sigma_1}{\sigma_k}$ where $k = 100$ for the well conditioned and bad conditioned submatrices in the above were 80.53 ± 0.31 and 124.95 ± 2.00. The average evaluation loss (mean ± standard deviation) across training steps and the difference in the average steps required to reach various loss thresholds are shown in the table 2.

Eval Steps	Well-Cond Loss	Bad-Cond Loss
0	0.6931 ± 0.0000	0.6931 ± 0.0000
1000	0.6872 ± 0.0015	0.6905 ± 0.0013
2000	0.6817 ± 0.0044	0.6861 ± 0.0019
3000	0.6787 ± 0.0045	0.6829 ± 0.0019
3840	0.6783 ± 0.0046	0.6822 ± 0.0021

(a) Average evaluation loss (mean $\pm$ std).

Threshold	Well-cond	Bad-cond
0.693	58	256
0.690	699	1096
0.688	911	1459
0.685	1340	2249
0.683	1714	2939
0.680	2395	>3800

(b) Steps to reach thresholds.

Table 2: Comparison of well-conditioned vs bad-conditioned subsets for SHP dataset.

The relatively high variance in the well-conditioned subset stems from a single outlier. Further investigation is required to fully understand this behavior. Nonetheless, on average, subsets with lower condition numbers exhibit faster convergence when fine-tuned using DPO on the final layer. While these are small-scale experiments, they effectively seem to validate the theoretical dependencies and highlight the promise of extending this analysis to other real-world preference datasets.