MASS-DPO: MULTI-NEGATIVE ACTIVE SAMPLE SELECTION FOR DIRECT POLICY OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-negative preference optimization under the Plackett–Luce (PL) model extends Direct Preference Optimization (DPO) by leveraging comparative signals across one preferred and multiple rejected responses. However, optimizing over large pools of negatives is computationally prohibitive, and many candidates contribute redundant gradients due to their similar effects on policy updates. To address this, we introduce **MASS-DPO**, which derives the Fisher information matrix directly from the PL objective and shows that the problem of selecting negatives naturally reduces to a D-optimal design formulation. This formulation guarantees maximal informativeness and comprehensive coverage of the current policy's weaknesses. Moreover, the log-determinant criterion underlying D-optimal design admits a submodular structure, which we exploit through an incremental greedy algorithm that provides the natural computational realization of D-optimality, combining scalability with theoretical rigor. This incremental greedy strategy efficiently resolves the combinatorial complexity inherent in selecting a D-optimal negative set from large candidate pools. We establish convergence guarantees and finite-sample error bounds under this framework, and empirically demonstrate that MASS-DPO improves optimization efficiency and enhances downstream performance, achieving stronger alignment with substantially fewer negatives.

1 INTRODUCTION

Direct Preference Optimization (DPO) (Rafailov et al., 2023) aligns models directly with human preferences by optimizing pairwise comparisons without explicitly constructing reward functions (Christiano et al., 2017; Ouyang et al., 2022; Stiennon et al., 2020b). Recent works have generalized DPO using the Plackett-Luce (PL) model (Plackett, 1975; Luce et al., 1959) to accommodate multiple negative samples, enriching preference signals for more robust alignment. However, current multi-negative approaches such as Softmax-DPO (S-DPO) (Chen et al., 2024) and Direct Multi-Preference Optimization (DMPO) (Bai et al., 2024)typically select negatives randomly or heuristically, leading to redundant gradient information and computational inefficiencies.

To address these limitations, we propose MASS-DPO (**M**ulti-negative **A**ctive **S**ample **S**election for **D**irect **P**reference **O**ptimization), a theoretically-grounded active negative selection setting derived from the multi-negative Plackett-Luce preference optimization objective (Plackett, 1975; Luce et al., 1959). MASS-DPO formulates negative sample selection as a D-optimal design problem (Pukelsheim, 2006b; Kiefer, 1959; Pukelsheim, 2006a), leveraging the Fisher information matrix to measure the informativeness of each negative candidate (Fisher & Russell, 1922; Chaloner & Verdinelli, 1995; Flaherty et al., 2005; Kirsch et al., 2019). This selection is non-trivial, as many negative responses produce highly similar gradients under the PL objective, resulting in overlapping optimization signals that fail to introduce novel information for improving the policy. Without careful selection, the model repeatedly updates toward already-learned directions, leading to inefficient learning and slower convergence. To overcome this, MASS-DPO actively selects a diverse and informative subset of negatives by optimizing for coverage and signal diversity through a D-optimal design formulation.

Although the D-optimal formulation provides theoretical guarantees of optimal negative sampling, it introduces significant combinatorial complexity when selecting from a large candidate pool (Krause & Guestrin, 2012; Kirsch et al., 2019). To efficiently overcome this challenge, we further pro-

pose an incremental greedy algorithm that efficiently identifies a compact subset of negatives, provably equivalent in optimality to the full combinatorial selection (Nemhauser et al., 1978; Krause & Guestrin, 2012; Kirsch et al., 2019). This incremental approach effectively balances theoretical rigor with computational practicality, aligning with previous successful applications of greedy algorithms in information-theoretic sample selection (Sener & Savarese, 2017; Kirsch et al., 2019; Kveton et al., 2025).

We provide comprehensive theoretical analyses, establishing finite-sample estimation error bounds and convergence guarantees under our proposed selection framework. Empirically, we demonstrate that MASS-DPO significantly enhances optimization efficiency and alignment quality, achieving superior downstream task performance using substantially fewer negatives compared to existing methods across diverse benchmarks in language modeling and recommendation tasks. We summarize our contributions as follows:

- We propose MASS-DPO, an active negative sample selection method formulated as a D-optimal design problem, theoretically derived from the multi-negative Plackett-Luce optimization objective.
- We further introduce an incremental greedy selection algorithm, ensuring theoretical equivalence to the global optimal solution while significantly reducing computational overhead.
- We establish rigorous theoretical guarantees, including finite-sample estimation error bounds and convergence properties for MASS-DPO.
- Extensive empirical evaluations demonstrate that MASS-DPO outperforms baseline methods in optimization efficiency and downstream task performance across multiple language models and both recommendation and multiple-choice QA tasks.

2 Related Works

Direct Preference Optimization. DPO (Rafailov et al., 2023) aligns language models with human preferences by optimizing likelihood ratios of preferred over dispreferred responses, avoiding explicit reward modeling and associated complexities such as reward misgeneralization seen in RLHF (Christiano et al., 2017; Ouyang et al., 2022; Stiennon et al., 2020b). Recent extensions, such as introducing dynamic margins in ODPO (Amini et al., 2024) and computational optimizations via prefix sharing (Wang & Hegde, 2024), have further improved DPO's effectiveness and efficiency. However, standard DPO methods are typically restricted to binary preference pairs, which limits the diversity of supervision and often results in inefficient use of available preference data. In contrast, our approach extends beyond binary comparisons by leveraging actively selected, informative multi-negative samples, enabling more efficient and robust alignment.

Multi-negative Preference Optimization. Recent work has extended standard DPO's binary preference pairs to leverage multiple negatives for richer comparative signals and enhanced alignment. Softmax-DPO (S-DPO) (Chen et al., 2024) generalizes the pairwise Bradley–Terry loss (Bradley & Terry, 1952) to Plackett–Luce ranking (Plackett, 1975; Luce et al., 1959), providing richer gradient signals. Direct Multi-Preference Optimization (DMPO) (Bai et al., 2024) averages over multiple negatives to promote diverse negative learning. Multi Pair-wise Preference Optimization (MPPO) (Xie et al., 2024) extends DPO by directly modeling multi-negative feedback with average-likelihood loss, removing the need for a reference model and enabling flexible use of negative samples. Tree Preference Optimization (TPO) (Liao et al., 2024) structures multi-negative alignment through hierarchical preference decomposition. Despite these advances in multi-negative preference optimization, current methods still largely depend on heuristic or random negative selection strategies. Our work addresses this limitation by proposing MASS-DPO, which leverages D-optimal design for theoretically grounded, strategic negative sample selection.

3 Preliminaries

3.1 Direct Preference Optimization

Direct Preference Optimization (DPO) (Rafailov et al., 2023) aligns a learned policy with human pairwise judgments (Christiano et al., 2017; Stiennon et al., 2020a; Ouyang et al., 2022) without

an explicit reward model. Under the Bradley-Terry-Luce framework (Bradley & Terry, 1952), two responses y_1, y_2 to prompt x with latent scores $r(x, y_1), r(x, y_2)$ satisfy

$$p^*(y_1 \succ y_2 \mid x) = \sigma(r(x, y_1) - r(x, y_2)), \tag{1}$$

where $\sigma(z) = 1/(1 + e^{-z})$. Rearranging the optimal-policy relation gives an implicit reward decomposition up to an additive normalizer $Z(x)$:

$$r(x, y) = \beta \log \frac{\pi^*(y \mid x)}{\pi_{\text{ref}}(y \mid x)} + Z(x), \quad Z(x) = \sum_y \pi_{\text{ref}}(y \mid x) \cdot \exp\Big(\tfrac{1}{\beta}\, r(x, y)\Big) \tag{2}$$

Substituting equation 2 into equation 1 and simplifying leads to the DPO training objective

$$\mathcal{L}_{\text{DPO}}(\theta) = -\mathbb{E}_{(x,y_1,y_2)\sim D}\Big[\log \sigma\Big(\beta \log \tfrac{\pi_\theta(y_1|x)}{\pi_{\text{ref}}(y_1|x)} - \log \tfrac{\pi_\theta(y_2|x)}{\pi_{\text{ref}}(y_2|x)}\Big)\Big]. \tag{3}$$

3.2 Multi-negative Preference Optimization

Multi-negative preference optimization generalizes the Direct Preference Optimization framework (Rafailov et al., 2023) to better align language models with multiple negative preferences. While traditional DPO employs the Bradley-Terry (BT) model (Bradley & Terry, 1952) to capture pairwise comparisons, multi-negative preference optimization leverages the Plackett-Luce (PL) model (Plackett, 1975; Luce et al., 1959) to accommodate the ranking of a preferred item against multiple disfavored items.

Consider a user prompt x_u that is formed from historical interactions, along with a preferred item e_p and a set of dispreferred items E_d. The aim is to maximize the probability that the preferred item e_p is ranked above every item in E_d, as described by

$$p^*(e_p \succ E_d \mid x_u) = \frac{\exp(r(x_u, e_p))}{\sum_{e_d \in \{e_p\} \cup E_d} \exp(r(x_u, e_d))}, \tag{4}$$

where $r(x_u, e)$ is the latent reward function defined over the prompt-response pairs in the RLHF framework (Ouyang et al., 2022). From Eq. equation 4, we obtain the following multi-negative preference loss:

$$\mathcal{L}(\theta) = -\mathbb{E}_{(x_u,e_p,E_d)\sim D}\Big[\log \sigma\Big(-\log \sum_{e_d \in E_d} \exp\big(\beta\, \Delta(x_u, e_d, e_p)\big)\Big)\Big], \tag{5}$$

with $\sigma(\cdot)$ denoting the sigmoid function and $\Delta(x_u, e_d, e_p) = \log \frac{\pi_\theta(e_d|x_u)}{\pi_{\text{ref}}(e_d|x_u)} - \log \frac{\pi_\theta(e_p|x_u)}{\pi_{\text{ref}}(e_p|x_u)}$. Notably, this formulation reverts to the original DPO setup when the set of dispreferred items contains just a single element (i.e., $|E_d| = 1$). This naturally extends DPO by incorporating multi-negative preference alignment into language model training for recommendation tasks.

4 MASS-DPO: Multi-negative Active Sample Selection

In multi-negative preference optimization tasks (*e.g.*, recommendation, multiple-choice QA, information retrieval), the selection of negative samples significantly influences alignment efficiency and effectiveness. *Uninformative* negatives, already well-separated from preferred responses, waste gradient computations and hinder convergence (Yang et al., 2023; Kalantidis et al., 2020; Robinson et al., 2020; Zhang et al., 2022). Thus, the key challenge is strategically selecting a compact yet informative subset of negatives to highlight the policy's weaknesses while maintaining numerical stability (Ma et al., 2024; Kirsch & Gal, 2022; Fan et al., 2023). To address this, we propose MASS-DPO (Figure 1), an active negative selection method formulated as a D-optimal design problem (Pukelsheim, 2006b; Cohn, 1993; Kirsch et al., 2019), maximizing a Fisher-information surrogate (Fisher & Russell, 1922; Jung & Lee, 2021; Liu et al., 2024; Neilsen et al., 2018; Sourati et al., 2017; Chaloner & Verdinelli, 1995; Ash et al., 2021). By maximizing this surrogate, MASS-DPO effectively minimizes the volume of the confidence ellipsoid of policy parameters connecting computational efficiency with robust statistical guarantees (Sec. 5). We outline the core assumptions and derive gradient and curvature expressions central to our analysis and optimization approach.

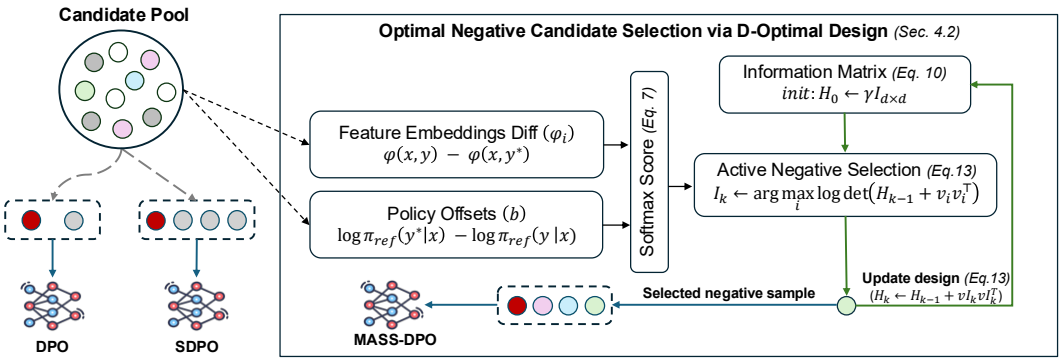

Figure 1: Overview of MASS-DPO's D-optimal selection. Each candidate is scored by its feature-difference and policy offset via softmax (Equation (8)). The green loop denotes the incremental greedy update: starting from H_0, we incrementally pick the negative that maximally increases $\log\det H$, then update H accordingly until n samples are selected.

4.1 Setting

Following (Kveton et al., 2025; Riquelme et al., 2018; Das et al., 2024; Mukherjee et al., 2024; Liu et al., 2024; Thekumparampil et al., 2024) in regret minimization and reward-model active learning, we linearize the policy's final layer to obtain a tractable Fisher-information objective. We assume:

Assumption 4.1 *To enable tractable analysis and algorithmic design, we assume the policy under consideration takes a log-linear form:*

$$\pi(y \mid x;\theta) \propto \exp\big(\phi(x,y)^\top \theta\big), \tag{6}$$

where $\phi(x,y) \in \mathbb{R}^d$ denotes the feature embedding of the context–response pair (x,y), and $\theta \in \mathbb{R}^d$ the model parameters.

Under Assumption 4.1, we can represent the relative preference between a preferred response y^* and a set of negative responses $\{y_i\}_{i=1}^n$ through feature differences. Specifically, defining

$$\phi_i = \phi(x,y_i) - \phi(x,y^*), \quad b_i = \log \frac{\pi_{\text{ref}}(y^* \mid x)}{\pi_{\text{ref}}(y_i \mid x)},$$

allows us to write the multi-negative DPO loss compactly in terms of the log-sum-exp operator.

$$L(\theta; S_n) = -\log\sigma\Big(-\log \sum_{i\in S_n} \exp\big(\beta\,(\phi_i^\top\theta + b_i)\big)\Big), \tag{7}$$

where $S_n = \{y_1,\dots,y_n\}$ is the chosen negative set, $\sigma(\cdot)$ the sigmoid. Then, we propose the following lemmas to quantify how each candidate negative alters the gradient and curvature. These statistics show that negatives whose feature differences are *diverse* and *orthogonal*, enlarge the information matrix the most, while redundant examples leave their volume almost unchanged.

Lemma 4.1 *Definition of auxiliary terms: normalization factor Z_n and softmax weights p_j for gradient computation*

$$p_j = \frac{\exp\big[\beta(\phi_j^\top\theta + b_j)\big]}{\sum_{k\in S_n}\exp\big[\beta(\phi_k^\top\theta + b_k)\big]}, \quad Z_n = -\log\sum_{i\in S_n}\exp\Big[\beta(\phi_i^\top\theta + b_i)\Big] \tag{8}$$

Then the gradient of equation 7 with respect to θ is given by

$$\nabla_\theta L(\theta; S_n) = \beta\,(1-\sigma(Z_n))\sum_{j\in S_n} p_j\,\phi_j. \tag{9}$$

The detailed derivation is provided in Appendix A.1. This result shows that the gradient is a weighted combination of feature differences, scaled by the probability of misranking, thus facilitating intuitive interpretations in terms of correction signals. We observe that p_j emphasises negatives whose score margin is small, which are borderline mistakes that have the greatest influence on the policy, corroborating the need to focus selection on *hard yet informative* examples.

Lemma 4.2 *Let $\phi = \sum_{j \in S_n} p_j \, \phi_j$ denote the expected feature difference under the softmax distribution. The Hessian of equation 7 is then*

$$\nabla^2 L(\theta; S_n) = \beta^2 (1 - \sigma(Z_n)) \Big[\sigma(Z_n) \, \phi \phi^\top + \sum_{j=1}^{n} p_j (\phi_j - \phi)(\phi_j - \phi)^\top \Big]$$

$$\succeq \beta^2 (1 - \sigma(Z_n)) \sum_{j=1}^{n} p_j (\phi_j - \phi)(\phi_j - \phi)^\top, \tag{10}$$

which is positive semi-definite and captures both the low-rank and dispersion contributions to curvature. The detailed derivation is provided in Appendix A.2.

Based on the lower bound of the Hessian matrix of the multi-negative DPO objective, we directly maximize the latter motivates a determinant objective that prefers sets to spread along orthogonal directions in feature space. These gradient and Hessian expressions form the basis for our multi-negative active sampling strategies, enabling principled optimization of negative sets under budget constraints while controlling estimation uncertainty and convergence behavior.

4.2 Negative Selection via D-Optimal Design

When selecting from a large-scale negative pool in multi-negative DPO, more negatives can improve parameter estimates, but those samples can add little beyond what is already conveyed by a smaller, well-chosen subset. MASS-DPO enables negative selection as a ***D-optimal design*** (Kiefer, 1959; Pukelsheim, 2006b; Kirsch et al., 2019) problem that explicitly maximizes the information gain (Chaloner & Verdinelli, 1995) about the policy parameters.

Fisher-information objective. For a candidate negative $j \in \mathcal{D}$ let $v_j = \sqrt{p_j}\big(\phi_j - \phi\big)$ be its Fisher-information contribution, where p_j is the softmax weight derived in Section 4.1. Given a subset $S \subseteq \mathcal{D}$ we define the regularised information matrix

$$H(S) = \gamma I + \beta^2 (1 - \sigma(Z_n)) \sum_{j \in S} v_j v_j^\top, \quad \gamma > 0. \tag{11}$$

The D-optimal criterion seeks to optimize the following objective,

$$S_n^* \;=\; \arg \max_{S \subset \mathcal{D},\, |S| = n} \log \det H(S), \tag{12}$$

which maximizes the volume of the confidence ellipsoid for the policy parameters and promotes better convergence of DPO. Problem equation 12 is however NP-hard (Welch, 1982; Allen-Zhu et al., 2021), as it is a combinatorial optimization over $\binom{|\mathcal{D}|}{n}$ subsets. To overcome this computational challenge, we further propose a greedy and iterative sample-selection strategy to incrementally optimize information gain (Nemhauser & Wolsey, 1978; Krause et al., 2008).

Incremental Greedy Information Maximization.; To overcome the combinatorial optimization problem, we exploit the matrix-determinant identity

$$\log \det\big(H + v v^\top\big) = \log \det H + \log\big(1 + v^\top H^{-1} v\big), \tag{13}$$

which is valid for any positive-definite matrix H and vector v. We initialize the design matrix by $H_0 = \gamma I$, and the incremental greedy algorithm adds one negative at a time. At iteration k it selects

$$i_k = \arg \max_{i \notin S_{k-1}} v_i^\top H_{k-1}^{-1} v_i, \quad S_k = S_{k-1} \cup \{i_k\}, \quad H_k = H_{k-1} + v_{i_k} v_{i_k}^\top, \tag{14}$$

and updates H_k^{-1} via the Sherman-Morrison formula (Sherman & Morrison, 1950) in $\mathcal{O}(d^2)$ time. The term $v_i^\top H_{k-1}^{-1} v_i$ is the covariance matrix H_{k-1} induced norm of v_i, so each step chooses the negative probing the *least explored* direction of the parameter space (Kveton et al., 2025). We illustrate this algorithm in detail in Algorithm 1. We formally demonstrate that this greedy Algorithm 1 achieves the same objective value as the intractable global optimum in Equation (12).

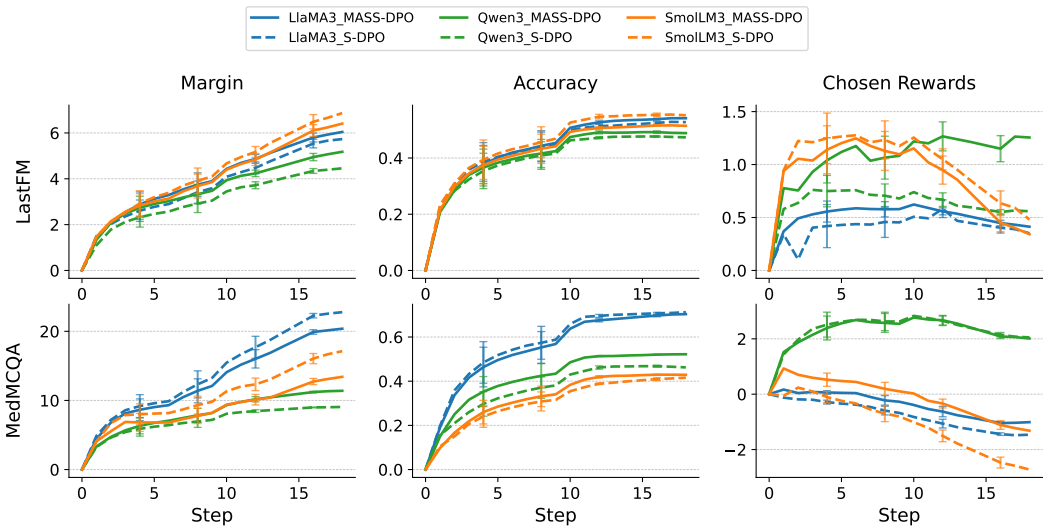

Figure 2: Margin, accuracy, and chosen reward comparisons on LastFM and MedMCQA datasets. MASS-DPO maintains consistently larger margins and higher accuracies, clearly illustrating the advantage of optimal negative selection over the random weighting approach used in SDPO.

Lemma 4.3 (Optimality of the Incremental Greedy Algorithm) *With $H_0 = \gamma I$ and $\gamma > 0$, the subset S_n produced by the above procedure satisfies*

$$\log\det H(S_n) = \log\det H(S_n^*), \tag{15}$$

where $S_n^\star$ is the maximizer in equation 12.

We provide the proof of Lemma 4.3 in Appendix A.3. Lemma 4.3 ensures that the incremental selection mechanism embedded in MASS-DPO realizes the maximal Fisher information attainable with exactly n negatives, justifying the finite-sample error bounds and convergence rates established in Section 5. Empirically, this property translates into faster alignment with a fraction of the computational cost required by exhaustive negative processing.

5 Theoretical Analysis

We analyze the generalization performance of MASS-DPO under the linearized setting introduced above. Our goal is to quantify how the error in the estimated policy parameters affects the quality of logit predictions across all possible negative samples. To do so, we rely on a few standard assumptions necessary for our analysis (feature structure, boundedness, diversity, and culminate in a finite-sample generalization guarantee); the full statements are deferred to Appendix B. The analysis then proceeds to our main finite-sample guarantees.

Theorem 5.1 (Maximum Logit Error Bound)

Let $\hat{\theta}_n = \arg\min_{\theta\in\Theta} L(\theta; \mathcal{S}_n)$. Then the maximum logit error under is

$$\mathcal{E}(\hat{\theta}_n, \theta_*) = \tilde{O}(d\sqrt{\log(1/\delta)/n})$$

with probability at least $1-\delta$, where $\tilde{O}$ hides all logarithmic factors but those in δ.

Theorem 5.2 (Batch Design Estimation Error)
With probability at least $1-\delta$, given the total dataset $S_{k,n}$ of k prompts and selected n negative samples per prompt, the deviation of the estimated parameter from the true optimum is bounded in the $\Sigma_{k,n}$-norm:

$$\left\|\hat{\theta}_{k,n} - \theta_*\right\|_{\Sigma_{k,n}} \le \sqrt{\frac{d}{4}\log\left(\frac{1}{\delta} + \frac{k\cdot c_{\min}/\gamma}{(1 - c_{\min}\cdot k/\gamma)^{1/d}\cdot\delta}\right)} + 2\,\gamma^{1/2}. \tag{16}$$

where $\gamma > 0$ is the regularization constant used to define $\Sigma_{k,n} = \gamma I + \nabla^2 L(\theta^*; S_{k,n})$, and the weighted sum of the features is $\phi = \sum_{i \in S_n} p_i \phi_i$.

This follows (Abbasi-Yadkori et al., 2011; Kveton et al., 2025) by treating the S-DPO loss as a generalized linear model and applying self-normalized concentration bounds to the stochastic gradients. In practice, Theorem 5.1 suggests that with only a small number of negatives, MASS-DPO can already achieve bounded logit error, which translates into faster convergence in training. Theorem A.1 and Theorem 5.2 further imply that the selected negatives ensure stable generalization and better margin improvements across prompts, which we will verify in our experiments Section 6.

6 EXPERIMENTS

Datasets. Following recent DPO-based recommendation work (Chen et al., 2024), (Sun et al., 2024), and (He et al., 2025) we utilize two widely adopted real-world recommendation benchmarks: LastFM (Bertin-Mahieux et al., 2011) and MovieLens (Harper & Konstan, 2015). For QA tasks, we adopt two challenging multiple-choice QA datasets: MedMCQA (Pal et al., 2022), a medical-domain QA benchmark, and QASC (Khot et al., 2020), a scientific reasoning QA dataset. Following prior works (Rafailov et al., 2023), we report Accuracy, Margin, Chosen Rewards and several additional utility metrics, with detailed methodology available in Appendix B.2.

Methods. We benchmark MASS-DPO against established preference alignment approaches, categorized into pairwise methods, DPO (Rafailov et al., 2023) and its multi-negative extension DPO-k, and multi-negative methods, Softmax-DPO (SDPO) (Chen et al., 2024) and DMPO (Bai et al., 2024). To maintain fairness and manage computational costs, the number of negative candidates during training is set to 3 for all multi-negative methods (DPO-k, DMPO, SDPO, MASS-DPO) and 1 for DPO. However, for evaluation and test sets, we include all available negative candidates to better assess the model's ability to select the best sample from a larger pool (e.g., 20 candidates), thereby increasing the search space and providing a more robust measure of real-world performance.

Implementation details are provided in Appendix B.3. Our experiments are designed to validate the theoretical insights in Sections 4 and 5. In particular:

LLM Usage: We used large language models solely for grammar refinement and minor wording edits in drafting parts of this paper.

6.1 HOW WELL DOES MASS-DPO OPTIMIZE THE MULTI-NEGATIVE PREFERENCE LEARNING OBJECTIVE?

We compare MASS-DPO's active negative selection to the softmax-based random selection in SDPO on recommendation (LastFM) and QA (MedMCQA). Figure 2 tracks three alignment metrics during training: *margin* (logit gap between preferred vs. rejected), *accuracy*, and *chosen rewards*. Additional results for Movielens and QASC appear in Appendix B.2. Across experiments, MASS-DPO (solid) achieves larger margins and faster early gains than SDPO (dashed) on both datasets, with the gap emerging early and persisting through training. *Accuracy* follows the same pattern: curves for MASS-DPO rise more quickly and attain consistently higher plateaus. Finally, *chosen-reward trajectories* under MASS-DPO are smoother and more stable across steps, while SDPO exhibits noticeably noisier dynamics. Taken together, these trends indicate that actively selecting informative negatives leads to more efficient optimization of the multi-negative preference objective than random softmax selection.

6.2 HOW DOES MASS-DPO IMPROVE DOWNSTREAM POLICY PERFORMANCE COMPARED TO EXISTING PREFERENCE OPTIMIZATION METHODS?

To evaluate MASS-DPO's effectiveness in downstream policy performance we benchmark its performance against established preference optimization methods: DPO, DMPO, DPO-k, and SDPO on four datasets (MedMCQA, QASC, LastFM, MovieLens) using *Accuracy*. Results are reported for three base models in Table 1. MASS-DPO consistently outperform baselines across datasets and language models. Notably, MASS-DPO outperforms prior methods on average for Qwen3 and SmolLM3, and is a close second to S-DPO for Llama3. Per-dataset, MASS-DPO wins the majority

Model	Setting	Medmcqa	QASC	LastFM	MovieLens	Avg↑
Qwen3	DPO	43.49	68.43	45.75	31.96	47.41
	DMPO	28.91	66.78	43.40	25.66	41.19
	DPO-k	55.56	<u>71.96</u>	<u>51.10</u>	44.56	<u>55.80</u>
	S-DPO	<u>52.56</u>	71.08	50.25	**48.19**	55.52
	MASS-DPO	**56.66**	**72.19**	**52.30**	<u>47.58</u>	**57.18**
SmolLM3	DPO	33.27	67.00	51.90	37.60	47.44
	DMPO	25.50	65.23	50.10	28.68	42.38
	DPO-k	44.09	<u>69.98</u>	55.70	51.36	55.28
	S-DPO	**44.99**	69.43	<u>55.90</u>	**55.70**	<u>56.50</u>
	MASS-DPO	<u>44.19</u>	**71.63**	**57.25**	<u>54.03</u>	**56.78**
Llama3	DPO	52.25	71.08	54.60	33.52	52.86
	DMPO	25.70	69.87	49.95	28.18	43.42
	DPO-k	71.04	<u>73.95</u>	55.65	44.46	61.27
	S-DPO	**72.19**	**74.61**	<u>56.55</u>	<u>49.55</u>	**63.23**
	MASS-DPO	<u>71.29</u>	73.62	**57.35**	**49.70**	<u>62.99</u>

Table 1: Accuracy (%) on four tasks across three base models. **Bold** = best, <u>underlined</u> = second best.

Model	Method	MedMCQA MRR/Margin	QASC MRR/Margin	LastFM MRR/Margin	MovieLens MRR/Margin	Average↑ MRR/Margin
Qwen3	S-DPO	69.74 / 9.33	81.87 / 5.29	64.12 / 4.67	61.29 / **4.98**	69.26 / 6.07
	MASS-DPO	**73.30** / **11.76**	**82.71** / **6.22**	**66.28** / **5.51**	**61.61** / 4.94	**70.97** / **7.11**
SmolLM3	S-DPO	**66.64** / **19.10**	81.64 / **8.07**	70.13 / **7.62**	**68.73** / 7.32	71.78 / 10.53
	MASS-DPO	66.30 / 14.42	**82.73** / 7.91	**70.79** / 7.29	68.13 / **7.41**	**71.99** / 9.26
Llama3	S-DPO	**83.44** / **23.93**	**84.13** / **7.26**	70.58 / 6.42	63.92 / 5.20	75.52 / **10.70**
	MASS-DPO	82.86 / 21.36	84.06 / 7.24	**70.71** / **6.53**	**65.58** / **5.75**	**75.80** / 10.22

Table 2: MRR and Margin across four datasets. Each cell shows *MRR / Margin*.

of tasks for Qwen3 and matches or surpasses the strongest baseline on two tasks for both SmolLM3 and Llama3, while remaining top–2 in every dataset.

By contrast, simpler pairwise methods (DPO, DMPO, DPO-k) consistently underperform across tasks, emphasizing the limitations of these methods in effectively exploiting multiple negative samples. Although SDPO integrates multiple negatives through a softmax-based weighting, its lower performance relative to MASS-DPO demonstrates the critical importance of strategic negative selection rather than random or heuristic selection. These empirical results directly confirm our theoretical insights, maximizing the D-optimal objective (Eq. 12) selects complementary negatives that expand Fisher-information coverage, yielding higher downstream accuracy than random/softmax weighting (S-DPO) and pairwise methods that do not strategically use multiple negatives.

6.3 How does MASS-DPO achieve better negative selection?

We assess negative-selection quality using downstream utility metrics, MRR and Margin (Table 2), and ranking quality on recommendation and QA via Recall/NDCG at $k \in \{1, 3\}$ (Tables 3 and 5). Across base models and datasets, MASS-DPO consistently improves ***MRR*** over S-DPO, while delivering higher or comparable ***Margins***. On ranking metrics, MASS-DPO attains the most or tied-best scores in a majority of {R@1, N@1} cells and remains competitive at {R@3, N@3}, indicating better performance on both recommendation and QA. These results demonstrate that active negative selection via MASS-DPO effectively highlights and corrects policy weaknesses, enhancing downstream alignment and utility beyond standard optimization methods.

6.4 Ablation Studies

MASS-DPO's behavior is governed by the Fisher-information structure (Equation (11)) and the D-optimal selection objective (Equation (12)). We therefore ablate two key knobs predicted by theory to matter most: the KL regularization scale β and the number of selected negatives k.

Model	Method	LastFM				MovieLens			
		R@1	R@3	N@1	N@3	R@1	R@3	N@1	N@3
Qwen3	DPO	46.15	72.60	46.15	61.60	29.64	59.48	29.64	46.89
	DMPO	44.50	72.05	44.50	60.57	24.50	56.30	24.50	42.88
	DPO-k	49.50	76.45	49.50	65.36	41.63	68.95	41.63	57.71
	S-DPO	48.55	75.10	48.55	64.14	45.92	71.47	45.92	60.86
	MASS-DPO	**51.10**	**77.20**	**51.10**	**66.48**	**45.97**	**71.52**	**45.97**	**61.10**
SmolLM3	DPO	51.70	78.15	51.70	67.29	37.25	65.68	37.25	53.77
	DMPO	50.30	77.90	50.30	66.54	28.23	60.43	28.23	47.02
	DPO-k	56.30	80.55	56.30	70.71	51.01	75.71	51.01	65.41
	S-DPO	55.60	**81.35**	55.60	70.84	**55.09**	**78.18**	**55.09**	**68.64**
	MASS-DPO	**57.05**	80.70	**57.05**	**71.08**	54.18	77.57	54.18	68.03
Llama3	DPO	55.15	80.35	55.15	70.06	34.48	63.56	34.48	51.31
	DMPO	49.95	78.35	49.95	66.76	27.82	58.72	27.82	45.69
	DPO-k	56.05	80.30	56.05	70.41	43.95	70.77	43.95	59.69
	S-DPO	56.50	80.85	56.50	70.95	48.94	73.39	48.94	63.25
	MASS-DPO	**56.60**	**81.15**	**56.60**	**71.17**	**50.66**	**76.01**	**50.66**	**65.57**

Table 3: Recall (R) and NDCG (N) at k={1,3} on LastFM and MovieLens.

(a) β ablation

Model	β	Medmcqa	QASC	LastFM	MovieLens
Qwen3	0.1	**56.66**0.77	**72.19**1.05	**52.30**0.79	**47.58**0.80
	0.5	46.29 0.79	71.41 1.06	48.15 0.81	39.82 0.79
	1.0	43.49 0.77	69.65 1.06	44.15 0.80	34.12 0.74
SmolLM3	0.1	**44.19**0.79	**71.63**1.07	**57.25**0.79	54.03 0.77
	0.5	39.73 0.76	**71.63**1.03	54.75 0.79	**56.30**0.79
	1.0	35.42 0.75	68.98 1.06	52.00 0.80	52.12 0.80
Llama3	0.1	**71.29**0.74	**73.62**1.03	**57.35**0.81	49.70 0.80
	0.5	69.69 0.74	73.51 1.01	55.75 0.80	**51.06**0.78
	1.0	66.28 0.76	72.19 1.02	52.25 0.81	45.92 0.78

(b) Negatives k ablation

Model	k	Medmcqa	QASC	LastFM	MovieLens
Qwen3	1	50.95 0.78	68.21 1.13	47.80 0.80	32.86 0.73
	3	56.66 0.77	72.19 1.05	52.30 0.79	47.58 0.80
	5	**57.31**0.75	**73.73**1.03	**54.50**0.79	**58.11**0.78
SmolLM3	1	29.26 0.72	65.67 1.09	50.70 0.81	34.58 0.75
	3	44.19 0.79	**71.63**1.07	57.25 0.79	54.03 0.77
	5	**46.59**0.79	**71.63**1.04	**59.50**0.77	**65.07**0.74
Llama3	1	46.99 0.78	71.96 1.03	52.70 0.81	32.71 0.73
	3	71.29 0.74	73.62 1.03	57.35 0.81	49.70 0.80
	5	**73.55**0.71	**74.94**0.98	**60.05**0.80	**60.99**0.77

Table 4: MASS-DPO ablations on two hyperparameters. (a) Varying the scale β (0.1, 0.5, 1.0) while holding the number of negatives k fixed. (b) Varying k (1, 3, 5) while holding β fixed.

6.4.1 Effect of β

The coefficient β tunes the pull toward the reference model and, through the softmax weights p_j, scales each candidate's Fisher contribution $v_j = \sqrt{p_j}\,\tilde{\phi}_j$ in the D-optimal criterion. Larger β sharpens p_j and can better separate informative from redundant negatives, but too much regularization restricts useful policy updates. Sweeping $\beta \in \{0.1, 0.5, 1.0\}$ across three model families (Table 4a), we find $\beta = 0.1$ consistently yields the strongest results.

6.4.2 Number of Negatives (k)

D-optimal design predicts that adding more negatives improves parameter estimation until coverage of the information space saturates. Varying $k \in \{1, 3, 5\}$ with our greedy selector (Table 4b) shows monotonic gains from $k{=}1 \to 3 \to 5$ across models and datasets. These results indicate the greedy procedure reliably assembles complementary negatives that expand $\log\det$ of the information matrix, aligning empirical improvements with our D-optimal design analysis.

7 Conclusion

In this work, we introduced MASS-DPO, a theoretically grounded approach to active negative sample selection for multi-negative direct preference optimization. By formulating negative sampling as a D-optimal design problem, we effectively addressed redundancy and computational inefficiencies inherent in existing methods. Our incremental greedy algorithm ensures computational feasibility while retaining theoretical optimality. Theoretical analyses confirm the efficiency and convergence guarantees of MASS-DPO, and comprehensive experiments illustrate its superior performance and scalability across diverse language modeling and both recommendation and QA tasks.

REFERENCES

Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. *Advances in neural information processing systems*, 24, 2011.

Zeyuan Allen-Zhu, Yuanzhi Li, Aarti Singh, and Yining Wang. Near-optimal discrete optimization for experimental design: A regret minimization approach. *Mathematical Programming*, 186: 439–478, 2021.

Afra Amini, Tim Vieira, and Ryan Cotterell. Direct preference optimization with an offset. *arXiv preprint arXiv:2402.10571*, 2024.

Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, and Sham Kakade. Gone fishing: Neural active learning with fisher embeddings. *Advances in Neural Information Processing Systems*, 34:8927–8939, 2021.

Zhuoxi Bai, Ning Wu, Fengyu Cai, Xinyi Zhu, and Yun Xiong. Finetuning large language model for personalized ranking. *arXiv preprint arXiv:2405.16127*, 2024.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3, 2025.

Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2334029.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical science*, pp. 273–304, 1995.

Yuxin Chen, Junfei Tan, An Zhang, Zhengyi Yang, Leheng Sheng, Enzhi Zhang, Xiang Wang, and Tat-Seng Chua. On softmax direct preference optimization for recommendation. *arXiv preprint arXiv:2406.09215*, 2024.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 4302–4310, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

David Cohn. Neural network exploration using optimal experiment design. *Advances in neural information processing systems*, 6, 1993.

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

Lu Fan, Jiashu Pu, Rongsheng Zhang, and Xiao-Ming Wu. Neighborhood-based hard negative mining for sequential recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 2042–2046, 2023.

R. A. Fisher and Edward John Russell. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222(594-604):309–368, 1922. doi: 10.1098/rsta.1922.0009. URL https://royalsocietypublishing.org/doi/abs/10.1098/rsta.1922.0009.

Patrick Flaherty, Adam Arkin, and Michael Jordan. Robust design of biological experiments. *Advances in neural information processing systems*, 18, 2005.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *Acm transactions on interactive intelligent systems (tiis)*, 5(4):1–19, 2015.

Xiaoxin He, Nurendra Choudhary, Jieyi Jiang, Edward W Huang, Bryan Hooi, Xavier Bresson, and Karthik Subbian. Reclaif: Reinforcement learning from ai feedback for recommendation systems. 2025.

Yongsu Jung and Ikjin Lee. Optimal design of experiments for optimization-based model calibration using fisher information matrix. *Reliability Engineering & System Safety*, 216:107968, 2021.

Yannis Kalantidis, Mert Bulent Sariyildiz, Noe Pion, Philippe Weinzaepfel, and Diane Larlus. Hard negative mixing for contrastive learning. *Advances in neural information processing systems*, 33: 21798–21809, 2020.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8082–8090, 2020.

Jack Kiefer. Optimum experimental designs. *Journal of the Royal Statistical Society: Series B (Methodological)*, 21(2):272–304, 1959.

Andreas Kirsch and Yarin Gal. Unifying approaches in active learning and active sampling via fisher information and information-theoretic quantities. *arXiv preprint arXiv:2208.00549*, 2022.

Andreas Kirsch, Joost Van Amersfoort, and Yarin Gal. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Advances in neural information processing systems*, 32, 2019.

Andreas Krause and Carlos E Guestrin. Near-optimal nonmyopic value of information in graphical models. *arXiv preprint arXiv:1207.1394*, 2012.

Andreas Krause, Ajit Singh, and Carlos Guestrin. Near-optimal sensor placements in gaussian processes: Theory, efficient algorithms and empirical studies. *Journal of Machine Learning Research*, 9(2), 2008.

Branislav Kveton, Xintong Li, Julian McAuley, Ryan Rossi, Jingbo Shang, Junda Wu, and Tong Yu. Active learning for direct preference optimization. *arXiv preprint arXiv:2503.01076*, 2025.

Weibin Liao, Xu Chu, and Yasha Wang. Tpo: Aligning large language models with multi-branch & multi-step preference trees. *arXiv preprint arXiv:2410.12854*, 2024.

Pangpang Liu, Chengchun Shi, and Will Wei Sun. Dual active learning for reinforcement learning from human feedback. *arXiv preprint arXiv:2410.02504*, 2024.

R Duncan Luce et al. *Individual choice behavior*, volume 4. Wiley New York, 1959.

Haokai Ma, Ruobing Xie, Lei Meng, Fuli Feng, Xiaoyu Du, Xingwu Sun, Zhanhui Kang, and Xiangxu Meng. Negative sampling in recommendation: A survey and future directions. *arXiv preprint arXiv:2409.07237*, 2024.

Subhojyoti Mukherjee, Anusha Lalitha, Kousha Kalantari, Aniket Anand Deshmukh, Ge Liu, Yifei Ma, and Branislav Kveton. Optimal design for human preference elicitation. *Advances in Neural Information Processing Systems*, 37:90132–90159, 2024.

Tracianne B Neilsen, David F Van Komen, Mark K Transtrum, Makenzie B Allen, and David P Knobles. Optimal experimental design for machine learning using the fisher information. In *Proceedings of Meetings on Acoustics*, volume 35. AIP Publishing, 2018.

G. L. Nemhauser and L. A. Wolsey. Best algorithms for approximating the maximum of a submodular set function. *Mathematics of Operations Research*, 3(3):177–188, 1978. ISSN 0364765X, 15265471. URL http://www.jstor.org/stable/3689488.

George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical programming*, 14:265–294, 1978.

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

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In Gerardo Flores, George H Chen, Tom Pollard, Joyce C Ho, and Tristan Naumann (eds.), *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pp. 248–260. PMLR, 07–08 Apr 2022. URL https://proceedings.mlr.press/v174/pal22a.html.

Robin L Plackett. The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 24(2):193–202, 1975.

Friedrich Pukelsheim. *Optimal Design of Experiments*. Society for Industrial and Applied Mathematics, 2006a. doi: 10.1137/1.9780898719109. URL https://epubs.siam.org/doi/abs/10.1137/1.9780898719109.

Friedrich Pukelsheim. *Optimal design of experiments*. SIAM, 2006b.

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

Carlos Riquelme, George Tucker, and Jasper Snoek. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. *arXiv preprint arXiv:1802.09127*, 2018.

Joshua Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. *arXiv preprint arXiv:2010.04592*, 2020.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

Jack Sherman and Winifred J. Morrison. Adjustment of an inverse matrix corresponding to a change in one element of a given matrix. *The Annals of Mathematical Statistics*, 21(1):124–127, 1950. ISSN 00034851. URL http://www.jstor.org/stable/2236561.

Jamshid Sourati, Murat Akcakaya, Todd K Leen, Deniz Erdogmus, and Jennifer G Dy. Asymptotic analysis of objectives based on fisher information in active learning. *Journal of Machine Learning Research*, 18(34):1–41, 2017.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. Learning to summarize from human feedback. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020a. Curran Associates Inc. ISBN 9781713829546.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in neural information processing systems*, 33:3008–3021, 2020b.

Chao Sun, Yaobo Liang, Yaming Yang, Shilin Xu, Tianmeng Yang, and Yunhai Tong. Direct preference optimization for llm-enhanced recommendation systems. *arXiv preprint arXiv:2410.05939*, 2024.

Qwen Team. Qwen3, April 2025. URL https://qwenlm.github.io/blog/qwen3/.

Kiran Koshy Thekumparampil, Gaurush Hiranandani, Kousha Kalantari, Shoham Sabach, and Branislav Kveton. Comparing few to rank many: Active human preference learning using randomized frank-wolfe. *arXiv preprint arXiv:2412.19396*, 2024.

Franklin Wang and Sumanth Hegde. Accelerating direct preference optimization with prefix sharing. *arXiv preprint arXiv:2410.20305*, 2024.

William J. Welch. Branch-and-bound search for experimental designs based on d optimality and other criteria. *Technometrics*, 24(1):41–48, 1982. ISSN 00401706. URL http://www.jstor.org/stable/1267576.

Shuo Xie, Fangzhi Zhu, Jiahui Wang, Lulu Wen, Wei Dai, Xiaowei Chen, Junxiong Zhu, Kai Zhou, and Bo Zheng. Mppo: Multi pair-wise preference optimization for llms with arbitrary negative samples. *arXiv preprint arXiv:2412.15244*, 2024.

Zhi Yang, Jiwei Qin, Chuan Lin, Yanping Chen, Ruizhang Huang, and Yongbin Qin. Ganrec: A negative sampling model with generative adversarial network for recommendation. *Expert Syst. Appl.*, 214(C), March 2023. ISSN 0957-4174. doi: 10.1016/j.eswa.2022.119155. URL https://doi.org/10.1016/j.eswa.2022.119155.

Zhaoyang Zhang, Xuying Wang, Xiaoming Mei, Chao Tao, and Haifeng Li. False: False negative samples aware contrastive learning for semantic segmentation of high-resolution remote sensing image. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022.

A Appendix

Lemma A.1 (Gradient Derivation) *Consider the loss for a single sample*

$$L(\theta) = -\log \sigma\Big(Z(\theta)\Big), \quad \textit{with} \quad Z(\theta) = -\log\left(\sum_{j=1}^{n} \exp\Big[\beta\left(\phi_j^\top \theta - b_j\right)\Big]\right), \tag{17}$$

$$\frac{d}{dz}\big[-\log\sigma(Z(\theta))\big] = -\frac{1}{\sigma(Z(\theta))}\cdot\sigma'(Z(\theta)) = -\frac{\sigma(Z(\theta))(1-\sigma(Z(\theta)))}{\sigma(Z(\theta))} = -(1-\sigma(Z(\theta))).$$

$$\frac{\partial L}{\partial Z(\theta)} = -(1-\sigma(Z(\theta))). \tag{18}$$

$$A(\theta) = \sum_{j=1}^{n} \exp\Big[\beta\left(\phi_j^\top\theta - b_j\right)\Big],$$

so that $Z(\theta) = -\log A(\theta)$. *Then,*

$$\frac{\partial Z(\theta)}{\partial\theta} = -\frac{1}{A(\theta)}\frac{\partial A(\theta)}{\partial\theta},$$

$$\frac{\partial A(\theta)}{\partial\theta} = \sum_{j=1}^{n}\exp\Big[\beta\left(\phi_j^\top\theta - b_j\right)\Big]\beta\phi_j,$$

$$\frac{\partial Z(\theta)}{\partial\theta} = -\beta\sum_{j=1}^{n}\frac{\exp\Big[\beta\left(\phi_j^\top\theta - b_j\right)\Big]}{A(\theta)}\phi_j = -\beta\sum_{j=1}^{n}p_j\phi_j,$$

where the softmax weights are defined as

$$p_j = \frac{\exp\Big[\beta\left(\phi_j^\top\theta - b_j\right)\Big]}{A(\theta)}.$$

$$\frac{\partial L}{\partial\theta} = \frac{\partial L}{\partial Z(\theta)}\cdot\frac{\partial Z(\theta)}{\partial\theta} = -(1-\sigma(Z(\theta)))\cdot\Big[-\beta\sum_{j=1}^{n}p_j\phi_j\Big] = \beta(1-\sigma(Z(\theta)))\sum_{j=1}^{n}p_j\phi_j.$$

Thus, the gradient of the loss is

$$\nabla_\theta L = \beta(1-\sigma(Z(\theta)))\sum_{j=1}^{n}p_j\phi_j. \tag{19}$$

Lemma A.2 (Hessian Derivation) *Recall the multi-negative DPO loss:*

$$L(\theta; S_n) = -\log \sigma \left(-\log \sum_{i \in S_n} \exp(\beta(\phi_i^\top \theta + b_i)) \right),$$

where $\sigma(\cdot)$ denotes the sigmoid function, and we define the shorthand

$$Z_n = -\log \sum_{i \in S_n} \exp(\beta(\phi_i^\top \theta + b_i)), \quad p_j = \frac{\exp(\beta(\phi_j^\top \theta + b_j))}{\sum_{k \in S_n} \exp(\beta(\phi_k^\top \theta + b_k))}, \quad \phi = \sum_{j \in S_n} p_j \phi_j.$$

Starting from the gradient Equation (9)*,*

$$\nabla_\theta L(\theta; S_n) = \beta(1 - \sigma(Z_n)) \sum_{j \in S_n} p_j \phi_j,$$

we derive the Hessian by differentiating again with respect to θ:

$$\nabla_\theta^2 L(\theta; S_n) = \beta \nabla_\theta \left[(1 - \sigma(Z_n)) \sum_{j \in S_n} p_j \phi_j \right] \tag{20}$$

$$= \beta(1 - \sigma(Z_n)) \nabla_\theta \sum_{j \in S_n} p_j \phi_j + \beta \sigma(Z_n)(1 - \sigma(Z_n)) \sum_{j \in S_n} p_j \phi_j \nabla_\theta Z_n^\top. \tag{21}$$

Expanding the first term using the definition of p_j gives:

$$\nabla_\theta \sum_{j \in S_n} p_j \phi_j = \beta \sum_{j \in S_n} p_j \phi_j \phi_j^\top - \beta \left(\sum_{j \in S_n} p_j \phi_j \right) \left(\sum_{j \in S_n} p_j \phi_j \right)^\top \tag{22}$$

$$= \beta \sum_{j \in S_n} p_j (\phi_j - \phi)(\phi_j - \phi)^\top. \tag{23}$$

Note also that:

$$\nabla_\theta Z_n = \beta \sum_{j \in S_n} p_j \phi_j = \beta \phi.$$

Thus, substituting back, the Hessian becomes:

$$\nabla_\theta^2 L(\theta; S_n) = \beta^2 (1 - \sigma(Z_n)) \sum_{j \in S_n} p_j (\phi_j - \phi)(\phi_j - \phi)^\top + \beta^2 \sigma(Z_n)(1 - \sigma(Z_n)) \phi \phi^\top \tag{24}$$

$$= \beta^2 (1 - \sigma(Z_n)) \left[\sigma(Z_n)\, \phi \phi^\top + \sum_{j \in S_n} p_j (\phi_j - \phi)(\phi_j - \phi)^\top \right]. \tag{25}$$

This demonstrates how the Hessian measures curvature based on the variance of feature differences under the softmax weights p_j, capturing essential geometric insights into policy optimization.

Theorem A.1 (Single Design Estimation Error) *Following (Kveton et al., 2025), to bound $\left\| \hat{\theta}_n - \theta_* \right\|_{\Sigma_n}$, we also show that $\|\nabla L(\theta_*; \mathcal{S}_n)\|_{\Sigma_n^{-1}}$ is small with high probability. We recall the from Lemma A.1,*

$$\nabla_\theta L(\theta_*; S_n) = -\beta \left(1 - \sigma(Z_n(\theta_*))\right) \sum_{j \in S_n}^{n} p_j \phi_j,$$

where $Z_n(\theta_*) = \sum_{k \in S_n} \exp\left[\beta(\phi_k^\top \theta_* - b_k)\right]$. Since the covariant matrix is lower-bounded as

$$\Sigma_n \succeq V_n = c_{\min} \left(\frac{\gamma}{c_{\min}} I_d - (1 - \sigma(Z_n)) \phi \phi^\top + \sum_{j \in \mathcal{S}_n} p_j \phi_j \phi_j^\top \right).$$

Then we have

$$\|\nabla L(\theta_*;\mathcal{S}_n)\|_{\Sigma_n^{-1}} \le \frac{\beta(1-\sigma(Z_n))}{\sqrt{c_{\min}}}\left\|\sum_{j\in\mathcal{S}_n} p_j\phi_j\right\|_{V_n^{-1}}.$$

Following (Kveton et al., 2025), we have

$$\begin{aligned}\left\|\sum_{i\in\mathcal{S}_n} p_i\phi_i\right\|_{V_n^{-1}} &\le \sqrt{\frac{d}{4}\log\left(\frac{1+(1-c_{\min}(1-\sigma(Z_n))\|\phi\|^2/\gamma)^{-1/d}\cdot\sum_{j\in\mathcal{S}_n} p_j\cdot c_{\min}\|\phi_j\|^2/\gamma}{\delta}\right)}\\ &\le \sqrt{\frac{d}{4}\log\left(\frac{1+(1-c_{\min}\|\phi\|^2/\gamma)^{-1/d}\cdot c_{\min}/\gamma}{\delta}\right)}\end{aligned}$$

Then, with all equations, we have

$$\begin{aligned}\left\|\hat{\theta}_n-\theta_*\right\|_{\Sigma_n} &\le \|\nabla L(\theta_*;\mathcal{S}_n)\|_{\Sigma_n^{-1}} + 2\gamma^{\frac{1}{2}}\\ &\le \sqrt{\frac{\beta^2 d}{c_{\min}}\log\left(\frac{1+(1-c_{\min}\|\phi\|^2/\gamma)^{-1/d}\cdot c_{\min}/\gamma}{\delta}\right)} + 2\gamma^{\frac{1}{2}},\end{aligned} \tag{26}$$

holds with probability at least $1-\delta$.

Theorem A.2 (Batch Design Estimation Error) *Now we consider learning with all collected samples* $\mathcal{S}_{k,n}$*, where each* i*-th prompt is corresponded with* n *negative samples actively collected* $\mathcal{S}_{i,n}$*. Following Lemma A.1, to bound* $\left\|\hat{\theta}_{k,n}-\theta_*\right\|_{\Sigma_{k,n}}$*, we also show that* $\|\nabla L(\theta_*;\mathcal{S}_{k,n})\|_{\Sigma_{k,n}^{-1}}$ *is small with high probability. We recall the from Lemma A.1,*

$$\nabla_\theta L(\theta_*;S_{k,n}) = -\beta\sum_{i=1}^{k}(1-\sigma(Z_i(\theta_*)))\sum_{j\in S_{i,n}}^{n} p_{i,j}\phi_{i,j},$$

where $Z_i(\theta_*)=\sum_{j\in S_{i,n}}\exp\left[\beta(\phi_{i,j}^\top\theta_* - b_{i,j})\right]$*. Since the covariant matrix is lower-bounded as*

$$\Sigma_{k,n}\succeq V_{k,n} = c_{\min}\left(\frac{\gamma}{c_{\min}}I_d - \sum_{i=1}^{k}(1-\sigma(Z_{i,n}))\phi_i\phi_i^\top + \sum_{i=1}^{k}\sum_{j\in\mathcal{S}_{i,n}} p_{i,j}\phi_{i,j}\phi_{i,j}^\top\right).$$

Then we have

$$\|\nabla L(\theta_*;\mathcal{S}_{k,n})\|_{\Sigma_{k,n}^{-1}} \le \frac{\beta\sum_{i=1}^{k}(1-\sigma(Z_{i,n}))}{\sqrt{c_{\min}}}\left\|\sum_{i=1}^{k}\sum_{j\in\mathcal{S}_{i,n}} p_{i,j}\phi_{i,j}\phi_{i,j}^\top\right\|_{V_{k,n}^{-1}}.$$

Following (Kveton et al., 2025), we have

$$\begin{aligned}\left\|\sum_{i\in\mathcal{S}_{k,n}} p_i\phi_i\right\|_{V_{k,n}^{-1}} &\le \sqrt{\frac{d}{4}\log\left(1/\delta + \frac{\sum_{i=1}^{k}\sum_{j\in\mathcal{S}_{i,n}} p_{i,j}\cdot c_{\min}\|\phi_{i,j}\|^2/\gamma}{\left(1-c_{\min}\sum_{i=1}^{k}(1-\sigma(Z_{i,n}))\|\phi\|^2/\gamma\right)^{1/d}\cdot\delta}\right)}\\ &\le \sqrt{\frac{d}{4}\log\left(1/\delta + \frac{k\cdot c_{\min}/\gamma}{(1-c_{\min}\cdot k/\gamma)^{1/d}\cdot\delta}\right)}\end{aligned}$$

Then, with all equations, we have

$$\begin{aligned}\left\|\hat{\theta}_{k,n}-\theta_*\right\|_{\Sigma_{k,n}} &\le \|\nabla L(\theta_*;\mathcal{S}_{k,n})\|_{\Sigma_{k,n}^{-1}} + 2\gamma^{\frac{1}{2}}\\ &\le \sqrt{\frac{d}{4}\log\left(1/\delta + \frac{k\cdot c_{\min}/\gamma}{(1-c_{\min}\cdot k/\gamma)^{1/d}\cdot\delta}\right)} + 2\gamma^{\frac{1}{2}},\end{aligned} \tag{27}$$

holds with probability at least $1-\delta$.

Lemma A.3 (Optimality of the Incremental Greedy Algorithm) *Let $(i_1, \ldots, i_n)$ be the greedy indices and $H_k = \gamma I + \sum_{t=1}^{k} v_{i_t} v_{i_t}^\top$. Iterating equation 13 yields*

$$\det H_n = \det(\gamma I) \prod_{k=1}^{n} \big(1 + v_{i_k}^\top H_{k-1}^{-1} v_{i_k}\big). \tag{28}$$

Now consider any other subset $S = \{j_1, \ldots, j_n\}$ (arbitrary order) and define $\tilde{H}_k$ analogously. Because i_k maximises $v^\top H_{k-1}^{-1} v$ among the remaining candidates and $H_{k-1} \succeq \tilde{H}_{k-1}$, we have $v_{i_k}^\top H_{k-1}^{-1} v_{i_k} \geq v_{j_k}^\top \tilde{H}_{k-1}^{-1} v_{j_k}$ for every k. Applying equation 28 to both sequences and multiplying the n inequalities delivers $\det H_n \geq \det \tilde{H}_n$, and hence $\det H(S_n) \geq \det H(S)$ for all admissible S. Taking logarithms completes the argument.

Algorithm 1 Greedy D-Optimal Multi-negative Active Sample Selection

Input: context x, preferred response y^*, candidate set $\mathcal{D} = \{y_i\}_{i=1}^N$, policy parameter θ, scale β, number of negatives n
Compute feature differences and offsets, for each $i \in [N]$,
  $\phi_i \leftarrow \phi(x, y_i) - \phi(x, y^*), \quad b_i \leftarrow \log \pi_{\text{ref}}(y^* \mid x) - \log \pi_{\text{ref}}(y_i \mid x)$
Compute scores and softmax weights, for each $i \in [N]$,
  $s_i \leftarrow \beta\big(\phi_i^\top \theta + b_i\big), \quad p_i \leftarrow \exp(s_i) / \sum_{k=1}^{N} \exp(s_k)$
Center and weight features, for all $i \in [N]$
  $\phi \leftarrow \sum_{j=1}^{N} p_j \, \phi_j, \quad \tilde{\phi}_i \leftarrow \phi_i - \phi, \quad v_i \leftarrow \sqrt{p_i}\, \tilde{\phi}_i$
Initialize matrices and set: $H_0 \leftarrow \gamma \mathbf{I}_{d\times d}$, $S_0 \leftarrow \emptyset$
for $k = 1, \ldots, n$ **do**
  Incremental selection $I_k \leftarrow \arg\max_{i \in [N] \setminus S_{k-1}} \log\det\big(H_{k-1} + v_i v_i^\top\big)$
  Incremental update selection and design: $S_k \leftarrow S_{k-1} \cup \{I_k\}, \quad H_k \leftarrow H_{k-1} + v_{I_k} v_{I_k}^\top$
end for
Output: selected negatives set S_n

B Technical Assumptions

B.1 Assumption Details

Assumption B.1 (Bounded Feature and Bias) *For any (x, y) pair, the feature vectors and reference policy log-ratio are bounded:*

$$\|\phi(x, y)\|_2 \leq 1, \quad |b_i| \leq 1.$$

Additionally, we constrain the parameter space to a unit ball: $\|\theta\|_2 \leq 1$.

Assumption B.2 (Bounded Design Weights) *Let $p_i = \exp[\beta(\phi_i^\top \theta - b_i)] / \sum_j \exp[\beta(\phi_j^\top \theta - b_j)]$. Then there exist constants $0 < c_{\min} \leq c_{\max} \leq \frac{1}{4}\beta^2$ such that:*

$$c_{\min} \leq \beta^2 (1 - \sigma(Z(\theta))) p_i \leq c_{\max}, \quad \forall i.$$

B.2 Experimental Settings

To further manage computational costs, we cap the number of response candidates at 20 for the LastFM, MovieLens, and MedMCQA datasets, and at 8 for QASC. Although MedMCQA natively provides only four options per question, we expand this to 20 by pooling all candidates with the same `subject_name` field. We also subsample each dataset to 20k training samples, 200 samples for online evaluation, and 2,000 samples for testing. All prompts are formatted using each model's provided chat template to ensure consistent input structure across tasks.

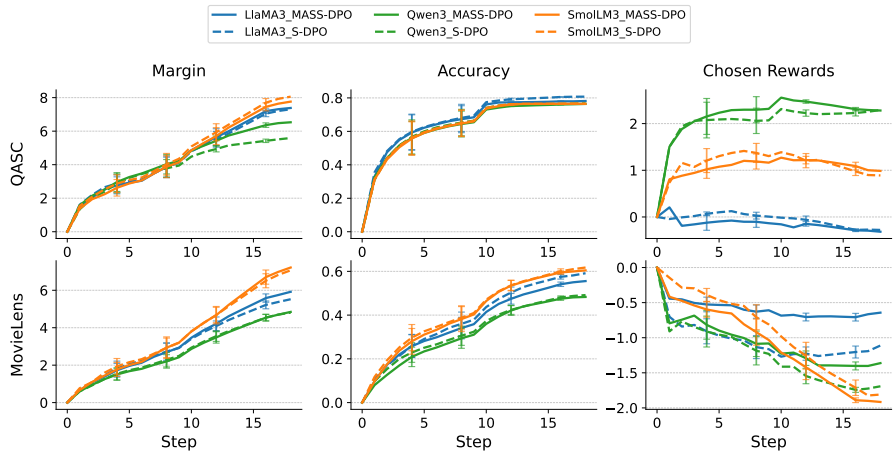

Figure 3: Comparison of MASS-DPO and SDPO on the MovieLens and QASC datasets. MASS-DPO consistently achieves higher margins, superior accuracy, and stable improvements in chosen rewards, highlighting the benefits of active negative sample selection. The x-axis (Step) counts the on-the-fly evaluations during training.

Model	Method	MedMCQA				QASC			
		R@1	R@3	N@1	N@3	R@1	R@3	N@1	N@3
Qwen3	DPO	39.25	84.51	39.25	65.37	67.77	90.62	67.77	81.29
	DMPO	26.02	74.89	26.02	53.60	68.21	90.51	68.21	81.40
	DPO-k	54.59	89.67	54.59	74.86	71.08	**92.72**	71.08	**83.95**
	S-DPO	51.03	86.52	51.03	71.54	70.42	91.50	70.42	83.04
	MASS-DPO	**56.34**	**89.72**	**56.34**	**75.62**	**71.85**	91.61	**71.85**	83.73
SmolLM3	DPO	33.33	81.75	33.33	61.01	67.11	90.07	67.11	80.70
	DMPO	26.02	75.39	26.02	53.93	66.11	88.74	66.11	79.54
	DPO-k	44.16	85.91	44.16	68.09	70.31	90.18	70.31	82.06
	S-DPO	**45.46**	87.22	**45.46**	**69.54**	70.53	91.39	70.53	82.80
	MASS-DPO	44.81	**87.47**	44.81	69.40	**72.52**	**91.83**	**72.52**	**83.82**
Llama3	DPO	51.48	87.82	51.48	72.30	71.30	91.72	71.30	83.41
	DMPO	24.36	74.99	24.36	52.84	69.32	91.94	69.32	82.67
	DPO-k	71.13	93.88	71.13	84.51	73.95	92.38	73.95	84.93
	S-DPO	**72.33**	94.34	**72.33**	**85.20**	**74.17**	92.38	**74.17**	**85.13**
	MASS-DPO	71.23	**94.49**	71.23	84.84	73.84	**92.60**	73.84	**85.13**

Table 5: Recall (R) and NDCG (N) at k={1,3} on MedMCQA and QASC.

B.3 Implementation Details

We implement our experiments using PyTorch, leveraging three widely used pre-trained LLMs: *LlaMA-3.2-3B-Instruct* (Grattafiori et al., 2024), *SmolLM3* (Bakouch et al., 2025), and *Qwen3-4B* (Team, 2025). Each model is fine-tuned on 8 NVIDIA A100 GPUs with a per-device batch size of 2, gradient accumulation steps of 8, learning rate of 10^{-5}, a cosine learning-rate scheduler with warmup ratio 0.05, and the Paged AdamW optimizer for 3 epochs with a fixed KL penalty coefficient at $\beta = 0.1$ across all experiments. More details included in We enable gradient checkpointing, gradient clipping is applied with a maximum norm of 0.3, and evaluation uses a batch size of 2. We extract representation vectors by mean-pooling the final hidden states, using either (a) all tokens from the concatenated prompt–response sequence or (b) only the response tokens, where prompt positions are masked out. Both strategies use the same pretrained LLM and tokenization pipeline.

B.4 Results