How Sampling Shapes LLM Alignment: From One-Shot Optima to Iterative Dynamics

Yurong Chen[1], Yu He[2], Michael I. Jordan[1,3], Fan Yao[4]
[1]Inria, École Normale Supérieure, PSL Research University
[2]Northwestern University
[3]University of California, Berkeley
[4]University of North Carolina at Chapel Hill
yurong.chen@inria.fr, yuhe2030@u.northwestern.edu,
jordan@cs.berkeley.edu, fanyao@unc.edu

Current version: August 2026

Abstract

Standard methods for aligning large language models with human preferences learn from pairwise comparisons among sampled candidate responses and regularize toward a reference policy. Despite their effectiveness, the effects of sampling and reference choices are poorly understood theoretically. We investigate these effects through Ψ-Preference Optimization, a widely used preference alignment framework including Direct Preference Optimization and Identity Preference Optimization as its special cases. We show that proper instance-dependent samplers can yield stronger ranking guarantees, while skewed on-policy samplers can induce excessive concentration under structured preferences. We then analyze iterative alignment dynamics in which the learned policy feeds back into future sampling and reference policies, reflecting common practices of using model-generated preference data and regularly refreshing the reference policy. We prove that these dynamics can exhibit oscillatory instability or entropy collapse for certain parameter choices, and characterize regimes that guarantee stability. Our results yield practical guiding principles for more robust iterative alignment: avoid fully closed-loop training, and jointly tune the strengths of on-policy sampling and reference model updates. Simulations and experiments on Qwen2.5-1.5B on real-world preference data validate our findings.

1 Introduction

Modern large language models (LLMs) need to be aligned with human preferences, social values, and normative standards to remain reliable [Askell et al., 2021, Bai et al., 2022, Guan et al., 2025]. A standard approach to alignment is Preference Optimization (PO), which learns from sampled pairwise preferences between candidate responses while regularizing the policy toward a reference model [Ouyang et al., 2022, Rafailov et al., 2023, Azar et al., 2024]. A growing line of work studies how to design samplers to improve PO efficiency, including downstream task performance [Liu et al., 2023, Dong et al., 2023, Kim et al., 2025, Pan et al., 2025] and convergence rates [Feng et al., 2025, Shi et al., 2024]. At the same time, empirical and theoretical studies have identified failure modes in PO, including robustness issues [Chowdhury et al., 2024], overoptimization [Zhu et al., 2024, Wolf et al., 2025], and diversity loss [Guo et al., 2025, Murthy et al., 2025]. These findings show that sampler choices, as central parameters that shape the objective and hence the solution, may improve certain performance metrics, but leave open a more basic question: how do they alter the learned policy itself, and can they help explain the failure modes observed in practice?

This question becomes even more urgent in iterative PO practice, where comparison samplers and reference policies are no longer fixed exogenous choices, but are periodically updated using deployed models [Liu et al., 2023, Dong et al., 2023, Gorbatovski et al., 2024, Hugging Face, 2026, Kim et al., 2025]. Such workflows create a self-reinforcing loop: the current model shapes the data distribution and reference policy used in the next update, and seemingly minor design choices can compound across rounds and thereby steer the future learned policy. Related concerns have been studied under the broader theme of model collapse, often in settings where new models are trained on fully model-generated data or under stylized distributional assumptions [Shumailov et al., 2024, Ferbach et al., 2024]. In contrast, practical PO involves more general feedback mechanisms: a mixture of on- and off-policy sampling, a partly refreshed reference, and the PO rules jointly determine the next policy. How these design choices affect policy collapse and stability remains poorly understood.

In this work, we take a first step toward theoretically understanding these effects in LLM alignment within the algorithmic framework of ΨPO [Azar et al., 2024], a broad class of PO objectives representative of modern alignment methods. Our contributions can be summarized as follows:

1. We characterize the critical role of sampling in shaping the *one-shot ΨPO solution* (Section 3): any fixed sampler can fail to achieve natural ranking desiderata, while instance-dependent samplers can recover them. We show skewed samplers amplify policy concentration under natural preference structures.

2. We study *iterative PO dynamics* where the current policy affects both future sampling and reference policies (Section 4). The dynamics are governed by the strengths of on-policy sampling, reference refresh, and regularization. We prove that aggressive closed-loop updates can lead to oscillatory instability under cyclic preferences and to entropy collapse under transitive preferences.

3. We identify regimes that *guarantee convergence* and *mitigate collapse* of iterative PO, yielding *practical guidance* for experimental design: avoid fully closed-loop training, preserve off-policy coverage, and tune reference refresh jointly with on-policy sampling and update strength. Simulations and experiments on real-world preference data support the theoretical predictions (Section 5).

1.1 Related work

Preference Optimization (PO). Direct PO (DPO) [Rafailov et al., 2023] is among the most widely used methods for LLM alignment. It fits policies directly from comparisons under the Bradley–Terry (BT) model [Bradley and Terry, 1952], providing an efficient implementation of reinforcement learning with human feedback (RLHF) Ouyang et al. [2022]. Azar et al. [2024] introduced the ΨPO framework unifying a family of transformed preference objectives, including Identity PO (IPO) and DPO. We study sampling and reference effects within the general ΨPO framework.

Sampling effects in PO. Some empirical works study the design of preference samplers to improve PO outcomes. For example, Liu et al. [2023], Dong et al. [2023] show that rejection-style sampling can improve performance by concentrating comparisons on more informative candidates. Kim et al. [2025], Pan et al. [2025] emphasize improving coverage or quality of sampled responses to enhance performance. Feng et al. [2025], Shi et al. [2024] design samplers to accelerate the training convergence rate. Although these results are informative, they stop short of providing a systematic understanding of how sampling *shapes the optimizer itself* and how this dependence propagates under iterative deployment.

Online and iterative PO. Recent work studies online preference learning, where each alignment round uses newly sampled data from the current or previous policy, and shows that this approach can outperform purely offline training [Dong et al., 2024, Tajwar et al., 2024, Ye et al., 2024, Guo et al., 2024, Pan et al., 2025]. Other work develops exploration mechanisms for online PO [Xiong et al., 2023, Calandriello et al., 2024]. Beyond sampling, several methods update the reference policy during training: TR-DPO refreshes the reference model [Gorbatovski et al., 2024], Xu et al. [2024] use moving-average references, and Kim et al. [2025] study online DPO with previous policies as references. Similar reference-sync mechanisms also appear in widely used toolkits [Hugging Face, 2026]. We complement these empirical and algorithmic studies with a theoretical analysis of the closed-loop dynamics induced by iterative sampling and reference refresh, establishing conditions for convergence, oscillatory instability, and collapse.

Social choice properties of alignment methods. A growing line of research connects LLM alignment to social choice, asking whether methods such as reinforcement learning from human feedback (RLHF) [Ge et al., 2024], DPO, and Nash learning from human feedback (NLHF) [Liu et al., 2025] satisfy basic axioms when viewed as aggregating pairwise preferences. However, these analyses neglect the effect of sampling, typically assuming an arbitrary given dataset [Noothigattu et al., 2020], or (implicitly) a fixed uniform sampling distribution [Xiao et al., 2025]. Besides, little is known about IPO. We study how sampling affects the social choice behavior of the IPO solution.

Model collapse in iterative training. Our work is also related to the growing literature on model collapse and instability in recursively trained generative models. A large body of recent theoretical work studies collapse under iterative training on model-generated data, often through stylized distribution-matching settings such as Gaussian, linear, or other simplified generative models [Shumailov et al., 2024, Alemohammad et al., 2023, Bertrand et al., 2023, Gerstgrasser et al., 2024, Yi et al., 2025]. These works identify fundamental risks of recursive data generation, but their mechanisms are relatively far from modern preference-optimization pipelines. In contrast, we study collapse and instability in iterative preference alignment, where the feedback loop arises through pairwise sampling, KL regularization, and reference-policy updates.

Recent empirical studies have observed related failure modes in LLM alignment. For example, Lanchantin et al. [2025] report diversity collapse in DPO, especially in online settings, while Wolf et al. [2025] show that iterated RLHF can be highly sensitive to initialization and feedback transfer. Our contribution is to turn these empirical observations into a predictive theoretical object: we characterize when iterative preference optimization converges,when it oscillates under cyclic preferences, and when it collapses under strongly transitive preferences.

A further distinction is the role of the reference policy. Recent methods such as TR-DPO explicitly update or learn the reference model and show that doing so can materially affect alignment performance [Gorbatovski et al., 2024]. Our framework captures this practice through a self-referential parameter and analyzes how reference refresh interacts with on-policy sampling. This interaction is absent from prior model-collapse analyses, but is central to modern alignment workflows: the current model can shape both the future data distribution and the future regularization target. Thus, our results provide a theoretical explanation of how sampling, reference choice, and iterative deployment jointly produce collapse, cycling, or stability in preference-based alignment.

2 Preliminaries

We formalize the LLM alignment model. Following Azar et al. [2024], we fix a prompt s and a finite, possibly large, set of candidate responses $[K] := \{1, \dots, K\}$.[1] We work with the following objects:

- An underlying *preference matrix* $P^* \in [0,1]^{K \times K}$, where P^*_{ij} is the probability that response i is preferred over response j. Thus, $P^*_{ij} + P^*_{ji} = 1$ for all i, j. We use P^*_i to denote the i-th row of P^*. Denote the set of all valid preference matrices by $\mathcal{P}$. Note that we do not assume P^*_{ij} conforms to the BT model, i.e., $P^*_{ij} = \sigma(r_i - r_j)$ for some latent rewards r_i, where σ is the sigmoid function.
- A *sampling distribution* $\boldsymbol{\mu} \in \Delta_K^\circ$ over responses.[2] We assume that each comparison pair (i, j) is generated by sampling $i \sim \boldsymbol{\mu}$ and $j \sim \boldsymbol{\mu}$ independently, so the probability of observing (i, j) is $\mu_i \mu_j$. This models the common regime where preference data are collected from responses generated by some behavior policy, such as a uniform sampler or a deployed LLM.
- A *learned policy* $\boldsymbol{\pi} = \mathrm{softmax}(\boldsymbol{\theta}) \in \Delta_K^\circ$ with logits $\boldsymbol{\theta} \in \mathbb{R}^K$, regularized toward a *reference policy* $\boldsymbol{\pi}_{\mathrm{ref}} \in \Delta_K^\circ$ to favor solutions close to $\boldsymbol{\pi}_{\mathrm{ref}}$.

Our goal is to understand how P^*, $\boldsymbol{\mu}$, and $\boldsymbol{\pi}_{\mathrm{ref}}$ jointly shape $\boldsymbol{\pi}$, and how this dependence evolves when the learned policy is redeployed and used to generate future preference data or reference models.

2.1 Ψ-Preference Optimization (ΨPO) as a family of KL-regularized PO objectives

ΨPO [Azar et al., 2024] is a family of PO objectives whose population form is given by

$$\max_{\boldsymbol{\pi} \in \Delta_K} \mathbb{E}_{i \sim \boldsymbol{\pi},\, j \sim \boldsymbol{\mu}} \big[\Psi(P^*_{ij})\big] - \frac{1}{\beta} \mathrm{KL}(\boldsymbol{\pi} \,\|\, \boldsymbol{\pi}_{\mathrm{ref}}), \tag{1}$$

where $\Psi : (0,1) \to \mathbb{R}$ is a continuous, strictly increasing link function, and the inverse-temperature parameter $\beta > 0$ controls the strength of the KL regularization. Preference probabilities equal to 0 or 1 are allowed only when Ψ has finite continuous extensions at these endpoints; throughout, every entry of $P = \Psi(P^*)$ is assumed finite.[3] Equivalently, by defining the transformed preference matrix $P := \Psi(P^*)$, the objective can be written as $\max_{\boldsymbol{\pi} \in \Delta_K} \mathbb{E}_{i \sim \boldsymbol{\pi},\, j \sim \boldsymbol{\mu}}[P_{ij}] - \frac{1}{\beta} \mathrm{KL}(\boldsymbol{\pi} \,\|\, \boldsymbol{\pi}_{\mathrm{ref}})$. Treating the learner as a row player with payoff P, (1) can be viewed as maximizing expected utility against the column strategy $\boldsymbol{\mu}$, while regularizing toward $\boldsymbol{\pi}_{\mathrm{ref}}$. This family includes widely used PO methods, encompassing DPO [Rafailov et al., 2023] under the BT model with logit link $\Psi(q) = \log[q/(1-q)]$,[4] and IPO with the identity link $\Psi(q) = q$, which was proposed to mitigate over-optimization issues in DPO and RLHF.

For some parts of our theory, we impose the following symmetry condition on the link function.

Condition 2.1 (Symmetric ΨPO)**.** *For every $q \in (0,1)$, it holds that $\Psi(q) + \Psi(1-q) = 2\Psi(1/2)$.*

[1] With a finite vocabulary and a bounded output length, an LLM's response space is finite, though enormous. Our prompt-conditional formulation isolates policy-level effects, abstracting away from neural parameterization and optimization.

[2] For an integer K, denote the K-dimensional probability simplex by $\Delta_K := \{\boldsymbol{\mu} \in \mathbb{R}^K_{\geq 0} : \sum_{k=1}^K \mu_k = 1\}$. Given a set S, S° denotes the interior of S. Unless stated otherwise, we consider $\mu_i > 0$ for all i to avoid degenerate identifiability issues.

[3] In practice, P^*_{ij} is unknown, and one observes noisy binary outcomes indicating the preferred response with expectation P^*_{ij}. We analyze the population objective with known P^* to reveal the *structural* dependence of the learned policy on $\boldsymbol{\mu}$ and $\boldsymbol{\pi}_{\mathrm{ref}}$, making the effects analytically transparent.

[4] For generic DPO that does not assume the BT model, we analyze sampling dependence and iterative dynamics in Appendix E.

Under Condition 2.1, the transformed matrix P preserves a constant-sum structure $P_{ij} + P_{ji} = 2\Psi(1/2)$. This property is useful in several of our theoretical results. Both IPO and DPO belong to this symmetric ΨPO family.

A known key property of the ΨPO objective (1) is that it admits a closed-form optimizer as a function of the transformed preference matrix P, the sampling strategy $\boldsymbol{\mu}$, and the reference policy $\boldsymbol{\pi}_{\text{ref}}$. This closed form, which we present in the following Proposition, is the cornerstone of our subsequent analysis. We provide its proof in Appendix A for completeness.

Proposition 2.2. *Given an underlying preference matrix $P^* \in \mathcal{P}$ and its transformed matrix $P = \Psi(P^*)$, the ΨPO problem with sampling strategy $\boldsymbol{\mu} \in \Delta_K^\circ$ and reference policy $\boldsymbol{\pi}_{\text{ref}} \in \Delta_K^\circ$ admits a unique solution $\boldsymbol{\pi}^\star \in \Delta_K^\circ$. Moreover, this solution has the closed form*

$$\boldsymbol{\pi}^\star \propto \boldsymbol{\pi}_{\text{ref}} \odot \exp(\beta P \boldsymbol{\mu}), \tag{2}$$

where $\odot$ denotes element-wise multiplication.

The explicit mapping $(P, \boldsymbol{\mu}, \boldsymbol{\pi}_{\text{ref}}) \mapsto \boldsymbol{\pi}^\star$ in (2) makes ΨPO well suited to our goals: it separates the roles of the transformed preference matrix, the sampling strategy, and the reference policy. This separation enables sharp comparisons across sampling schemes and a principled analysis of self-reinforcing deployment dynamics.

3 Sampling Shapes ΨPO Optima

In this section, we consider the one-shot, offline setting and study the properties of the ΨPO optimizer. Since $\boldsymbol{\pi}_{\text{ref}}$ enters the solution only via a multiplicative shift, we focus on the sampling effect and fix a uniform reference, $\boldsymbol{\pi}_{\text{ref}} = \frac{1}{K}\mathbf{1}$, where $\mathbf{1} \in \mathbb{R}^K$ is the all-one vector. Our results extend immediately to general reference policies by replacing π_i with its increments over the reference policy, $\pi_i/\pi_{\text{ref},i}$.

3.1 (Im)possibility of axiom satisfiability under sampling

We begin by examining whether ΨPO satisfies certain natural desiderata, treating it as a preference aggregation method that maps preference matrix to a distribution over responses. For a sampling distribution $\boldsymbol{\mu} \in \Delta_K^\circ$, (2) defines a rule $F_{\boldsymbol{\mu}}(P^*) := \boldsymbol{\pi}^\star(P^*; \boldsymbol{\mu})$ induced by ΨPO. We say that *ΨPO under sampling $\boldsymbol{\mu}$ satisfies an axiom* if $F_{\boldsymbol{\mu}}$ satisfies it.

Since ΨPO's KL term induces full-support solutions, we state axioms for full-support distributions. The first and most natural set formalizes the idea that "better responses should receive higher probabilities in a good policy." We first recall classic structural notions from social choice theory.

Definition 3.1. *Let $P^* \in \mathcal{P}$ be a preference matrix. We make the following definitions:*

- *P^* admits a* Condorcet winner *if there exists $c \in [K]$ such that $P^*_{cj} > \frac{1}{2}$ for all $j \neq c$.*
- *A nonempty set $S \subseteq [K]$ is* dominant *if $P^*_{ij} > \frac{1}{2}$ for all $i \in S$ and $j \notin S$. The* Smith set *of P^*, $\text{Smith}(P^*)$, is the unique inclusion-minimal dominant set.*
- *P^* is* transitive *if there exists an ordering $a_1 \succ a_2 \succ \cdots \succ a_K$ such that $P^*_{a_i a_j} \geq \frac{1}{2}$ for all $i < j$. When all inequalities hold strictly, P^* is* strictly transitive.

When P^* is transitive, we relabel the responses so that $1 \succ 2 \succ \cdots \succ K$. Now we define the corresponding social choice axioms.

Definition 3.2. *Let $F : \mathcal{P} \to \Delta_K$ be an aggregation rule. A rule F is*

- Condorcet-top *if for every P^* that admits a Condorcet winner c, F ranks c strictly above all others, i.e., $F(P^*)_c > F(P^*)_j$ for all $j \neq c$.*

- Smith-top *if every element in* $\mathrm{Smith}(P^*)$ *receives strictly larger probability than all elements outside* $\mathrm{Smith}(P^*)$*, i.e.,* $F(P^*)_i > F(P^*)_j$ *for all* $i \in \mathrm{Smith}(P^*), j \notin \mathrm{Smith}(P^*)$.
- Order-preserving *if for every (strictly) transitive* P^* *with order* $1 \succ 2 \succ \cdots \succ K$*, the output distribution preserves the order (strictly), i.e.,* $F(P^*)_1 \geq (>) \cdots \geq (>) F(P^*)_K$.

Our first proposition establishes that fixed sampling can be incompatible with these axioms.

Proposition 3.3. *Let* $K \geq 3$*. For any* $\boldsymbol{\mu}$*,* ΨPO *under* $\boldsymbol{\mu}$ *is not* Condorcet-top, Smith-top, *or* Order-preserving.

In contrast, the next theorem shows that allowing $\boldsymbol{\mu}$ to depend on the preference instance provides additional leverage and can restore the axioms.

Theorem 3.4. *For any* P^* *with Condorcet winner* c*, there exists a* $\boldsymbol{\mu} \in \Delta_K^\circ$ *such that* $F_{\boldsymbol{\mu}}(P^*)$ *assigns* c *the unique highest probability. Moreover, if Condition 2.1 holds, then for any* P^**, there exists a* $\boldsymbol{\mu} \in \Delta_K^\circ$ *such that* $F_{\boldsymbol{\mu}}(P^*)_i > F_{\boldsymbol{\mu}}(P^*)_j$ *for any* $i \in \mathrm{Smith}(P^*)$ *and* $j \notin \mathrm{Smith}(P^*)$.

Remark 3.5. *We note that these are structural results: they identify the scope and limits of sampler design, establishing sampling as part of the preference-aggregation rule and a direct lever for shaping the learned policy while keeping the alignment method fixed. The proof of Theorem 3.4 is constructive: given* P^**, a sampler* $\boldsymbol{\mu}$ *satisfying the axioms can be computed via linear programs. Moreover, such a sampler is practically plausible, as it upweights responses preferred under* P*, mirroring the behavior of human users and well-trained LLMs. Altogether, this perspective motivates designing samplers to achieve desired aggregation properties.*

When P *is unknown, the same construction can be applied to an empirical preference* $\widehat{P}$ *estimated from pairwise comparisons. In Section B.4, we show that if* $\widehat{P}$ *is uniformly close to* P*, the resulting sampler still yields the desirable properties with high probability. Thus, instance-dependent sampling is not merely existential, but statistically robust and learnable from finite comparison data. These guarantees provide a starting point for sampler design; developing scalable, adaptive sampling methods remains an important direction for future work.*

Remark 3.6. *Liu et al. [2025] study non-full-support analogues of Condorcet-top and Smith-top, and show that unregularized Nash learning from human feedback (NLHF) satisfies them. Their guarantee has a different game-theoretic flavor: NLHF can be viewed as choosing a policy that maximizes worst-case expected utility over sampling strategies, which ensures that the learned support lies in the Smith set. By contrast, our instance-dependent sampling result gives full-support separation: every Smith alternative receives a strictly larger probability than every non-Smith alternative. Thus, our result preserves all Smith-set alternatives while separating them from non-Smith ones, rather than merely excluding the latter from the support.*

However, order preservation is subtler: changing $\boldsymbol{\mu}$ can drastically alter the ranking, and some transitive instances cannot be made order-preserving by any full-support sampler. We state this result in Section B.3. Nevertheless, the following subclass of transitive preferences is robust to sampling.

Definition 3.7. *A transitive matrix* P^* *with order* $1 \succ 2 \succ \cdots \succ K$ *is* strongly transitive *(ST) if for all* $i < j < k$*,* $P^*_{ik} \geq \max\{P^*_{ij}, P^*_{jk}\}$.

ST captures a monotone strengthening of preferences along an order and is a strict generalization of the BT and Thurstone models [Thurstone, 2017], which are widely used in the alignment literature [Ouyang et al., 2022]. For any $\boldsymbol{\mu}$, ΨPO under $\boldsymbol{\mu}$ is order-preserving over ST matrices. See Section B.5 for details.

3.2 Skewed sampling amplifies skewness

We now study how sampling shapes the *concentration* of the ΨPO solution, quantified by *pairwise logit gaps*. Our main finding is that, under ST and a natural structural condition on P^*, a sampler more aligned with the ranking of P^* necessarily induces a policy skewed toward dominant responses. Notably, this result does not rely on the symmetry of Ψ, but only requires the basic monotonicity condition. Thus, the amplification effect reflects a structural limitation of PO objectives rather than an artifact of a particular link choice. To formally present our result, we first introduce the definition of majorization used in this paper.

Definition 3.8. *A vector* $\boldsymbol{a} = (a_1, \ldots, a_K) \in \mathbb{R}^K$ *is* majorized *(Maj) if* $\frac{1}{m}\sum_{k=1}^{m} a_k \geq \frac{1}{K}\sum_{k=1}^{K} a_k$, $\forall m \in [K-1]$. *We say that* $\boldsymbol{a}$ *is* strictly Maj *if all the inequalities above are strict. Further, we write* $\boldsymbol{a} \succeq \boldsymbol{b}$ *if* $\boldsymbol{a} - \boldsymbol{b}$ *is Maj, and* $\boldsymbol{a} \succ \boldsymbol{b}$ *if* $\boldsymbol{a} - \boldsymbol{b}$ *is strictly Maj.*

Saying that a vector $\boldsymbol{a}$ is Maj means that every prefix average of $\boldsymbol{a}$ is at least as large as the global average. In our setting, the relevant vector is $P_i - P_j$. Intuitively, $P_i \succ P_j$ indicates that the advantage of response i over response j is concentrated toward higher-ranked columns.

Theorem 3.9. *Let* $P^* \in \mathcal{P}$ *be ST with order* $1 \succ \cdots \succ K$, *and let* $P = \Psi(P^*)$. *Let* $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \Delta_K^\circ$ *with* $\boldsymbol{\mu} \neq \boldsymbol{\mu}'$ *and* $\boldsymbol{\mu} - \boldsymbol{\mu}'$ *non-increasing. Define* $\boldsymbol{\theta}(\boldsymbol{\mu}) := \beta^{-1}\big(\log \boldsymbol{\pi}^\star(P; \boldsymbol{\mu}) - \log \boldsymbol{\pi}_{\mathrm{ref}}\big)$, *where* $\boldsymbol{\pi}^\star(P; \boldsymbol{\mu})$ *is the population ΨPO optimizer in (2). Then, for any* $i < j$ *such that* $P_i \succeq P_j$,

$$\theta_i(\boldsymbol{\mu}) - \theta_j(\boldsymbol{\mu}) \geq \theta_i(\boldsymbol{\mu}') - \theta_j(\boldsymbol{\mu}') \geq 0. \tag{3}$$

The second inequality is strict if and only if $P_i \neq P_j$. *The first inequality is strict whenever* $P_i \succ P_j$.

Under ST preferences, every sampler preserves the response order, yet Theorem 3.9 isolates how strongly that order is expressed in the learned probabilities, a distinction that ranking axioms alone leave unresolved. It shows that sampling can amplify concentration even in a one-shot setting: if $P_i \succ P_j$, then shifting sampling mass toward higher-ranked responses increases the learned logit gap between i and j. A concrete comparison takes $\boldsymbol{\mu}' = K^{-1}\mathbf{1}$ and a nonuniform full-support sampler with $\mu_1 \geq \cdots \geq \mu_K$: then $\boldsymbol{\mu} - \boldsymbol{\mu}'$ is automatically non-increasing, so the theorem applies to every pair satisfying the stated Maj condition.

A simple case where the required majorization condition holds is the Head–Tail Separation (HTS) preference structure, in which the response set contains a small head set $\mathcal{H}$ well separated from a long dominated tail $\mathcal{T}$, and the head scores have bounded span. For IPO with BT preferences $P_{ij} = \sigma(r_i - r_j)$, we can show that $P_i - P_j$ is strictly Maj for every $i < j$ in $\mathcal{H}$, so such a sampling shift toward higher-ranked responses strictly sharpens the learned gaps among head responses. Formal HTS conditions and the proof of our claim are given in Appendix B.8.

4 Long-Term Effects of Iterative ΨPO

We now study iterative ΨPO pipelines, where training and deployment are interleaved: the current policy generates candidate responses, preference data are collected, the policy is updated, and the updated policy is redeployed. In practice, the reference model may also be refreshed over time to track newer checkpoints. Thus, both the sampling distribution and the reference policy can evolve endogenously, creating a feedback loop across rounds. We model this loop as a discrete-time dynamical system over policies. At round t, the incumbent policy is $\boldsymbol{\pi}_t \in \Delta_K^\circ$. The reference policy is a geometric mixture of the current policy and a fixed base reference:

$$\boldsymbol{\pi}_{\mathrm{ref}}^{(t)} \;\propto\; \boldsymbol{\pi}_t^{\alpha}\boldsymbol{\pi}_{\mathrm{ref}}^{1-\alpha}, \tag{4}$$

where $\alpha \in [0,1]$ controls the degree of reference refresh. The sampler is an affine mixture of the current policy and a fixed base sampler:

$$\boldsymbol{\mu}_t \coloneqq \lambda \boldsymbol{\pi}_t + (1-\lambda)\boldsymbol{\pi}_0, \tag{5}$$

where $\lambda \in [0,1]$ controls how on-policy the preference data are. This captures common practices such as dynamic reference updates [Gorbatovski et al., 2024] and mixing on-policy with replay or offline data [Calandriello et al., 2024].

Assuming that the policy is fully optimized at each round and applying Proposition 2.2 with transformed payoff matrix $P = \Psi(P^*)$ gives the following closed-form iterative update.

Definition 4.1. *The* mixed reference/sampling ΨPO (MRS-ΨPO) dynamics*, initialized at* $\boldsymbol{\pi}_1 \in \Delta_K^\circ$*, are defined by*

$$\boldsymbol{\pi}_{t+1} = \mathrm{MRS}(\boldsymbol{\pi}_t; P, \alpha, \beta, \lambda, \boldsymbol{\pi}_{\mathrm{ref}}, \boldsymbol{\pi}_0) \coloneqq \mathrm{softmax}\Big(\log \boldsymbol{\pi}_{\mathrm{ref}}^{(t)} + \beta P \boldsymbol{\mu}_t\Big), \qquad t \geq 1, \tag{6}$$

where $\boldsymbol{\pi}_{\mathrm{ref}}^{(t)}$ *and* $\boldsymbol{\mu}_t$ *are defined in* (4) *and* (5)*.*

The parameters α, λ, and β are the key feedback knobs: $\alpha = 0$ keeps the reference fixed, while $\alpha = 1$ makes it fully self-referential; $\lambda = 1$ uses fully on-policy sampling, while $\lambda < 1$ preserves off-policy coverage; and β controls the strength of each ΨPO update.

The sampler and KL reference generally change with the current policy, so successive rounds need not optimize the same objective. Consequently, exact optimization at every round does not by itself guarantee the long-term behaviors of the pipelines. We identify sharply different long-run behaviors of these natural dynamics. Under cyclic preferences, they can become locally unstable and oscillatory (Proposition 4.2); under strongly transitive preferences, they can collapse toward extreme policies (Theorem 4.4). On the positive side, we identify explicit parameter regimes that guarantee convergence (Theorem 4.3). Together, these results show that iterative PO should be viewed as a closed-loop system whose stability depends on how reference refresh, on-policy sampling, and update strength are tuned, inspiring guidance for practical experimental design.

4.1 Condorcet cycles can induce policy oscillation

We begin with a Condorcet cycle (e.g., rock–paper–scissors), a preference structure that violates transitivity. This pattern arises when pairwise comparisons cannot be reconciled into a global ranking, which has been observed in various experimental studies [Tversky, 1969] and is common in practice whenever preference data arise from heterogeneous annotators.

The following result shows that non-transitive preferences can make iterative MRS-ΨPO unstable. We give a three-response construction and identify an explicit instability regime.

Proposition 4.2. *Let* $K = 3$ *and fix* $a \in (0, 1/2)$*. Consider the following preference matrix* $P^* = (\frac{1}{2}, \frac{1}{2}+a, \frac{1}{2}-a; \frac{1}{2}-a, \frac{1}{2}, \frac{1}{2}+a; \frac{1}{2}+a, \frac{1}{2}-a, \frac{1}{2})$ *so that* $1 \succ 2$*,* $2 \succ 3$*, and* $3 \succ 1$ *with probability* $1/2 + a$*. Let* $P = \Psi(P^*)$ *with symmetric* Ψ *in Condition 2.1, define* $a_\Psi \coloneqq (\Psi(1/2+a) - \Psi(1/2-a))/2$*, and set* $\boldsymbol{\pi}_{\mathrm{ref}} = \boldsymbol{\pi}_0 = (1/3, 1/3, 1/3)$*. Then the uniform policy* $\boldsymbol{\pi}^\star = (1/3, 1/3, 1/3)$ *is a fixed point of the MRS-ΨPO dynamics* (6)*. Moreover, if*

$$\alpha^2 + \frac{\beta^2 \lambda^2 a_\Psi^2}{3} > 1, \tag{7}$$

then $\boldsymbol{\pi}^\star$ *is linearly unstable. In particular, nearby trajectories have an unstable oscillatory mode and do not converge locally to* $\boldsymbol{\pi}^\star$*.*

Condition (7) shows how the feedback knobs affect stability. Larger reference refresh α, stronger on-policy sampling λ, larger update strength β, or a larger transformed cycle margin a_Ψ all make oscillation more likely. In particular, when $\alpha = 1$, any nonzero cyclic signal with $\beta\lambda a_\Psi > 0$ destabilizes the uniform point. Thus, under non-transitive, cyclic preferences, a fully self-referential iterative pipeline can chase a rotating preference signal rather than settle. Reducing α or $\beta\lambda$ dampens this feedback. The proof is in Appendix C.

Our next result shows that policy oscillation is not inevitable and gives a global contraction regime for *arbitrary* transformed payoff matrices P, including transitive and non-transitive preferences.

Theorem 4.3. *For any $\boldsymbol{\pi}_{\text{ref}}, \boldsymbol{\pi}_0 \in \Delta_K^\circ$, consider the MRS-ΨPO dynamics $\boldsymbol{\pi}_{t+1} = \text{MRS}(\boldsymbol{\pi}_t; \alpha, \beta, \lambda)$ defined in (6). Let $\Pi := I - \frac{1}{K}\mathbf{1}\mathbf{1}^\top$, $\widetilde{P} := \Pi P$. If*

$$\alpha + \frac{\beta\lambda}{2}\|\widetilde{P}\|_2 < 1, \tag{8}$$

then there exists a unique fixed point $\boldsymbol{\pi}^\star \in \Delta_K^\circ$, and for any initialization $\boldsymbol{\pi}_1 \in \Delta_K^\circ$, the iterates $\{\boldsymbol{\pi}_t\}$ converge to $\boldsymbol{\pi}^\star$.

Theorem 4.3 gives a simple stability rule. The matrix $\widetilde{P} = \Pi P$ is the part of the transformed payoff matrix that affects relative logits; additive constants are removed because they do not change the softmax. Stability is guaranteed when the total feedback gain is small: the reference is not refreshed too aggressively, the sampler is not too on-policy, and the PO update is not too strong.

The norm term also shows that stability depends on the effective comparison structure rather than only on the ambient number of responses. For example, when Ψ is symmetric and $\widetilde{P}$ has at most d nonzero entries in each row and column with bounded entries, $\|\widetilde{P}\|_2 = O(d)$. Thus, any valid upper bound on $\|\widetilde{P}\|_2$ can be substituted into (8) to obtain a sufficient convergence guarantee and guide the choice of α, λ, and β. Parameter settings that do not satisfy this sufficient condition may still converge. The proof is in Appendix C.

4.2 Strong transitivity can lead to entropy collapse

We next show that iterative alignment can amplify the one-shot concentration effect from Section 3.2. Under ST preferences, a preference-aligned sampler can arise endogenously from the deployed policy, and MRS-ΨPO may therefore become increasingly concentrated over time.

Throughout this subsection, we assume that P^* is *strictly strongly transitive* (SST) with order $1 \succ \cdots \succ K$: it is ST and, for every $i \in [K-1]$, there exists some $k \in [K]$ such that $P^*_{ik} > P^*_{i+1,k}$.[5] Since Ψ is strictly increasing, the transformed matrix $P = \Psi(P^*)$ is also SST. We say that $\boldsymbol{\pi}$ is *ε-collapsed* if $\pi_1 \geq 1 - \varepsilon$. The next Theorem 4.4 gives an eventual concentration bound for $\alpha < 1$ and proves entropy collapse for $\alpha = 1$:

Theorem 4.4. *Assume P^* is SST, and let $\boldsymbol{\pi}_0, \boldsymbol{\pi}_{\text{ref}} \in \Delta_K^\circ$. Consider the MRS-ΨPO dynamics $\boldsymbol{\pi}_{t+1} = \text{MRS}(\boldsymbol{\pi}_t; \alpha, \beta, \lambda)$. These dynamics exhibit the following two forms of collapse:*

- ***ε-collapse when $\alpha < 1$.*** *If $\alpha, \lambda \in [0, 1)$, then there exists T such that for all $t \geq T$,*

$$\pi_{t,1} \geq 1 - \left(\exp\left(\underline{r}_{\text{ref}} + \frac{\beta\lambda\delta}{2(1-\alpha)}\right) - 1\right)^{-1},$$

whenever the exponent is positive, where $\underline{r}_{\text{ref}} := \min_{i\in[K-1]} \log \frac{\pi_{\text{ref},i}}{\pi_{\text{ref},i+1}}$, and $\delta := \min_{i\in[K-1]}\{(1-\lambda)\sum_{j=1}^K \pi_{0,j}(P_{ij} - P_{i+1,j})\} > 0$.

[5] The SST assumption is used for clarity. The same qualitative conclusions hold under weaker ST conditions with heavier notation; see Section C.3.

- ***True collapse when*** $\alpha = 1$**.** *If* $\alpha = 1$ *and* $\lambda < 1$*, then there exists* $T < \infty$ *such that* $H(\boldsymbol{\pi}_{t+1}) < H(\boldsymbol{\pi}_t), \forall t \geq T$*, where* $H(\boldsymbol{\pi}) := -\sum_i \pi_i \log \pi_i$*. Moreover,* $\boldsymbol{\pi}_t \to \boldsymbol{e}_1, H(\boldsymbol{\pi}_t) \to 0$*, as* $t \to \infty$*.*

Theorem 4.4 shows that self-reinforcement can drive concentration even when both $\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_{\text{ref}}$ have full support. If $\alpha = 1$, preference advantages accumulate across iterations and yield winner-take-all collapse. If $\alpha < 1$, the fixed reference prevents exact collapse to a vertex, but the policy can still become arbitrarily concentrated as $\beta\lambda\delta/(1-\alpha)$ grows.

Even when preferences are fixed and each pipeline converges to a unique policy, the limiting distribution can depend on pipeline design. In particular, feedback can drive the policy to the top response despite full-support sampling and per-round KL regularization. The zero-entropy limit is a convergence guarantee when selecting the top response is intended, and a potential failure mode when diversity or continued exploration matters. For fixed $\lambda \in (0, 1)$, larger reference refresh α or stronger update strength β strengthens the concentration guarantee. Thus, full-support sampling and per-round KL regularization alone do not guarantee continued diversity. We summarize and provide useful practical guidance in the following remark.

Remark 4.5 (Practical implications for stabilizing iterative ΨPO)**.** *The iterative results suggest two practical guidelines:* ***[Avoid fully self-referential updates.]*** *Fully self-referential updates* ($\alpha = 1$) *create the strongest feedback loop: they amplify cyclic components and small advantages in the underlying preference structure. One should always keep* $\alpha < 1$ *to partially anchor the reference to a fixed baseline.* ***[Tune*** α ***and*** $\beta\lambda$ ***jointly.]*** *Theorems 4.3 and 4.4 show that stability and diversity depend on the combined feedback strength of the iterative update. The contraction condition requires* $\alpha + \frac{\beta\lambda}{2}\|\Pi P\|_2$ *to be small, while the concentration bound depends on* $\beta\lambda\delta/(1-\alpha)$*, with* δ *itself depending on* λ*. When entropy decays rapidly, reduce* α *or* β*; when the cyclic pattern grows, reduce* α*,* $\beta\lambda$*, or both.*

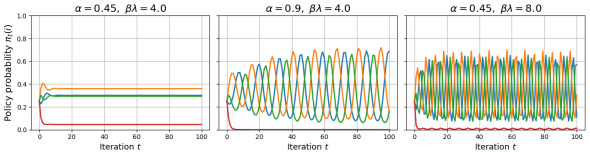
As a final remark, although our results in this section rely on known $P = \Psi(P^*)$, they are also robust to noisy preference observations in practice. In Section D, we show that these phenomena admit stochastic analogues if each round uses a fresh unbiased estimate $\widehat{P}_t$ of P. Our finding is that the contraction and cyclic-instability regimes persist with a greater variance, and the entropy-collapse conclusion continues to hold with high probability.

5 Experiments

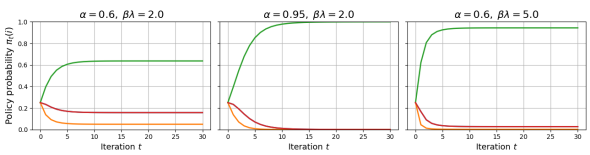
We empirically study the self-reinforcing dynamics predicted by our theory at two complementary levels. First, we simulate the population-level MRS dynamics induced by empirical preference matrices, where the update rule can be applied directly and the roles of the feedback parameters can be isolated. Second, we run iterative preference optimization on real LLMs, where the policy is represented by a neural model and updated through stochastic fine-tuning. Together, these experiments test both the theoretical mechanism and its manifestation in parametric LLM training. All omitted experimental details and robustness checks are provided in Sections F to H.

5.1 Population-level MRS dynamics

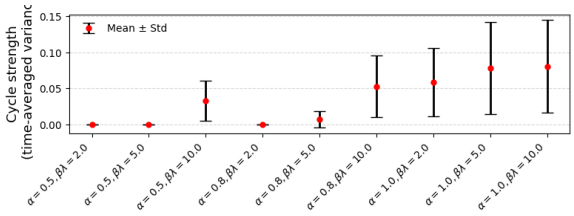
Setup. We first evaluate the MRS dynamics on empirical preference matrices constructed from the NVIDIA HelpSteer dataset [Wang et al., 2024], which consists of human preference feedback for LLM alignment. We use the training split, which contains 35,331 response-level records, each including a prompt paired with a model-generated response that is independently scored along multiple qualitative attributes. Since most prompts have at least four associated responses, we construct preference matrices with $K = 4$ responses per prompt. For each unordered response pair (i, j), we randomly select an evaluation attribute as the comparison



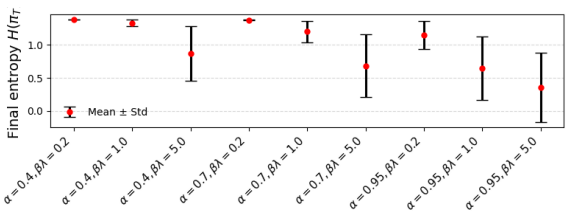
(a) Cyclic instance: stronger feedback induces oscillations.



(b) Strongly transitive instance: stronger feedback induces concentration.



(c) Aggregated oscillation strength across cyclic matrices.



(d) Aggregated entropy across strongly transitive matrices.

Figure 1: Population-level MRS-IPO simulations. The top row shows representative trajectories on cyclic and ST preference matrices; the bottom row aggregates the corresponding behaviors across empirical preferences.

criterion and map the two responses' scores (s_i, s_j) with respect to that attribute to a preference probability via $P_{ij} = \sigma(s_i - s_j)$. Applying this procedure independently across response pairs yields preference matrices that may be either cyclic or ST. Following the theoretical distinction in Section 4, we classify the resulting matrices into cyclic instances and ST instances. After filtering, the dataset contains 118 cyclic instances and 4,924 ST instances. Full construction details are given in Section F.2.

We run the MRS-IPO dynamics in (6) on these instances and report the main population-level results as functions of $(\alpha, \beta\lambda)$. All runs are initialized from the uniform policy, with $\boldsymbol{\pi}_{\text{ref}} = \boldsymbol{\pi}_1 = \frac{1}{K}\mathbf{1}$, and are run for $T = 3000$ iterations. Robustness to finite-sample preference noise and complementary MRS-DPO simulations are deferred to Sections F.3 and G, respectively.

Population-level results. Figure 1 summarizes the MRS-IPO dynamics on empirical cyclic and ST preference matrices. On a representative cyclic matrix, weaker updates lead to policy convergence, while increasing either the self-referential weight α or the effective update-and-sampling strength $\beta\lambda$ induces persistent oscillations, consistent with the instability predicted by Proposition 4.2 and theorem 4.3. On a representative ST matrix, all runs converge, but stronger feedback drives the limiting policy toward a more extreme, nearly deterministic distribution on the top response. The aggregate results show the same pattern across the constructed matrices: for cyclic preferences, the time-averaged coordinate variance increases with α and $\beta\lambda$, indicating stronger oscillations; for ST preferences, the terminal entropy $H(\boldsymbol{\pi}_T)$ decreases toward zero as the update becomes more aggressive. Thus, the population experiments support the main theoretical message: self-reference and model-dependent sampling amplify preference dynamics, producing oscillation under cyclic structure and concentration under ST structure.

5.2 Iterative preference optimization on real LLMs

5.2.1 Transitive preferences

Setup. We run real-LLM instantiations of MRS-IPO and MRS-DPO on Qwen2.5-1.5B with LoRA [Hu et al., 2022], using fixed offline pairwise comparisons constructed from the NVIDIA HelpSteer dataset [Wang et al., 2024]. For the transitive experiments, we retain prompts

with exactly four distinct responses, rank the responses by their mean HelpSteer attribute score, and use all within-prompt comparisons. To approximate model-dependent sampling without repeatedly collecting new on-policy preference data, the current model dynamically reweights this fixed pair pool according to the MRS sampler. At each outer iteration we draw two comparisons per sampled prompt, yielding 1,000 weighted pairs for IPO and 500 for DPO; α controls how strongly the reference is refreshed toward the current policy, λ controls model-dependent reweighting, and β is the training-loss parameter defined in Section H.1.

We evaluate on a fixed panel of 500 prompts, each with the same four-response support throughout training. This panel is sampled from the prepared HelpSteer prompt universe and is used for repeatable measurement rather than as a disjoint validation split. At the beginning of every outer iteration, we apply softmax to the current model's length-normalized log-probabilities over that support, and we report the prompt-averaged entropy. Thus, the metric measures concentration over fixed candidate responses rather than the entropy of open-ended generation. We sweep $\alpha \in \{0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.97, 0.99\}$ and $\lambda \in \{0.5, 0.9\}$ at $\beta = 1$, and compare both objectives over the same 100 outer iterations, indexed by $t = 0, \ldots, 99$. Each configuration is a single run with seed 0. Implementation details are provided in Section H.

Results. Figure 2 reports all 16 transitive settings for each objective. For MRS-IPO, increasing α produces an ordered transition from mild concentration to near collapse. At $\lambda = 0.5$, the terminal entropy falls from 1.19 at $\alpha = 0.5$ to 0.11 at $\alpha = 0.99$; at $\lambda = 0.9$, it falls from 1.17 to 0.09. The similar ordering across both values of λ supports the feedback-amplification mechanism predicted by Theorem 4.4, while the slightly lower curves at larger λ show that more on-policy sampling further strengthens concentration.

MRS-DPO is markedly more aggressive: even $\alpha = 0.5$ reaches entropy 0.48 for $\lambda = 0.5$ and 0.32 for $\lambda = 0.9$ at $t = 99$, and by $\alpha = 0.8$ the two settings reach 0.19 and 0.17, respectively. For the most self-referential DPO runs, entropy remains low but the trajectories become visibly noisier and terminal entropy is not strictly monotone in α, consistent with continued policy movement near the collapsed boundary. Overall, the denser sweep sharpens the original comparison: self-reference systematically drives concentration, on-policy reweighting usually amplifies it, and DPO reaches low-entropy policies substantially earlier than IPO. Complete terminal entropy values and setup details are reported in Sections H.2 and H.4.

5.2.2 Training under cyclic preferences

To test whether the real-LLM feedback loop also reflects cyclic preference structure, we construct a separate fixed offline pool on four responses per prompt. Its six directed comparisons form a full tournament containing the cycle $y_0 \succ y_1 \succ y_2 \succ y_3 \succ y_0$, with the two remaining edges oriented as $y_0 \succ y_2$ and $y_1 \succ y_3$. This is a *fixed-pool* protocol: the binary labels are fixed before training, and the current model only changes which comparisons receive more weight. For a direct IPO–DPO comparison, we use the three completed configurations shared by both objectives,

$$(\alpha, \lambda, \beta) \in \{(0.1, 0.1, 0.01),\ (0.5, 0.5, 5),\ (0.99, 0.8, 1)\},$$

with $\tau = 1$, seed 0, and the same fixed-support evaluation rule as above. The cyclic construction, sampling rule, and optimization hyperparameters are detailed in Section H.3.

Results. Figure 3 shows that cyclic labels do not produce a single universal entropy pattern. The weak-feedback setting $(0.1, 0.1, 0.01)$ concentrates gradually, reaching entropy 0.150 for IPO and 0.219 for DPO at $t = 99$. By contrast, for $(0.5, 0.5, 5)$, entropy remains high and nearly constant, ending at 1.282 and 1.234, respectively. The strongly self-referential setting $(0.99, 0.8, 1)$ rapidly enters a low-entropy regime; its terminal entropy is 0.096 for IPO and 0.177 for DPO, with visibly larger late-iteration excursions for DPO. Thus, even with the same

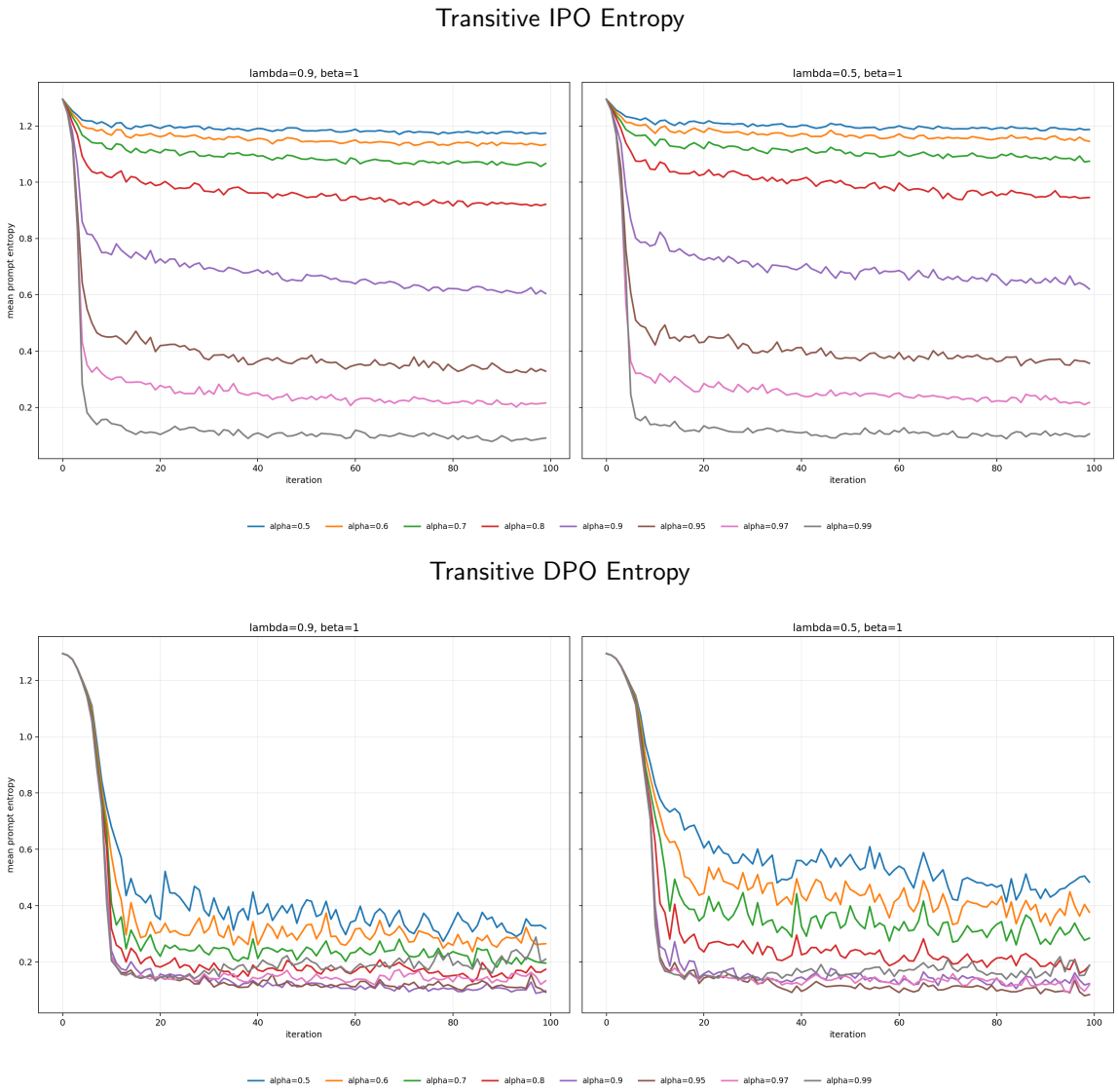


Figure 2: Mean fixed-support entropy during iterative fine-tuning of Qwen2.5-1.5B on transitive HelpSteer preferences. MRS-IPO is shown above and MRS-DPO below; within each image, the left and right panels correspond to $\lambda = 0.9$ and $\lambda = 0.5$, respectively, at $\beta = 1$. Both objectives are shown over the same 100 outer iterations, $t = 0, \ldots, 99$, with one curve per α.

cyclic tournament, feedback strength and objective choice determine whether the induced fixed-support policy stays diffuse, concentrates smoothly, or continues to move within a concentrated regime.

6 Conclusions

We have presented a theoretical investigation of how sampling and reference choices shape both one-shot PO solutions and iterative PO dynamics under the general ΨPO framework. Our results show that careless choices of these factors can compromise ranking guarantees, induce policy oscillations, or lead to entropy collapse. We also identify hyperparameter regimes that recover stability and mitigate collapse, providing practical guidance for designing robust iterative preference-alignment pipelines.

Our findings point to a key open problem: how to design alignment pipelines with principled, adaptive control over sampling and reference updates—otherwise, instability is a predictable

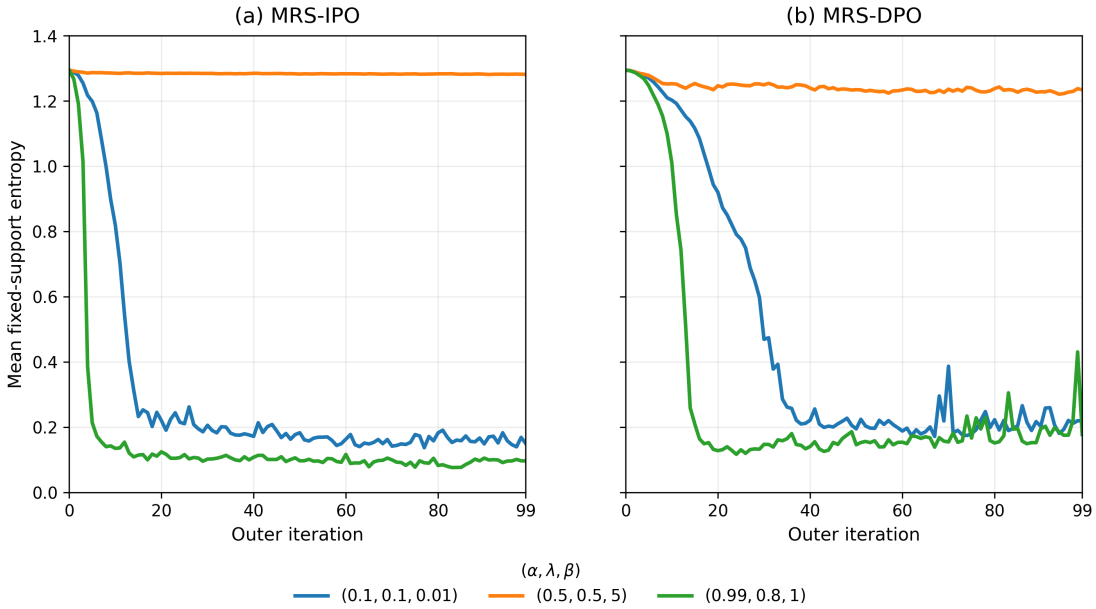


Figure 3: Mean fixed-support entropy for the three completed cyclic configurations shared by MRS-IPO and MRS-DPO. Tuples in the legend denote (α, λ, β). Both panels show $t = 0, \ldots, 99$; the response support and cyclic labels are fixed throughout training.

outcome of feedback-driven alignment. Addressing this challenge requires treating preference alignment as a stochastic online learning system rather than a static offline optimization problem. We view this shift in perspective as essential for building reliable alignment methods at scale.

References

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. Self-consuming generative models go mad. In *The Twelfth International Conference on Learning Representations*, 2023.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*, 2021.

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

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

Quentin Bertrand, Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. On the stability of iterative retraining of generative models on their own data. In *The Twelfth International Conference on Learning Representations*, 2023.

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.

Daniele Calandriello, Daniel Guo, Remi Munos, Mark Rowland, Yunhao Tang, Bernardo Avila Pires, Pierre Harvey Richemond, Charline Le Lan, Michal Valko, Tianqi Liu, et al. Human alignment of large language models through online preference optimisation. *arXiv preprint arXiv:2403.08635*, 2024.

Sayak Ray Chowdhury, Anush Kini, and Nagarajan Natarajan. Provably robust dpo: Aligning language models with noisy feedback. *arXiv preprint arXiv:2403.00409*, 2024.

Hanze Dong, Wei Xiong, Deepanshu Goyal, Yihan Zhang, Winnie Chow, Rui Pan, Shizhe Diao, Jipeng Zhang, Kashun Shum, and Tong Zhang. Raft: Reward ranked finetuning for generative foundation model alignment. *arXiv preprint arXiv:2304.06767*, 2023.

Hanze Dong, Wei Xiong, Bo Pang, Haoxiang Wang, Han Zhao, Yingbo Zhou, Nan Jiang, Doyen Sahoo, Caiming Xiong, and Tong Zhang. RLHF workflow: From reward modeling to online RLHF. *arXiv preprint arXiv:2405.07863*, 2024.

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.

Damien Ferbach, Quentin Bertrand, Avishek Joey Bose, and Gauthier Gidel. Self-consuming generative models with curated data provably optimize human preferences. *arXiv preprint arXiv:2407.09499*, 2024.

Luise Ge, Daniel Halpern, Evi Micha, Ariel D. Procaccia, Itai Shapira, Yevgeniy Vorobeychik, and Junlin Wu. Axioms for AI alignment from human feedback. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 80439–80465. Curran Associates, Inc., 2024. doi: 10.52202/079017-2557. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9328208f88ec69420031647e6ff97727-Paper-Conference.pdf.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, et al. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *arXiv preprint arXiv:2404.01413*, 2024.

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

Melody Y. Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, Hyung Won Chung, Sam Toyer, Johannes Heidecke, Alex Beutel, and Amelia Glaese. Deliberative alignment: Reasoning enables safer language models, 2025. URL https://arxiv.org/abs/2412.16339.

Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, et al. Direct language model alignment from online AI feedback. *arXiv preprint arXiv:2402.04792*, 2024.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. Benchmarking linguistic diversity of large language models. *Transactions of the Association for Computational Linguistics*, 13:1507–1526, 2025.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2): 3, 2022.

Hugging Face. TRL documentation: DPO trainer. https://huggingface.co/docs/trl/v0.27.0/dpo_trainer, 2026. Version v0.27.0, accessed 2026-01-23.

Kyung Rok Kim, Yumo Bai, Chonghuan Wang, and Guanting Chen. Understanding the impact of sampling quality in direct preference optimization. *arXiv preprint arXiv:2506.04272*, 2025.

Jack Lanchantin, Angelica Chen, Shehzaad Dhuliawala, Ping Yu, Jason Weston, Sainbayar Sukhbaatar, and Ilia Kulikov. Diverse preference optimization. *arXiv preprint arXiv:2501.18101*, 2025.

Kaizhao Liu, Qi Long, Zhekun Shi, Weijie J Su, and Jiancong Xiao. Statistical impossibility and possibility of aligning LLMs with human preferences: From Condorcet paradox to Nash equilibrium. *arXiv preprint arXiv:2503.10990*, 2025.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization. *arXiv preprint arXiv:2309.06657*, 2023.

Sonia Krishna Murthy, Tomer Ullman, and Jennifer Hu. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11241–11258, 2025.

Ritesh Noothigattu, Dominik Peters, and Ariel D Procaccia. Axioms for learning from pairwise comparisons. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 17745–17754. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/cdaa9b682e10c291d3bbadca4c96f5de-Paper.pdf.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Yu Pan, Zhongze Cai, Guanting Chen, Huaiyang Zhong, and Chonghuan Wang. What matters in data for DPO? *arXiv preprint arXiv:2508.18312*, 2025.

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.

Ruizhe Shi, Runlong Zhou, and Simon Shaolei Du. The crucial role of samplers in online direct preference optimization. In *The Thirteenth International Conference on Learning Representations*, 2024.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631(8022): 755–759, 2024.

Fahim Tajwar, Anikait Singh, Archit Sharma, Rafael Rafailov, Jeff Schneider, Tengyang Xie, Stefano Ermon, Chelsea Finn, and Aviral Kumar. Preference fine-tuning of LLMs should leverage suboptimal, on-policy data. *arXiv preprint arXiv:2404.14367*, 2024.

Louis L Thurstone. A law of comparative judgment. In *Scaling*, pages 81–92. Routledge, 2017.

Amos Tversky. Intransitivity of preferences. *Psychological review*, 76(1):31, 1969.

Zhilin Wang, Yi Dong, Jiaqi Zeng, Virginia Adams, Makesh Narsimhan Sreedhar, Daniel Egert, Olivier Delalleau, Jane Scowcroft, Neel Kant, Aidan Swope, et al. Helpsteer: Multi-attribute helpfulness dataset for SteerLM. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3371–3384, 2024.

Lorenz Wolf, Robert Kirk, and Mirco Musolesi. Reward model overoptimisation in iterated RLHF. *arXiv preprint arXiv:2505.18126*, 2025.

Jiancong Xiao, Zhekun Shi, Kaizhao Liu, Qi Long, and Weijie J Su. Theoretical tensions in RLHF: Reconciling empirical success with inconsistencies in social choice theory. *arXiv preprint arXiv:2506.12350*, 2025.

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. *arXiv preprint arXiv:2312.11456*, 2023.

Wenda Xu, Jiachen Li, William Yang Wang, and Lei Li. Bpo: Staying close to the behavior llm creates better online llm alignment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11125–11139, 2024.

Chenlu Ye, Wei Xiong, Yuheng Zhang, Hanze Dong, Nan Jiang, and Tong Zhang. Online iterative reinforcement learning from human feedback with general preference model. *Advances in Neural Information Processing Systems*, 37:81773–81807, 2024.

Bingji Yi, Qiyuan Liu, Yuwei Cheng, and Haifeng Xu. Escaping model collapse via synthetic data verification: Near-term improvements and long-term convergence. *arXiv preprint arXiv:2510.16657*, 2025.

Banghua Zhu, Michael I Jordan, and Jiantao Jiao. Iterative data smoothing: Mitigating reward overfitting and overoptimization in rlhf. *arXiv preprint arXiv:2401.16335*, 2024.

A Proof of Proposition 2.2

Proposition 2.2. *Given an underlying preference matrix $P^* \in \mathcal{P}$ and its transformed matrix $P = \Psi(P^*)$, the ΨPO problem with sampling strategy $\boldsymbol{\mu} \in \Delta_K^\circ$ and reference policy $\boldsymbol{\pi}_{\text{ref}} \in \Delta_K^\circ$ admits a unique solution $\boldsymbol{\pi}^\star \in \Delta_K^\circ$. Moreover, this solution has the closed form*

$$\boldsymbol{\pi}^\star \propto \boldsymbol{\pi}_{\text{ref}} \odot \exp(\beta P \boldsymbol{\mu}), \tag{2}$$

where $\odot$ denotes element-wise multiplication.

Proof. Let $v := P\boldsymbol{\mu}$, so that $v_i = (P\boldsymbol{\mu})_i = \sum_{j=1}^K \mu_j \Psi(P^*_{ij})$. The population ΨPO objective in (1) can be written as

$$J(\boldsymbol{\pi}) = \langle \boldsymbol{\pi}, v \rangle - \frac{1}{\beta} \mathrm{KL}(\boldsymbol{\pi} \| \boldsymbol{\pi}_{\text{ref}}), \qquad \boldsymbol{\pi} \in \Delta_K, \tag{9}$$

where

$$\mathrm{KL}(\boldsymbol{\pi} \| \boldsymbol{\pi}_{\text{ref}}) := \sum_{i=1}^K \pi_i \log \frac{\pi_i}{\pi_{\text{ref},i}}.$$

Since $\boldsymbol{\pi}_{\text{ref}} \in \Delta_K^\circ$, this quantity is well-defined on Δ_K under the convention $0 \log 0 = 0$.

Define

$$Z := \sum_{\ell=1}^K \pi_{\text{ref},\ell} \exp(\beta v_\ell),$$

and define a probability vector $q = (q_1, \ldots, q_K) \in \Delta_K^\circ$ by

$$q_i := \frac{\pi_{\text{ref},i} \exp(\beta v_i)}{Z}, \qquad i \in [K].$$

Because $\boldsymbol{\pi}_{\text{ref}} \in \Delta_K^\circ$ and v_i is finite for every i, we have $q_i > 0$ for every i and $\sum_i q_i = 1$.

We now rewrite the objective by completing the KL term. For any $\boldsymbol{\pi} \in \Delta_K$,

$$\begin{aligned} \mathrm{KL}(\boldsymbol{\pi} \| q) &:= \sum_{i=1}^K \pi_i \log \frac{\pi_i}{q_i} \\ &= \sum_{i=1}^K \pi_i \log \frac{\pi_i}{\pi_{\text{ref},i}} - \beta \sum_{i=1}^K \pi_i v_i + \log Z. \end{aligned}$$

Rearranging gives

$$\langle \boldsymbol{\pi}, v \rangle - \frac{1}{\beta} \mathrm{KL}(\boldsymbol{\pi} \| \boldsymbol{\pi}_{\text{ref}}) = \frac{1}{\beta} \log Z - \frac{1}{\beta} \mathrm{KL}(\boldsymbol{\pi} \| q).$$

Therefore,

$$J(\boldsymbol{\pi}) = \frac{1}{\beta} \log Z - \frac{1}{\beta} \mathrm{KL}(\boldsymbol{\pi} \| q).$$

Since $\mathrm{KL}(\boldsymbol{\pi} \| q) \geq 0$, with equality if and only if $\boldsymbol{\pi} = q$, the unique maximizer of (9) is $\boldsymbol{\pi}^\star = q$. Hence

$$\pi_i^\star = \frac{\pi_{\text{ref},i} \exp(\beta (P\boldsymbol{\mu})_i)}{\sum_{\ell=1}^K \pi_{\text{ref},\ell} \exp(\beta (P\boldsymbol{\mu})_\ell)}, \qquad i \in [K].$$

Equivalently,

$$\boldsymbol{\pi}^\star = \mathrm{softmax}(\log \boldsymbol{\pi}_{\text{ref}} + \beta P \boldsymbol{\mu}) \propto \boldsymbol{\pi}_{\text{ref}} \odot \exp(\beta P \boldsymbol{\mu}).$$

Finally, in relative-logit coordinates

$$\theta_i := \frac{1}{\beta} \log \frac{\pi_i}{\pi_{\text{ref},i}},$$

the optimizer satisfies

$$\theta_i^\star = (P\boldsymbol{\mu})_i - \frac{1}{\beta}\log Z.$$

Thus, modulo an additive constant, the optimal relative logits are

$$\boldsymbol{\theta}^\star \equiv P\boldsymbol{\mu}.$$

This completes the proof. □

B Omitted Proofs in Section 3

B.1 Proof of Proposition 3.3

Proposition 3.3. *Let $K \geq 3$. For any $\boldsymbol{\mu}$, ΨPO under $\boldsymbol{\mu}$ is not* Condorcet-top, Smith-top, *or* Order-preserving.

Proof. Fix any $\boldsymbol{\mu} \in \Delta_K^\circ$. Since every strictly transitive preference matrix has a unique Condorcet winner, and its Smith set is the singleton containing that winner, it suffices to construct a strictly transitive matrix P^* such that the unique Condorcet winner is not ranked top by $F_{\boldsymbol{\mu}}(P^*)$.

Choose $\rho \in (1/2, 1)$ such that $\Psi(\rho) > \Psi(1/2)$. Fix any $\tau \in (1/2, 1)$. For $\varepsilon \in (0, 1/2)$ to be specified later, define the underlying preference matrix P^* by

$$P^*_{12} = P^*_{13} = \frac{1}{2} + \varepsilon, \qquad P^*_{23} = \rho,$$

and for every remaining pair $1 \leq i < j \leq K$ not among $(1,2), (1,3), (2,3)$, set

$$P^*_{ij} = \tau.$$

Finally impose

$$P^*_{ji} = 1 - P^*_{ij}, \qquad P^*_{ii} = \frac{1}{2}, \qquad \forall i, j \in [K],$$

to make P^* a valid preference matrix.

By construction, every upper-triangular entry of P^* is strictly larger than $1/2$, so P^* is strictly transitive with order $1 \succ 2 \succ 3 \succ \cdots \succ K$. Hence response 1 is the unique Condorcet winner of P^*, and $\mathrm{Smith}(P^*) = \{1\}$.

Let $P := \Psi(P^*)$ be the transformed matrix induced by ΨPO. (2) implies,

$$F_{\boldsymbol{\mu}}(P^*)_2 > F_{\boldsymbol{\mu}}(P^*)_1 \quad \Longleftrightarrow \quad (P\boldsymbol{\mu})_2 > (P\boldsymbol{\mu})_1.$$

Now compare $(P\boldsymbol{\mu})_2$ and $(P\boldsymbol{\mu})_1$. For every $k \geq 4$, we have $P^*_{1k} = P^*_{2k} = \tau$, which means $P_{1k} = P_{2k} = \Psi(\tau)$. Therefore, columns $k \geq 4$ cancel in the difference $(P\boldsymbol{\mu})_2 - (P\boldsymbol{\mu})_1$. Thus

$$\begin{aligned}(P\boldsymbol{\mu})_2 - (P\boldsymbol{\mu})_1 &= \mu_1 (P_{21} - P_{11}) + \mu_2 (P_{22} - P_{12}) + \mu_3 (P_{23} - P_{13}) \\ &= \mu_1 \left[\Psi\left(\frac{1}{2} - \varepsilon\right) - \Psi\left(\frac{1}{2}\right)\right] + \mu_2 \left[\Psi\left(\frac{1}{2}\right) - \Psi\left(\frac{1}{2} + \varepsilon\right)\right] \\ &\quad + \mu_3 \left[\Psi(\rho) - \Psi\left(\frac{1}{2} + \varepsilon\right)\right].\end{aligned}$$

Since Ψ is continuous at $1/2$,

$$\lim_{\varepsilon \downarrow 0} \left((P\boldsymbol{\mu})_2 - (P\boldsymbol{\mu})_1\right) = \mu_3 \left(\Psi(\rho) - \Psi\left(\frac{1}{2}\right)\right).$$

Because $\boldsymbol{\mu} \in \Delta_K^\circ$, we have $\mu_3 > 0$, and by assumption

$$\Psi(\rho) - \Psi\left(\frac{1}{2}\right) > 0.$$

Hence

$$\mu_3 \left(\Psi(\rho) - \Psi\left(\frac{1}{2}\right)\right) > 0.$$

It follows that for all sufficiently small $\varepsilon > 0$,

$$(P\boldsymbol{\mu})_2 - (P\boldsymbol{\mu})_1 > 0.$$

Therefore,

$$F_{\boldsymbol{\mu}}(P^*)_2 > F_{\boldsymbol{\mu}}(P^*)_1.$$

Since response 1 is the unique Condorcet winner of P^*, $F_{\boldsymbol{\mu}}$ is not Condorcet-top. Since $\mathrm{Smith}(P^*) = \{1\}$, $F_{\boldsymbol{\mu}}$ is not Smith-top either. Finally, because P^* is strictly transitive with order $1 \succ 2 \succ \cdots \succ K$ while the rule ranks response 2 above response 1, $F_{\boldsymbol{\mu}}$ is not order-preserving.

This completes the proof. □

B.2 Proof of Theorem 3.4

Theorem 3.4. *For any P^* with Condorcet winner c, there exists a $\boldsymbol{\mu} \in \Delta_K^\circ$ such that $F_{\boldsymbol{\mu}}(P^*)$ assigns c the unique highest probability. Moreover, if Condition 2.1 holds, then for any P^*, there exists a $\boldsymbol{\mu} \in \Delta_K^\circ$ such that $F_{\boldsymbol{\mu}}(P^*)_i > F_{\boldsymbol{\mu}}(P^*)_j$ for any $i \in \mathrm{Smith}(P^*)$ and $j \notin \mathrm{Smith}(P^*)$.*

Proof. Let

$$P \coloneqq \Psi(P^*), \qquad c_\Psi \coloneqq \Psi(1/2).$$

Similarly to the proof of Proposition 3.3, by (2), to compare the probability masses of $F_{\boldsymbol{\mu}}$, it suffices to compare the coordinates of $P\boldsymbol{\mu}$.

We first prove statement 1. Let c be a Condorcet winner of P^*. Consider the non-full-support sampling distribution $\boldsymbol{\mu}' = \boldsymbol{e}_c$ that put all mass on response c, where $\boldsymbol{e}_c$ denotes the one-hot vector with c-th entry equal to 1 and all other entries equal to 0. Then

$$(P\boldsymbol{\mu}')_c = P_{cc} = c_\Psi.$$

For any $j \neq c$, since c is a Condorcet winner,

$$P^*_{jc} < 1/2.$$

Since Ψ is strictly increasing,

$$(P\boldsymbol{\mu}')_j = P_{jc} = \Psi(P^*_{jc}) < c_\Psi = (P\boldsymbol{\mu}')_c.$$

Therefore,

$$F_{\boldsymbol{\mu}'}(P^*)_c > F_{\boldsymbol{\mu}'}(P^*)_j, \qquad \forall j \neq c.$$

One can then mix it with a uniform sampling to obtain a full-support sampling $\boldsymbol{\mu} = \varepsilon\boldsymbol{\mu}_0 + (1-\varepsilon)\boldsymbol{\mu}'$, where $\boldsymbol{\mu}_0 = \frac{1}{K}\mathbf{1}$ is the uniform sampling distribution and $\mathbf{1}$ is the all-one vector. Since the inequality is strict, it will preserve for a sufficiently small ε.

Now we prove statement 2. Let $S = \mathrm{Smith}(P^*)$ and $N = [K] \setminus S$. W.L.O.G.,

$$S = \{1, \dots, L\}.$$

By definition of the Smith set, for every $i \in S$ and $j \in N$,

$$P^*_{ij} > 1/2, \qquad P^*_{ji} < 1/2.$$

Let $P_{SS} = \Psi(P^*_{SS})$, and define

$$A_{SS} \coloneqq P_{SS} - c_\Psi \mathbf{1}\mathbf{1}^\top.$$

By Condition 2.1, for every $i \neq j$,

$$(A_{SS})_{ji} = \Psi(P^*_{ji}) - c_\Psi = \Psi(1 - P^*_{ij}) - c_\Psi = -(\Psi(P^*_{ij}) - c_\Psi) = -(A_{SS})_{ij},$$

and $(A_{SS})_{ii} = 0$. Hence A_{SS} is skew-symmetric, so the corresponding zero-sum game has value 0. By the minimax theorem, there exists $\boldsymbol{\mu}_S \in \Delta(S)$ such that

$$A_{SS}\boldsymbol{\mu}_S \geq \mathbf{0},$$

equivalently,

$$P_{SS}\boldsymbol{\mu}_S \geq c_\Psi \mathbf{1}_S.$$

Extend $\boldsymbol{\mu}_S$ to $\boldsymbol{\mu}'$ on $[K]$ by setting zero mass on N. Then for every $i \in S$,

$$(P\boldsymbol{\mu}')_i \geq c_\Psi.$$

On the other hand, for every $j \in N$ and $k \in S$, we have $P^*_{jk} < 1/2$. Since Ψ is strictly increasing,

$$P_{jk} = \Psi(P^*_{jk}) < c_\Psi.$$

Therefore,

$$(P\boldsymbol{\mu}')_j < c_\Psi, \qquad \forall j \in N.$$

Thus

$$(P\boldsymbol{\mu}')_i > (P\boldsymbol{\mu}')_j, \qquad \forall i \in S,\ j \in N,$$

which implies

$$F_{\boldsymbol{\mu}'}(P^*)_i > F_{\boldsymbol{\mu}'}(P^*)_j, \qquad \forall i \in S,\ j \in N.$$

Again, one can then mix it with a uniform sampling to obtain a full-support sampling $\boldsymbol{\mu} = \varepsilon\boldsymbol{\mu}_0 + (1-\varepsilon)\boldsymbol{\mu}'$. Since the inequality is strict, $F_{\boldsymbol{\mu}}(P^*)_i > F_{\boldsymbol{\mu}}(P^*)_j$ will still hold for $i \in S$ and $j \in N$ for sufficiently small $\varepsilon > 0$. □

B.3 Fragility of order preservation under sampling

Proposition B.1. *For any $K \geq 4$, there exists a transitive P^*, such that for any $i \in [K-1]$, there exists a $\boldsymbol{\mu}$ for which $F_{\boldsymbol{\mu}}(P^*)$ ranks the best response in the i-th position. There exists a transitive P^* such that, for every $\boldsymbol{\mu} \in \Delta^\circ_K$, the output $F_{\boldsymbol{\mu}}(P^*)$ is not order-preserving.*

Proof. Let $c_\Psi \coloneqq \Psi(1/2)$. For $q \in (1/2, 1)$, define

$$\alpha(q) \coloneqq \Psi(q) - c_\Psi, \qquad \beta(q) \coloneqq c_\Psi - \Psi(1-q).$$

Since Ψ is continuous and strictly increasing, both α and β are continuous, strictly increasing, positive on $(1/2, 1)$, and satisfy

$$\alpha(q) \to 0, \qquad \beta(q) \to 0 \qquad \text{as } q \downarrow 1/2.$$

Choose any points

$$\frac{1}{2} < q_2 < q_3 < q_4 < 1.$$

Since

$$\beta(q_2) > 0, \qquad \alpha(q_3) > 0, \qquad \alpha(q_4) - \alpha(q_2) > 0, \qquad \beta(q_4) - \beta(q_3) > 0,$$

we can choose $q_1 \in (1/2, q_2)$ sufficiently close to $1/2$ so that

$$2\beta(q_1) < \beta(q_2), \quad 2\alpha(q_1) < \alpha(q_3), \quad \beta(q_1) < \alpha(q_4) - \alpha(q_2), \quad \alpha(q_1) < \beta(q_4) - \beta(q_3). \tag{10}$$

For $r = 1, 2, 3, 4$, write

$$\alpha_r \coloneqq \alpha(q_r) = \Psi(q_r) - c_\Psi, \qquad \beta_r \coloneqq \beta(q_r) = c_\Psi - \Psi(1-q_r).$$

We first prove statement 1. Fix any $K \geq 4$. Define an underlying preference matrix P^* by

$$\begin{aligned} P^*_{1j} &= q_1, \qquad \forall j > 1, \\ P^*_{ij} &= q_2, \qquad \forall\, 2 \leq i < j \leq K, \end{aligned}$$

together with

$$P^*_{ji} = 1 - P^*_{ij}, \qquad P^*_{ii} = \frac{1}{2}, \qquad \forall i, j \in [K].$$

Since every upper-triangular entry of P^* is strictly larger than $1/2$, the matrix P^* is transitive with order $1 \succ 2 \succ \cdots \succ K$.

Let $P \coloneqq \Psi(P^*)$. Then

$$\begin{aligned} P_{1j} &= c_\Psi + \alpha_1, \qquad \forall j > 1, \\ P_{ij} &= c_\Psi + \alpha_2, \qquad \forall\, 2 \leq i < j \leq K, \\ P_{j1} &= c_\Psi - \beta_1, \qquad \forall j > 1, \\ P_{ji} &= c_\Psi - \beta_2, \qquad \forall\, 2 \leq i < j \leq K, \end{aligned}$$

and $P_{ii} = c_\Psi$ for all i. Below is an example matrix with $K = 4$, together with the transformed matrix.

$$P^* = \begin{pmatrix} \frac{1}{2} & q_1 & q_1 & q_1 \\ 1-q_1 & \frac{1}{2} & q_2 & q_2 \\ 1-q_1 & 1-q_2 & \frac{1}{2} & q_2 \\ 1-q_1 & 1-q_2 & 1-q_2 & \frac{1}{2} \end{pmatrix}, \qquad P = \begin{pmatrix} c_\Psi & c_\Psi + \alpha_1 & c_\Psi + \alpha_1 & c_\Psi + \alpha_1 \\ c_\Psi - \beta_1 & c_\Psi & c_\Psi + \alpha_2 & c_\Psi + \alpha_2 \\ c_\Psi - \beta_1 & c_\Psi - \beta_2 & c_\Psi & c_\Psi + \alpha_2 \\ c_\Psi - \beta_1 & c_\Psi - \beta_2 & c_\Psi - \beta_2 & c_\Psi \end{pmatrix}.$$

Since $\alpha_2 > \alpha_1 > 0$ and $\beta_2 > \beta_1 > 0$, for every fixed $j \in \{2, \dots, K\}$, the j-th column of P has the form

$$\begin{aligned} P_{1j} &= c_\Psi + \alpha_1, \\ P_{ij} &= c_\Psi + \alpha_2, \qquad \forall\, 2 \le i < j, \\ P_{jj} &= c_\Psi, \\ P_{ij} &= c_\Psi - \beta_2, \qquad \forall\, i > j. \end{aligned}$$

Hence in the ranking induced by the j-th column, response 1 is exactly in position $j-1$.

Now let

$$\boldsymbol{\mu}^{(\delta,j)} \coloneqq (1-\delta)\boldsymbol{e}_j + \frac{\delta}{K}\mathbf{1},$$

where $\boldsymbol{e}_j$ is the j-th standard basis vector and $\delta \in (0,1)$. Then, denote the j-th column of P as $P_{\cdot j}$. We have,

$$\boldsymbol{\mu}^{(\delta,j)} \in \Delta_K^\circ \qquad \text{and} \qquad P\boldsymbol{\mu}^{(\delta,j)} \to P\boldsymbol{e}_j = P_{\cdot j} \quad \text{as } \delta \to 0.$$

Since the score of response 1 is strictly separated from every other score in the j-th column, its position is preserved for all sufficiently small $\delta > 0$. By (2), $F_{\boldsymbol{\mu}^{(\delta,j)}}(P^*)$ ranks response 1 in position $j-1$ for all sufficiently small $\delta > 0$. Since $j-1$ ranges over $[K-1]$, statement 1 follows.

We now prove statement 2. Fix any $K \ge 4$. Define $P^* \in \mathcal{P}$ as follows. On the first four responses, set

$$\begin{aligned} P^*_{12} &= q_1, \qquad P^*_{13} = q_2, \qquad P^*_{14} = q_1, \\ P^*_{23} &= q_4, \qquad P^*_{24} = q_3, \qquad P^*_{34} = q_1. \end{aligned}$$

For all remaining upper-triangular entries involving indices at least 5, set

$$\begin{aligned} P^*_{ik} &= q_1, \qquad \forall\, 1 \le i \le 4 < k \le K, \\ P^*_{ij} &= q_1, \qquad \forall\, 5 \le i < j \le K. \end{aligned}$$

Finally impose

$$P^*_{ji} = 1 - P^*_{ij}, \qquad P^*_{ii} = \frac{1}{2}, \qquad \forall i, j \in [K].$$

Again, every upper-triangular entry is strictly larger than $1/2$, so P^* is transitive with order $1 \succ 2 \succ 3 \succ \cdots \succ K$.

Let $P \coloneqq \Psi(P^*)$. Below is an example matrix with $K = 5$, together with the transformed matrix.

$$P^* = \begin{pmatrix} \frac{1}{2} & q_1 & q_2 & q_1 & q_1 \\ 1-q_1 & \frac{1}{2} & q_4 & q_3 & q_1 \\ 1-q_2 & 1-q_4 & \frac{1}{2} & q_1 & q_1 \\ 1-q_1 & 1-q_3 & 1-q_1 & \frac{1}{2} & q_1 \\ 1-q_1 & 1-q_1 & 1-q_1 & 1-q_1 & \frac{1}{2} \end{pmatrix},$$

and,

$$P = \begin{pmatrix} c_\Psi & c_\Psi+\alpha_1 & c_\Psi+\alpha_2 & c_\Psi+\alpha_1 & c_\Psi+\alpha_1 \\ c_\Psi-\beta_1 & c_\Psi & c_\Psi+\alpha_4 & c_\Psi+\alpha_3 & c_\Psi+\alpha_1 \\ c_\Psi-\beta_2 & c_\Psi-\beta_4 & c_\Psi & c_\Psi+\alpha_1 & c_\Psi+\alpha_1 \\ c_\Psi-\beta_1 & c_\Psi-\beta_3 & c_\Psi-\beta_1 & c_\Psi & c_\Psi+\alpha_1 \\ c_\Psi-\beta_1 & c_\Psi-\beta_1 & c_\Psi-\beta_1 & c_\Psi-\beta_1 & c_\Psi \end{pmatrix}.$$

On the first four responses, the transformed matrix is

$$P_{1:4,1:4} = \begin{pmatrix} c_\Psi & c_\Psi + \alpha_1 & c_\Psi + \alpha_2 & c_\Psi + \alpha_1 \\ c_\Psi - \beta_1 & c_\Psi & c_\Psi + \alpha_4 & c_\Psi + \alpha_3 \\ c_\Psi - \beta_2 & c_\Psi - \beta_4 & c_\Psi & c_\Psi + \alpha_1 \\ c_\Psi - \beta_1 & c_\Psi - \beta_3 & c_\Psi - \beta_1 & c_\Psi \end{pmatrix}.$$

Moreover, for every $k \geq 5$,

$$P_{1k} = P_{2k} = P_{3k} = P_{4k} = c_\Psi + \alpha_1.$$

Hence, for any $\boldsymbol{\mu} \in \Delta_K^\circ$, the contributions of columns $k \geq 5$ cancel when comparing rows 1 and 2, and also when comparing rows 3 and 4. Therefore,

$$(P\boldsymbol{\mu})_1 - (P\boldsymbol{\mu})_2 = \beta_1\mu_1 + \alpha_1\mu_2 - (\alpha_4 - \alpha_2)\mu_3 - (\alpha_3 - \alpha_1)\mu_4,$$

and

$$(P\boldsymbol{\mu})_3 - (P\boldsymbol{\mu})_4 = -(\beta_2 - \beta_1)\mu_1 - (\beta_4 - \beta_3)\mu_2 + \beta_1\mu_3 + \alpha_1\mu_4.$$

Suppose, for contradiction, that there exists $\boldsymbol{\mu} \in \Delta_K^\circ$ such that $F_{\boldsymbol{\mu}}(P^*)$ is order-preserving. Order-preservation implies

$$F_{\boldsymbol{\mu}}(P^*)_1 \geq F_{\boldsymbol{\mu}}(P^*)_2 \qquad \text{and} \qquad F_{\boldsymbol{\mu}}(P^*)_3 \geq F_{\boldsymbol{\mu}}(P^*)_4.$$

By (2), this yields

$$(P\boldsymbol{\mu})_1 - (P\boldsymbol{\mu})_2 \geq 0 \qquad \text{and} \qquad (P\boldsymbol{\mu})_3 - (P\boldsymbol{\mu})_4 \geq 0.$$

Adding the two inequalities gives

$$0 \leq (2\beta_1 - \beta_2)\mu_1 + \big(\alpha_1 - (\beta_4 - \beta_3)\big)\mu_2 + \big(\beta_1 - (\alpha_4 - \alpha_2)\big)\mu_3 + (2\alpha_1 - \alpha_3)\mu_4.$$

However, by (10), every coefficient on the right-hand side is strictly negative. Since $\boldsymbol{\mu} \in \Delta_K^\circ$, we have $\mu_i > 0$ for $i = 1, 2, 3, 4$, and therefore the right-hand side is strictly negative, a contradiction.

Hence no full-support sampling distribution can make $F_{\boldsymbol{\mu}}(P^*)$ order-preserving. This proves statement 2 and completes the proof. □

B.4 Learning instance-dependent samplers from comparisons

We show that the instance-dependent samplers in Theorem 3.4 remain valid under estimation error and can be learned from finitely many pairwise comparisons. Throughout this subsection, fix $\beta > 0$ and let $K \geq 2$. Fix a known interval $\mathcal{I} = [\ell, 1 - \ell]$, with $0 \leq \ell < 1/2$, containing all entries of P^*, and assume Ψ is continuous and strictly increasing on $\mathcal{I}$. For a link continuous on $[0, 1]$, one may take $\ell = 0$; for the logit link, require $\ell > 0$. Write $P = \Psi(P^*)$ entrywise and define

$$c_\Psi := \Psi(1/2), \qquad R_\Psi := \Psi(1 - \ell) - \Psi(\ell). \tag{11}$$

For $t \geq 0$ and $a > 0$, let

$$\omega_\Psi(t) := \sup_{\substack{x,y \in \mathcal{I} \\ |x-y| \leq t}} |\Psi(x) - \Psi(y)|,$$

$$\omega_\Psi^{-1}(a) := \sup\{t \in [0, 1 - 2\ell] : \omega_\Psi(t) \leq a\}.$$

Here, $\omega_\Psi(t)$ is the largest possible change in Ψ when its input changes by at most t, while $\omega_\Psi^{-1}(a)$ gives the largest input tolerance that keeps this change at most a. Here and below, $\|M\|_\infty = \max_{i,j} |M_{ij}|$.

For every unordered pair $i < j$, observe n independent comparisons $X_{ij}^{(1)}, \dots, X_{ij}^{(n)} \sim \text{Bernoulli}(P_{ij}^*)$. With $\text{clip}_{\mathcal{I}}$ denoting projection onto $\mathcal{I}$, set

$$\widehat{P}_{ij}^* = \text{clip}_{\mathcal{I}}\left(\frac{1}{n}\sum_{m=1}^{n} X_{ij}^{(m)}\right), \qquad i < j,$$
$$\widehat{P}_{ji}^* = 1 - \widehat{P}_{ij}^*, \qquad \widehat{P}_{ii}^* = 1/2, \qquad \widehat{P} = \Psi(\widehat{P}^*).$$

Clipping preserves reciprocity and cannot increase the estimation error. It also ensures that the transformed estimates are finite for the logit link.

Let $\emptyset \neq \Gamma \subseteq \{(i,j) \in [K]^2 : i \neq j\}$ be the desired ranking constraints. For $0 < \zeta < 1$, define

$$\Delta_K^\zeta \coloneqq \{\boldsymbol{\mu} \in \Delta_K : \mu_k \geq \zeta/K \text{ for every } k\},$$
$$G_\Gamma^\Psi(P^*, \boldsymbol{\mu}) \coloneqq \min_{(i,j)\in\Gamma} (P_i - P_j)^\top \boldsymbol{\mu},$$
$$\rho_{\Gamma,\zeta}^{\Psi,\star}(P^*) \coloneqq \max_{\boldsymbol{\mu}\in\Delta_K^\zeta} G_\Gamma^\Psi(P^*, \boldsymbol{\mu}).$$

Compute the empirical max-margin sampler by the linear program

$$(\widehat{\boldsymbol{\mu}}, \widehat{\rho}) \in \arg \max_{\boldsymbol{\mu}\in\Delta_K^\zeta,\, \rho\in\mathbb{R}} \rho \quad \text{s.t.} \quad (\widehat{P}_i - \widehat{P}_j)^\top \boldsymbol{\mu} \geq \rho, \qquad (i,j) \in \Gamma. \tag{12}$$

Proposition B.2. *If $\|\widehat{P} - P\|_\infty \leq \varepsilon_\Psi$, then*

$$G_\Gamma^\Psi(P^*, \widehat{\boldsymbol{\mu}}) \geq \widehat{\rho} - 2\varepsilon_\Psi \geq \rho_{\Gamma,\zeta}^{\Psi,\star}(P^*) - 4\varepsilon_\Psi.$$

In particular, if $\rho_{\Gamma,\zeta}^{\Psi,\star}(P^*) > 4\varepsilon_\Psi$, the population ΨPO solution with sampler $\widehat{\boldsymbol{\mu}}$ satisfies $\pi_i^\star(P; \widehat{\boldsymbol{\mu}}) > \pi_j^\star(P; \widehat{\boldsymbol{\mu}})$ for every $(i,j) \in \Gamma$.

Proof. For every $\boldsymbol{\mu} \in \Delta_K$ and every pair (i,j),

$$\left|(P_i - P_j)^\top \boldsymbol{\mu} - (\widehat{P}_i - \widehat{P}_j)^\top \boldsymbol{\mu}\right| \leq \sum_k \mu_k \big(|P_{ik} - \widehat{P}_{ik}| + |P_{jk} - \widehat{P}_{jk}|\big) \leq 2\varepsilon_\Psi.$$

Taking minima over Γ gives the same bound for the difference between the population and empirical margins. Applied to $\widehat{\boldsymbol{\mu}}$, this proves the first inequality. Let $\boldsymbol{\mu}^\star \in \Delta_K^\zeta$ attain the population optimum $\rho_{\Gamma,\zeta}^{\Psi,\star}(P^*)$. Since $\boldsymbol{\mu}^\star$ is feasible for the empirical LP, empirical optimality and the preceding margin bound give

$$\begin{aligned}\widehat{\rho} &\geq \min_{(i,j)\in\Gamma} (\widehat{P}_i - \widehat{P}_j)^\top \boldsymbol{\mu}^\star \\ &\geq G_\Gamma^\Psi(P^*, \boldsymbol{\mu}^\star) - 2\varepsilon_\Psi \\ &= \rho_{\Gamma,\zeta}^{\Psi,\star}(P^*) - 2\varepsilon_\Psi.\end{aligned}$$

Subtracting $2\varepsilon_\Psi$ from both sides gives the second inequality. By (2), every strictly positive score difference gives the corresponding strict policy ranking. □

Corollary B.3. *Suppose one of the following cases holds, and define γ_{raw}, γ_Ψ, and Γ accordingly.*

1. Condorcet-top. *The matrix P^* has a Condorcet winner c. Set*

$$\gamma_{\mathrm{raw}} = \min_{j \neq c}(P^*_{cj} - 1/2) > 0,$$
$$\gamma_{\Psi} = \min_{j \neq c}(c_{\Psi} - P_{jc}) > 0, \qquad \Gamma = \{(c, j) : j \neq c\}.$$

2. Smith-top. *Assume Condition 2.1, let $S = \mathrm{Smith}(P^*) \neq [K]$, and suppose there are no off-diagonal ties. Set*

$$\gamma_{\mathrm{raw}} = \min_{i \neq j} |P^*_{ij} - 1/2| > 0,$$
$$\gamma_{\Psi} = \min_{j \notin S,\, k \in S}(c_{\Psi} - P_{jk}) > 0, \qquad \Gamma = S \times ([K] \setminus S).$$

For $\delta \in (0, 1)$, suppose

$$0 < \zeta \leq \frac{\gamma_{\Psi}}{2(\gamma_{\Psi} + R_{\Psi})},$$
$$n \geq \max\left\{\frac{1}{2\gamma^2_{\mathrm{raw}}}, \frac{1}{2\big(\omega^{-1}_{\Psi}(\gamma_{\Psi}/8)\big)^2}\right\} \log \frac{K(K-1)}{\delta}.$$

Then, with probability at least $1 - \delta$, the Condorcet winner or Smith set of $\widehat{P}^*$ equals c or S, respectively. Solving (12) with $\widehat{\Gamma}$ obtained by replacing c or S by its empirical counterpart yields a full-support sampler whose population ΨPO solution is Condorcet-top or Smith-top, respectively.

Proof. We use one concentration event to identify the constraints and control their transformed margins. Set $L = 1 - 2\ell$ and $\varepsilon_n = \sqrt{\log(K(K-1)/\delta)/(2n)}$. Since $P^*_{ij} \in \mathcal{I}$, clipping cannot increase the absolute estimation error. Thus, for each observed pair $i < j$, Hoeffding's inequality gives

$$\Pr\big(|\widehat{P}^*_{ij} - P^*_{ij}| \geq \varepsilon_n\big) \leq \Pr\left(\left|\frac{1}{n}\sum_{m=1}^{n} X^{(m)}_{ij} - P^*_{ij}\right| \geq \varepsilon_n\right)$$
$$\leq 2e^{-2n\varepsilon^2_n}.$$

The diagonal errors are zero, and the errors for (i, j) and (j, i) have the same absolute value by reciprocity. A union bound over the $\binom{K}{2}$ unordered pairs therefore yields

$$\Pr\big(\|\widehat{P}^* - P^*\|_{\infty} \geq \varepsilon_n\big) \leq \sum_{i<j} \Pr\big(|\widehat{P}^*_{ij} - P^*_{ij}| \geq \varepsilon_n\big)$$
$$\leq 2\binom{K}{2} e^{-2n\varepsilon^2_n}$$
$$= K(K-1)e^{-2n\varepsilon^2_n} = \delta.$$

Work on the complementary event. The sample-size assumption implies

$$\|\widehat{P}^* - P^*\|_{\infty} < \varepsilon_n \leq \gamma_{\mathrm{raw}}, \qquad \varepsilon_n \leq r := \omega^{-1}_{\Psi}(\gamma_{\Psi}/8) \leq L.$$

In the Condorcet case, every estimated comparison involving c retains its strict majority direction, so the empirical winner is c. In the Smith case, every majority direction is recovered, so the empirical Smith set is S. Thus $\widehat{\Gamma} = \Gamma$ in either case.

Having shown that $\widehat{\Gamma} = \Gamma$, we next prove that $\|\widehat{P} - P\|_{\infty} < \gamma_{\Psi}/8$. To obtain this bound, we show that ω_{Ψ} is continuous and strictly increasing on $[0, L]$. For $r = \omega^{-1}_{\Psi}(\gamma_{\Psi}/8)$, these properties

will give $\omega_\Psi(r) \le \gamma_\Psi/8$ and $\omega_\Psi(\|\widehat{P}^* - P^*\|_\infty) < \omega_\Psi(r)$, using the previously established bound $\|\widehat{P}^* - P^*\|_\infty < r$.

Since Ψ is continuous on the compact interval $\mathcal{I}$, Ψ is uniformly continuous on $\mathcal{I}$. By the definition of ω_Ψ, uniform continuity of Ψ implies that $\omega_\Psi(h) \to 0$ as $h \downarrow 0$. For $0 \le s < t \le L$, any interval of length at most t can be split into two consecutive pieces of lengths at most s and $t - s$. The triangle inequality therefore gives

$$0 \le \omega_\Psi(t) - \omega_\Psi(s) \le \omega_\Psi(t - s).$$

Thus, ω_Ψ is continuous on $[0, L]$.

To show that ω_Ψ is strictly increasing, fix $0 \le s < t \le L$. Continuity of Ψ and compactness ensure that the supremum defining $\omega_\Psi(s)$ is attained by an interval $[x, y] \subseteq \mathcal{I}$ of length at most s, with $\Psi(y) - \Psi(x) = \omega_\Psi(s)$. Since $t \le L$, this interval can be extended within $\mathcal{I}$ to length t. At least one endpoint moves, and strict monotonicity of Ψ makes the output difference strictly larger. Hence $\omega_\Psi(t) > \omega_\Psi(s)$. Consequently, $\omega_\Psi(r) \le \gamma_\Psi/8$ and, on the same event,

$$e := \|\widehat{P} - P\|_\infty \le \omega_\Psi(\|\widehat{P}^* - P^*\|_\infty) < \omega_\Psi(r) \le \gamma_\Psi/8.$$

It remains to bound the population optimal margin. In the Condorcet case, $\boldsymbol{\nu} = \boldsymbol{e}_c$ gives $(P_c - P_j)^\top \boldsymbol{\nu} \ge \gamma_\Psi$ for every $j \ne c$. In the Smith case, apply the minimax theorem to the skew-symmetric payoff matrix $-(P_{SS} - c_\Psi \mathbf{1}\mathbf{1}^\top)$. Its value is zero, so a column minimax strategy $\boldsymbol{\nu} \in \Delta(S)$ satisfies $P_{SS}\boldsymbol{\nu} \ge c_\Psi \mathbf{1}$. After extending $\boldsymbol{\nu}$ by zero outside S, every $j \notin S$ has $P_j^\top \boldsymbol{\nu} \le c_\Psi - \gamma_\Psi$. Hence $G_\Gamma^\Psi(P^*, \boldsymbol{\nu}) \ge \gamma_\Psi$ in either case. With $\boldsymbol{\mu}_0 = K^{-1}\mathbf{1}$, the full-support mixture $\boldsymbol{\mu} = (1 - \zeta)\boldsymbol{\nu} + \zeta\boldsymbol{\mu}_0 \in \Delta_K^\zeta$ satisfies

$$G_\Gamma^\Psi(P^*, \boldsymbol{\mu}) \ge (1 - \zeta)\gamma_\Psi - \zeta R_\Psi \ge \gamma_\Psi/2.$$

Since $\boldsymbol{\mu} \in \Delta_K^\zeta$, the definition of the population optimum gives

$$\rho_{\Gamma,\zeta}^{\Psi,\star}(P^*) \ge G_\Gamma^\Psi(P^*, \boldsymbol{\mu}) \ge \gamma_\Psi/2.$$

On the event under consideration, $\widehat{\Gamma} = \Gamma$, so $\widehat{\boldsymbol{\mu}}$ solves the empirical LP with these same constraints. Applying Proposition B.2 with $\varepsilon_\Psi = e = \|\widehat{P} - P\|_\infty$ therefore gives

$$\begin{aligned} G_\Gamma^\Psi(P^*, \widehat{\boldsymbol{\mu}}) &\ge \rho_{\Gamma,\zeta}^{\Psi,\star}(P^*) - 4e \\ &\ge \gamma_\Psi/2 - 4e > 0, \end{aligned}$$

where the last inequality uses $e < \gamma_\Psi/8$. The ranking conclusion of the proposition yields $\pi_i^\star(P; \widehat{\boldsymbol{\mu}}) > \pi_j^\star(P; \widehat{\boldsymbol{\mu}})$ for every $(i, j) \in \Gamma$, proving Condorcet-top or Smith-top on this same event. □

B.5 Robustness of order preservation against sampling under strongly transitive preferences

Theorem B.4. *For any $\boldsymbol{\mu} \in \Delta_K^\circ$, ΨPO under $\boldsymbol{\mu}$ is order-preserving over ST matrices.*

Proof. Strong transitivity and reciprocity imply $P_{ik}^* \ge P_{jk}^*$ for every $i < j$ and every k. Since Ψ is strictly increasing, $P_{ik} \ge P_{jk}$. For any sampling distribution $\boldsymbol{\mu}$, and any $i < j$,

$$\sum_{k=1}^K P_{ik}\mu_k \ge \sum_{k=1}^K P_{jk}\mu_k$$

follows from these coordinatewise inequalities. If P^* is strictly transitive, $P_{ii} > P_{ji}$, so the inequality is strict because $\mu_i > 0$. The conclusion follows from (2). □

B.6 Additional social choice axioms that ΨPO satisfies under arbitrary sampling

In this appendix, we record two additional axiomatic properties of the aggregation rule induced by ΨPO: data-pooling consistency and pairwise monotonicity.

For two sampling distributions $\boldsymbol{\mu}^{(1)}, \boldsymbol{\mu}^{(2)} \in \Delta_K^\circ$, define the pooled sampling distribution with mixture weight $\lambda \in (0,1)$

$$\boldsymbol{\mu}^{(\lambda)} := \lambda \boldsymbol{\mu}^{(1)} + (1-\lambda)\boldsymbol{\mu}^{(2)}.$$

Definition B.5. *A family of aggregation rules $F_{\boldsymbol{\mu}} : \mathcal{P} \to \Delta_K$ satisfies* data-pooling consistency *if for any P^* and samplings $\boldsymbol{\mu}^{(1)}$, $\boldsymbol{\mu}^{(2)}$, $F_{\boldsymbol{\mu}^{(1)}}(P^*) = F_{\boldsymbol{\mu}^{(2)}}(P^*)$ implies $F_{\boldsymbol{\mu}^{(\lambda)}}(P^*) = F_{\boldsymbol{\mu}^{(1)}}(P^*)$ for all $\lambda \in (0,1)$.*

Definition B.6. *An aggregation rule $F : \mathcal{P} \to \Delta_K$ satisfies* pairwise monotonicity *if for any preferences P^* and Q^* that differ only on the pair (i,j) and $P^*_{ij} \geq Q^*_{ij}$, the following holds*

$$\frac{F(P^*)_i}{F(P^*)_j} \geq \frac{F(Q^*)_i}{F(Q^*)_j}.$$

If $P^*_{ij} > Q^*_{ij}$ always yields a strict inequality above, we say that F satisfies strict pairwise monotonicity.

Intuitively, pooling consistency captures the idea that aggregating indistinguishable sources of data should not change the outcome: if two sampling distributions lead to the same solution, then any mixture of them should do so as well. Pairwise monotonicity formalizes a basic responsiveness requirement: strengthening the preference of i over j, while leaving all else unchanged, should not decrease the relative weight assigned to i over j.

Theorem B.7. *ΨPO satisfies data-pooling consistency. For any sampling distribution $\boldsymbol{\mu} \in \Delta_K^\circ$, the rule $F_{\boldsymbol{\mu}}$ induced by ΨPO satisfies strict pairwise monotonicity.*

Proof. We first prove data-pooling consistency. Fix $P^* \in \mathcal{P}$, let $P = \Psi(P^*)$, and suppose

$$F_{\boldsymbol{\mu}^{(1)}}(P^*) = F_{\boldsymbol{\mu}^{(2)}}(P^*).$$

It suffices to show that all pairwise log-odds under $\boldsymbol{\mu}^{(\lambda)}$ agree with those under $\boldsymbol{\mu}^{(1)}$.

Notice that for every pair $i, j \in [K]$,

$$\log \frac{F_{\boldsymbol{\mu}^{(\lambda)}}(P^*)_i}{F_{\boldsymbol{\mu}^{(\lambda)}}(P^*)_j} = \beta(P_i - P_j)^\top \boldsymbol{\mu}^{(\lambda)}.$$

Using the definition of $\boldsymbol{\mu}^{(\lambda)}$, this equals

$$\lambda\beta(P_i - P_j)^\top \boldsymbol{\mu}^{(1)} + (1-\lambda)\beta(P_i - P_j)^\top \boldsymbol{\mu}^{(2)}.$$

Therefore,

$$\log \frac{F_{\boldsymbol{\mu}^{(\lambda)}}(P^*)_i}{F_{\boldsymbol{\mu}^{(\lambda)}}(P^*)_j} = \lambda \log \frac{F_{\boldsymbol{\mu}^{(1)}}(P^*)_i}{F_{\boldsymbol{\mu}^{(1)}}(P^*)_j} + (1-\lambda) \log \frac{F_{\boldsymbol{\mu}^{(2)}}(P^*)_i}{F_{\boldsymbol{\mu}^{(2)}}(P^*)_j}.$$

Since $F_{\boldsymbol{\mu}^{(1)}}(P^*) = F_{\boldsymbol{\mu}^{(2)}}(P^*)$, the two log-odds on the right-hand side are equal. Hence

$$\log \frac{F_{\boldsymbol{\mu}^{(\lambda)}}(P^*)_i}{F_{\boldsymbol{\mu}^{(\lambda)}}(P^*)_j} = \log \frac{F_{\boldsymbol{\mu}^{(1)}}(P^*)_i}{F_{\boldsymbol{\mu}^{(1)}}(P^*)_j}$$

for every $i, j \in [K]$. Thus all pairwise odds of $F_{\boldsymbol{\mu}^{(\lambda)}}(P^*)$ and $F_{\boldsymbol{\mu}^{(1)}}(P^*)$ agree. Since both are probability distributions, they are identical:

$$F_{\boldsymbol{\mu}^{(\lambda)}}(P^*) = F_{\boldsymbol{\mu}^{(1)}}(P^*).$$

This proves data-pooling consistency.

We now prove pairwise monotonicity. Let $P^*, Q^* \in \mathcal{P}$ differ only on the pair (i, j), and assume $P^*_{ij} \geq Q^*_{ij}$. Since $P^*, Q^* \in \mathcal{P}$, this also implies

$$P^*_{ji} = 1 - P^*_{ij} \leq 1 - Q^*_{ij} = Q^*_{ji}.$$

Let $P \coloneqq \Psi(P^*)$ and $Q \coloneqq \Psi(Q^*)$. By (2),

$$\log \frac{F_{\boldsymbol{\mu}}(P^*)_i}{F_{\boldsymbol{\mu}}(P^*)_j} - \log \frac{F_{\boldsymbol{\mu}}(Q^*)_i}{F_{\boldsymbol{\mu}}(Q^*)_j} = \beta \left[(P_i - P_j)^\top \boldsymbol{\mu} - (Q_i - Q_j)^\top \boldsymbol{\mu} \right].$$

Because P^* and Q^* differ only on entries (i, j) and (j, i), all other terms cancel, the right-hand side becomes

$$\beta \left\{ \mu_j \left[\Psi(P^*_{ij}) - \Psi(Q^*_{ij}) \right] + \mu_i \left[\Psi(Q^*_{ji}) - \Psi(P^*_{ji}) \right] \right\}.$$

Since Ψ is non-decreasing,

$$\Psi(P^*_{ij}) - \Psi(Q^*_{ij}) \geq 0$$

and

$$\Psi(Q^*_{ji}) - \Psi(P^*_{ji}) \geq 0.$$

Since $\mu_i, \mu_j > 0$, it follows that

$$\log \frac{F_{\boldsymbol{\mu}}(P^*)_i}{F_{\boldsymbol{\mu}}(P^*)_j} \geq \log \frac{F_{\boldsymbol{\mu}}(Q^*)_i}{F_{\boldsymbol{\mu}}(Q^*)_j}.$$

Exponentiating both sides gives

$$\frac{F_{\boldsymbol{\mu}}(P^*)_i}{F_{\boldsymbol{\mu}}(P^*)_j} \geq \frac{F_{\boldsymbol{\mu}}(Q^*)_i}{F_{\boldsymbol{\mu}}(Q^*)_j}.$$

Thus $F_{\boldsymbol{\mu}}$ satisfies pairwise monotonicity.

Finally, suppose that $P^*_{ij} > Q^*_{ij}$. Since Ψ is strictly increasing, we have

$$\Psi(P^*_{ij}) - \Psi(Q^*_{ij}) > 0$$

and

$$\Psi(Q^*_{ji}) - \Psi(P^*_{ji}) > 0.$$

Consequently,

$$\frac{F_{\boldsymbol{\mu}}(P^*)_i}{F_{\boldsymbol{\mu}}(P^*)_j} > \frac{F_{\boldsymbol{\mu}}(Q^*)_i}{F_{\boldsymbol{\mu}}(Q^*)_j}.$$

This completes the proof.

□

B.7 Proof of Theorem 3.9

Theorem 3.9. *Let $P^* \in \mathcal{P}$ be ST with order $1 \succ \cdots \succ K$, and let $P = \Psi(P^*)$. Let $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \Delta_K^\circ$ with $\boldsymbol{\mu} \neq \boldsymbol{\mu}'$ and $\boldsymbol{\mu} - \boldsymbol{\mu}'$ non-increasing. Define $\boldsymbol{\theta}(\boldsymbol{\mu}) := \beta^{-1}\big(\log \boldsymbol{\pi}^\star(P; \boldsymbol{\mu}) - \log \boldsymbol{\pi}_{\mathrm{ref}}\big)$, where $\boldsymbol{\pi}^\star(P; \boldsymbol{\mu})$ is the population ΨPO optimizer in (2). Then, for any $i < j$ such that $P_i \succeq P_j$,*

$$\theta_i(\boldsymbol{\mu}) - \theta_j(\boldsymbol{\mu}) \geq \theta_i(\boldsymbol{\mu}') - \theta_j(\boldsymbol{\mu}') \geq 0. \tag{3}$$

The second inequality is strict if and only if $P_i \neq P_j$. The first inequality is strict whenever $P_i \succ P_j$.

Proof. Fix any pair $1 \leq i < j \leq K$ and define the transformed row-difference vector

$$\boldsymbol{a} := P_i - P_j \in \mathbb{R}^K, \qquad a_k := P_{ik} - P_{jk}.$$

By the closed-form ΨPO solution in (2), the optimizer under sampling distribution $\boldsymbol{\mu}$ satisfies

$$\log \pi_\ell^\star(P; \boldsymbol{\mu}) - \log \pi_{\mathrm{ref},\ell} = \beta(P\boldsymbol{\mu})_\ell + c(\boldsymbol{\mu}), \qquad \forall \ell \in [K],$$

where $c(\boldsymbol{\mu}) \in \mathbb{R}$ is a normalization constant independent of ℓ. Therefore,

$$\theta_\ell(\boldsymbol{\mu}) = (P\boldsymbol{\mu})_\ell + \frac{c(\boldsymbol{\mu})}{\beta}.$$

Taking the difference between coordinates i and j, the normalization constant cancels and gives

$$\theta_i(\boldsymbol{\mu}) - \theta_j(\boldsymbol{\mu}) = (P_i - P_j)^\top \boldsymbol{\mu} = \boldsymbol{a}^\top \boldsymbol{\mu}. \tag{13}$$

Analogously,

$$\theta_i(\boldsymbol{\mu}') - \theta_j(\boldsymbol{\mu}') = \boldsymbol{a}^\top \boldsymbol{\mu}'. \tag{14}$$

Let $\boldsymbol{x} := \boldsymbol{\mu} - \boldsymbol{\mu}'$. Since $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$ are probability distributions, $\sum_{k=1}^K x_k = 0$. By assumption, $x_1 \geq x_2 \geq \cdots \geq x_K$. We use the following equivalent characterization of the Maj condition.

Lemma B.8 (Equivalent characterization of Maj)**.** *For any $\boldsymbol{a} = (a_1, \ldots, a_K) \in \mathbb{R}^K$, the following statements are equivalent:*

(i) $\boldsymbol{a}$ is Maj, i.e.,

$$\frac{1}{m}\sum_{k=1}^{m} a_k \geq \frac{1}{K}\sum_{k=1}^{K} a_k, \qquad \forall m = 1, \ldots, K-1.$$

(ii) For every $\boldsymbol{x} \in \mathbb{R}^K$ satisfying

$$x_1 \geq x_2 \geq \cdots \geq x_K, \qquad \sum_{k=1}^{K} x_k = 0,$$

one has $\boldsymbol{a}^\top \boldsymbol{x} \geq 0$.

Moreover, if $\boldsymbol{a}$ is strictly Maj, then $\boldsymbol{a}^\top \boldsymbol{x} > 0$ for every nonzero such $\boldsymbol{x}$. More generally, for a fixed $\boldsymbol{x}$, strictness holds if and only if $\boldsymbol{a}^\top \boldsymbol{x} > 0$.

Proof of Lemma B.8. For $m = 1, \ldots, K-1$, define

$$\boldsymbol{x}^{(m)} := (\underbrace{K-m, \ldots, K-m}_{m \text{ times}}, \underbrace{-m, \ldots, -m}_{K-m \text{ times}}).$$

Each $\boldsymbol{x}^{(m)}$ is non-increasing and has zero sum. Every non-increasing zero-sum vector $\boldsymbol{x}$ admits the decomposition

$$\boldsymbol{x} = \frac{1}{K}\sum_{m=1}^{K-1}(x_m - x_{m+1})\boldsymbol{x}^{(m)}, \tag{15}$$

where each coefficient $x_m - x_{m+1}$ is nonnegative. Furthermore,

$$\boldsymbol{a}^\top \boldsymbol{x}^{(m)} = (K-m)\sum_{k=1}^{m} a_k - m\sum_{k=m+1}^{K} a_k = K\sum_{k=1}^{m} a_k - m\sum_{k=1}^{K} a_k.$$

Thus, $\boldsymbol{a}^\top \boldsymbol{x}^{(m)} \geq 0$ is exactly the Maj inequality for prefix m. This proves both directions: if all Maj inequalities hold, then (15) implies $\boldsymbol{a}^\top \boldsymbol{x} \geq 0$ for all non-increasing zero-sum $\boldsymbol{x}$; conversely, applying condition (ii) to each $\boldsymbol{x}^{(m)}$ yields every Maj inequality.

If $\boldsymbol{a}$ is strictly Maj, then $\boldsymbol{a}^\top \boldsymbol{x}^{(m)} > 0$ for all m. Hence, for any nonzero $\boldsymbol{x}$ in the cone, at least one coefficient $x_m - x_{m+1}$ is positive in (15), which implies $\boldsymbol{a}^\top \boldsymbol{x} > 0$. □

Since $P_i \succeq P_j$ by assumption, Lemma B.8 gives

$$\boldsymbol{a}^\top(\boldsymbol{\mu} - \boldsymbol{\mu}') \geq 0.$$

Combining this with (13)–(14), we obtain

$$\theta_i(\boldsymbol{\mu}) - \theta_j(\boldsymbol{\mu}) \geq \theta_i(\boldsymbol{\mu}') - \theta_j(\boldsymbol{\mu}').$$

If $(P_i - P_j)^\top(\boldsymbol{\mu} - \boldsymbol{\mu}') > 0$, the inequality is strict. In particular, by Lemma B.8, this strictness is guaranteed whenever $P_i - P_j$ is strictly Maj and $\boldsymbol{\mu} \neq \boldsymbol{\mu}'$.

It remains to show that the baseline gap is nonnegative. Since P^* is ST with order $1 \succ \cdots \succ K$ and Ψ is non-decreasing, the transformed matrix $P = \Psi(P^*)$ satisfies $P_{ik} \geq P_{jk}$ for all k whenever $i < j$. Hence $\boldsymbol{a} \geq 0$ coordinatewise. If $P_i \neq P_j$, at least one coordinate of $\boldsymbol{a}$ is strictly positive. Since $\boldsymbol{\mu}' \in \Delta_K^\circ$, we have

$$\boldsymbol{a}^\top \boldsymbol{\mu}' \geq 0.$$

Equality holds if and only if $P_i = P_j$. Using (14), this gives

$$\theta_i(\boldsymbol{\mu}') - \theta_j(\boldsymbol{\mu}') \geq 0.$$

Combining the preceding displays proves (3) and, under the additional strict-alignment condition, (3) holds strictly. □

B.8 General discussion of majorization and head-tail separation

Corollary B.9. *Consider IPO ($P = P^*$) and let P be induced by the BT model $P_{ij} = \sigma(r_i - r_j)$, $r_1 > r_2 > \cdots > r_K$. Given an integer $2 \leq H \leq K/2$ and $\delta > 3\ln 2$, we say P satisfies* Head–Tail Separation (HTS) *with parameters (H, δ) if $[K]$ decomposes into a head set $\mathcal{H} = \{1, \ldots, H\}$ and a tail set $\mathcal{T} = \{H+1, \ldots, K\}$ such that:*
1. (heads dominate tails) for all $k \in \mathcal{H}$, $r_k \geq 0$, while for all $k \in \mathcal{T}$, $r_k \leq -\delta$;
2. (bounded head span) $r_1 - r_H \leq 1$.
Then for every $i < j$ in $\mathcal{H}$, $P_i - P_j$ is strictly Maj. Consequently, for any distinct $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \Delta_K^\circ$ with $\boldsymbol{\mu} - \boldsymbol{\mu}'$ non-increasing, both inequalities in (3) are strict for every $i < j$ in $\mathcal{H}$.

Corollary B.9 captures a common large-response regime: only a small head set contains meaningful competing responses, while the long tail is uniformly dominated. The bounded head span controls every head coordinate of each row difference. No minimum gap between head scores is required.

We first give a coordinate lemma that converts head–tail bounds into the Maj condition required by Theorem 3.9.

Lemma B.10 (Head–tail bounds imply Maj)**.** *Let* $\boldsymbol{a} = (a_1, \ldots, a_K) \in \mathbb{R}_+^K$*. Fix* $H \in \{1, \ldots, K-1\}$*. Suppose there exist constants* $B \geq \epsilon \geq 0$ *such that*

(L1) ***(Bounded element)*** $0 \leq a_k \leq B$ *for all* $k \in [K]$*;*

(L2) ***(Tail upper bound and monotonicity)*** $\epsilon \geq a_{H+1} \geq a_{H+2} \geq \cdots \geq a_K$*;*

(L3) ***(Head lower bound)*** $a_k \geq \alpha$ *for all* $k \leq H$*, where*

$$\alpha \geq \frac{HB + (K-H)\epsilon}{K}.$$

Then $\boldsymbol{a}$ *is Maj. If the inequality in (L3) is strict, i.e.,*

$$a_k > \frac{HB + (K-H)\epsilon}{K}, \qquad \forall k \leq H,$$

then $\boldsymbol{a}$ *is strictly Maj.*

Proof of Lemma B.10. Let

$$S := \sum_{k=1}^{K} a_k, \qquad A_m := \sum_{k=1}^{m} a_k.$$

By (L1) on the head and (L2) on the tail,

$$S \leq HB + (K-H)\epsilon. \tag{16}$$

First consider $m \leq H$. By (L3) and (16),

$$A_m \geq m\alpha \geq \frac{m}{K}\{HB + (K-H)\epsilon\} \geq \frac{m}{K}S.$$

Thus the Maj inequality holds for every $m \leq H$.

Now consider $m > H$. Let

$$R_m := \sum_{k=m+1}^{K} a_k$$

be the suffix mass after index m, and let

$$\bar{a}_{\mathcal{T}} := \frac{1}{K-H} \sum_{k=H+1}^{K} a_k$$

be the average tail value. Since the tail sequence is non-increasing by (L2), every suffix average is at most the full tail average:

$$\frac{R_m}{K-m} \leq \bar{a}_{\mathcal{T}}.$$

Moreover, (L2) gives $\bar{a}_{\mathcal{T}} \leq \epsilon$, while (L3) and $B \geq \epsilon$ imply the head average is at least $\alpha \geq \epsilon$. Hence the global average S/K, which is a weighted average of the head and tail averages, is at least the tail average:

$$\bar{a}_{\mathcal{T}} \leq \frac{S}{K}.$$

Therefore,

$$R_m \leq (K-m)\frac{S}{K}.$$

Since $A_m = S - R_m$, we obtain

$$A_m \geq S - (K-m)\frac{S}{K} = \frac{m}{K}S.$$

Thus the Maj inequality holds for every $m > H$ as well.

If (L3) is strict, then the inequalities for $m \le H$ are strict. In addition, the head average is strictly larger than the tail average, so $\bar{a}_{\mathcal{T}} < S/K$ and the suffix argument above is strict for every $m > H$. Hence all Maj inequalities are strict. □

Now we prove Corollary B.9.

Proof. We verify the three conditions of Lemma B.10. Fix $i < j$ in $\mathcal{H}$, let $\Delta := r_i - r_j > 0$, and put $\boldsymbol{a} := P_i - P_j$. For every $k \in [K]$,

$$a_k = \int_{r_j - r_k}^{r_i - r_k} \sigma'(u)\, du.$$

Since $0 < \sigma'(u) \le 1/4$, we have $0 < a_k \le \Delta/4$ for all k. For $k \in \mathcal{H}$, the integration interval lies in $[-1, 1]$ because $r_1 - r_H \le 1$, so $a_k \ge \Delta\sigma'(1)$. For $k \in \mathcal{T}$, its left endpoint is at least δ, so $\sigma'(u) \le e^{-u}$ gives $a_k \le \Delta e^{-\delta}$. Moreover, since σ' is non-increasing on $[0, \infty)$, the derivative of $\sigma(u+\Delta) - \sigma(u)$ is $\sigma'(u+\Delta) - \sigma'(u) \le 0$. Since $r_j - r_k$ increases along the tail, $a_{H+1} \ge \cdots \ge a_K$.

Apply the lemma with

$$B = \frac{\Delta}{4}, \qquad \epsilon = \Delta e^{-\delta}, \qquad \alpha = \Delta\sigma'(1).$$

We have $B \ge \epsilon \ge 0$. Writing $p := H/K \le 1/2$ and using $e^{-\delta} < 1/8$,

$$\begin{aligned} \frac{HB + (K-H)\epsilon}{K} &= \Delta\left(\frac{p}{4} + (1-p)e^{-\delta}\right) < \Delta\left(\frac{p}{4} + \frac{1-p}{8}\right) \\ &\le \frac{3\Delta}{16} < \Delta\sigma'(1) = \alpha, \end{aligned}$$

where $\sigma'(1) = e/(1+e)^2 > 3/16$. Thus the strict part of Lemma B.10 gives that $P_i - P_j$ is strictly Maj. Finally, BT preferences are ST and $a_k > 0$ implies $P_i \ne P_j$, so Theorem 3.9 gives both strict inequalities.

□

C Omitted Proofs in Section 4

C.1 Proof of Proposition 4.2

Proposition 4.2. *Let $K = 3$ and fix $a \in (0, 1/2)$. Consider the following preference matrix $P^* = (\frac{1}{2}, \frac{1}{2}+a, \frac{1}{2}-a; \frac{1}{2}-a, \frac{1}{2}, \frac{1}{2}+a; \frac{1}{2}+a, \frac{1}{2}-a, \frac{1}{2})$ so that $1 \succ 2$, $2 \succ 3$, and $3 \succ 1$ with probability $1/2 + a$. Let $P = \Psi(P^*)$ with symmetric Ψ in Condition 2.1, define $a_\Psi := (\Psi(1/2+a) - \Psi(1/2-a))/2$, and set $\boldsymbol{\pi}_{\mathrm{ref}} = \boldsymbol{\pi}_0 = (1/3, 1/3, 1/3)$. Then the uniform policy $\boldsymbol{\pi}^\star = (1/3, 1/3, 1/3)$ is a fixed point of the MRS-ΨPO dynamics (6). Moreover, if*

$$\alpha^2 + \frac{\beta^2\lambda^2 a_\Psi^2}{3} > 1, \tag{7}$$

then $\boldsymbol{\pi}^\star$ is linearly unstable. In particular, nearby trajectories have an unstable oscillatory mode and do not converge locally to $\boldsymbol{\pi}^\star$.

Proof. Let $c_\Psi := \Psi(1/2)$ and

$$a_\Psi := \frac{\Psi(1/2+a) - \Psi(1/2-a)}{2}.$$

By the assumed symmetry condition $\Psi(1/2+a)+\Psi(1/2-a)=2\Psi(1/2)$, the transformed payoff matrix $P=\Psi(P^*)$ can be written as

$$P=\begin{pmatrix} c_\Psi & c_\Psi+a_\Psi & c_\Psi-a_\Psi \\ c_\Psi-a_\Psi & c_\Psi & c_\Psi+a_\Psi \\ c_\Psi+a_\Psi & c_\Psi-a_\Psi & c_\Psi \end{pmatrix} = c_\Psi \mathbf{1}\mathbf{1}^\top + A_\Psi,$$

where

$$A_\Psi=\begin{pmatrix} 0 & a_\Psi & -a_\Psi \\ -a_\Psi & 0 & a_\Psi \\ a_\Psi & -a_\Psi & 0 \end{pmatrix}, \qquad A_\Psi^\top=-A_\Psi.$$

Let $\boldsymbol{\pi}^\star:=(1/3,1/3,1/3)$ and assume $\boldsymbol{\pi}_{\mathrm{ref}}=\boldsymbol{\pi}_0=\boldsymbol{\pi}^\star$. Define the MRS-ΨPO map $T:\Delta_3^\circ\to\Delta_3^\circ$ by

$$T(\boldsymbol{\pi})=\mathrm{softmax}\Big(\alpha\log\boldsymbol{\pi}+(1-\alpha)\log\boldsymbol{\pi}^\star+\beta\,P\big(\lambda\boldsymbol{\pi}+(1-\lambda)\boldsymbol{\pi}^\star\big)\Big),$$

so that the dynamics is $\boldsymbol{\pi}_{t+1}=T(\boldsymbol{\pi}_t)$.

First of all, it is straightforward to see $\boldsymbol{\pi}^\star$ is a fixed point: each row of P sums to $3c_\Psi$, hence $P\boldsymbol{\pi}^\star=c_\Psi\mathbf{1}$. Therefore, $P(\lambda\boldsymbol{\pi}^\star+(1-\lambda)\boldsymbol{\pi}^\star)=P\boldsymbol{\pi}^\star=c_\Psi\mathbf{1}$ and

$$T(\boldsymbol{\pi}^\star)=\mathrm{softmax}\Big(\log\boldsymbol{\pi}^\star+\beta c_\Psi\mathbf{1}\Big)=\boldsymbol{\pi}^\star,$$

since softmax is invariant to additive constants.

Next we analyze the stability of $\boldsymbol{\pi}^\star$. We compute the Jacobian of T at $\boldsymbol{\pi}^\star$: write $T(\boldsymbol{\pi})=\mathrm{softmax}(g(\boldsymbol{\pi}))$ where

$$g(\boldsymbol{\pi})=\alpha\log\boldsymbol{\pi}+(1-\alpha)\log\boldsymbol{\pi}^\star+\beta\,P\big(\lambda\boldsymbol{\pi}+(1-\lambda)\boldsymbol{\pi}^\star\big).$$

The Jacobian of softmax is $D[\mathrm{softmax}(\boldsymbol{z})]=J(\boldsymbol{p}):=\mathrm{Diag}(\boldsymbol{p})-\boldsymbol{p}\boldsymbol{p}^\top$ for $\boldsymbol{p}=\mathrm{softmax}(\boldsymbol{z})$. By the chain rule, $D[T(\boldsymbol{\pi})]=J(T(\boldsymbol{\pi}))\,D[g(\boldsymbol{\pi})]$. At $\boldsymbol{\pi}^\star$, $T(\boldsymbol{\pi}^\star)=\boldsymbol{\pi}^\star$, hence $D[T(\boldsymbol{\pi}^\star)]=J(\boldsymbol{\pi}^\star)\,D[g(\boldsymbol{\pi}^\star)]$. We compute

$$D[g(\boldsymbol{\pi}^\star)]=\alpha\,\mathrm{Diag}(1/\boldsymbol{\pi}^\star)+\beta\lambda P=3\alpha I+\beta\lambda P,$$

since $\pi_i^\star=1/3$. Also,

$$J(\boldsymbol{\pi}^\star)=\mathrm{Diag}(\boldsymbol{\pi}^\star)-\boldsymbol{\pi}^\star(\boldsymbol{\pi}^\star)^\top=\frac{1}{3}I-\frac{1}{9}\mathbf{1}\mathbf{1}^\top.$$

Restricting to the tangent space $\mathbf{1}^\perp:=\{\boldsymbol{v}:\mathbf{1}^\top\boldsymbol{v}=0\}$, we have $J(\boldsymbol{\pi}^\star)\boldsymbol{v}=\frac{1}{3}\boldsymbol{v}$ and $P\boldsymbol{v}=A_\Psi\boldsymbol{v}$. Therefore, for all $\boldsymbol{v}\in\mathbf{1}^\perp$,

$$D[T(\boldsymbol{\pi}^\star)]\boldsymbol{v}=\frac{1}{3}(3\alpha I+\beta\lambda P)\boldsymbol{v}=\Big(\alpha I+\frac{\beta\lambda}{3}A_\Psi\Big)\boldsymbol{v}.$$

Hence the relevant linearization on the simplex is

$$M:=D[T(\boldsymbol{\pi}^\star)]\big|_{\mathbf{1}^\perp}=\alpha I+\frac{\beta\lambda}{3}A_\Psi\big|_{\mathbf{1}^\perp}.$$

A direct multiplication yields $A_\Psi^2=-3a_\Psi^2\left(I-\frac{1}{3}\mathbf{1}\mathbf{1}^\top\right)$, so on $\mathbf{1}^\perp$ we have $A_\Psi^2=-3a_\Psi^2 I$. Thus the eigenvalues of $A_\Psi|_{\mathbf{1}^\perp}$ are $\pm i\sqrt{3}\,a_\Psi$, and the eigenvalues of M are

$$\lambda_\pm=\alpha\pm i\frac{\beta\lambda}{3}\sqrt{3}\,a_\Psi=\alpha\pm i\frac{\beta\lambda a_\Psi}{\sqrt{3}},$$

with modulus

$$|\lambda_\pm|=\sqrt{\alpha^2+\frac{a_\Psi^2\beta^2\lambda^2}{3}}.$$

Therefore, if (7) holds, $|\lambda_\pm| > 1$ and the spectral radius of $D[T(\boldsymbol{\pi}^\star)]|_{\mathbf{1}^\perp}$ exceeds 1, implying that $\boldsymbol{\pi}^\star$ is a linearly unstable fixed point.

Since $\lambda_\pm$ are complex with nonzero imaginary part whenever $\beta\lambda a_\Psi \neq 0$, the linearized map on $\mathbf{1}^\perp$ is a rotation combined with expansion by factor $\rho := |\lambda_\pm| > 1$. Thus, trajectories starting sufficiently close to $\boldsymbol{\pi}^\star$, but not at $\boldsymbol{\pi}^\star$, initially move away with a rotational component and do not converge to $\boldsymbol{\pi}^\star$. This completes the proof. □

C.2 Proof of Theorem 4.3

Theorem 4.3. *For any $\boldsymbol{\pi}_{\text{ref}}, \boldsymbol{\pi}_0 \in \Delta_K^\circ$, consider the MRS-ΨPO dynamics $\boldsymbol{\pi}_{t+1} = \text{MRS}(\boldsymbol{\pi}_t; \alpha, \beta, \lambda)$ defined in (6). Let $\Pi := I - \frac{1}{K}\mathbf{1}\mathbf{1}^\top, \widetilde{P} := \Pi P$. If*

$$\alpha + \frac{\beta\lambda}{2}\|\widetilde{P}\|_2 < 1, \tag{8}$$

then there exists a unique fixed point $\boldsymbol{\pi}^\star \in \Delta_K^\circ$, and for any initialization $\boldsymbol{\pi}_1 \in \Delta_K^\circ$, the iterates $\{\boldsymbol{\pi}_t\}$ converge to $\boldsymbol{\pi}^\star$.

In addition, for symmetric Ψ and $A_\Psi := P - \Psi(1/2)\mathbf{1}\mathbf{1}^\top, R_\Psi := \max_{i,j} |(A_\Psi)_{ij}|$. If A_Ψ has at most d nonzero entries in each row, then (8) is implied by

$$\alpha \; + \; \frac{\beta\lambda d R_\Psi}{2} \; < \; 1.$$

For IPO, $R_\Psi \leq 1/2$, yielding a condition $\alpha + \beta\lambda d/4 < 1$.

Proof. Define the MRS-ΨPO update map $T : \Delta_K^\circ \to \Delta_K^\circ$ by

$$T(\boldsymbol{\pi}) := \text{softmax}\Big(\alpha \log \boldsymbol{\pi} + (1-\alpha)\log \boldsymbol{\pi}_{\text{ref}} + \beta\, P\big(\lambda\boldsymbol{\pi} + (1-\lambda)\boldsymbol{\pi}_0\big)\Big),$$

where $P = \Psi(P^*)$ is the transformed payoff matrix. The dynamics is $\boldsymbol{\pi}_{t+1} = T(\boldsymbol{\pi}_t)$. We will show that under (8), T is a contraction in centered logit coordinates, which implies a unique fixed point and global convergence.

Let $\mathbf{1} \in \mathbb{R}^K$ be the all-ones vector and let $\Pi := I - \frac{1}{K}\mathbf{1}\mathbf{1}^\top$ be the orthogonal projector onto $\mathbf{1}^\perp := \{x \in \mathbb{R}^K : \mathbf{1}^\top x = 0\}$. For any $\boldsymbol{\pi} \in \Delta_K^\circ$, define the centered logit

$$\boldsymbol{z}(\boldsymbol{\pi}) := \Pi \log \boldsymbol{\pi} \in \mathbf{1}^\perp.$$

Since $\log \boldsymbol{\pi} = \boldsymbol{z}(\boldsymbol{\pi}) + c(\boldsymbol{\pi})\mathbf{1}$ for some scalar $c(\boldsymbol{\pi})$ and $\text{softmax}(\boldsymbol{x} + c\mathbf{1}) = \text{softmax}(\boldsymbol{x})$, we have the bijection

$$\boldsymbol{\pi} = \text{softmax}(\boldsymbol{z}(\boldsymbol{\pi})), \qquad \boldsymbol{z}(\text{softmax}(\boldsymbol{x})) = \boldsymbol{x} \quad \text{for all } \boldsymbol{x} \in \mathbf{1}^\perp.$$

Hence it suffices to study the induced map on $\mathbf{1}^\perp$.

Let $\boldsymbol{x}_t := \boldsymbol{z}(\boldsymbol{\pi}_t) \in \mathbf{1}^\perp$ so that $\boldsymbol{\pi}_t = \text{softmax}(\boldsymbol{x}_t)$. Taking centered logits of $\boldsymbol{\pi}_{t+1} = T(\boldsymbol{\pi}_t)$ gives

$$\begin{aligned}
\boldsymbol{x}_{t+1} = \Pi \log \boldsymbol{\pi}_{t+1} &= \Pi\Big(\alpha \log \boldsymbol{\pi}_t + (1-\alpha)\log \boldsymbol{\pi}_{\text{ref}} + \beta\, P\big(\lambda\boldsymbol{\pi}_t + (1-\lambda)\boldsymbol{\pi}_0\big)\Big) \\
&= \alpha\Pi \log \boldsymbol{\pi}_t + \Pi\Big((1-\alpha)\log \boldsymbol{\pi}_{\text{ref}} + \beta(1-\lambda)P\boldsymbol{\pi}_0\Big) + \beta\lambda\, \Pi P\, \boldsymbol{\pi}_t.
\end{aligned}$$

Define the constant vector

$$\mathbf{b} := \Pi\Big((1-\alpha)\log \boldsymbol{\pi}_{\text{ref}} + \beta(1-\lambda)P\boldsymbol{\pi}_0\Big) \in \mathbf{1}^\perp.$$

Thus the centered-logit dynamics becomes

$$\boldsymbol{x}_{t+1} = F(\boldsymbol{x}_t) := \alpha\boldsymbol{x}_t + \beta\lambda\, \Pi P\, \text{softmax}(\boldsymbol{x}_t) + \mathbf{b}, \qquad \boldsymbol{x}_t \in \mathbf{1}^\perp.$$

Let $\boldsymbol{\pi} = \mathrm{softmax}(\boldsymbol{x})$. The Jacobian of softmax is

$$D\left[\mathrm{softmax}(\boldsymbol{x})\right] = J(\boldsymbol{\pi}) := \mathrm{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^\top,$$

which satisfies the uniform spectral norm bound $\|J(\boldsymbol{\pi})\|_2 \le \frac{1}{2}$ for all $\boldsymbol{\pi} \in \Delta_K$. Indeed, for any unit vector $\boldsymbol{v}$,

$$\boldsymbol{v}^\top J(\boldsymbol{\pi})\boldsymbol{v} = \sum_{i=1}^K \pi_i v_i^2 - \Big(\sum_{i=1}^K \pi_i v_i\Big)^2 = \mathrm{Var}(v_I), \qquad I \sim \boldsymbol{\pi},$$

and $\mathrm{Var}(v_I) \le \frac{(\max_i v_i - \min_i v_i)^2}{4} \le \frac{(\sqrt{2})^2}{4} = \frac{1}{2}$. Taking the supremum over $\|\boldsymbol{v}\|_2 = 1$ gives $\|J(\boldsymbol{\pi})\|_2 \le \frac{1}{2}$.

Differentiating F yields, for $\boldsymbol{x} \in \mathbf{1}^\perp$ and $\boldsymbol{\pi} = \mathrm{softmax}(\boldsymbol{x})$,

$$D[F(\boldsymbol{x})] = \alpha I + \beta\lambda\,\Pi P\, J(\boldsymbol{\pi}).$$

Using submultiplicativity,

$$\|D[F(\boldsymbol{x})]\|_2 \le \alpha + \beta\lambda\,\|\Pi P\|_2\,\|J(\boldsymbol{\pi})\|_2 \le \alpha + \frac{\beta\lambda}{2}\|\widetilde{P}\|_2, \tag{17}$$

where $\widetilde{P} = \Pi P$. Therefore, if (8) holds, then F is a contraction on the complete metric space $(\mathbf{1}^\perp, \|\cdot\|_2)$:

$$\|F(\boldsymbol{x}) - F(\boldsymbol{y})\|_2 \le L\|\boldsymbol{x} - \boldsymbol{y}\|_2, \qquad \forall\, \boldsymbol{x}, \boldsymbol{y} \in \mathbf{1}^\perp,$$

for some $L < 1$.

By the Banach fixed-point theorem, F admits a unique fixed point $\boldsymbol{x}^\star \in \mathbf{1}^\perp$, and for any initialization $\boldsymbol{x}_1 \in \mathbf{1}^\perp$, the iterates $\boldsymbol{x}_{t+1} = F(\boldsymbol{x}_t)$ satisfy $\boldsymbol{x}_t \to \boldsymbol{x}^\star$. Define

$$\boldsymbol{\pi}^\star := \mathrm{softmax}(\boldsymbol{x}^\star) \in \Delta_K^\circ.$$

Since the map $\boldsymbol{x} \mapsto \mathrm{softmax}(\boldsymbol{x})$ is continuous, we have $\boldsymbol{\pi}_t = \mathrm{softmax}(\boldsymbol{x}_t) \to \mathrm{softmax}(\boldsymbol{x}^\star) = \boldsymbol{\pi}^\star$. Finally, because $\boldsymbol{x}^\star = F(\boldsymbol{x}^\star)$ is equivalent to $\boldsymbol{\pi}^\star = T(\boldsymbol{\pi}^\star)$ via the bijection between $\mathbf{1}^\perp$ and Δ_K°, the limit $\boldsymbol{\pi}^\star$ is precisely the unique fixed point of the original MRS-ΨPO map T on Δ_K°.

It remains to prove the explicit sparse sufficient condition. Suppose Ψ satisfies $\Psi(q) + \Psi(1-q) = 2\Psi(1/2)$ on the entries of P^* and let $c_\Psi := \Psi(1/2)$. Then

$$A_\Psi := P - c_\Psi \mathbf{1}\mathbf{1}^\top$$

is skew-symmetric, because for $i \ne j$,

$$(A_\Psi)_{ji} = \Psi(P^*_{ji}) - c_\Psi = \Psi(1 - P^*_{ij}) - c_\Psi = -(\Psi(P^*_{ij}) - c_\Psi) = -(A_\Psi)_{ij},$$

and $(A_\Psi)_{ii} = 0$. Since $\Pi(c_\Psi \mathbf{1}\mathbf{1}^\top) = 0$, we have $\Pi P = \Pi A_\Psi$ and therefore $\|\Pi P\|_2 \le \|A_\Psi\|_2$. If A_Ψ has at most d nonzero entries in each row and $R_\Psi = \max_{i,j} |(A_\Psi)_{ij}|$, then

$$\|A_\Psi\|_\infty \le dR_\Psi, \qquad \|A_\Psi\|_1 = \|A_\Psi\|_\infty \le dR_\Psi,$$

where the equality follows from skew-symmetry. Using $\|M\|_2 \le \sqrt{\|M\|_1\|M\|_\infty}$ gives

$$\|\Pi P\|_2 \le \|A_\Psi\|_2 \le dR_\Psi.$$

Thus (8) is implied by $\alpha + \frac{\beta\lambda d R_\Psi}{2} < 1$. For IPO, $\Psi(q) = q$ and hence $R_\Psi \le 1/2$, giving the earlier condition $\alpha + \beta\lambda d/4 < 1$. □

C.3 Proof of Theorem 4.4

We show an extended version of Theorem 4.4 under a relaxed assumption on the transformed payoff matrix $P = \Psi(P^*)$. We first introduce these definitions for P as follows. Since Ψ is strictly increasing, the monotonicity assumptions/definitions in C.1 of P is equivalent to the same set of assumptions/definitions on the true underlying preference matrix P^* as well.

Definition C.1. *Suppose $P \in \mathbb{R}^{K\times K}$ is a strongly transitive (ST) transformed payoff matrix with total order $1 \succ 2 \succ \cdots \succ K$. Then we say*

1. *P is SST, if for any $i \in [K-1]$, $P_{ik} > P_{i+1,k}$ holds for at least one $k \in [K]$;*
2. *P is ST$^+$, if $P_{1k} > P_{2k}$ holds for at least one $k \in [K]$.*

We can see SST means that for any $1 \le i < j \le K$, P_i weakly dominates P_j element-wise with at least one strict inequality, and ST$^+$ only requires it holds for $(i,j) = (1,2)$ — there can be many tied responses within $\{2, \cdots, K\}$ with identical P_i. The next full version of Theorem 4.4 shows collapse guarantees under ST$^+$ transformed payoff structures.

Theorem C.2. *Let $\boldsymbol{\pi}_0 \in \Delta_K^\circ$ and $\boldsymbol{\pi}_{\text{ref}} \in \Delta_K^\circ$. Consider the MRS-ΨPO dynamics $\boldsymbol{\pi}_{t+1} = \text{MRS}(\boldsymbol{\pi}_t; \alpha, \beta, \lambda)$ with transformed payoff matrix $P = \Psi(P^*)$, and let $r_{\text{ref},i} := \log \frac{\pi_{\text{ref},i}}{\pi_{\text{ref},i+1}}$.*

- ***(ε-collapse when $\alpha < 1$).*** *If $\alpha, \lambda \in [0,1)$, there exists T such that for all $t \ge T$,*
  - *if P is SST, then*

$$\boldsymbol{\pi}_{t,1} \;\ge\; 1 - \left(\exp\left(\underline{r}_{\text{ref}} + \frac{\beta\lambda\delta}{2(1-\alpha)}\right) - 1\right)^{-1},$$

    whenever the exponent is positive, where $\underline{r}_{\text{ref}} := \min_{i\in[K-1]} r_{\text{ref},i}$;
  - *if P is ST$^+$, then*

$$\boldsymbol{\pi}_{t,1} \;\ge\; 1 - (K-1)\exp\left(-r_{\text{ref},1} - \frac{\beta\lambda\delta_1}{2(1-\alpha)}\right),$$

    provided the adjacent tail log-ratios are eventually nonnegative, which is guaranteed for instance when $r_{\text{ref},i} \ge 0$ and $\pi_{1,i} \ge \pi_{1,i+1}$ for all $i \ge 2$.

  Here

$$\delta_i := (1-\lambda)\sum_{j=1}^{K} \pi_{0,j}\,(P_{ij} - P_{i+1,j}), \qquad \delta := \min_{i\in[K-1]} \delta_i,$$

  and these constants are positive under SST. In particular, if $\boldsymbol{\pi}_{\text{ref}}$ is aligned with the total order and $\lambda > 0$, then $r_{\text{ref},i} \ge 0$ and the SST bound reduces to

$$\boldsymbol{\pi}_{t,1} \;\ge\; 1 - \left(\exp\left(\frac{\beta\lambda\delta}{2(1-\alpha)}\right) - 1\right)^{-1}.$$

- *If $\alpha = 1$, $\lambda < 1$, and P is SST, there exists a finite T such that for all $t \ge T$,*

$$H(\boldsymbol{\pi}_{t+1}) < H(\boldsymbol{\pi}_t),$$

  where $H(\boldsymbol{\pi}) := -\sum_{i=1}^K \pi_i \log \pi_i$ is the Shannon entropy. In addition, $\boldsymbol{\pi}_t \to \boldsymbol{e}_1$ and $H(\boldsymbol{\pi}_t) \to 0$ as $t \to +\infty$.

Next, we prove this theorem. For the convenience of notation in its proof, we define log-ratios w.r.t. any policy $\boldsymbol{\pi}_t \in \Delta_K^\circ$ and rewrite the MRS-ΨPO dynamics in the following lemma.

Lemma C.3 (Log-ratio recursion). *For $i = 1, \dots, K-1$, define adjacent log-ratios*

$$r_{t,i} := \log \frac{\pi_{t,i}}{\pi_{t,i+1}}, \qquad r_{\mathrm{ref},i} := \log \frac{\pi_{\mathrm{ref},i}}{\pi_{\mathrm{ref},i+1}}.$$

Let $\boldsymbol{\mu}_t = \lambda \boldsymbol{\pi}_t + (1-\lambda)\boldsymbol{\pi}_0$. Then the MRS-$\Psi$PO dynamics imply the exact recursion

$$r_{t+1,i} = \alpha r_{t,i} + (1-\alpha) r_{\mathrm{ref},i} + \beta\Big((P\boldsymbol{\mu}_t)_i - (P\boldsymbol{\mu}_t)_{i+1}\Big). \tag{18}$$

Proof. For any vector $\boldsymbol{z}$, $\mathrm{softmax}(\boldsymbol{z})_i / \mathrm{softmax}(\boldsymbol{z})_{i+1} = \exp(z_i - z_{i+1})$. Apply this to the logits inside (6) to obtain (18). □

The next lemma shows how entropy collapse can be measured by large log-ratios.

Lemma C.4 (From large log-ratios to ε-collapse). *Let $\boldsymbol{\pi} \in \Delta_K^\circ$ satisfy $r_i := \log(\pi_i/\pi_{i+1}) \geq R > 0$ for all $i \in [K-1]$, then*

$$1 - \pi_1 \;\leq\; \frac{1}{e^R - 1}. \tag{19}$$

In particular, if $R \geq \log\big(1 + 1/\varepsilon\big)$ then $\boldsymbol{\pi}$ is ε-collapsed. If $r_1 \geq R > 0$ and $r_i \geq 0$ for all $i \geq 2$, then $1 - \pi_1 \leq (K-1)e^{-R}$.

Proof. The condition $r_i \geq R$ implies $\pi_{i+1} \leq e^{-R}\pi_i$ for each i. Thus $\pi_k \leq e^{-(k-1)R}\pi_1$ for all $k \geq 2$, and hence

$$1 - \pi_1 = \sum_{k=2}^{K} \pi_k \leq \sum_{k=2}^{K} e^{-(k-1)R}\pi_1 < \sum_{k=2}^{+\infty} e^{-(k-1)R} = \frac{1}{e^R - 1}.$$

As a result, when $R \geq \log\big(1 + 1/\varepsilon\big)$, $1 - \pi_1 < \varepsilon$. If $r_1 \geq R$ and $r_i \geq 0$ for all $i \geq 2$, then $\pi_k \leq \pi_2 \leq e^{-R}\pi_1$ for all $k \geq 2$, and so

$$1 - \pi_1 = \sum_{k=2}^{K} \pi_k \leq (K-1)e^{-R}\pi_1 \leq (K-1)e^{-R}.$$

□

In addition, we need the following lemma to leverage the assumption of P being SST or ST^+.

Lemma C.5 (Uniform drift induced by off-policy mixing). *Assume P is SST and suppose $\boldsymbol{\pi}_0 \in \Delta_K^\circ$. Fix any $\lambda \in [0, 1)$ and define*

$$\delta_i := (1-\lambda) \sum_{j=1}^{K} \pi_{0,j}\,(P_{ij} - P_{i+1,j}), \qquad i = 1, \dots, K-1. \tag{20}$$

Then $\delta_i > 0$ for all i. Moreover, for all $t \geq 0$,

$$(P\boldsymbol{\mu}_t)_i - (P\boldsymbol{\mu}_t)_{i+1} \;\geq\; \delta_i, \qquad \boldsymbol{\mu}_t := \lambda\boldsymbol{\pi}_t + (1-\lambda)\boldsymbol{\pi}_0. \tag{21}$$

If P is ST^+ and not SST, we have $\delta_1 > 0$ and $\delta_i \geq 0$ for all $i \geq 2$.

Proof. Let $d^{(i)} \in \mathbb{R}^K$ be defined by $d_j^{(i)} := P_{ij} - P_{i+1,j}$. By SST, $d_j^{(i)} \geq 0$ for all j and there exists j with $d_j^{(i)} > 0$. Since $\boldsymbol{\pi}_0 \in \Delta_K^\circ$ has $\pi_{0,j} > 0$ for all j, we have $\langle d^{(i)}, \boldsymbol{\pi}_0 \rangle > 0$ and thus $\delta_i = (1-\lambda)\langle d^{(i)}, \boldsymbol{\pi}_0 \rangle > 0$.

Finally,

$$(P\boldsymbol{\mu}_t)_i - (P\boldsymbol{\mu}_t)_{i+1} = \sum_{j=1}^{K} \mu_{t,j}\, d_j^{(i)} \geq (1-\lambda) \sum_{j=1}^{K} \pi_{0,j}\, d_j^{(i)} = \delta_i,$$

proving (21). If P is ST^+, the same argument holds for $i = 1$, while weak monotonicity gives $\delta_i \geq 0$ for the remaining adjacent pairs. □

Based on Lemmas C.3 to C.5, we next prove a full version of Theorem C.2, as shown in the following Theorem C.6.

Theorem C.6 (True collapse for $\alpha = 1$ and ε-collapse for $\alpha < 1$)**.** *Assume P is ST^+. Let $\boldsymbol{\pi}_0 \in \Delta_K^\circ$ and $\boldsymbol{\pi}_{\mathrm{ref}} \in \Delta_K^\circ$. Fix any $\lambda \in [0, 1)$ and define δ_i as in (20). Consider the MRS-ΨPO dynamics $\boldsymbol{\pi}_{t+1} = \mathrm{MRS}(\boldsymbol{\pi}_t; P, \alpha, \beta, \lambda, \boldsymbol{\pi}_{\mathrm{ref}}, \boldsymbol{\pi}_0)$ given by (6). The following claims hold:*

1. ***(True collapse when $\alpha = 1$).*** *If $\alpha = 1$ and P is SST, then for every $i = 1, \ldots, K-1$,*

$$r_{t,i} \;\geq\; r_{1,i} + (t-1)\,\beta\,\delta_i, \tag{22}$$

so $r_{t,i} \to +\infty$ for all i and therefore $\boldsymbol{\pi}_t \to \boldsymbol{e}_1$. Moreover, there exists a finite T such that for all $t \geq T$,

$$H(\boldsymbol{\pi}_{t+1}) < H(\boldsymbol{\pi}_t) \quad \text{unless } \boldsymbol{\pi}_t \text{ is already a vertex,}$$

and consequently $H(\boldsymbol{\pi}_t) \to 0$.

2. ***(No exact collapse when $\alpha < 1$).*** *If $\alpha \in [0, 1)$, adjacent log-ratios remain uniformly bounded. More precisely, let*

$$G_P := \max_{i \in [K-1]} \max_{\eta \in \Delta_K} |(P\eta)_i - (P\eta)_{i+1}| < \infty.$$

Then

$$|r_{t,i} - r_{\mathrm{ref},i}| \leq \alpha^{t-1} |r_{1,i} - r_{\mathrm{ref},i}| + \frac{\beta G_P}{1-\alpha} \qquad \forall t \geq 1.$$

In particular, $\min_i \pi_{t,i}$ is bounded below by a positive constant depending on $(K, \alpha, \beta, \lambda, P, \boldsymbol{\pi}_{\mathrm{ref}}, \boldsymbol{\pi}_0, \boldsymbol{\pi}_1)$, and hence $H(\boldsymbol{\pi}_t)$ is bounded away from 0; thus $\boldsymbol{\pi}_t$ does not converge to any vertex of Δ_K.

3. ***(ε-collapse when $\alpha < 1$).*** *If $\alpha \in [0, 1)$, there exists a finite $T(P, \boldsymbol{\pi}_0, \boldsymbol{\pi}_{\mathrm{ref}}, \boldsymbol{\pi}_1, \alpha, \beta, \lambda)$ such that for every i with $\delta_i > 0$ and all $t \geq T$,*

$$r_{t,i} > r_{\mathrm{ref},i} + \frac{\beta\lambda\delta_i}{2(1-\alpha)}. \tag{23}$$

Consequently, when P is SST, the iterates become ε-collapsed with

$$\varepsilon = \left(\exp\left(\underline{r}_{\mathrm{ref}} + \frac{\beta\lambda\delta}{2(1-\alpha)} \right) - 1 \right)^{-1}, \qquad \underline{r}_{\mathrm{ref}} := \min_i r_{\mathrm{ref},i}, \tag{24}$$

provided the exponent is positive, where $\delta := \min_i \delta_i > 0$. When P is ST^+, the same argument gives

$$\varepsilon = (K-1) \exp\left(-r_{\mathrm{ref},1} - \frac{\beta\lambda\delta_1}{2(1-\alpha)} \right)$$

provided the adjacent tail log-ratios are eventually nonnegative.

Proof. We prove the three claims one by one.

1. Claim 1: We first prove $\alpha = 1$ implies true collapse. When $\alpha = 1$, the recursion (18) becomes

$$r_{t+1,i} = r_{t,i} + \beta\big((P\boldsymbol{\mu}_t)_i - (P\boldsymbol{\mu}_t)_{i+1}\big).$$

By Lemma C.5, $(P\boldsymbol{\mu}_t)_i - (P\boldsymbol{\mu}_t)_{i+1} \geq \delta_i$, hence $r_{t+1,i} \geq r_{t,i} + \beta\delta_i$ and (22) follows by induction. For SST P, $\delta_i > 0$ for every i, and therefore every adjacent log-ratio diverges to $+\infty$. This forces $\boldsymbol{\pi}_t \to \boldsymbol{e}_1$.

To obtain strict entropy decrease after some finite time, note that $r_{t,i} \to +\infty$ for all i implies that for all sufficiently large t, $\pi_{t,1} \geq \pi_{t,2} \geq \cdots \geq \pi_{t,K}$. Also, under the SST assumption, the score vector $\boldsymbol{s}_t := P\boldsymbol{\mu}_t$ is aligned for every t:

$$s_{t,i} - s_{t,i+1} = (P\boldsymbol{\mu}_t)_i - (P\boldsymbol{\mu}_t)_{i+1} \geq \delta_i > 0.$$

Therefore the weights $w_{t,i} := \exp(\beta s_{t,i})$ satisfy $w_{t,1} > w_{t,2} > \cdots > w_{t,K}$. When $\alpha = 1$, the update is a monotone exponential tilt $\boldsymbol{\pi}_{t+1} \propto \boldsymbol{\pi}_t \odot \boldsymbol{w}_t$. Standard majorization arguments for monotone tilting imply $\boldsymbol{\pi}_{t+1}$ majorizes $\boldsymbol{\pi}_t$ for all large t, and the majorization is strict unless $\boldsymbol{\pi}_t$ is already a vertex. Since Shannon entropy is strictly Schur-concave, $H(\boldsymbol{\pi}_{t+1}) < H(\boldsymbol{\pi}_t)$ for all sufficiently large t unless $\boldsymbol{\pi}_t$ is already a vertex. Finally, $\boldsymbol{\pi}_t \to \boldsymbol{e}_1$ implies $H(\boldsymbol{\pi}_t) \to 0$.

2. Claim 2: $\alpha < 1$ prevents exact collapse. From (18) and the definition of G_P,

$$\begin{aligned} |r_{t+1,i} - r_{\mathrm{ref},i}| &= \left|\alpha(r_{t,i} - r_{\mathrm{ref},i}) + \beta\big((P\boldsymbol{\mu}_t)_i - (P\boldsymbol{\mu}_t)_{i+1}\big)\right| \\ &\le \alpha|r_{t,i} - r_{\mathrm{ref},i}| + \beta G_P. \end{aligned}$$

Unrolling this recursion yields

$$|r_{t,i} - r_{\mathrm{ref},i}| \le \alpha^{t-1}|r_{1,i} - r_{\mathrm{ref},i}| + \beta G_P(1 + \alpha + \cdots + \alpha^{t-2}) \le \alpha^{t-1}|r_{1,i} - r_{\mathrm{ref},i}| + \frac{\beta G_P}{1-\alpha}.$$

Hence all adjacent log-ratios are uniformly bounded, which implies that all coordinates of $\boldsymbol{\pi}_t$ remain bounded away from 0. Thus $\boldsymbol{\pi}_t$ cannot converge to a vertex and $H(\boldsymbol{\pi}_t)$ is bounded below by a positive constant.

3. Claim 3: Finally we prove ε-collapse when $\alpha < 1$. By Lemma C.5, $(P\boldsymbol{\mu}_t)_i - (P\boldsymbol{\mu}_t)_{i+1} \ge \delta_i$. Fix an adjacent pair i with $\delta_i > 0$. Substitute this into (18) to obtain the lower-bound recursion

$$r_{t+1,i} \ge \alpha r_{t,i} + (1-\alpha) r_{\mathrm{ref},i} + \beta\delta_i.$$

Solving this affine recursion gives

$$r_{t,i} \ge \alpha^{t-1} r_{1,i} + (1 - \alpha^{t-1}) r_{\mathrm{ref},i} + \frac{1-\alpha^{t-1}}{1-\alpha}\beta\delta_i.$$

Since $\delta_i > 0$ and $0 \le \lambda < 1$, for all sufficiently large t,

$$r_{t,i} > r_{\mathrm{ref},i} + \frac{\beta\lambda\delta_i}{2(1-\alpha)},$$

which proves (23).
When P is SST, $\delta_i > 0$ for all $i \in [K-1]$. Let

$$R := \underline{r}_{\mathrm{ref}} + \frac{\beta\lambda\delta}{2(1-\alpha)}, \qquad \underline{r}_{\mathrm{ref}} := \min_i r_{\mathrm{ref},i}, \qquad \delta := \min_i \delta_i.$$

If $R > 0$, then Lemma C.4 implies

$$\pi_{t,1} \ge 1 - \left(e^R - 1\right)^{-1}$$

for all sufficiently large t. By (18), $r_{1,i} \ge 0$ and $r_{\mathrm{ref},i} \ge 0$ imply $r_{t,i} \ge 0$ for all t when $i \ge 2$. When P is ST^+, the same argument gives a lower bound on $r_{t,1}$ with $\delta_1 > 0$; if the remaining adjacent tail log-ratios are eventually nonnegative, then Lemma C.4 gives

$$1 - \pi_{t,1} \le (K-1)\exp\left(-r_{\mathrm{ref},1} - \frac{\beta\lambda\delta_1}{2(1-\alpha)}\right).$$

□

D Robustness of Section 4 under Stochastic Pairwise Observations

The population results in Section 4 assume that the transformed payoff matrix $P = \Psi(P^*)$ is known. In practice, preference information is often observed through noisy pairwise labels. This appendix shows that the three conclusions in Section 4 admit stochastic analogues when the algorithm uses fresh noisy estimates across iterations.

Stochastic MRS-ΨPO dynamics. Let $\mathcal{F}_t$ denote the history before round t. Thus the current policy $\boldsymbol{\pi}_t$ and the mixed sampling distribution

$$\boldsymbol{\mu}_t := \lambda \boldsymbol{\pi}_t + (1-\lambda)\boldsymbol{\pi}_0 \tag{25}$$

are $\mathcal{F}_t$-measurable. At round t, instead of using the population matrix P, the update uses a freshly drawn random matrix $\widehat{P}_t$:

$$\boldsymbol{\pi}_{t+1} = \mathrm{softmax}\Big(\alpha \log \boldsymbol{\pi}_t + (1-\alpha)\log \boldsymbol{\pi}_{\mathrm{ref}} + \beta \widehat{P}_t \boldsymbol{\mu}_t\Big). \tag{26}$$

Throughout the main part of this appendix, we assume that $\widehat{P}_t$ is a conditionally unbiased estimator of the transformed payoff matrix:

$$\mathbb{E}[\widehat{P}_t \mid \mathcal{F}_t] = P. \tag{27}$$

This assumption is exact for IPO, where $\Psi(q) = q$, if each entry is estimated by the empirical mean of Bernoulli comparison labels. It also holds for affine links. For nonlinear plug-in links, (27) can fail; we discuss the resulting bias briefly at the end of this appendix section.

High-level message. Under (27), the stochastic dynamics are a martingale perturbation of the population MRS-ΨPO dynamics. Therefore, the population results in Section 4 should be interpreted statistically rather than deterministically:

- **Contraction.** The population contraction result in Theorem 4.3 becomes stochastic stability around the population fixed point. If the observation noise vanishes, the stochastic policy and sampling iterates converge to the population fixed point; if the noise persists, they remain in a variance-controlled neighborhood.
- **Cyclic instability.** The Condorcet-cycle instability in Proposition 4.2 persists in the conditional mean dynamics. The unstable cyclic mode is the same as in the population analysis, and fresh observation noise is amplified along this mode.
- **Entropy collapse.** The collapse result in Theorem 4.4 does not require each realized matrix $\widehat{P}_t$ to satisfy SST. If the population transformed matrix P has positive adjacent drift under the sampling distributions $\boldsymbol{\mu}_t$, then realized SST violations enter only as martingale noise. For $\alpha = 1$, collapse occurs almost surely under controlled noise; for $\alpha < 1$, the policy is concentrated with high probability whenever the population drift dominates the stochastic fluctuation scale.

D.1 Weighted noise bounds from pairwise observations

The stochastic update depends on $\widehat{P}_t - P$ only through the weighted vector $(\widehat{P}_t - P)\boldsymbol{\mu}_t$. Thus we do not need a uniform matrix perturbation bound such as $\|\widehat{P}_t - P\|_2$. We instead control the actual noisy directions used by the dynamics.

For one round t, and for each ordered pair (a, j), let $n_{aj,t}$ be the number of independent observations used to estimate the transformed payoff entry P_{aj}. We denote these observations by

$$U_{aj,t}^{(1)}, \ldots, U_{aj,t}^{(n_{aj,t})}.$$

The estimator is

$$\widehat{P}_{t,aj} := \frac{1}{n_{aj,t}} \sum_{\ell=1}^{n_{aj,t}} U_{aj,t}^{(\ell)}. \tag{28}$$

In the unbiased transformed-observation model,

$$\mathbb{E}[U_{aj,t}^{(\ell)} \mid \mathcal{F}_t] = P_{aj},$$

so that $\mathbb{E}[\widehat{P}_{t,aj} \mid \mathcal{F}_t] = P_{aj}$. Assume moreover that the centered observations are conditionally sub-Gaussian with variance proxy $s^2_{aj,t}$: for every $u \in \mathbb{R}$,

$$\mathbb{E}\Big[\exp\Big(u(U^{(\ell)}_{aj,t} - P_{aj})\Big) \mid \mathcal{F}_t\Big] \le \exp\left(\frac{u^2 s^2_{aj,t}}{2}\right). \tag{29}$$

Here $s^2_{aj,t}$ is a per-observation noise proxy; for a Bernoulli observation, Hoeffding's lemma gives $s^2_{aj,t} \le 1/4$.

We first bound the centered-logit update noise. Let

$$Z_{t+1} := \beta\Pi(\widehat{P}_t - P)\boldsymbol{\mu}_t, \qquad \Pi := I - \frac{1}{K}\mathbf{1}\mathbf{1}^\top. \tag{30}$$

Lemma D.1 (Conditional variance of stochastic MRS-ΨPO noise)**.** *Suppose* (27) *holds and the entry noises are conditionally independent given* $\mathcal{F}_t$*. Define*

$$v_{aj,t} := \mathrm{Var}(\widehat{P}_{t,aj} \mid \mathcal{F}_t).$$

Then

$$\mathbb{E}[\|Z_{t+1}\|_2^2 \mid \mathcal{F}_t] \le \beta^2 \sum_{a=1}^{K}\sum_{j=1}^{K} \mu^2_{t,j} v_{aj,t}. \tag{31}$$

Under the observation model (28)–(29)*,*

$$v_{aj,t} \le \frac{s^2_{aj,t}}{n_{aj,t}},$$

and therefore

$$\mathbb{E}[\|Z_{t+1}\|_2^2 \mid \mathcal{F}_t] \le \beta^2 \sum_{a=1}^{K}\sum_{j=1}^{K} \frac{\mu^2_{t,j} s^2_{aj,t}}{n_{aj,t}}. \tag{32}$$

If the estimated support is $S \subseteq [K] \times [K]$*, the sum in* (32) *is restricted to* $(a, j) \in S$*. In particular, if* $n_{aj,t} \ge n_{\min,S,t}$ *and* $s^2_{aj,t} \le \bar{s}^2_t$ *on* S*, and if the maximum column degree of* S *is* d_{col}*, then*

$$\mathbb{E}[\|Z_{t+1}\|_2^2 \mid \mathcal{F}_t] \le \frac{\beta^2 \bar{s}^2_t d_{\mathrm{col}}}{n_{\min,S,t}} \|\boldsymbol{\mu}_t\|_2^2. \tag{33}$$

Proof. Let $E_t := \widehat{P}_t - P$. Since $\|\Pi v\|_2 \le \|v\|_2$,

$$\mathbb{E}[\|Z_{t+1}\|_2^2 \mid \mathcal{F}_t] \le \beta^2 \mathbb{E}[\|E_t\boldsymbol{\mu}_t\|_2^2 \mid \mathcal{F}_t].$$

For coordinate a,

$$(E_t\boldsymbol{\mu}_t)_a = \sum_{j=1}^{K} \mu_{t,j} E_{t,aj}.$$

By conditional independence and centering,

$$\mathrm{Var}((E_t\boldsymbol{\mu}_t)_a \mid \mathcal{F}_t) = \sum_{j=1}^{K} \mu^2_{t,j} v_{aj,t}.$$

Summing over a gives (31). Under (28)–(29), averaging $n_{aj,t}$ independent sub-Gaussian observations gives $v_{aj,t} \le s^2_{aj,t}/n_{aj,t}$, yielding (32). The support and degree bound follows from

$$\sum_{(a,j)\in S} \mu^2_{t,j} = \sum_{j=1}^{K} \mu^2_{t,j} \,|\{a : (a, j) \in S\}| \le d_{\mathrm{col}} \|\boldsymbol{\mu}_t\|_2^2.$$

□

For entropy collapse, we need a sharper scalar noise bound. For $i \in [K-1]$, define the adjacent-drift noise

$$\zeta_{t,i} := ((\widehat{P}_t - P)\boldsymbol{\mu}_t)_i - ((\widehat{P}_t - P)\boldsymbol{\mu}_t)_{i+1}. \tag{34}$$

Lemma D.2 (Weighted adjacent noise)**.** *Under the assumptions of Lemma D.1, $\zeta_{t,i}$ is conditionally sub-Gaussian with variance proxy*

$$\sigma_{t,i}^2(\boldsymbol{\mu}_t) \le \sum_{j=1}^{K} \mu_{t,j}^2 \left(\frac{s_{ij,t}^2}{n_{ij,t}} + \frac{s_{i+1,j,t}^2}{n_{i+1,j,t}} \right). \tag{35}$$

In particular, if $\sigma_{t,i}^2(\boldsymbol{\mu}_t) \le \sigma_i^2$ uniformly over t, then

$$\mathbb{E}[\exp(u\zeta_{t,i}) \mid \mathcal{F}_t] \le \exp\left(\frac{u^2\sigma_i^2}{2} \right), \qquad \forall u \in \mathbb{R}.$$

For IPO with Bernoulli observations, one may take $s_{aj,t}^2 \le 1/4$, so

$$\sigma_{t,i}^2(\boldsymbol{\mu}_t) \le \frac{1}{4} \sum_{j=1}^{K} \mu_{t,j}^2 \left(\frac{1}{n_{ij,t}} + \frac{1}{n_{i+1,j,t}} \right). \tag{36}$$

Proof. By definition,

$$\zeta_{t,i} = \sum_{j=1}^{K} \mu_{t,j} \left[(\widehat{P}_{t,ij} - P_{ij}) - (\widehat{P}_{t,i+1,j} - P_{i+1,j}) \right].$$

Conditioning on $\mathcal{F}_t$, $\boldsymbol{\mu}_t$ is fixed. The entry estimates are conditionally independent, and each averaged entry has sub-Gaussian proxy $s_{aj,t}^2/n_{aj,t}$. A weighted sum of conditionally independent sub-Gaussian variables is sub-Gaussian with variance proxy equal to the sum of the squared weights times the individual proxies. This proves (35). The Bernoulli case follows from Hoeffding's lemma. □

D.2 Stochastic perturbation of the MRS-ΨPO map

The stochastic sampling iterate is $\boldsymbol{\mu}_t$. For analysis, we use centered logits only as an auxiliary coordinate. For $h \in \mathbf{1}^\perp$, write

$$\boldsymbol{\pi}(h) := \mathrm{softmax}(h), \qquad \boldsymbol{\mu}(h) := \lambda\boldsymbol{\pi}(h) + (1-\lambda)\boldsymbol{\pi}_0.$$

The population MRS-ΨPO map in centered-logit coordinates is

$$F_P(h) := \alpha h + \Pi\left((1-\alpha)\log \boldsymbol{\pi}_{\mathrm{ref}} + \beta P\boldsymbol{\mu}(h)\right). \tag{37}$$

Let

$$h_t := \Pi \log \boldsymbol{\pi}_t.$$

Taking centered logits in (26) gives the exact recursion

$$h_{t+1} = F_P(h_t) + Z_{t+1}, \qquad Z_{t+1} := \beta\Pi(\widehat{P}_t - P)\boldsymbol{\mu}_t. \tag{38}$$

By (27),

$$\mathbb{E}[Z_{t+1} \mid \mathcal{F}_t] = 0. \tag{39}$$

Thus the stochastic dynamics are not generated by a fixed empirical map $F_{\widehat{P}}$; they are a martingale-difference perturbation of the population map F_P.

The same recursion is especially transparent in adjacent log-ratio coordinates. For $i \in [K-1]$, define

$$r_{t,i} := \log \frac{\pi_{t,i}}{\pi_{t,i+1}}, \qquad r_{\mathrm{ref},i} := \log \frac{\pi_{\mathrm{ref},i}}{\pi_{\mathrm{ref},i+1}},$$

$$\Delta_i(\boldsymbol{\mu}_t) := (P\boldsymbol{\mu}_t)_i - (P\boldsymbol{\mu}_t)_{i+1}, \qquad \zeta_{t,i} := ((\widehat{P}_t - P)\boldsymbol{\mu}_t)_i - ((\widehat{P}_t - P)\boldsymbol{\mu}_t)_{i+1}.$$

Then

$$r_{t+1,i} = \alpha r_{t,i} + (1-\alpha) r_{\mathrm{ref},i} + \beta \Delta_i(\boldsymbol{\mu}_t) + \beta \zeta_{t,i}. \tag{40}$$

This identity is exact and does not require the realized matrix $\widehat{P}_t$ to satisfy any structural property such as SST.

Derivation of (38) and (40). Taking logs in (26) gives

$$\log \boldsymbol{\pi}_{t+1} = \alpha \log \boldsymbol{\pi}_t + (1-\alpha) \log \boldsymbol{\pi}_{\mathrm{ref}} + \beta \widehat{P}_t \boldsymbol{\mu}_t - c_t \mathbf{1}$$

for a normalizing scalar c_t. Applying Π removes the scalar term and yields

$$h_{t+1} = \alpha h_t + \Pi\left((1-\alpha) \log \boldsymbol{\pi}_{\mathrm{ref}} + \beta P \boldsymbol{\mu}_t\right) + \beta \Pi (\widehat{P}_t - P)\boldsymbol{\mu}_t = F_P(h_t) + Z_{t+1}.$$

Conditional unbiasedness gives $\mathbb{E}[Z_{t+1} \mid \mathcal{F}_t] = 0$ because $\boldsymbol{\mu}_t$ is $\mathcal{F}_t$-measurable. For log-ratios, use $\mathrm{softmax}(z)_i / \mathrm{softmax}(z)_{i+1} = \exp(z_i - z_{i+1})$ in (26). Subtracting coordinate $i+1$ from coordinate i gives (40). □

D.3 Stochastic contraction

We first revisit the positive contraction result. When fresh noise is injected at every round, persistent noise generally prevents convergence to the deterministic population fixed point. Instead, the policy and sampling iterates remain concentrated around it. If the per-round noise vanishes, they converge to it.

Let

$$L_P := \alpha + \frac{\beta\lambda}{2} \|\Pi P\|_2. \tag{41}$$

When $L_P < 1$, the population map F_P is a contraction in centered-logit coordinates.

Theorem D.3 (Stochastic stability in the contraction regime)**.** *Assume $L_P < 1$, and let $h^\star \in \mathbf{1}^\perp$ be the unique fixed point of F_P. Define*

$$\boldsymbol{\pi}^\star := \mathrm{softmax}(h^\star), \qquad \boldsymbol{\mu}^\star := \lambda \boldsymbol{\pi}^\star + (1-\lambda)\boldsymbol{\pi}_0.$$

Consider the stochastic recursion (38). Suppose

$$\mathbb{E}[Z_{t+1} \mid \mathcal{F}_t] = 0, \qquad \mathbb{E}[\|Z_{t+1}\|_2^2 \mid \mathcal{F}_t] \le \nu_t^2,$$

where $\{\nu_t^2\}_{t \ge 1}$ is deterministic. Then, for every $t \ge 1$,

$$\mathbb{E}\|h_t - h^\star\|_2^2 \le L_P^{2(t-1)} \|h_1 - h^\star\|_2^2 + \sum_{s=1}^{t-1} L_P^{2(t-1-s)} \nu_s^2. \tag{42}$$

Consequently,

$$\mathbb{E}\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^\star\|_2^2 \le \frac{\lambda^2}{4} \mathbb{E}\|h_t - h^\star\|_2^2. \tag{43}$$

If $\sup_s \nu_s \le \bar{\nu}$, then

$$\limsup_{t \to \infty} \mathbb{E}\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^\star\|_2^2 \le \frac{\lambda^2 \bar{\nu}^2}{4(1 - L_P^2)}. \tag{44}$$

If $\nu_t \to 0$, then $\boldsymbol{\mu}_t \to \boldsymbol{\mu}^\star$ in mean square. If, moreover, $\sum_t \nu_t^2 < \infty$, then $\boldsymbol{\mu}_t \to \boldsymbol{\mu}^\star$ almost surely.

Proof. The derivative of the population map is

$$DF_P(h) = \alpha I + \beta\lambda \Pi P J(\boldsymbol{\pi}(h)),$$

where $J(\boldsymbol{\pi}) = \mathrm{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^\top$ is the softmax Jacobian. Since $\|J(\boldsymbol{\pi})\|_2 \le 1/2$,

$$\|DF_P(h)\|_2 \le \alpha + \frac{\beta\lambda}{2}\|\Pi P\|_2 = L_P < 1.$$

Thus F_P is a contraction and has a unique fixed point $h^\star$.

Using $F_P(h^\star) = h^\star$, the stochastic recursion gives

$$h_{t+1} - h^\star = F_P(h_t) - F_P(h^\star) + Z_{t+1}.$$

Taking conditional second moments and using $\mathbb{E}[Z_{t+1} \mid \mathcal{F}_t] = 0$,

$$\begin{aligned}\mathbb{E}[\|h_{t+1} - h^\star\|_2^2 \mid \mathcal{F}_t] &= \|F_P(h_t) - F_P(h^\star)\|_2^2 + \mathbb{E}[\|Z_{t+1}\|_2^2 \mid \mathcal{F}_t] \\ &\le L_P^2 \|h_t - h^\star\|_2^2 + \nu_t^2.\end{aligned}$$

Taking expectations and unrolling gives (42). Since $\|J(\boldsymbol{\pi})\|_2 \le 1/2$, softmax is $1/2$-Lipschitz in ℓ_2, hence

$$\|\boldsymbol{\mu}_t - \boldsymbol{\mu}^\star\|_2 = \lambda\|\boldsymbol{\pi}_t - \boldsymbol{\pi}^\star\|_2 \le \frac{\lambda}{2}\|h_t - h^\star\|_2.$$

This proves (43). The bounded-noise and vanishing-noise claims follow from the geometric series. If $\sum_t \nu_t^2 < \infty$, the same supermartingale argument applied to the displayed conditional recursion gives almost-sure convergence. □

D.4 Stochastic Condorcet-cycle instability

We next revisit the RPS instability result. Since $\mathbb{E}[Z_{t+1} \mid \mathcal{F}_t] = 0$, the local conditional mean dynamics have the same Jacobian as the population dynamics. Thus cyclic instability persists in the mean, and persistent observation noise is amplified along the unstable rotational mode.

Let $u = (1/3, 1/3, 1/3)$, and assume the RPS setting of Proposition 4.2. Let $h = 0$ denote the centered-logit coordinate of the uniform policy. Write the linearized stochastic recursion on $\mathbf{1}^\perp$ as

$$y_{t+1} = My_t + \xi_{t+1}, \tag{45}$$

where

$$M := DF_P(0), \qquad \xi_{t+1} := \beta\Pi(\widehat{P}_t - P)\boldsymbol{\mu}_t$$

projected to the tangent space. In the symmetric-link RPS case,

$$\rho_{\mathrm{cyc}} := \sqrt{\alpha^2 + a_\Psi^2 \frac{\beta^2\lambda^2}{3}}$$

is the modulus of the nontrivial eigenvalues of M.

Proposition D.4 (Stochastic persistence of cyclic instability)**.** *Assume $\rho_{\mathrm{cyc}} > 1$. Then the uniform fixed point is linearly unstable for the conditional mean dynamics. Moreover, suppose the linearized noise satisfies*

$$\mathbb{E}[\xi_{t+1} \mid \mathcal{F}_t] = 0, \qquad \mathrm{Cov}(\xi_{t+1} \mid \mathcal{F}_t) \succeq \sigma_{\mathrm{cyc}}^2 I$$

on the two-dimensional tangent space. Then the linearized stochastic recursion (45) satisfies

$$\mathbb{E}\|y_t\|_2^2 \ge \rho_{\mathrm{cyc}}^{2(t-1)}\|y_1\|_2^2 + 2\sigma_{\mathrm{cyc}}^2 \frac{\rho_{\mathrm{cyc}}^{2(t-1)} - 1}{\rho_{\mathrm{cyc}}^2 - 1}. \tag{46}$$

Thus the unstable cyclic mode persists under fresh noisy observations, and stochastic fluctuations are amplified along that mode.

Proof. Conditional unbiasedness gives

$$\mathbb{E}[y_{t+1} \mid \mathcal{F}_t] = My_t.$$

Therefore the conditional mean linearization is exactly the population linearization. If $\rho_{\mathrm{cyc}} > 1$, this linear map has an unstable mode.

In the RPS case, M is normal on the two-dimensional tangent space and satisfies $M^*M = \rho_{\mathrm{cyc}}^2 I$. Thus $\|M^k v\|_2 = \rho_{\mathrm{cyc}}^k \|v\|_2$. Unrolling (45) gives

$$y_t = M^{t-1}y_1 + \sum_{s=1}^{t-1} M^{t-1-s}\xi_{s+1}.$$

Orthogonality of martingale increments in second moment gives

$$\mathbb{E}\|y_t\|_2^2 = \rho_{\mathrm{cyc}}^{2(t-1)}\|y_1\|_2^2 + \sum_{s=1}^{t-1} \mathbb{E}\|M^{t-1-s}\xi_{s+1}\|_2^2.$$

The covariance lower bound gives

$$\mathbb{E}\|M^{t-1-s}\xi_{s+1}\|_2^2 \geq 2\sigma_{\mathrm{cyc}}^2 \rho_{\mathrm{cyc}}^{2(t-1-s)}.$$

Summing the geometric series yields (46). □

D.5 Stochastic entropy collapse

We now show that entropy collapse does not require the empirical matrix used at each round to satisfy SST. It is enough that the population transformed matrix $P = \Psi(P^*)$ induces positive adjacent drift under the sampling distributions $\boldsymbol{\mu}_t$.

Lemma D.5 (Population SST implies positive effective drift)**.** *Assume the population transformed matrix $P = \Psi(P^*)$ is SST with order $1 \succ \cdots \succ K$. For each adjacent pair $i \in [K-1]$, define*

$$d_j^{(i)} := P_{ij} - P_{i+1,j}, \qquad j \in [K].$$

Then $d_j^{(i)} \geq 0$ for all j. Moreover,

$$\Delta_i(\boldsymbol{\mu}_t) := (P\boldsymbol{\mu}_t)_i - (P\boldsymbol{\mu}_t)_{i+1} = \lambda\langle d^{(i)}, \boldsymbol{\pi}_t\rangle + (1-\lambda)\langle d^{(i)}, \boldsymbol{\pi}_0\rangle. \tag{47}$$

Consequently, if

$$\delta_i^{\mathrm{eff}} := \inf_{t\geq 1}\left\{\lambda\langle d^{(i)}, \boldsymbol{\pi}_t\rangle + (1-\lambda)\langle d^{(i)}, \boldsymbol{\pi}_0\rangle\right\} > 0, \tag{48}$$

then $\Delta_i(\boldsymbol{\mu}_t) \geq \delta_i^{\mathrm{eff}}$ for all t. In particular, if $\lambda < 1$, $\boldsymbol{\pi}_0 \in \Delta_K^\circ$, and

$$\delta_i^{\mathrm{base}} := \langle d^{(i)}, \boldsymbol{\pi}_0\rangle > 0,$$

then

$$\delta_i^{\mathrm{eff}} \geq (1-\lambda)\delta_i^{\mathrm{base}}. \tag{49}$$

Proof. SST with order $1 \succ \cdots \succ K$ implies that row i weakly dominates row $i+1$, hence $P_{ij} - P_{i+1,j} \geq 0$ for every j. Using $\boldsymbol{\mu}_t = \lambda\boldsymbol{\pi}_t + (1-\lambda)\boldsymbol{\pi}_0$, we obtain (47). The definition of δ_i^{eff} then gives $\Delta_i(\boldsymbol{\mu}_t) \geq \delta_i^{\mathrm{eff}}$. The baseline lower bound follows because $\lambda\langle d^{(i)}, \boldsymbol{\pi}_t\rangle \geq 0$. □

Theorem D.6 (Stochastic collapse under noisy comparison observations)**.** *Consider the stochastic MRS-ΨPO dynamics (26). Suppose that for every adjacent pair $i \in [K-1]$, the following conditions hold.*

1. ***Positive effective population drift.*** *There exists* $\delta_i^{\mathrm{eff}} > 0$ *such that*

$$\Delta_i(\boldsymbol{\mu}_t) = (P\boldsymbol{\mu}_t)_i - (P\boldsymbol{\mu}_t)_{i+1} \geq \delta_i^{\mathrm{eff}} \qquad \text{for all } t.$$

2. ***Martingale noise.***

$$\mathbb{E}[\zeta_{t,i} \mid \mathcal{F}_t] = 0 \qquad \text{for all } t.$$

3. ***Conditional sub-Gaussianity.*** *There exists* $\sigma_i^2 < \infty$ *such that, for every* $u \in \mathbb{R}$,

$$\mathbb{E}[\exp(u\zeta_{t,i}) \mid \mathcal{F}_t] \leq \exp\left(\frac{u^2\sigma_i^2}{2}\right).$$

Then the following claims hold.

1. ***True collapse for*** $\alpha = 1$***.*** *If* $\alpha = 1$*, then for every* $i \in [K-1]$,

$$\liminf_{t\to\infty} \frac{r_{t,i}}{t} \geq \beta\delta_i^{\mathrm{eff}} \qquad \text{almost surely.}$$

Consequently,

$$\boldsymbol{\pi}_t \to \mathbf{e}_1, \qquad \boldsymbol{\mu}_t \to \lambda\mathbf{e}_1 + (1-\lambda)\boldsymbol{\pi}_0, \qquad H(\boldsymbol{\pi}_t) \to 0 \qquad \text{almost surely.}$$

Moreover, for every fixed $t \geq 2$,

$$\mathbb{P}\left(r_{t,i} < r_{1,i} + \frac{\beta\delta_i^{\mathrm{eff}}}{2}(t-1)\right) \leq \exp\left(-\frac{(\delta_i^{\mathrm{eff}})^2(t-1)}{8\sigma_i^2}\right).$$

2. ***High-probability*** ε***-collapse for*** $\alpha < 1$***.*** *If* $\alpha \in [0,1)$*, then for every fixed* $t \geq 2$*, every* $u > 0$*, and every adjacent pair* $i \in [K-1]$,

$$\mathbb{P}\left[r_{t,i} < \alpha^{t-1}r_{1,i} + (1-\alpha^{t-1})r_{\mathrm{ref},i} + \beta\delta_i^{\mathrm{eff}}\frac{1-\alpha^{t-1}}{1-\alpha} - u\right] \leq \exp\left(-\frac{u^2(1-\alpha^2)}{2\beta^2\sigma_i^2}\right). \tag{50}$$

Consequently, for any confidence level $\xi \in (0,1)$*, define*

$$R_t(\xi) := \min_{i\in[K-1]} \left\{\alpha^{t-1}r_{1,i} + (1-\alpha^{t-1})r_{\mathrm{ref},i} + \beta\delta_i^{\mathrm{eff}}\frac{1-\alpha^{t-1}}{1-\alpha} - \beta\sigma_i\sqrt{\frac{2\log((K-1)/\xi)}{1-\alpha^2}}\right\}.$$

If $R_t(\xi) > 0$*, then with probability at least* $1-\xi$,

$$1 - \pi_{t,1} \leq \frac{1}{\exp(R_t(\xi)) - 1}.$$

Proof. The proof follows from the exact adjacent log-ratio recursion (40):

$$r_{t+1,i} = \alpha r_{t,i} + (1-\alpha)r_{\mathrm{ref},i} + \beta\Delta_i(\boldsymbol{\mu}_t) + \beta\zeta_{t,i}.$$

This recursion only requires positive population drift; the realized matrix $\widehat{P}_t$ need not satisfy SST.

Case 1: $\alpha = 1$. When $\alpha = 1$, unrolling gives

$$r_{t,i} = r_{1,i} + \beta \sum_{s=1}^{t-1} \Delta_i(\boldsymbol{\mu}_s) + \beta \sum_{s=1}^{t-1} \zeta_{s,i}.$$

By the positive effective drift assumption,

$$\sum_{s=1}^{t-1} \Delta_i(\boldsymbol{\mu}_s) \geq \delta_i^{\text{eff}}(t-1).$$

Let $M_{t-1,i} := \sum_{s=1}^{t-1} \zeta_{s,i}$. Since $\mathbb{E}[\zeta_{s,i} \mid \mathcal{F}_s] = 0$, $M_{t,i}$ is a martingale. The conditional sub-Gaussian assumption implies uniformly bounded conditional second moments, so the martingale strong law gives $M_{t-1,i}/t \to 0$ almost surely. Therefore,

$$\liminf_{t\to\infty} \frac{r_{t,i}}{t} \geq \beta\delta_i^{\text{eff}} \qquad \text{almost surely.}$$

Thus every adjacent log-ratio diverges to $+\infty$ almost surely. Since $\pi_{t,i+1}/\pi_{t,i} = \exp(-r_{t,i})$, we obtain $\boldsymbol{\pi}_t \to \mathbf{e}_1$, and hence $\boldsymbol{\mu}_t \to \lambda\mathbf{e}_1 + (1-\lambda)\boldsymbol{\pi}_0$. By continuity of entropy, $H(\boldsymbol{\pi}_t) \to 0$.

For the finite-time bound, conditional sub-Gaussianity and the tower property imply that $M_{t-1,i}$ is sub-Gaussian with variance proxy $(t-1)\sigma_i^2$. Hence

$$\mathbb{P}(M_{t-1,i} \leq -a) \leq \exp\left(-\frac{a^2}{2(t-1)\sigma_i^2}\right).$$

Taking $a = \delta_i^{\text{eff}}(t-1)/2$ gives the stated inequality.

Case 2: $\alpha < 1$. For $\alpha \in [0, 1)$, unrolling (40) yields

$$\begin{aligned} r_{t,i} = \alpha^{t-1} r_{1,i} + (1-\alpha^{t-1}) r_{\text{ref},i} + \beta \sum_{s=1}^{t-1} \alpha^{t-1-s} \Delta_i(\boldsymbol{\mu}_s) \\ + \beta \sum_{s=1}^{t-1} \alpha^{t-1-s} \zeta_{s,i}. \end{aligned}$$

Using $\Delta_i(\boldsymbol{\mu}_s) \geq \delta_i^{\text{eff}}$,

$$r_{t,i} \geq \alpha^{t-1} r_{1,i} + (1-\alpha^{t-1}) r_{\text{ref},i} + \beta\delta_i^{\text{eff}} \frac{1-\alpha^{t-1}}{1-\alpha} + \beta W_{t,i},$$

where

$$W_{t,i} := \sum_{s=1}^{t-1} \alpha^{t-1-s} \zeta_{s,i}.$$

By recursive conditioning and conditional sub-Gaussianity, $W_{t,i}$ is sub-Gaussian with variance proxy

$$\sigma_i^2 \sum_{s=1}^{t-1} \alpha^{2(t-1-s)} \leq \frac{\sigma_i^2}{1-\alpha^2}.$$

Therefore, for every $u > 0$,

$$\mathbb{P}(\beta W_{t,i} < -u) \leq \exp\left(-\frac{u^2(1-\alpha^2)}{2\beta^2\sigma_i^2}\right),$$

which proves (50).

Now choose

$$u_i = \beta\sigma_i\sqrt{\frac{2\log((K-1)/\xi)}{1-\alpha^2}}.$$

Applying (50) to every adjacent pair and taking a union bound gives $r_{t,i} \geq R_t(\xi)$ for all $i \in [K-1]$ with probability at least $1-\xi$. If $R_t(\xi) > 0$, then

$$\pi_{t,i+1} \leq e^{-R_t(\xi)}\pi_{t,i}, \qquad i \in [K-1].$$

Thus $\pi_{t,k} \leq e^{-(k-1)R_t(\xi)}\pi_{t,1}$ for every $k \geq 2$, and hence

$$1 - \pi_{t,1} = \sum_{k=2}^{K} \pi_{t,k} \leq \sum_{m=1}^{\infty} e^{-mR_t(\xi)} = \frac{1}{\exp(R_t(\xi)) - 1}.$$

This completes the proof. □

The dependence on λ is contained in the effective drift δ_i^{eff}. Expanding $\boldsymbol{\mu}_t = \lambda\boldsymbol{\pi}_t + (1-\lambda)\boldsymbol{\pi}_0$ gives

$$\delta_i^{\text{eff}} \geq \lambda\delta_i^{\text{on}} + (1-\lambda)\delta_i^{\text{base}},$$

where

$$\delta_i^{\text{on}} := \inf_{t\geq 1}\langle d^{(i)}, \boldsymbol{\pi}_t\rangle, \qquad \delta_i^{\text{base}} := \langle d^{(i)}, \boldsymbol{\pi}_0\rangle.$$

Thus the collapse scale for $\alpha < 1$ is

$$\frac{\beta}{1-\alpha}\left[\lambda\delta_i^{\text{on}} + (1-\lambda)\delta_i^{\text{base}}\right].$$

The conservative baseline-only argument sets $\delta_i^{\text{on}} = 0$, yielding the weaker lower bound $\beta(1-\lambda)\delta_i^{\text{base}}/(1-\alpha)$. This bound becomes vacuous as $\lambda \to 1$, but this does not mean collapse disappears in the on-policy regime. It only means that one must use an on-policy or effective-gap lower bound.

D.6 Non-unbiased plug-in links

The main stochastic results above assume the transformed estimator is conditionally unbiased for $P = \Psi(P^*)$. This is exact for IPO and affine links. For nonlinear plug-in estimators, however, the estimator

$$\widehat{P}_{t,ij} = \Psi(\widehat{P}^*_{t,ij})$$

is generally biased because $\mathbb{E}[\Psi(\widehat{P}^*_{t,ij}) \mid \mathcal{F}_t] \neq \Psi(P^*_{ij})$. For example, if $\Psi(q) = q^2$ and $\widehat{P}^*_{t,ij}$ is the empirical mean of n Bernoulli labels with mean p, then

$$\mathbb{E}[(\widehat{P}^*_{t,ij})^2] = p^2 + \frac{p(1-p)}{n},$$

so the plug-in estimator is upward biased for p^2. For the logit link $\Psi(q) = \log(q/(1-q))$, clipping is needed to avoid boundary singularities, and the clipped plug-in estimator is also biased in general.

In such cases one can decompose

$$\widehat{P}_t - P = B_t + M_t, \qquad B_t := \mathbb{E}[\widehat{P}_t \mid \mathcal{F}_t] - P, \qquad \mathbb{E}[M_t \mid \mathcal{F}_t] = 0.$$

The martingale part M_t is handled by the results above. The bias part B_t shifts the population drift. For contraction, it adds a deterministic perturbation term $\beta\Pi B_t\boldsymbol{\mu}_t$. For collapse, define

$$b_i := \sup_t \left|(B_t\boldsymbol{\mu}_t)_i - (B_t\boldsymbol{\mu}_t)_{i+1}\right|.$$

Then the conclusions of Theorem D.6 remain valid after replacing δ_i^{eff} by $\delta_i^{\text{eff}} - b_i$, provided $\delta_i^{\text{eff}} > b_i$. Thus nonlinear plug-in links remain analyzable, but the cleanest stochastic analogues are obtained in the conditionally unbiased transformed-observation model.

E Additional Discussion of DPO

In this section, we present a parallel set of theoretical results for generic DPO without the BT model assumption. In contrast to ΨPO, generic DPO does not admit a closed-form expression for the optimal policy $\boldsymbol{\pi}^\star$; instead, $\boldsymbol{\pi}^\star$ is characterized implicitly through a first-order optimality condition (FOC) given by the following Proposition E.1. Such an implicit characterization makes extending the ΨPO analysis to DPO substantially more challenging. Nevertheless, in this section, we develop DPO analogues of the main paper ΨPO results wherever possible, noting that some statements must be presented in a weaker form due to additional technical obstacles. Since the link function Ψ in ΨPO does not enter the formulation of generic DPO, in this section, we use P (instead of P^*) to denote the underlying preference matrix for simplicity. Throughout this appendix, we assume $0 < P_{ij} < 1$ for every $i \neq j$.

E.1 Generic DPO optima under general sampling distributions

Similar to Proposition 2.2, which characterizes the optimizer of the ΨPO population loss, we have the existence and uniqueness result for the generic DPO solution as follows:

Proposition E.1 (Existence and Uniqueness of the Generic DPO Optimum)**.** *Given preference matrix $P \in \mathcal{P}$, the DPO problem with sampling strategy $\boldsymbol{\mu} \in \Delta_K^\circ$ and reference policy $\boldsymbol{\pi}_{\text{ref}} \in \Delta_K^\circ$ is equivalent to solving the optimization problem*

$$\min_{\boldsymbol{\pi} \in \Delta_K^\circ} \left[- \sum_{1 \le i,j \le K} \mu_i \mu_j \, P_{ij} \, \log \sigma \left(\beta^{-1} \log \frac{\pi_i}{\pi_{\text{ref},i}} - \beta^{-1} \log \frac{\pi_j}{\pi_{\text{ref},j}} \right) \right], \tag{51}$$

which admits a convex reformulation in logits and a unique solution $\boldsymbol{\pi}^\star \in \Delta_K^\circ$. Moreover, let $\boldsymbol{\theta}^\star = \beta^{-1}(\log \boldsymbol{\pi}^\star - \log \boldsymbol{\pi}_{\text{ref}})$ and let Q be the Bradley–Terry matrix induced by $\boldsymbol{\theta}^\star$, i.e. $Q_{ij} = \sigma(\theta_i^\star - \theta_j^\star)$, then the first-order optimality condition for its solution $\boldsymbol{\pi}^\star$ is

$$\operatorname{Diag}(\boldsymbol{\mu}) \, (P - Q)^\top \boldsymbol{\mu} = \mathbf{0}. \tag{52}$$

Proof. Define the reparametrization

$$\theta_i \;=\; \beta^{-1} \Big(\log \pi_i - \log \pi_{\text{ref},i} \Big), \qquad i \in [K].$$

Then $\pi_i = \pi_{\text{ref},i} e^{\beta \theta_i}$ and the DPO objective becomes

$$L(\boldsymbol{\theta}) \;=\; - \sum_{i,j=1}^{K} \mu_i \mu_j \, P_{ij} \, \log \sigma(\theta_i - \theta_j),$$

the simplex constraint $\sum_i \pi_i = 1$ is equivalent to the smooth equality constraint

$$\sum_{i=1}^{K} \pi_{\text{ref},i} e^{\beta \theta_i} = 1. \tag{53}$$

First of all, we show the convexity of $L(\boldsymbol{\theta})$. Let $\ell(x) := -\log \sigma(x) = \log(1 + e^{-x})$ which satisfies $\ell''(x) = \sigma(x)(1 - \sigma(x)) > 0, \forall x \in \mathbb{R}$ and is therefore strictly convex. The objective in (51) can be written as $L(\boldsymbol{\theta}) = \sum_{i,j=1}^{K} \mu_i \mu_j P_{ij} \, \ell(\theta_i - \theta_j)$, with the understanding that $\boldsymbol{\theta}$ is defined up to an additive constant (since the objective depends only on logit differences). Thus optimizing over $\boldsymbol{\pi} \in \Delta_K^\circ$ is equivalent to optimizing over a gauge-fixed affine subspace, e.g. $\{\boldsymbol{\theta} : \sum_i \theta_i = 0\}$. Each term $\ell(\theta_i - \theta_j)$ is the composition of a strictly convex function with a linear map, hence is convex in $\boldsymbol{\theta}$. Since $\mu_i \mu_j P_{ij} \geq 0$, $\mathcal{L}(\theta)$ is a nonnegative weighted sum of convex functions and is therefore convex.

In addition, we show strict convexity of L modulo the constant-shift invariance by computing its Hessian. Using $\nabla(\theta_i - \theta_j) = \boldsymbol{e}_i - \boldsymbol{e}_j$, we have

$$\nabla^2 \ell(\theta_i - \theta_j) = \ell''(\theta_i - \theta_j)\,(\boldsymbol{e}_i - \boldsymbol{e}_j)(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top.$$

Hence

$$\nabla^2 L(\boldsymbol{\theta}) = \sum_{i,j=1}^{K} \mu_i \mu_j P_{ij}\, \ell''(\theta_i - \theta_j)\,(\boldsymbol{e}_i - \boldsymbol{e}_j)(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top. \tag{54}$$

Using the dueling identity $P_{ij} + P_{ji} = 1$ and the symmetry $\ell''(-x) = \ell''(x)$, we may group terms (i, j) and (j, i) to obtain

$$\nabla^2 L(\boldsymbol{\theta}) = \sum_{1 \le i < j \le K} \mu_i \mu_j\, \ell''(\theta_i - \theta_j)\,(\boldsymbol{e}_i - \boldsymbol{e}_j)(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top. \tag{55}$$

For any $\boldsymbol{v} \in \mathbb{R}^K$,

$$\boldsymbol{v}^\top \nabla^2 L(\boldsymbol{\theta})\, \boldsymbol{v} = \sum_{i<j} \mu_i \mu_j\, \ell''(\theta_i - \theta_j)\,(v_i - v_j)^2.$$

If $\boldsymbol{\mu}$ has full support, then each coefficient $\mu_i \mu_j\, \ell''(\theta_i - \theta_j)$ is strictly positive. Therefore,

$$\boldsymbol{v}^\top \nabla^2 L(\boldsymbol{\theta})\, \boldsymbol{v} = 0 \quad \Longleftrightarrow \quad v_i = v_j \;\; \forall i, j,$$

i.e. $\boldsymbol{v} \in \operatorname{span}\{\mathbf{1}\}$. Thus the Hessian is positive semidefinite with nullspace exactly equal to the constant-shift direction, and $L(\boldsymbol{\theta})$ is strictly convex on any gauge-fixed subspace (e.g. $\sum_i \theta_i = 0$ or $\theta_K = 0$). To establish existence, let $c = \min_{i,j} \mu_i \mu_j P_{ij} > 0$. Taking a term from a minimum-logit response to a maximum-logit response gives

$$L(\boldsymbol{\theta}) \ge c\big(\max_i \theta_i - \min_i \theta_i\big).$$

On $\{\boldsymbol{\theta} : \sum_i \theta_i = 0\}$, the logit range is at least $\|\boldsymbol{\theta}\|_\infty$. Thus L is coercive on this subspace and attains its minimum; strict convexity makes the minimizer unique.

Finally, since the mapping

$$\boldsymbol{\theta} \;\longleftrightarrow\; \pi_i = \frac{\pi_{\mathrm{ref},i} e^{\beta\theta_i}}{\sum_k \pi_{\mathrm{ref},k} e^{\beta\theta_k}}$$

is a bijection between the gauge-fixed $\boldsymbol{\theta}$-space $\{\boldsymbol{\theta} \in \mathbb{R}^K, \sum_{i\in[K]} \theta_i = 0\}$ and Δ_K°, we conclude that OP 51 admits a unique optimum $\boldsymbol{\pi}^\star \in \Delta_K^\circ$.

Now we derive the first-order condition for $\boldsymbol{\pi}^\star$. Note that solving OP (51) is equivalent to minimizing $L(\boldsymbol{\theta})$ subject to (53), and we let $\boldsymbol{\theta}^\star = \beta^{-1}(\log \boldsymbol{\pi}^\star - \log \boldsymbol{\pi}_{\mathrm{ref}})$ and define the Bradley–Terry matrix Q by

$$Q_{ij} \;=\; \sigma(\theta_i^\star - \theta_j^\star), \qquad i, j \in [K].$$

For the first step, we derive the KKT condition, which reduces to $\nabla L(\boldsymbol{\theta}^\star) = \mathbf{0}$. Consider the Lagrangian

$$\mathcal{L}(\boldsymbol{\theta}, \lambda) \;=\; L(\boldsymbol{\theta}) + \lambda \left(\sum_{i=1}^{K} \pi_{\mathrm{ref},i} e^{\beta\theta_i} - 1 \right).$$

At an optimal solution $(\boldsymbol{\theta}^\star, \lambda^\star)$, the KKT stationarity conditions read

$$\nabla L(\boldsymbol{\theta}^\star) + \lambda^\star \nabla \left(\sum_{i=1}^{K} \pi_{\mathrm{ref},i} e^{\beta\theta_i} \right) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star} = \mathbf{0}. \tag{56}$$

Since $L(\boldsymbol{\theta})$ depends only on differences $\theta_i - \theta_j$, it is invariant to common shifts: $L(\boldsymbol{\theta}+c\mathbf{1}) = L(\boldsymbol{\theta})$ for all $c \in \mathbb{R}$. Differentiating with respect to c at $c = 0$ gives

$$\mathbf{1}^\top \nabla L(\boldsymbol{\theta}) = 0 \qquad \text{for all } \boldsymbol{\theta}. \tag{57}$$

On the other hand,

$$\nabla \left(\sum_{i=1}^{K} \pi_{\mathrm{ref},i} e^{\beta\theta_i} \right) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^\star} = \beta\, \boldsymbol{\pi}^\star,$$

and $\mathbf{1}^\top(\beta\boldsymbol{\pi}^\star) = \beta > 0$. Taking the inner product of (56) with $\mathbf{1}$ and using (57) yields $\lambda^\star \beta = 0$, hence $\lambda^\star = 0$ and therefore

$$\nabla L(\boldsymbol{\theta}^\star) = \mathbf{0}. \tag{58}$$

Next we compute $\nabla L(\boldsymbol{\theta})$ and identify the matrix form. Recall that $\frac{d}{dx} \log \sigma(x) = \sigma(-x)$. For each $k \in [K]$, only the terms with $i = k$ or $j = k$ depend on θ_k, so

$$\begin{aligned}
\frac{\partial L}{\partial \theta_k}(\boldsymbol{\theta}) &= -\sum_{j=1}^{K} \mu_k \mu_j\, P_{kj}\, \sigma\big(-(\theta_k - \theta_j)\big) + \sum_{i=1}^{K} \mu_i \mu_k\, P_{ik}\, \sigma\big(-(\theta_i - \theta_k)\big) \\
&= -\sum_{j=1}^{K} \mu_k \mu_j\, P_{kj}\, \sigma(\theta_j - \theta_k) + \sum_{i=1}^{K} \mu_i \mu_k\, P_{ik}\, \sigma(\theta_k - \theta_i).
\end{aligned}$$

Evaluating at $\boldsymbol{\theta}^\star$ and using $Q_{jk} = \sigma(\theta_j^\star - \theta_k^\star)$ and $Q_{ki} = \sigma(\theta_k^\star - \theta_i^\star)$ gives

$$\frac{\partial L}{\partial \theta_k}(\boldsymbol{\theta}^\star) = -\sum_{j=1}^{K} \mu_k \mu_j\, P_{kj}\, Q_{jk} + \sum_{i=1}^{K} \mu_i \mu_k\, P_{ik}\, Q_{ki}.$$

Now use the defining property of preference matrices $P \in \mathcal{P}$, namely $P_{ik} = 1 - P_{ki}$, together with the Bradley–Terry symmetry $Q_{ki} = 1 - Q_{ik}$. Then, for each i,

$$P_{ik}Q_{ki} - P_{ki}Q_{ik} = (1 - P_{ki})(1 - Q_{ik}) - P_{ki}Q_{ik} = 1 - P_{ki} - Q_{ik} = P_{ik} - Q_{ik}.$$

Hence

$$\frac{\partial L}{\partial \theta_k}(\boldsymbol{\theta}^\star) = \mu_k \sum_{i=1}^{K} \mu_i\, (P_{ik} - Q_{ik}).$$

Stacking these K equalities gives the vector identity

$$\nabla L(\boldsymbol{\theta}^\star) = \mathrm{Diag}(\boldsymbol{\mu})\, (P - Q)^\top \boldsymbol{\mu}.$$

Combining the above expression with (58) yields

$$\mathrm{Diag}(\boldsymbol{\mu})\, (P - Q)^\top \boldsymbol{\mu} = \mathbf{0},$$

which is exactly (52). □

E.2 The dependency of DPO solutions on sampling

To study how DPO solution depends on sampling strategy, we assume the reference policy in the DPO population loss $\boldsymbol{\pi}_{\mathrm{ref}} = \boldsymbol{\mu}_0$ be the uniform distribution on Δ_K without loss of generality and let $\boldsymbol{\mu}$ have full support. For any logit vector $\boldsymbol{\theta} \in \mathbb{R}^K$, let $Q(\boldsymbol{\theta})$ denote the Bradley–Terry matrix with entries $Q_{ij}(\boldsymbol{\theta}) = \sigma(\theta_i - \theta_j)$. Then, the first-order condition (FOC) (52) given in Proposition E.1 can be further simplified as

$$F(\boldsymbol{\theta}^\star, \boldsymbol{\mu}) := (P - Q(\boldsymbol{\theta}^\star))^\top \boldsymbol{\mu} = \mathbf{0}. \tag{59}$$

The new FOC (59) gives an implicit mapping from any sampling strategy $\boldsymbol{\mu} \in \Delta_K^\circ$ to the unique DPO solution $\boldsymbol{\pi}^\star(P;\boldsymbol{\mu})$ by specifying its logits $\boldsymbol{\theta}^\star(P;\boldsymbol{\mu}) = \beta^{-1}\log \boldsymbol{\pi}^\star(P;\boldsymbol{\mu})$ up to a constant shift. The following result shows that the unique solution is invariant to sampling when and only when the underlying preference structure P is derived from a BT model.

Proposition E.2 (Sampling Invariance Occurs *Only* Under True Bradley–Terry Preferences)**.** *Let $P \in \mathcal{P}$ be an arbitrary preference matrix. The solution $\boldsymbol{\pi}^\star(P;\boldsymbol{\mu})$ of OP (51) is independent of the sampling strategy $\boldsymbol{\mu}$ (i.e., $\boldsymbol{\pi}^\star(P;\boldsymbol{\mu})$ is a constant over all $\boldsymbol{\mu} \in \Delta_K$ with full support)* if and only if *P is exactly representable as a Bradley–Terry model, i.e.,*

$$P_{ij} = \sigma(r_i - r_j) \quad \textit{for some } r \in \mathbb{R}^K.$$

Proposition E.2 highlights a crucial but underappreciated aspect of RLHF-style preference learning:

- If the true pairwise preferences *do not* come from a BT model (which is almost always the case), then *DPO does not recover an intrinsic "ground truth" alignment policy.* Instead, it recovers a policy that also depends on the sampling strategy. Such a message is self-evident for ΨPO due to the weighted square-sum format of its loss function, but is not explicitly pointed out for DPO to the best of our knowledge.
- The only sampling-invariant situation is the perfectly idealized world where human preferences admit a single global utility function r_i, and where P_{ij} depends solely on $r_i - r_j$. This assumption is extremely restrictive: real-world preference judgments violate transitivity, context independence, and independence of irrelevant alternatives—all incompatible with the BT model.

Thus, the sampling policy is not merely a data-collection choice: it acts as an implicit regularizer that selects *which* among many possible BT approximations the DPO objective converges to. This has an important implication:

> *Even in the infinite-data limit, DPO learns a policy determined jointly by the preference structure P and the sampling distribution $\boldsymbol{\mu}$.*

In other words, "more data" does not wash out the effect of sampling; rather, the sampling strategy systematically biases the learned policy.

Proof. We prove the "if" and "only if" directions of the claim separately.

(If). Suppose P is BT-consistent, i.e., $\exists \bar{\boldsymbol{\theta}}$ such that $P_{ij} = \sigma(\bar{\theta}_i - \bar{\theta}_j), \forall i \neq j$. Then for any $\boldsymbol{\mu}$, the FOC (59) gives

$$F(\bar{\boldsymbol{\theta}}, \boldsymbol{\mu}) = (P - Q(\bar{\boldsymbol{\theta}}))^\top \boldsymbol{\mu} = \mathbf{0}.$$

By the strict convexity in logits and uniqueness established in Proposition E.1, the corresponding policy is the unique optimum for every $\boldsymbol{\mu}$.

(Only If). Conversely, suppose that $\boldsymbol{\pi}^\star(P;\boldsymbol{\mu})$ is the same policy for every full-support $\boldsymbol{\mu}$. Fix its reference-relative logits $\bar{\boldsymbol{\theta}}$, and write $R = P - Q(\bar{\boldsymbol{\theta}})$. The FOC gives $R^\top \boldsymbol{\mu} = \mathbf{0}$ for every $\boldsymbol{\mu} \in \Delta_K^\circ$. For each i, let $\boldsymbol{\mu}$ approach $\boldsymbol{e}_i$ through full-support distributions. Continuity gives $R^\top \boldsymbol{e}_i = \mathbf{0}$ for every i, hence $R = 0$ and $P = Q(\bar{\boldsymbol{\theta}})$. □

E.3 Long-term effects of iterative DPO

We first define the mixed reference/sampling dynamics for DPO as a counterpart to Definition 4.1.

Definition E.3. *The* mixed reference/sampling DPO (MRS-DPO) dynamics

$$\boldsymbol{\pi}_{t+1} = \mathrm{MRS}^{DPO}(\boldsymbol{\pi}_t; P, \alpha, \beta, \lambda, \boldsymbol{\pi}_{\mathrm{ref}}, \boldsymbol{\pi}_0),$$

initialized by $\boldsymbol{\pi}_t|_{t=1} = \boldsymbol{\pi}_1$ *is defined by*

$$\boldsymbol{\pi}_{t+1} = \mathrm{softmax}\Big(\log \boldsymbol{\pi}_{\mathrm{ref}}^{(t)} + \beta\, \boldsymbol{\theta}^\star(P, \boldsymbol{\mu}_t)\Big), t \geq 1, \tag{60}$$

where $\boldsymbol{\theta}^\star(P, \boldsymbol{\mu})$ *is the implicit mapping induced by the FOC condition*

$$F(\boldsymbol{\theta}^\star, \boldsymbol{\mu}) := (P - Q(\boldsymbol{\theta}^\star))^\top \boldsymbol{\mu} = \mathbf{0}, \tag{61}$$

$Q(\boldsymbol{\theta})$ *is the BT matrix induced by* $\boldsymbol{\theta}$, $\boldsymbol{\pi}_{\mathrm{ref}}^{(t)}$ *and* $\boldsymbol{\mu}_t$ *are defined in* (4), (5), $\alpha \in [0,1]$ *controls how strongly the* reference *model tracks the current deployment,* $\lambda \in [0,1]$ *controls how strongly the* sampling *is on-policy,* $\beta > 0$ *is the DPO inverse temperature, and* P *is the underlying pairwise preference matrix.*

Our next Proposition E.4 is a counterpart for Proposition 4.2, illustrating that under a preference structure with Condorcet cycles, the MRS-DPO dynamics can exhibit the same local instability as MRS-ΨPO.

Proposition E.4 (Local instability of MRS–DPO under a Condorcet cycle)**.** *Let* $K = 3$ *and let* $P = P_a^{\mathrm{RPS}} \in \mathcal{P}$ *be the rock–paper–scissors (Condorcet cycle) preference matrix*

$$P_a^{\mathrm{RPS}} = \begin{pmatrix} \frac{1}{2} & \frac{1}{2}+a & \frac{1}{2}-a \\ \frac{1}{2}-a & \frac{1}{2} & \frac{1}{2}+a \\ \frac{1}{2}+a & \frac{1}{2}-a & \frac{1}{2} \end{pmatrix}, \qquad a \in (0, \tfrac{1}{2}).$$

Assume $\boldsymbol{\pi}_{\mathrm{ref}} = \boldsymbol{\pi}_0 = \boldsymbol{\mu}_0 \triangleq (\frac{1}{3}, \frac{1}{3}, \frac{1}{3}) \in \Delta_3^\circ$ *and*

$$\boldsymbol{\pi}_{\mathrm{ref}}^{(t)} \propto \boldsymbol{\mu}_0^{1-\alpha} \odot \boldsymbol{\pi}_t^{\alpha}, \qquad \boldsymbol{\mu}_t = (1-\lambda)\boldsymbol{\mu}_0 + \lambda \boldsymbol{\pi}_t.$$

Consider the MRS–DPO map defined in Definition E.3:

$$\boldsymbol{\pi}_{t+1} = \mathrm{softmax}\Big(\log \boldsymbol{\pi}_{\mathrm{ref}}^{(t)} + \beta\, \boldsymbol{\theta}^\star(P, \boldsymbol{\mu}_t)\Big), \qquad \beta > 0.$$

Then $\boldsymbol{\mu}_0$ *is a fixed point. Moreover, if*

$$\alpha^2 + \frac{16a^2}{3}\beta^2\lambda^2 \;>\; 1, \tag{62}$$

then trajectories starting sufficiently close to $\boldsymbol{\mu}_0$*, but not at* $\boldsymbol{\mu}_0$*, initially move away with a rotational component and do not converge to* $\boldsymbol{\mu}_0$*.*

Proof. Reference logits in this proof are understood up to additive constants, which leave softmax unchanged. We first show $\boldsymbol{\mu}_0$ is a fixed point. Let $\boldsymbol{\mu} = \boldsymbol{\mu}_0$ and $\boldsymbol{\theta} = \mathbf{0}$. Then $Q(\mathbf{0})_{ij} = \sigma(0) = \frac{1}{2}$ for all i, j. Since $P_a^{\mathrm{RPS}} - \frac{1}{2}\mathbf{1}\mathbf{1}^\top$ is skew-symmetric with zero column sums, we have $(P_a^{\mathrm{RPS}} - Q(\mathbf{0}))^\top \boldsymbol{\mu}_0 = \mathbf{0}$. Hence $\boldsymbol{\theta}^\star(P, \boldsymbol{\mu}_0)$ may be chosen as $\mathbf{0}$. With $\boldsymbol{\pi}_{\mathrm{ref}}^{(t)} = \boldsymbol{\mu}_0$ at $\boldsymbol{\pi}_t = \boldsymbol{\mu}_0$, we obtain

$$\boldsymbol{\pi}_{t+1} = \mathrm{softmax}(\log \boldsymbol{\mu}_0 + \beta \mathbf{0}) = \boldsymbol{\mu}_0,$$

so $\boldsymbol{\mu}_0$ is a fixed point.

Next we derive the linearization of $\boldsymbol{\theta}^\star(P, \boldsymbol{\mu})$ at $\boldsymbol{\mu} = \boldsymbol{\mu}_0$. Define $F(\boldsymbol{\theta}, \boldsymbol{\mu}) \triangleq (P - Q(\boldsymbol{\theta}))^\top \boldsymbol{\mu}$. At $(\boldsymbol{\theta}, \boldsymbol{\mu}) = (\mathbf{0}, \boldsymbol{\mu}_0)$ we have $F(\mathbf{0}, \boldsymbol{\mu}_0) = \mathbf{0}$ as above. We compute the partial derivatives at $(\mathbf{0}, \boldsymbol{\mu}_0)$.

First, F is linear in $\boldsymbol{\mu}$, so

$$\frac{\partial F}{\partial \boldsymbol{\mu}}(\mathbf{0}, \boldsymbol{\mu}_0) = (P - Q(\mathbf{0}))^\top = (P - \tfrac{1}{2}\mathbf{1}\mathbf{1}^\top)^\top.$$

Second, for $Q_{ij}(\boldsymbol{\theta}) = \sigma(\theta_i - \theta_j)$ we have $\partial Q_{ij}/\partial\theta_k = \sigma'(\theta_i - \theta_j)(\delta_{ik} - \delta_{jk})$ and $\sigma'(0) = \frac{1}{4}$. Writing $F_m(\boldsymbol{\theta}, \boldsymbol{\mu}) = \sum_i (P_{im} - Q_{im}(\boldsymbol{\theta}))\mu_i$,

$$\frac{\partial F_m}{\partial \theta_k}(\mathbf{0}, \boldsymbol{\mu}_0) = -\sum_{i=1}^{3} \mu_i \frac{\partial Q_{im}}{\partial \theta_k}(\mathbf{0}) = -\frac{1}{4}\sum_{i=1}^{3} \mu_i(\delta_{ik} - \delta_{mk}) = -\frac{1}{4}(\mu_k - \delta_{mk}).$$

With $\mu_{0,k} = \frac{1}{3}$, this yields the 3×3 matrix

$$\frac{\partial F}{\partial \boldsymbol{\theta}}(\mathbf{0}, \boldsymbol{\mu}_0) = \frac{1}{6} I - \frac{1}{12}(\mathbf{1}\mathbf{1}^\top - I),$$

whose restriction to the tangent space T is simply $\frac{1}{4} I_T$ (because $\mathbf{1}\mathbf{1}^\top$ vanishes on T). Therefore, by the implicit function theorem (modulo the standard shift-invariance gauge, which is immaterial because softmax is shift-invariant),

$$D_{\boldsymbol{\mu}} \boldsymbol{\theta}^\star(P, \boldsymbol{\mu}_0)\big|_T = -\left(\frac{\partial F}{\partial \boldsymbol{\theta}}(\mathbf{0}, \boldsymbol{\mu}_0)\big|_T\right)^{-1} \left(\frac{\partial F}{\partial \boldsymbol{\mu}}(\mathbf{0}, \boldsymbol{\mu}_0)\big|_T\right) = -4\,(P - \tfrac{1}{2}\mathbf{1}\mathbf{1}^\top)^\top\big|_T.$$

For $P \in \mathcal{P}$, the matrix $S \triangleq P - \frac{1}{2}\mathbf{1}\mathbf{1}^\top$ is skew-symmetric, hence $S^\top = -S$, and thus

$$D_{\boldsymbol{\mu}} \boldsymbol{\theta}^\star(P, \boldsymbol{\mu}_0)\big|_T = 4\, S\big|_T. \tag{63}$$

Finally we linearize the full MRS–DPO map at $\boldsymbol{\mu}_0$. Let $\boldsymbol{\pi} = \boldsymbol{\mu}_0 + \boldsymbol{\delta}$ with $\delta \in T$ small. Then

$$\log \boldsymbol{\pi}_{\text{ref}}^{(t)} = (1 - \alpha)\log \boldsymbol{\mu}_0 + \alpha \log(\boldsymbol{\mu}_0 + \boldsymbol{\delta}), \qquad \boldsymbol{\mu}_t = \boldsymbol{\mu}_0 + \lambda\delta$$

Using elementwise Taylor expansion, $\log(\boldsymbol{\mu}_0 + \boldsymbol{\delta}) = \log \boldsymbol{\mu}_0 + 3\boldsymbol{\delta} + o(\|\boldsymbol{\delta}\|)$ (since each coordinate of $\boldsymbol{\mu}_0$ equals $1/3$). Using (63), $\boldsymbol{\theta}^\star(P, \boldsymbol{\mu}_t) = 4\lambda S\boldsymbol{\delta} + o(\|\boldsymbol{\delta}\|)$. Hence the softmax argument is

$$\log \boldsymbol{\pi}_{\text{ref}}^{(t)} + \beta \boldsymbol{\theta}^\star(P, \boldsymbol{\mu}_t) = \log \boldsymbol{\mu}_0 + \Big(3\alpha I + 4\beta\lambda S\Big)\boldsymbol{\delta} + o(\|\boldsymbol{\delta}\|).$$

The Jacobian of softmax at $\log \boldsymbol{\mu}_0$ is $\mathrm{Diag}(\boldsymbol{\mu}_0) - \boldsymbol{\mu}_0\boldsymbol{\mu}_0^\top$; restricted to T it equals $\frac{1}{3} I_T$. Therefore,

$$\boldsymbol{\delta}^+ \triangleq \boldsymbol{\pi}_{t+1} - \boldsymbol{\mu}_0 = \frac{1}{3}\Big(3\alpha I + 4\beta\lambda S\Big)\boldsymbol{\delta} + o(\|\boldsymbol{\delta}\|) = \Big(\alpha I + \frac{4\beta\lambda}{3} S\Big)\boldsymbol{\delta} + o(\|\boldsymbol{\delta}\|).$$

For the RPS matrix, $S = P - \frac{1}{2}\mathbf{1}\mathbf{1}^\top$ equals aA where

$$A = \begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}$$

has eigenvalues $0, \pm i\sqrt{3}$. Thus on T the eigenvalues of the linearized map are

$$\lambda_\pm = \alpha \pm i\,\frac{4\sqrt{3}}{3}\,\beta\lambda\, a.$$

Their modulus is

$$|\lambda_\pm| = \sqrt{\alpha^2 + \frac{16}{3}\beta^2\lambda^2 a^2}.$$

If (62) holds, then $|\lambda_\pm| > 1$, so the fixed point $\boldsymbol{\mu}_0$ is linearly unstable. Finally, if $\alpha = 1$ and $\lambda > 0$, then $|\lambda_\pm| > 1$ for every $\beta > 0$, so the regime holds. □

The next result is the counterpart of Theorem 4.3, recovering a hyperparameter regime of MRS-DPO to ensure stability for general preference structure P.

Theorem E.5 (A sufficient stability regime for MRS–DPO)**.** *Assume $P \in \mathcal{P}$ and fix $\alpha \in [0,1)$ and $\lambda \in [0,1)$, let $\boldsymbol{\pi}_{\mathrm{ref}}, \boldsymbol{\pi}_0 \in \Delta_K^\circ$. Define the on/off-policy sampling mixture*

$$\boldsymbol{\mu}(\boldsymbol{\pi}) \;:=\; (1-\lambda)\boldsymbol{\pi}_0 + \lambda\boldsymbol{\pi}, \qquad \underline{\mu} \;:=\; (1-\lambda)\min_i \pi_{0,i} \;>\; 0.$$

For each $\boldsymbol{\mu} \in \Delta_K^\circ$, let $\boldsymbol{\theta}^\star(P,\boldsymbol{\mu}) \in \mathbb{R}^K$ denote the (gauge-fixed) DPO optimizer in the θ-parameterization, i.e. $\mathbf{1}^\top\boldsymbol{\theta}^\star = 0$.

Assume the DPO optimizer is uniformly bounded on the sampling set induced by λ:

$$B \;:=\; \sup_{\boldsymbol{\pi}\in\Delta_K^\circ} \Big(\max_i \theta_i^\star(P,\boldsymbol{\mu}(\boldsymbol{\pi})) - \min_i \theta_i^\star(P,\boldsymbol{\mu}(\boldsymbol{\pi}))\Big) \;<\; \infty, \tag{64}$$

and let $s_B := \sigma'(B) > 0$ where $\sigma(t) = 1/(1+e^{-t})$.

Consider the MRS–DPO update defined by

$$\boldsymbol{\pi}_{t+1} = \mathrm{softmax}\Big((1-\alpha)\log\boldsymbol{\pi}_{\mathrm{ref}} + \alpha\log\boldsymbol{\pi}_t + \beta\,\boldsymbol{\theta}^\star(P,\boldsymbol{\mu}(\boldsymbol{\pi}_t))\Big), \qquad t \ge 1, \tag{65}$$

with any initialization $\boldsymbol{\pi}_1 \in \Delta_K^\circ$ and $\beta > 0$. Let $\Pi := I - \frac{1}{K}\mathbf{1}\mathbf{1}^\top$ and define the centered logit coordinate

$$\boldsymbol{x}(\boldsymbol{\pi}) := \Pi\log\boldsymbol{\pi} \in \mathbf{1}^\perp.$$

If

$$L \;:=\; \alpha \;+\; \frac{\beta\lambda}{\underline{\mu}^2\, s_B} \;<\; 1, \tag{66}$$

then the MRS–DPO dynamics is globally stable in centered-logit distance: there exists a unique fixed point $\boldsymbol{\pi}^\infty \in \Delta_K^\circ$ of (65), and for all $t \ge 1$,

$$\big\|\boldsymbol{x}(\boldsymbol{\pi}_t) - \boldsymbol{x}(\boldsymbol{\pi}^\infty)\big\|_2 \;\le\; L^{\,t-1}\,\big\|\boldsymbol{x}(\boldsymbol{\pi}_1) - \boldsymbol{x}(\boldsymbol{\pi}^\infty)\big\|_2.$$

Consequently, $\boldsymbol{\pi}_t \to \boldsymbol{\pi}^\infty$ as $t \to \infty$ (and in particular the dynamics cannot cycle).

Proof. First we rewrite the update as a map on centered logits. Let $\boldsymbol{z}_t := (1-\alpha)\log\boldsymbol{\pi}_{\mathrm{ref}} + \alpha\log\boldsymbol{\pi}_t + \beta\,\boldsymbol{\theta}^\star(P,\boldsymbol{\mu}(\boldsymbol{\pi}_t))$ so that $\boldsymbol{\pi}_{t+1} = \mathrm{softmax}(\boldsymbol{z}_t)$. Using $\log\mathrm{softmax}(\boldsymbol{z}) = \boldsymbol{z} - \log(\sum_i e^{z_i})\mathbf{1}$ and $\Pi\mathbf{1} = 0$, we obtain the exact identity

$$\boldsymbol{x}(\boldsymbol{\pi}_{t+1}) = \Pi\log\boldsymbol{\pi}_{t+1} = \Pi\boldsymbol{z}_t.$$

Moreover, since $\boldsymbol{x}(\boldsymbol{\pi}_t) = \Pi\log\boldsymbol{\pi}_t$, we have $\Pi(\alpha\log\boldsymbol{\pi}_t) = \alpha\boldsymbol{x}(\boldsymbol{\pi}_t)$, and therefore

$$\boldsymbol{x}(\boldsymbol{\pi}_{t+1}) = \alpha\,\boldsymbol{x}(\boldsymbol{\pi}_t) + (1-\alpha)\Pi\log\boldsymbol{\pi}_{\mathrm{ref}} + \beta\,\Pi\boldsymbol{\theta}^\star(P,\boldsymbol{\mu}(\boldsymbol{\pi}_t)). \tag{67}$$

Because we fix the gauge $\mathbf{1}^\top\boldsymbol{\theta}^\star = 0$, we have $\Pi\boldsymbol{\theta}^\star = \boldsymbol{\theta}^\star$ and we may drop Π on that term.

Define the induced map on $\mathbf{1}^\perp$:

$$F(\boldsymbol{x}) := \alpha\boldsymbol{x} + (1-\alpha)\Pi\log\boldsymbol{\pi}_{\mathrm{ref}} + \beta\,\boldsymbol{\theta}^\star\big(P,\boldsymbol{\mu}(\mathrm{softmax}(x))\big),$$

so that $\boldsymbol{x}(\boldsymbol{\pi}_{t+1}) = F(\boldsymbol{x}(\boldsymbol{\pi}_t))$.

For $\boldsymbol{\pi} = \mathrm{softmax}(\boldsymbol{x})$, the Jacobian is $J(\boldsymbol{\pi}) = \mathrm{Diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^\top$. For any $\boldsymbol{v} \in \mathbb{R}^K$,

$$\boldsymbol{v}^\top J(\boldsymbol{\pi})\boldsymbol{v} = \sum_i \pi_i v_i^2 - \Big(\sum_i \pi_i v_i\Big)^2 = \mathrm{Var}(v_I) \le \frac{(\max_i v_i - \min_i v_i)^2}{4} \le \frac{2\|\boldsymbol{v}\|_2^2}{4} = \frac{1}{2}\|\boldsymbol{v}\|_2^2,$$

where $I \sim \boldsymbol{\pi}$ and we used $\max_i v_i - \min_i v_i \le \sqrt{2}\|\boldsymbol{v}\|_2$. Thus $\|J(\boldsymbol{\pi})\|_{2\to2} \le 1/2$ uniformly, and by the mean value theorem,

$$\|\mathrm{softmax}(\boldsymbol{x}) - \mathrm{softmax}(\boldsymbol{y})\|_2 \le \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2. \tag{68}$$

Now we derive the Lipschitz continuity of the DPO optimizer map $\boldsymbol{\mu} \mapsto \boldsymbol{\theta}^\star(P, \boldsymbol{\mu})$. Fix $\boldsymbol{\mu} \in \Delta_K^\circ$ and let $Q(\boldsymbol{\theta})$ be the Bradley–Terry matrix induced by $\boldsymbol{\theta}$, $Q_{ij}(\boldsymbol{\theta}) = \sigma(\theta_i - \theta_j)$. Under the dueling identity, the DPO population objective in $\boldsymbol{\theta}$ has gradient

$$g(\boldsymbol{\theta}, \boldsymbol{\mu}) := \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{\mu}) = \mathrm{Diag}(\boldsymbol{\mu}) \left(Q(\boldsymbol{\theta}) - P\right) \boldsymbol{\mu},$$

and Hessian on $\mathbf{1}^\perp$ given by the Laplacian form

$$H(\boldsymbol{\theta}, \boldsymbol{\mu}) = \sum_{1 \le i < j \le K} \mu_i \mu_j \, \sigma'(\theta_i - \theta_j) \, (\boldsymbol{e}_i - \boldsymbol{e}_j)(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top.$$

By (64), all the optimizers lie in the convex set $\mathcal{C}_B := \{\boldsymbol{\theta} \in \mathbf{1}^\perp : \max_i \theta_i - \min_i \theta_i \le B\}$. For every $\boldsymbol{\theta} \in \mathcal{C}_B$, we have $\sigma'(\theta_i - \theta_j) \ge s_B$. Also, for $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{\pi})$ we have $\mu_i \ge \underline{\mu}$. Hence for any $\boldsymbol{\theta} \in \mathcal{C}_B$ and $\boldsymbol{v} \in \mathbf{1}^\perp$,

$$\boldsymbol{v}^\top H(\boldsymbol{\theta}, \boldsymbol{\mu}) \boldsymbol{v} \ge \underline{\mu}^2 s_B \sum_{i<j} (v_i - v_j)^2 = \underline{\mu}^2 s_B \, K \, \|\boldsymbol{v}\|_2^2,$$

where we used $\sum_{i<j}(v_i - v_j)^2 = K\|\boldsymbol{v}\|_2^2$ on $\mathbf{1}^\perp$. Therefore, $g(\cdot, \boldsymbol{\mu})$ is strongly monotone on $\mathcal{C}_B$ with modulus

$$m := \underline{\mu}^2 s_B K. \tag{69}$$

Next, fix $\boldsymbol{\theta}$ and write $M := Q(\boldsymbol{\theta}) - P$. Since $M_{ij} \in [-1, 1]$, we have $\|M\|_2 \le \|M\|_F \le K$. For any $\boldsymbol{\mu}, \boldsymbol{\nu} \in \Delta_K$,

$$g(\boldsymbol{\theta}, \boldsymbol{\mu}) - g(\boldsymbol{\theta}, \boldsymbol{\nu}) = \mathrm{Diag}(\boldsymbol{\mu} - \boldsymbol{\nu}) M \boldsymbol{\mu} + \mathrm{Diag}(\boldsymbol{\nu}) M (\boldsymbol{\mu} - \boldsymbol{\nu}),$$

and thus

$$\|g(\boldsymbol{\theta}, \boldsymbol{\mu}) - g(\boldsymbol{\theta}, \boldsymbol{\nu})\|_2 \le \|\boldsymbol{\mu} - \boldsymbol{\nu}\|_2 \, \|M\boldsymbol{\mu}\|_2 + \|M\|_2 \, \|\boldsymbol{\mu} - \boldsymbol{\nu}\|_2 \le K\|\boldsymbol{\mu} - \boldsymbol{\nu}\|_2 + K\|\boldsymbol{\mu} - \boldsymbol{\nu}\|_2 = 2K\|\boldsymbol{\mu} - \boldsymbol{\nu}\|_2.$$

For $\boldsymbol{\mu}, \boldsymbol{\nu}$ in the sampling set induced by λ, let $\boldsymbol{\theta}_\mu^\star := \boldsymbol{\theta}^\star(P, \boldsymbol{\mu})$ and $\boldsymbol{\theta}_\nu^\star := \boldsymbol{\theta}^\star(P, \boldsymbol{\nu})$ (both in $\mathbf{1}^\perp$), so $g(\boldsymbol{\theta}_\mu^\star, \boldsymbol{\mu}) = 0$ and $g(\boldsymbol{\theta}_\nu^\star, \boldsymbol{\nu}) = 0$. Since the segment joining these two optimizers lies in $\mathcal{C}_B$, strong monotonicity (69) gives

$$m\|\boldsymbol{\theta}_\mu^\star - \boldsymbol{\theta}_\nu^\star\|_2 \le \|g(\boldsymbol{\theta}_\mu^\star, \boldsymbol{\mu}) - g(\boldsymbol{\theta}_\nu^\star, \boldsymbol{\mu})\|_2 = \|g(\boldsymbol{\theta}_\nu^\star, \boldsymbol{\nu}) - g(\boldsymbol{\theta}_\nu^\star, \boldsymbol{\mu})\|_2 \le 2K\|\boldsymbol{\mu} - \boldsymbol{\nu}\|_2,$$

and hence

$$\|\boldsymbol{\theta}^\star(P, \boldsymbol{\mu}) - \boldsymbol{\theta}^\star(P, \boldsymbol{\nu})\|_2 \le \frac{2K}{m} \|\boldsymbol{\mu} - \boldsymbol{\nu}\|_2 = \frac{2}{\underline{\mu}^2 s_B} \|\boldsymbol{\mu} - \boldsymbol{\nu}\|_2. \tag{70}$$

Finally, we derive contraction of the centered-logit map. Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{1}^\perp$ and set $\boldsymbol{\pi} = \mathrm{softmax}(\boldsymbol{x})$, $\boldsymbol{\pi}' = \mathrm{softmax}(\boldsymbol{y})$. Since $\boldsymbol{\mu}(\boldsymbol{\pi}) = (1 - \lambda)\boldsymbol{\pi}_0 + \lambda\boldsymbol{\pi}$, we have

$$\|\boldsymbol{\mu}(\boldsymbol{\pi}) - \boldsymbol{\mu}(\boldsymbol{\pi}')\|_2 = \lambda\|\boldsymbol{\pi} - \boldsymbol{\pi}'\|_2 \le \frac{\lambda}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2$$

by (68). Combining with (70) gives

$$\|\boldsymbol{\theta}^\star(P, \boldsymbol{\mu}(\boldsymbol{\pi})) - \boldsymbol{\theta}^\star(P, \boldsymbol{\mu}(\boldsymbol{\pi}'))\|_2 \le \frac{2}{\underline{\mu}^2 s_B} \cdot \frac{\lambda}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2 = \frac{\lambda}{\underline{\mu}^2 s_B}\|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

Therefore,

$$\|F(\boldsymbol{x}) - F(\boldsymbol{y})\|_2 \le \alpha\|\boldsymbol{x} - \boldsymbol{y}\|_2 + \beta \cdot \frac{\lambda}{\underline{\mu}^2 s_B}\|\boldsymbol{x} - \boldsymbol{y}\|_2 = \left(\alpha + \frac{\beta\lambda}{\underline{\mu}^2 s_B}\right)\|\boldsymbol{x} - \boldsymbol{y}\|_2 = L\|\boldsymbol{x} - \boldsymbol{y}\|_2.$$

Under the condition (66), $L < 1$, so F is a contraction on the complete space $(\mathbf{1}^\perp, \|\cdot\|_2)$. By Banach's fixed point theorem, F admits a unique fixed point $\boldsymbol{x}^\infty$ and $\boldsymbol{x}(\boldsymbol{\pi}_t) \to \boldsymbol{x}^\infty$ geometrically. Let $\boldsymbol{\pi}^\infty := \mathrm{softmax}(\boldsymbol{x}^\infty)$; then $\boldsymbol{\pi}^\infty$ is the unique fixed point of (65) and $\boldsymbol{\pi}_t \to \boldsymbol{\pi}^\infty$. □

Finally, we present the long-term characterization of MRS-DPO dynamics under SST preference structures in the following Theorem, which corresponds to Theorem 4.4 and reveals a similar concern of entropy collapse.

Theorem E.6 (Entropy/peak-collapse of MRS–DPO under SST)**.** *Assume that $P \in \mathcal{P}$ is SST, and fix $\lambda \in [0,1)$. Let $\boldsymbol{\pi}_0, \boldsymbol{\pi}_{\text{ref}}, \boldsymbol{\pi}_1 \in \Delta_K^\circ$. Consider the MRS–DPO dynamics of Definition E.3,*

$$\boldsymbol{\pi}_{t+1} = \text{softmax}\Big(\log \boldsymbol{\pi}_{\text{ref}}^{(t)} + \beta\, \boldsymbol{\theta}^\star(P, \boldsymbol{\mu}_t) \Big), \qquad t \geq 1,$$

where $\boldsymbol{\theta}^\star(P,\boldsymbol{\mu})$ is defined implicitly by the DPO FOC $F(\boldsymbol{\theta}^\star, \boldsymbol{\mu}) = (P - Q(\boldsymbol{\theta}^\star))^\top \boldsymbol{\mu} = \mathbf{0}$, and $Q(\boldsymbol{\theta})$ is the BT matrix induced by $\boldsymbol{\theta}$.

Define the (SST) gap constant

$$\delta \;:=\; \min_{i \in [K-1]} \Big\{ (1-\lambda) \sum_{j=1}^{K} \pi_{0,j}\, (P_{ij} - P_{i+1,j}) \Big\} \;>\; 0, \tag{71}$$

(where $\delta > 0$ follows from $\lambda < 1$, $\boldsymbol{\pi}_0 \in \Delta_K^\circ$, and SST).

Then the MRS–DPO dynamics exhibits the following two forms of collapse.

- ***(ε-collapse for $\alpha, \lambda \in [0,1)$ when $\beta/(1-\alpha)$ is large).*** *Let*

$$\kappa \;:=\; \min_{i \in [K-1]} \log \frac{\pi_{\text{ref},i}}{\pi_{\text{ref},i+1}} \quad (\text{finite since } \boldsymbol{\pi}_{\text{ref}} \in \Delta_K^\circ), \quad c \;:=\; \frac{1}{2}\Big(\kappa + \frac{4\beta\delta}{1-\alpha}\Big), \quad \varepsilon \;:=\; \big(e^c - 1\big)^{-1}.$$

  If $\alpha, \lambda \in [0,1)$ and $c > 0$ (in particular, if $\beta\delta/(1-\alpha)$ is sufficiently large), then there exists a finite T such that for all $t \geq T$,

$$\boldsymbol{\pi}_{t,1} \;\geq\; 1 - \varepsilon. \tag{72}$$

  In particular, if $\boldsymbol{\pi}_{\text{ref}}$ is aligned *with the SST order (i.e. $\pi_{\text{ref},1} \geq \cdots \geq \pi_{\text{ref},K}$, so that $\kappa \geq 0$), then one may take $c = \frac{2\beta\delta}{1-\alpha}$ and hence $\varepsilon = (\exp(\frac{2\beta\delta}{1-\alpha}) - 1)^{-1}$.*

- ***(True entropy collapse when $\alpha = 1$).*** *If $\alpha = 1$, then $\boldsymbol{\pi}_t \to \boldsymbol{e}_1$ and $H(\boldsymbol{\pi}_t) \to 0$ as $t \to +\infty$. Moreover, there exists a finite T such that for all $t \geq T$,*

$$H(\boldsymbol{\pi}_{t+1}) < H(\boldsymbol{\pi}_t), \tag{73}$$

  where $H(\boldsymbol{\pi}) := -\sum_{i=1}^K \pi_i \log \pi_i$ is the Shannon entropy.

Proof. Throughout the proof, fix t and abbreviate

$$\boldsymbol{\mu} := \boldsymbol{\mu}_t, \qquad \boldsymbol{\theta}^\star := \boldsymbol{\theta}^\star(P,\boldsymbol{\mu}), \qquad Q := Q(\boldsymbol{\theta}^\star), \qquad \boldsymbol{\pi}^+ := \boldsymbol{\pi}_{t+1}, \qquad \boldsymbol{\pi} := \boldsymbol{\pi}_t.$$

We retain the superscript (t) on $\boldsymbol{\pi}_{\text{ref}}^{(t)}$ to distinguish it from the fixed reference $\boldsymbol{\pi}_{\text{ref}}$.

First we utilize the row-form of the DPO FOC and derive a gap lower bound for $\boldsymbol{\theta}^\star$. Since $P, Q \in \mathcal{P}$ we have $(P-Q)^\top = -(P-Q)$ (skew-symmetry), and the FOC $(P-Q)^\top \boldsymbol{\mu} = \mathbf{0}$ is equivalent to

$$(P-Q)\boldsymbol{\mu} = \mathbf{0} \quad \Longleftrightarrow \quad \sum_{j=1}^{K} \mu_j \big(P_{ij} - Q_{ij}\big) = 0, \;\; \forall i \in [K]. \tag{74}$$

Thus for each i,

$$\sum_{j=1}^{K} \mu_j P_{ij} = \sum_{j=1}^{K} \mu_j Q_{ij} = \sum_{j=1}^{K} \mu_j\, \sigma(\theta_i^\star - \theta_j^\star). \tag{75}$$

Fix any $i \in [K-1]$ and subtract (75) for i and $i+1$:

$$\sum_{j=1}^{K} \mu_j \big(P_{ij} - P_{i+1,j}\big) = \sum_{j=1}^{K} \mu_j \Big(\sigma(\theta_i^\star - \theta_j^\star) - \sigma(\theta_{i+1}^\star - \theta_j^\star) \Big). \tag{76}$$

By SST and full support of $\boldsymbol{\mu}$, the left-hand side is strictly positive. Since σ is strictly increasing, this implies $\theta_i^\star > \theta_{i+1}^\star$. Because σ is globally 1/4-Lipschitz (since $\sup_x \sigma'(x) = 1/4$), for each j,

$$\sigma(\theta_i^\star - \theta_j^\star) - \sigma(\theta_{i+1}^\star - \theta_j^\star) \le \frac{1}{4}(\theta_i^\star - \theta_{i+1}^\star).$$

Plugging this into (76) and using $\sum_j \mu_j = 1$ yields the key lower bound

$$\theta_i^\star - \theta_{i+1}^\star \;\ge\; 4\sum_{j=1}^{K} \mu_j \big(P_{ij} - P_{i+1,j}\big). \tag{77}$$

Moreover, since $\boldsymbol{\mu}_t = (1-\lambda)\boldsymbol{\pi}_0 + \lambda\boldsymbol{\pi}_t$ and $(P_{ij} - P_{i+1,j}) \ge 0$ for all j,

$$\sum_{j=1}^{K} \mu_{t,j}(P_{ij} - P_{i+1,j}) \;\ge\; (1-\lambda)\sum_{j=1}^{K} \pi_{0,j}(P_{ij} - P_{i+1,j}) \;\ge\; \delta,$$

so for all t and all $i \in [K-1]$,

$$\theta_i^\star - \theta_{i+1}^\star \;\ge\; 4\delta. \tag{78}$$

Next, we derive the log-ratio recursion formula for the policy iterates. By the softmax and multiplicative form of the update

$$\pi_i^+ = \frac{\pi_{\mathrm{ref},i}^{(t)}\, e^{\beta\theta_i^\star}}{\sum_{k=1}^{K} \pi_{\mathrm{ref},k}^{(t)}\, e^{\beta\theta_k^\star}},$$

hence for each $i \in [K-1]$,

$$\log \frac{\pi_i^+}{\pi_{i+1}^+} = \log \frac{\pi_{\mathrm{ref},i}^{(t)}}{\pi_{\mathrm{ref},i+1}^{(t)}} + \beta(\theta_i^\star - \theta_{i+1}^\star). \tag{79}$$

Assume $\alpha, \lambda \in [0,1)$. Taking coordinate differences in the reference update gives

$$\log \frac{\pi_{\mathrm{ref},i}^{(t)}}{\pi_{\mathrm{ref},i+1}^{(t)}} \;=\; \alpha \log \frac{\pi_i}{\pi_{i+1}} + (1-\alpha)\log \frac{\pi_{\mathrm{ref},i}}{\pi_{\mathrm{ref},i+1}} \;\ge\; \alpha \log \frac{\pi_i}{\pi_{i+1}} + (1-\alpha)\kappa. \tag{80}$$

Combining (79), (80), and (78) yields the uniform recursion

$$\log \frac{\pi_i^+}{\pi_{i+1}^+} \;\ge\; \alpha \log \frac{\pi_i}{\pi_{i+1}} + (1-\alpha)\kappa + 4\beta\delta, \qquad \forall i \in [K-1]. \tag{81}$$

Let the log ratio $r_{t,i} \coloneqq \log(\pi_{t,i}/\pi_{t,i+1})$. Then (81) implies

$$r_{t+1,i} \ge \alpha r_{t,i} + (1-\alpha)\kappa + 4\beta\delta.$$

Solving this affine recursion gives

$$r_{t,i} \;\ge\; \alpha^{t-1} r_{1,i} + (1-\alpha^{t-1})\Big(\kappa + \frac{4\beta\delta}{1-\alpha}\Big). \tag{82}$$

If $c = \frac{1}{2}(\kappa + \frac{4\beta\delta}{1-\alpha}) > 0$, then since $\alpha^{t-1} \to 0$ there exists a finite T such that for all $t \ge T$ and all $i \in [K-1]$,

$$\log \frac{\pi_{t,i}}{\pi_{t,i+1}} \;\ge\; c. \tag{83}$$

From (83), we have $\pi_{t,i+1} \le \pi_{t,i} e^{-c}$ and by induction $\pi_{t,k} \le \pi_{t,1} e^{-(k-1)c}$ for all $k \ge 2$. Hence

$$1 - \pi_{t,1} = \sum_{k=2}^{K} \pi_{t,k} \le \pi_{t,1} \sum_{m=1}^{K-1} e^{-mc} \le \frac{\pi_{t,1}}{e^c - 1},$$

which implies $1 \le \pi_{t,1}(1 + \frac{1}{e^c-1}) = \pi_{t,1} \frac{e^c}{e^c-1}$ and thus

$$\pi_{t,1} \;\ge\; \frac{e^c - 1}{e^c} \;=\; 1 - e^{-c} \;\ge\; 1 - (e^c - 1)^{-1} \;=\; 1 - \varepsilon.$$

This proves (72).

In addition, if $\boldsymbol{\pi}_{\text{ref}}$ is aligned with the order $1 \succ \cdots \succ K$ (nonincreasing), then $\kappa \ge 0$ and the stated simplification follows.

Finally we show the true collapse when $\alpha = 1$ and the induced entropy monotonicity. Since $\alpha = 1$, $\boldsymbol{\pi}_{\text{ref}}^{(t)} = \boldsymbol{\pi}_t$ and (79) becomes

$$\log \frac{\pi_{t+1,i}}{\pi_{t+1,i+1}} = \log \frac{\pi_{t,i}}{\pi_{t,i+1}} + \beta(\theta_i^\star - \theta_{i+1}^\star). \tag{84}$$

By (78), we have $\theta_i^\star - \theta_{i+1}^\star \ge 4\delta > 0$ for all i, hence from (84)

$$\log \frac{\pi_{t+1,i}}{\pi_{t+1,i+1}} \;\ge\; \log \frac{\pi_{t,i}}{\pi_{t,i+1}} + 4\beta\delta,$$

so each adjacent log-ratio diverges to $+\infty$ at least linearly in t. The same geometric-series argument as above then yields $\pi_{t,1} \to 1$, i.e. $\boldsymbol{\pi}_t \to \boldsymbol{e}_1$.

Since $\boldsymbol{\pi}_t \to \boldsymbol{e}_1$, we also have $H(\boldsymbol{\pi}_t) \to 0$. Finally, H is continuous on Δ_K° and strictly Schur-concave, and the update with $\alpha = 1$ is a strict reweighting toward action 1 with uniformly separated weights ($\theta_1^\star > \theta_2^\star > \cdots > \theta_K^\star$ and gaps $\ge 4\delta$). Therefore, there exists a neighborhood of $\boldsymbol{e}_1$ in which the MRS–DPO step strictly decreases entropy. Indeed, if $\pi_{t,1} > \max\{e^{-4\beta\delta}, 1/2\}$, every other coordinate decreases under the update, transferring mass to the already largest coordinate and thereby strictly decreasing entropy. Since $\boldsymbol{\pi}_t \to \boldsymbol{e}_1$, the trajectory eventually enters this neighborhood and remains there, which proves the existence of a finite T such that (73) holds for all $t \ge T$. □

F Experimental Details and Additional Results

F.1 Dataset description

All experiments were conducted using the NVIDIA HelpSteer dataset, which contains real human feedback collected for LLM alignment. We use 35,331 response-level records from the dataset. Each record is associated with a textual prompt and a model-generated response.

Available annotations. Each response is independently evaluated along multiple qualitative attributes. Specifically, the available fields include:

$$\{\texttt{helpfulness, correctness, coherence, complexity, verbosity}\}.$$

These scalar scores are used as the basis for constructing pairwise preference matrices.

An example entry from the dataset is shown below:

```
{ ‘‘prompt’’: ‘‘Explain why the sky is blue.’’, ‘‘response’’: ‘‘The sky appears
blue due to Rayleigh scattering...’’, ‘‘helpfulness’’: 3, ‘‘correctness’’: 4,
‘‘coherence’’: 4, ‘‘complexity’’: 2, ‘‘verbosity’’: 2 }
```

Data filtering and preprocessing. Our analysis focuses on preference dynamics over fixed-size sets of competing responses. To enable a uniform experimental protocol across prompts with varying numbers of responses, we restrict attention to prompts with at least $K = 4$ associated responses and construct preference instances by selecting $K = 4$ responses per prompt uniformly at random. This choice balances expressiveness and data coverage: most prompts in the dataset correspond to at most 4 responses, and preference sets of size 4 are sufficiently rich to admit both ST and cyclic structures.

Resulting structure. After preprocessing, each retained prompt induces a local preference learning instance consisting of a fixed-size set of $K = 4$ responses, together with their corresponding multi-dimensional evaluation scores. This collection of 7,497 local instances forms the basis from which all preference matrices used in the experiments are derived.

F.2 Construction of preference matrices

For learning instance after the preprocessing, each prompt associates with $K = 4$ responses, we construct a local probabilistic preference matrix

$$P \in [0, 1]^{K \times K},$$

where $P_{ij} \in [0, 1]$ represents the probability that response i is preferred over response j, and $P_{ij} + P_{ji} = 1$ for any $i, j \in [K]$.

Randomized attribute selection. Let $\mathcal{D}$ denote the set of available evaluation attributes. For each unordered pair of responses (i, j), we independently sample an attribute

$$d \sim \mathrm{Unif}(\mathcal{D}),$$

and denote by $s_i^{(d)}$ and $s_j^{(d)}$ the corresponding scalar scores.

Matrix Construction. For each (i, j), given the selected attribute d from the last step, we generate the probability related to (i, j) via sigmoid function:

$$P_{ij} = \mathrm{sigmoid}(s_i^{(d)} - s_j^{(d)}) = \frac{\exp\big(s_i^{(d)}\big)}{\exp\big(s_i^{(d)}\big) + \exp\big(s_j^{(d)}\big)}, \qquad P_{ji} = 1 - P_{ij},$$

with $P_{ii} = \frac{1}{2}$ by convention. By construction, the resulting matrix satisfies $P_{ij} + P_{ji} = 1$ for all $i \neq j$ and thus is a valid preference matrix.

While scores on the same attribute yield a ranking over responses, independent randomization of attributes across response pairs can introduce conflicting pairwise comparisons across different dimensions. As a result, the aggregated preference matrix need not correspond to any single latent scoring function, nor even be transitive. We classify them into two categories: ST matrices defined in Definition 3.7, and cyclic matrices that are not transitive.

Preference Structure Verification. To verify ST matrices in the experiments, for each matrix P, we first extract a candidate total order by sorting responses according to their aggregate preference scores $\sum_j P_{ij}$. We then verify whether P satisfies the strongly transitive condition in Definition 3.7 with respect to this order. Matrices that fail this verification are not classified as strongly transitive.

We test a matrix and verify it is cyclic if its induced dominance graph contains a directed cycle, i.e., there exist indices $i_1, \ldots, i_m$ with $m \geq 3$ such that $P_{i_\ell i_{\ell+1}} > \frac{1}{2}$ for all ℓ (indices modulo m).

Matrices that are neither ST nor cyclic are excluded from the analysis.

Statistics. After filtering, the dataset contains 4,924 ST preference matrices and 118 cyclic preference matrices.

F.3 Experimental setup in Section 5

The population-level experiments in Section 5.1 are conducted using preference matrices constructed from the dataset, without additional sampling noise. These matrices represent probabilistic pairwise comparisons aggregated from human evaluation scores, as described in Section F.2.

Unless otherwise stated, for each preference instance and parameter configuration $(\alpha, \beta\lambda)$, we fix $\lambda = 0.5$ and set $\beta = v/0.5$ for each value v of $\beta\lambda$. We initialize the dynamics with $\pi_{\text{ref}} = \pi_0$ equal to the uniform distribution over K responses, set the initial policy $\pi_1 = \pi_0$, and run the MRS-IPO dynamics for $T = 3000$ iterations. This horizon is sufficient to observe convergence, collapse, or persistent non-convergent behavior across all parameter regimes considered.

F.3.1 Preferences matrices for illustrative examples in Section 5.1 and Section F.4.1

We use the same preference matrices for the illustrative examples, both in the noiseless setting in Section 5.1 and the finite sample setting in Section F.4.

The strongly transitive preference matrix used throughout is given by

$$P_{\text{ST}} = \begin{pmatrix} 0.5 & 0.731 & 0.269 & 0.5 \\ 0.269 & 0.5 & 0.119 & 0.269 \\ 0.731 & 0.881 & 0.5 & 0.731 \\ 0.5 & 0.731 & 0.269 & 0.5 \end{pmatrix}.$$

This matrix admits a total order consistent with Definition 3.7, and all pairwise preferences satisfy the strong transitivity inequalities with respect to that order.

The cyclic preference matrix is given by

$$P_{\text{cyc}} = \begin{pmatrix} 0.5 & 0.731 & 0.269 & 0.5 \\ 0.269 & 0.5 & 0.731 & 0.881 \\ 0.731 & 0.269 & 0.5 & 0.731 \\ 0.5 & 0.119 & 0.269 & 0.5 \end{pmatrix},$$

with cycle $3 \succ 1 \succ 2 \succ 3$. The induced dominance graph of this matrix contains directed cycles.

F.3.2 Statistical metrics for aggregated experimental results in Section 5.1 and Section F.5

In Section 5.1 and Section F.5, we consider aggregated phenomena across all instances constructed from the HelpSteer dataset.

Specifically, for ST preferences, policy collapse is quantified using the Shannon entropy of the terminal policy:

$$H(\boldsymbol{\pi}_t) \;=\; -\sum_{i=1}^{K} \pi_{t,i} \log \pi_{t,i}.$$

Low entropy indicates more concentration on a small subset of responses, implying more serious policy collapse.

For cyclic preferences, we measure the oscillatory behavior of the dynamics using the time-averaged variance:

$$\text{CS}(\boldsymbol{\pi}) \;=\; \frac{1}{K} \sum_{i=1}^{K} \text{Var}_{t \geq T_0} \left[\pi_{t,i}\right],$$

computed using policies after a burn-in period T_0. This metric measures temporal variability after burn-in, and increases with the amplitude of persistent oscillations.

F.4 Finite sample preference noise under single parameter variations

We consider preference matrices constructed from the dataset as true underlying preference matrices. To account for the finite sample variability inherent in real-world preference data, we introduce controlled stochastic perturbations to these matrices before running the dynamics and conduct additional experiments examining the behavior of MRS-IPO dynamics under finite sample preference estimation. The goal of these experiments is not to study noise as a perturbation per se, but to assess whether the qualitative dynamical behaviors identified in Section 4 persists when preferences are instantiated from finite data.

Again, we consider the ST matrix and cyclic matrix in Section F.3.1. For each preference matrix $P \in [0,1]^{K\times K}$, we generate an empirical realization $\widehat{P}$ by independently sampling a fixed number n of pairwise comparisons for each unordered response pair. For each pair (i,j), the empirical preference value $\widehat{P}_{ij}$ is obtained by averaging the outcomes of n independent Bernoulli trials with success probability P_{ij}, yielding

$$\widehat{P}_{ij} \in \left\{0, \tfrac{1}{n}, \tfrac{2}{n}, \dots, 1\right\}, \qquad \widehat{P}_{ji} = 1 - \widehat{P}_{ij}.$$

Importantly, this finite sample preference realization $\widehat{P}$ is sampled *once per experimental run* and then held fixed throughout the entire execution of the MRS-IPO dynamics. That is, preference noise is introduced at the level of the preference instance, rather than being resampled over time or depending on the evolving policy π_t. This design isolates the effect of static estimation noise in the preference matrix and avoids confounding preference dynamics with on-policy resampling effects.

For each parameter of α, β, λ, we change its value with two other parameters fixed to observe how the dynamics change as one parameter varies. Unless otherwise stated, all noisy experiments use the same algorithmic initialization as in the noiseless setting. Specifically, we set $\pi_{\text{ref}} = \pi_0$ to be the uniform distribution over the K responses, initialize the policy with $\pi_1 = \pi_0$, and run the dynamics for $T = 3000$ iterations. Throughout these experiments, we set the number of finite sample realizations per preference instance to 50, which provides a stable estimate of the mean trajectory and variability without obscuring instance level effects and fix $n = 5$, which introduces moderate variability without overwhelming the underlying preference signal.

For parameter regimes in which finite sample realizations remain approximately phase aligned (e.g., convergent dynamics under strongly transitive preferences or weakly oscillatory regimes under cyclic preferences), we summarize behavior by reporting the mean trajectory across realizations, with shaded regions indicating $\pm$ one standard error (Figures 8–10 and Figures 4–6).

In contrast, in regimes where the dynamics depend strongly on the realized samples, such as collapse under strongly transitive preferences or strong oscillations under cyclic preferences, finite sample realizations can differ substantially in onset time, phase, or amplitude. In these cases, averaging trajectories can cancel the phenomena and obscure the underlying dynamical structure. Accordingly, we present representative individual finite sample realizations to illustrate the diversity of behaviors that arise in practice (Figure 11, 7).

Overall, these experiments indicate that the qualitative dynamical regimes predicted by the theory, convergence, collapse, and persistent cycling, remain observable even when the idealized structural assumptions (e.g., strict strong transitivity or perfectly cyclic preferences) are mildly violated due to finite sampling noise.

F.4.1 Experimental results

Cyclic preferences. Figures 4–6 present the corresponding experiments for cyclic preference matrices. In parameter regimes where the dynamics weakly oscillate, the oscillatory phases remain approximately aligned across realizations.

In these cases, averaged trajectories capture the qualitative behavior of the system and reveal how varying α, β, or λ modulates the magnitude of oscillations.

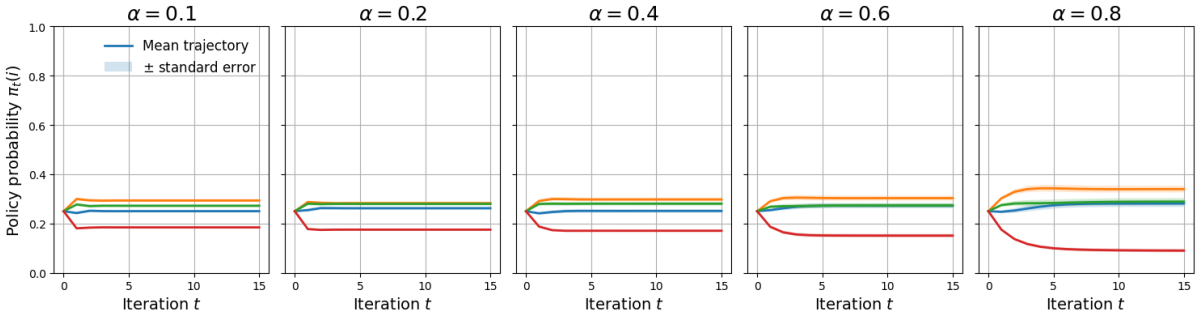


Figure 4: Effect of varying α under cyclic preferences, with $\beta = 2.0$ and $\lambda = 0.8$.

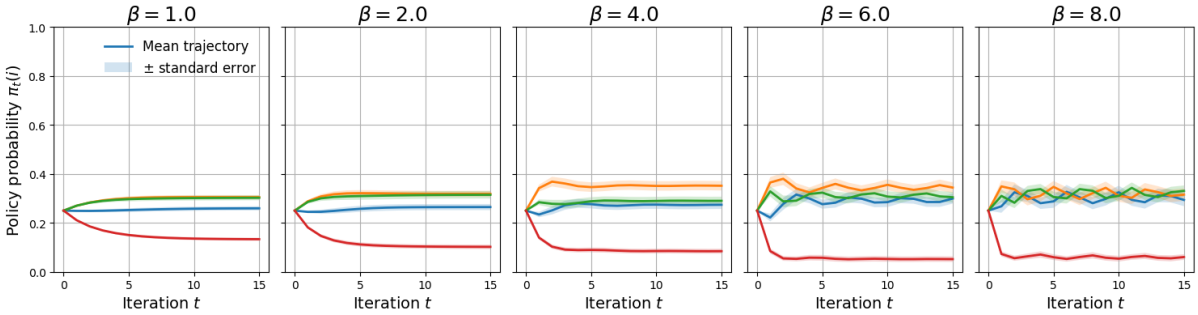


Figure 5: Effect of varying β under cyclic preferences, with $\alpha = 0.8$ and $\lambda = 0.8$.

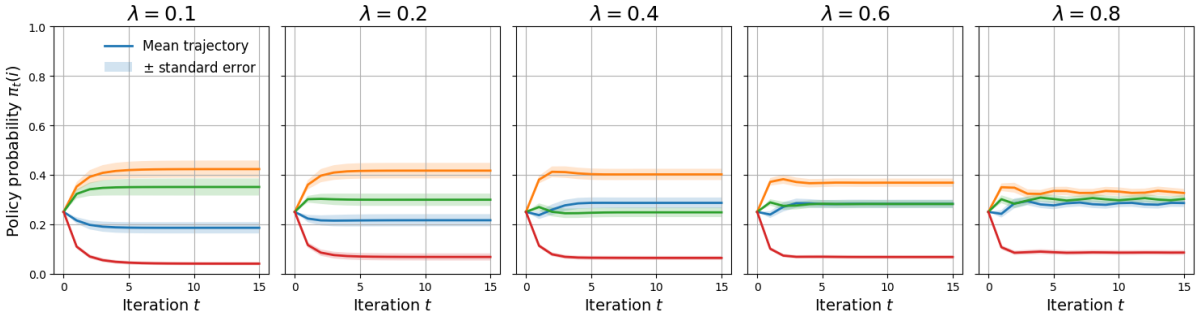


Figure 6: Effect of varying λ under cyclic preferences, with $\alpha = 0.6$ and $\beta = 5.0$.

However, finite-sample realizations can exhibit qualitatively different trajectories, so averaging them can obscure their individual behavior. Accordingly, we complement the averaged results with representative individual finite sample realizations. Figure 7 shows two such realizations generated from the same cyclic preference matrix under identical algorithmic parameters, illustrating the diversity of dynamical behaviors that can arise from finite sample variability.

ST preferences. Figures 8–10 report single parameter variation experiments under ST preferences in parameter regimes where the dynamics converge to a relatively diverse entropy. In these regimes, finite-sample realizations exhibit highly aligned trajectories, and averaging across instances yields a faithful representation of the dynamics. Varying α, β, or λ affects the speed of convergence and the concentration of the limiting policy in a manner consistent with the aggregated results analysis. Finite-sample estimation introduces only mild variability during the transient phase and does not qualitatively alter the convergent behavior.

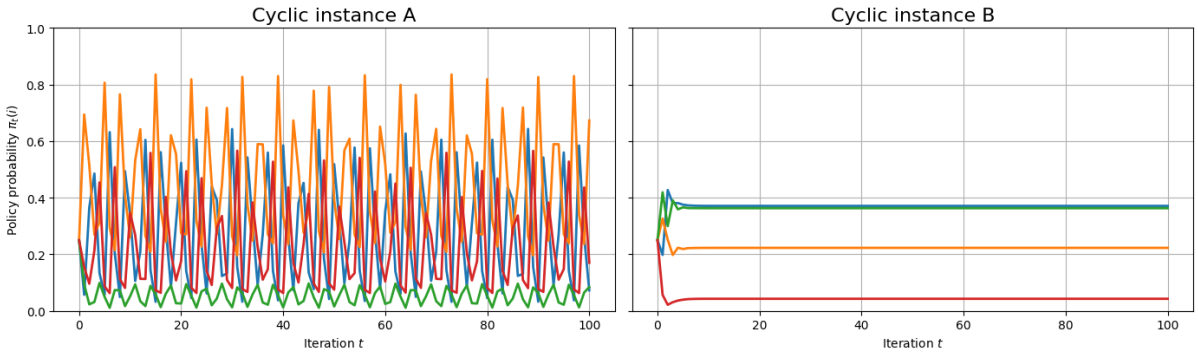


Figure 7: Two representative finite sample realizations generated from the same cyclic preference matrix. $\alpha = 0.9,\ \beta = 10.0,\ \lambda = 0.8$

.

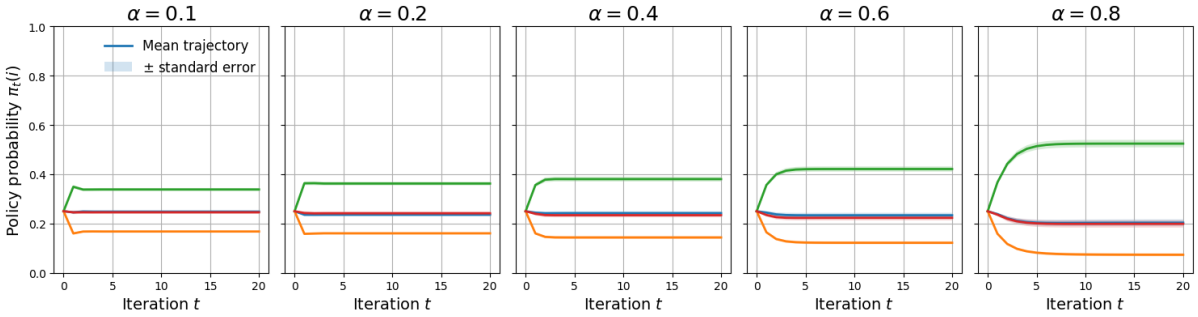


Figure 8: Effect of varying α under ST preferences, with $\beta = 2.0$ and $\lambda = 0.8$.

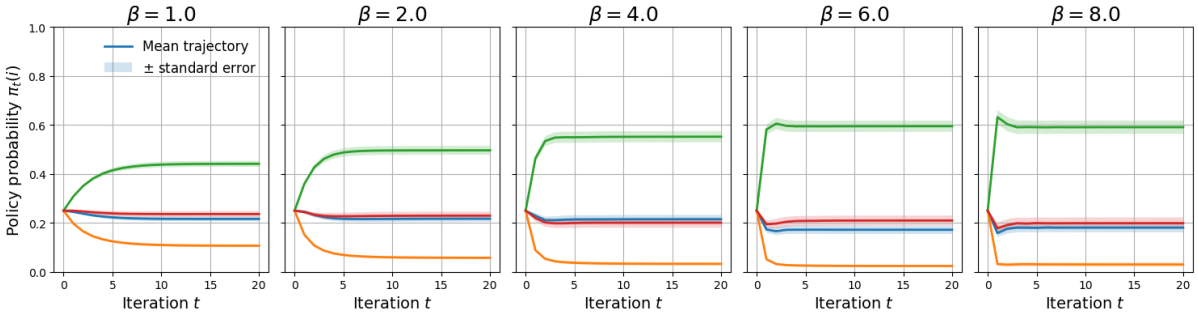


Figure 9: Effect of varying β under ST preferences, with $\alpha = 0.8$ and $\lambda = 0.8$.

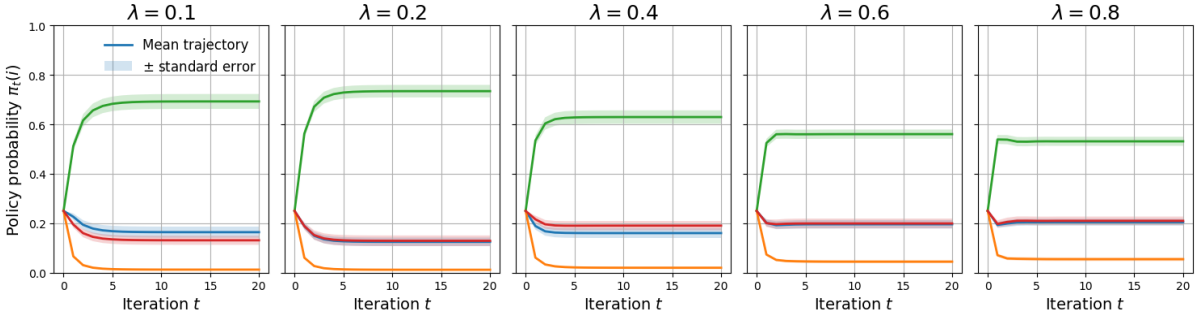


Figure 10: Effect of varying λ under ST preferences, with $\alpha = 0.6$ and $\beta = 5.0$.

The realizations exhibit qualitatively different trajectories, so averaging them can obscure their individual behavior. Therefore, for these parameter regimes, we illustrate representative individual finite-sample realizations rather than reporting aggregated averages. Figure 11 shows two such realizations generated from the same preference matrix, highlighting the diversity of dynamical behaviors induced by finite-sample variability.

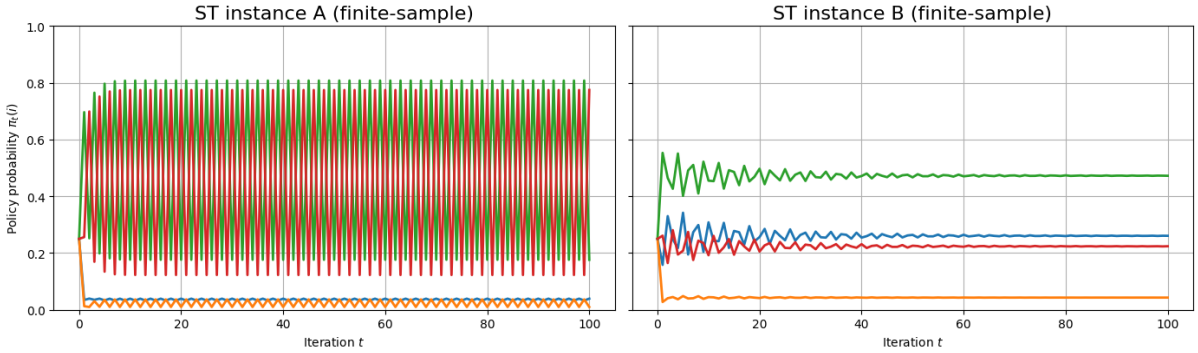


Figure 11: Two representative finite-sample realizations generated from the same ST population preference matrix. $\alpha = 0.5,\ \beta = 10.0,\ \lambda = 0.8$

.

F.5 Complementary aggregated results

This section reports aggregated statistics with β and λ treated separately and more possible parameter values (α, β, λ), as a complement to results in Section 5.1. The goal of these experiments is not to introduce new theoretical regimes, but to verify that the aggregated results predictions derived under the $\beta\lambda$ abstraction remain informative when the two parameters are separated. This separation allows us to examine how β (inverse temperature) and λ (on-policy sampling) influence convergence speed, collapse timing, and oscillation amplitude in practice, while preserving the qualitative structure predicted by the theory.

All statistics are computed over large collections of empirically constructed preference matrices derived from the HelpSteer dataset, consisting of 4924 strongly transitive matrices and 118 cyclic matrices. For each parameter configuration, the MRS-IPO dynamics are run independently on all preference matrices in the corresponding class, and we report the mean and standard deviation of the relevant statistic across matrices.

Convergence/Oscillation in cyclic preferences. Figure 12 reports the mean and standard deviation of cycle strength across all cyclic preference matrices for independently varied (α, β, λ). When β and λ are small, the cycle strength remains close to zero for most values of α, indicating convergence or only weak oscillations. As β or λ increases, persistent oscillatory behavior emerges, and the cycle strength rises. Larger values of α further amplify oscillations by reinforcing self-referential dynamics.

The large standard deviations observed in strong amplification regimes reflect substantial heterogeneity in oscillation amplitude, phase, and period across cyclic preference matrices.

Model collapse in ST preferences. For ST preferences, Figure 13 reports the mean and standard deviation of $H(\pi_T)$ across all strongly transitive preference matrices for a grid of parameter triples (α, β, λ). For small values of β and λ, the final entropy remains high across all α, indicating convergence to non-collapsed policies. Increasing β leads to systematically lower entropy, reflecting stronger preference amplification and faster concentration of mass on top-ranked actions. However, increasing λ leads to a slightly higher entropy. For the largest β values shown, the mean entropy is substantially lower, indicating stronger concentration.

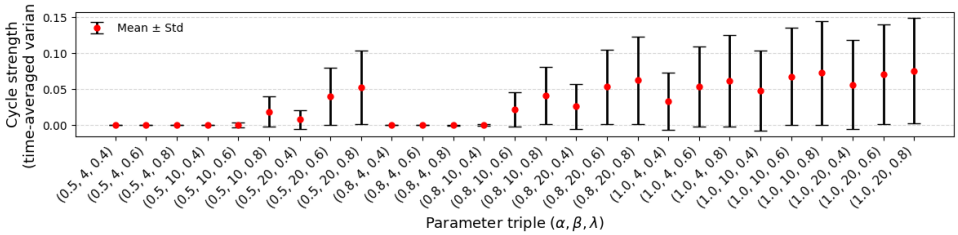


Figure 12: Mean ± standard deviation of cycle strength (time-averaged variance of $\{\pi_t\}$) across all 118 cyclic preference matrices as (α, β, λ) varies.

The standard deviation of entropy is small in weak amplification regimes, indicating consistent convergence behavior across preference matrices. In contrast, near the onset of collapse the variance increases substantially, indicating that the degree of model collapse relies crucially on the specific preference matrices.

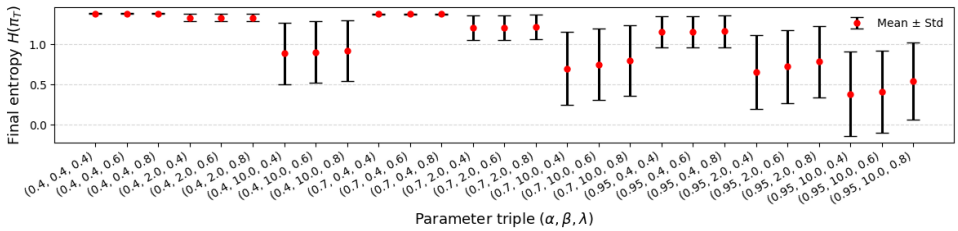


Figure 13: Mean ± standard deviation of the final policy entropy $H(\pi_T)$ across all 4924 strongly transitive preference matrices as (α, β, λ) varies.

Summary. Across thousands of empirically constructed preference instances, these statistics demonstrate that independently varying α, β, and λ produces systematic and interpretable changes in convergence, collapse, and oscillatory behavior.

G Experimental Results for MRS-DPO Dynamics

In this section, we provide additional simulations and evaluations for the Mixed Reference/Sampling DPO (MRS-DPO) dynamics defined in Definition E.3.

Experimental setup and protocol. We utilize the same dataset generation pipeline and preference matrices construction protocol as in MRS-IPO, described in Sections F.1 and F.2. To simulate the MRS-DPO update rule, we compute the implicit optimizer $\theta^\star$ using gradient-based optimization of the DPO loss in relative logits.

Overview of experiments. To validate the theoretical properties of MRS-DPO, we structure our analysis into two main components:

- **Parameter Sensitivity Analysis (Trajectories):** We first examine how the three hyperparameters—α, β, and λ—affect the optimization trajectory. We use the specific illustrative matrices (P_{ST} and P_{cyc}) defined in Section F.3.1 to observe asymptotic behavior in a noiseless setting with fixed preference probabilities.
- **Aggregate Experimental Results:** Finally, mirroring the analysis in Section F.3.2, we quantify the global convergence behavior by measuring the *Policy Entropy* (for ST preferences) and *Cycle Strength* (for Cyclic preferences) across a broader range of conditions.

G.1 Simulation of MRS-DPO under single parameter variations

Cyclic preferences. Figures 14 to 16 present the experimental results for MRS-DPO dynamics on a representative cyclic preference matrix P_{cyc}. In this idealized regime, the dynamics exhibit clear behavioral transitions governed by the stability conditions. Specifically, Figure 14 shows that as α increases, the system loses stability, shifting from convergence to a fixed point to a regime of persistent oscillations. Similarly, Figure 15 demonstrates that increasing β strengthens the preference update, which destabilizes the fixed point and drives the dynamics to oscillate strongly. Figure 16 further confirms this trend: while low values of λ allow sufficient exploration for convergence, increasing λ (greater strength of on-policy sampling) promotes instability, resulting in oscillations of increasing magnitude.

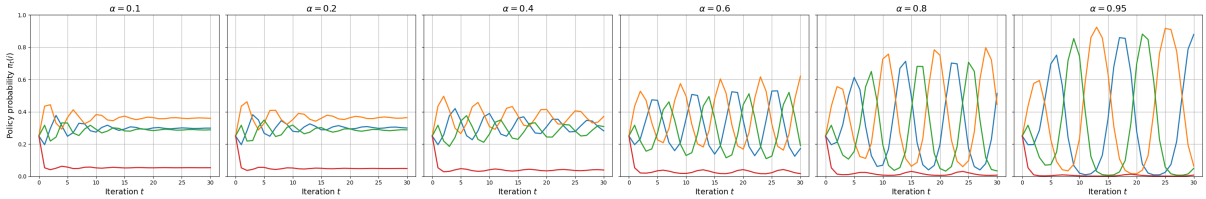


Figure 14: MRS-DPO: Effect of varying α under cyclic preferences (P_{cyc}), with $\beta = 2.0$ and $\lambda = 0.8$.

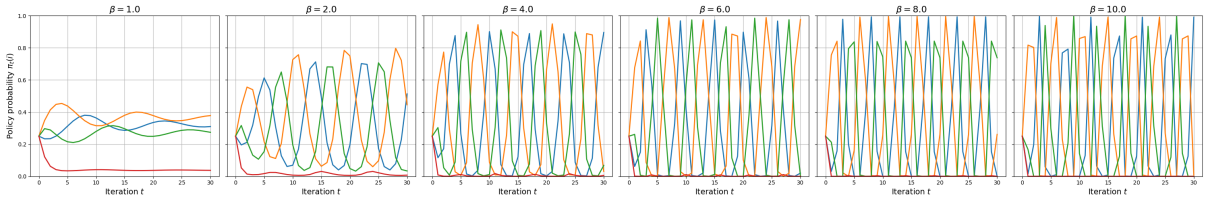


Figure 15: MRS-DPO: Effect of varying β under cyclic preferences (P_{cyc}), with $\alpha = 0.8$ and $\lambda = 0.8$.

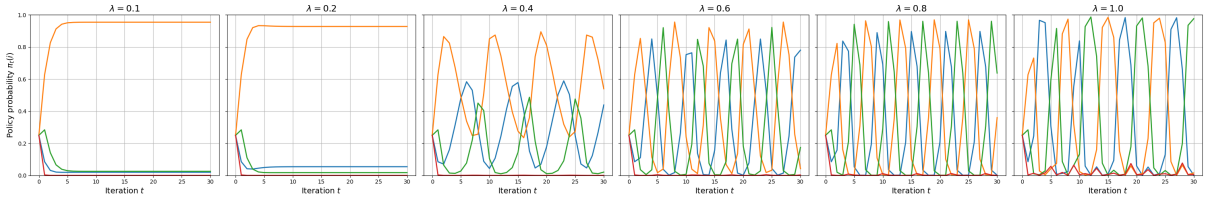


Figure 16: MRS-DPO: Effect of varying λ under cyclic preferences (P_{cyc}), with $\alpha = 0.6$ and $\beta = 5.0$.

ST preferences. We then examine the dynamics under a representative strongly transitive preference matrix P_{ST}. In this idealized regime, we observe clear trends in how the parameters influence policy concentration. Specifically, Figure 17 shows that increasing α yields a more concentrated limiting policy. Similarly, Figure 18 demonstrates that increasing the inverse temperature β leads to a sharper concentration of probability mass on the optimal response. In contrast, Figure 19 indicates that the collapse trend is relatively insensitive to the sampling parameter λ in this noiseless setting.

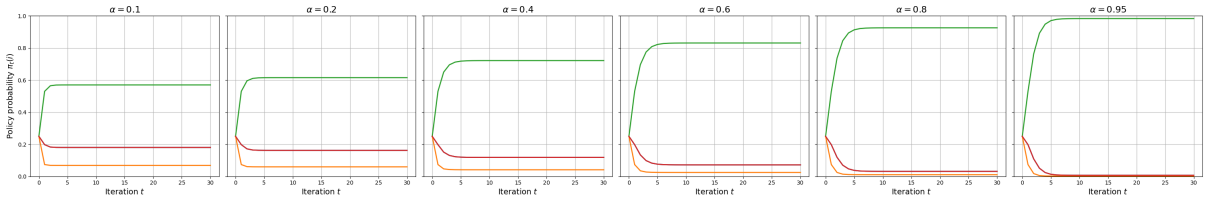


Figure 17: MRS-DPO: Effect of varying α under Strongly Transitive preferences (P_{ST}), with $\beta = 1.0$ and $\lambda = 0.6$.

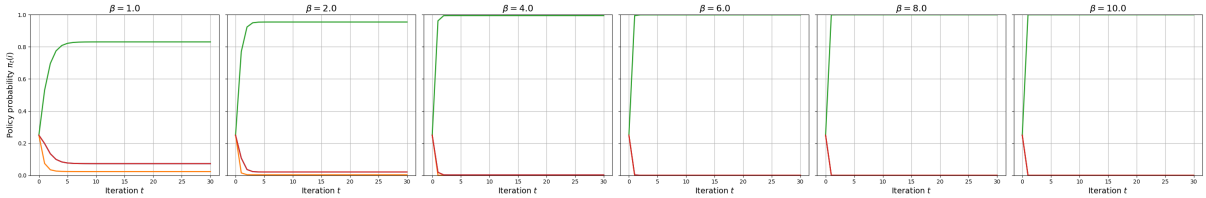


Figure 18: MRS-DPO: Effect of varying β under Strongly Transitive preferences (P_{ST}), with $\alpha = 0.6$ and $\lambda = 0.5$.

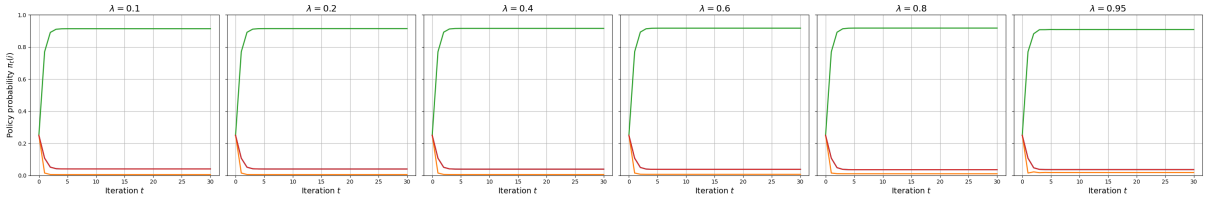


Figure 19: MRS-DPO: Effect of varying λ under Strongly Transitive preferences (P_{ST}), with $\alpha = 0.6$ and $\beta = 2.0$.

G.2 Complementary aggregated results for MRS-DPO

This section reports aggregated statistics for MRS-DPO dynamics with a broader grid of parameter values (α, β, λ), as a complement to the results in Section G.1. The goal of these experiments is to verify that the aggregated predictions derived under the theoretical abstraction remain informative when the parameters are separated in the DPO objective.

All statistics are computed over the same large collections of empirically constructed preference matrices derived from the HelpSteer dataset, consisting of 4924 strongly transitive matrices and 118 cyclic matrices. For each parameter configuration, the MRS-DPO dynamics are run independently on all preference matrices in the corresponding class, and we report the mean and standard deviation of the relevant statistic across matrices.

Convergence/Oscillation in cyclic preferences. Figure 20 reports the mean and standard deviation of cycle strength across all cyclic preference matrices for independently varied (α, β, λ). We observe a clear trend regarding the update and sampling parameters: increasing β or λ leads to a systematic increase in cycle strength, confirming that aggressive updates and skewed on-policy sampling are the primary drivers of instability in DPO. However, in contrast to the sensitivity observed with β and λ, the cycle strength appears relatively insensitive to the reference mixing parameter α. While the oscillation magnitude scales sharply with β and λ, varying α results in only marginal changes to the aggregate cycle strength. This suggests that for MRS-DPO, the global stability characteristics are dominated by the interaction between the inverse temperature β and the on-policy sampling λ, with the reference anchor playing a less significant role in dampening or amplifying the limit cycles compared to the IPO case.

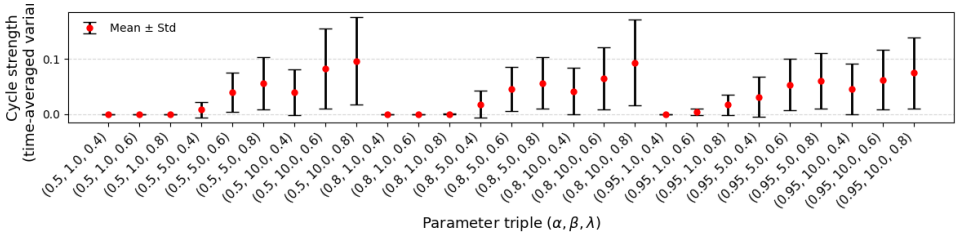


Figure 20: MRS-DPO: Mean ± standard deviation of cycle strength (time-averaged variance of $\{\boldsymbol{\pi}_t\}$) across all 118 cyclic preference matrices as (α, β, λ) varies.

Model collapse in ST preferences. For ST preferences, Figure 21 reports the mean and standard deviation of the final policy entropy $H(\boldsymbol{\pi}_T)$ across all strongly transitive preference matrices. Regarding parameter sensitivity, for small values of β and λ, the final entropy remains very high across all α. As the inverse temperature β increases, the entropy decreases significantly, identifying β as the primary driver of probability concentration in DPO. Increasing the reference weight α also contributes to entropy reduction, though the effect is weaker compared to β. However, increasing the on-policy sampling λ leads to a slight increase in the final entropy in this aggregate setting.

H Iterative ΨPO Experiments on Real LLMs

This section records the implementation used for the real-LLM experiments in Section 5.2. We first specify the shared ΨPO implementation, and then give the two protocols separately: transitive and cyclic. Both protocols train on fixed binary labels and evaluate fixed-support entropy.

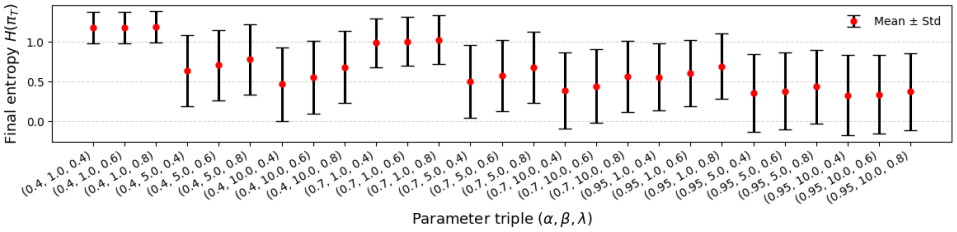


Figure 21: MRS-DPO: Mean ± standard deviation of the final policy entropy $H(\boldsymbol{\pi}_T)$ across all 4924 strongly transitive preference matrices as (α, β, λ) varies.

H.1 Shared model and ΨPO implementation

Model and LoRA parameterization. All runs use Qwen2.5-1.5B as the policy model. The backbone is frozen and we train LoRA adapters [Hu et al., 2022] with rank 16, LoRA scaling 32, and dropout 0.05 on the attention projections $\{\texttt{q_proj}, \texttt{k_proj}, \texttt{v_proj}, \texttt{o_proj}\}$ and MLP projections $\{\texttt{gate_proj}, \texttt{up_proj}, \texttt{down_proj}\}$. The runs use one NVIDIA H100 80GB GPU.

Mixed reference and losses. Let

$$\ell_\theta(y \mid x) = \frac{1}{|y|} \log \pi_\theta(y \mid x)$$

be the response-token-averaged log-probability under the trainable policy, and let $\ell_0(y \mid x)$ denote the corresponding score under a frozen copy of the initial policy π_0. Both IPO and DPO use the same mixed reference construction. In this implementation, the current-policy contribution is recomputed at each minibatch and detached:

$$\ell^{\text{ref}}(y \mid x) = (1-\alpha)\ell_0(y \mid x) + \alpha \, \text{sg}[\ell_\theta(y \mid x)] \,,$$

where sg denotes stop-gradient.

Thus no gradient passes through the reference-side score, while α determines how strongly that score follows the current policy.

For a labeled pair $p = (x, y^+, y^-)$, define

$$\Delta_\theta(p) = \Big(\ell_\theta(y^+ \mid x) - \ell^{\text{ref}}(y^+ \mid x)\Big) - \Big(\ell_\theta(y^- \mid x) - \ell^{\text{ref}}(y^- \mid x)\Big).$$

The pointwise objectives are

$$\ell_{\text{IPO}}(\theta; p) = \left(\Delta_\theta(p) - \frac{\beta}{2}\right)^2, \qquad \ell_{\text{DPO}}(\theta; p) = -\log \sigma\big(\beta^{-1}\Delta_\theta(p)\big) \,.$$

If $\widetilde{w}_t(p)$ denotes a pair weight, the common minibatch objective is

$$\mathcal{L}_t(\theta) = \frac{1}{|\mathcal{B}_t|} \sum_{p \in \mathcal{B}_t} \widetilde{w}_t(p) \, \ell_{\Psi\text{PO}}(\theta; p),$$

where $\ell_{\Psi\text{PO}}$ is the IPO or DPO loss above. The pair weights are specified below.

Optimization. Both objectives use AdamW with batch size 1, gradient accumulation 4, learning rate 10^{-5}, gradient-norm clipping at 1, linear warmup ratio 0.03, one epoch per outer iteration, and maximum sequence length 1537 tokens. Unless stated otherwise, the random seed is 0.

H.2 Transitive setup

Fixed transitive preference pool. We construct the offline pool from the NVIDIA HelpSteer response-level data [Wang et al., 2024]. For every response y, we average the five available attributes,

$$s(y) = \frac{1}{5}\left(s_{\text{help}}(y) + s_{\text{correct}}(y) + s_{\text{coherent}}(y) + s_{\text{complex}}(y) + s_{\text{verbose}}(y)\right).$$

After deduplicating responses, we retain prompts with exactly four responses. All six unordered response pairs are included, with the higher-scored response labeled y^+ and the lower-scored response labeled y^-; exact score ties follow the deterministic construction order. Because every orientation is induced by the same scalar score, the resulting labels are transitive. The pair pool and labels remain fixed throughout training.

Prompt-aware pair sampler. Let $\mathcal{P}_x$ be the six labeled pairs for prompt x. At the beginning of outer iteration t, the current policy assigns each pair the length-normalized margin

$$m_t(p) = \ell_t(y^+ \mid x) - \ell_t(y^- \mid x).$$

The model-induced distribution over pairs is

$$u_t(p \mid x) = \frac{\exp(\tau m_t(p))}{\sum_{p' \in \mathcal{P}_x} \exp(\tau m_t(p'))},$$

and the actual sampling distribution is

$$q_t(p \mid x) = (1-\varepsilon)\left[(1-\lambda)\frac{1}{|\mathcal{P}_x|} + \lambda u_t(p \mid x)\right] + \varepsilon\frac{1}{|\mathcal{P}_x|}.$$

We use $\tau = 1$ and $\varepsilon = 0.05$. Prompts are sampled uniformly without replacement within an outer iteration, and two pairs are then sampled with replacement from $q_t(\cdot \mid x)$ for each selected prompt. The sampled value $q_t(p \mid x)$ is used as the raw loss weight. Across the sampled training set, these weights are divided by their mean and then clipped to $[0.1, 10]$.

Each IPO outer iteration selects 500 prompts and therefore 1,000 pair draws. Each DPO outer iteration selects 250 prompts and therefore 500 pair draws. No reward model is queried, and neither the candidate responses nor their binary labels are regenerated.

H.2.1 Fixed-support entropy evaluation

The data-construction pipeline writes a reproducibly shuffled file of 1,000 prompts with their four dataset responses. For each run, 500 of these prompts are sampled with evaluation seed 123, and their four responses form a panel

$$\mathcal{Y}(x) = \{y_1, \ldots, y_4\}.$$

This panel is fixed for all outer iterations. It is sampled from the same prepared HelpSteer prompt universe as the pair pool and is therefore a fixed measurement panel, not a disjoint held-out training split.

At the beginning of outer iteration t, before the next update, we form

$$\rho_t(y_j \mid x) = \frac{\exp(\tau \ell_t(y_j \mid x))}{\sum_{y' \in \mathcal{Y}(x)} \exp(\tau \ell_t(y' \mid x))}$$

with $\tau = 1$. The reported metric is the mean finite-support entropy

$$\bar{H}_t = \frac{1}{500}\sum_x \left[-\sum_{y \in \mathcal{Y}(x)} \rho_t(y \mid x) \log \rho_t(y \mid x)\right].$$

This is entropy over four fixed candidate responses and should not be interpreted as the entropy of the full open-ended generation distribution.

Transitive sweep. For both MRS-IPO and MRS-DPO, we report

$$\alpha \in \{0.5, 0.6, 0.7, 0.8, 0.9, 0.95, 0.97, 0.99\}, \qquad \lambda \in \{0.5, 0.9\}, \qquad \beta = 1.$$

All curves and table entries use the same 100 pre-update snapshots $t = 0, \ldots, 99$, with one run per configuration and seed 0.

H.3 Cyclic setup

The cyclic protocol uses the same HelpSteer source, exact-four-response filter, Qwen/LoRA implementation, pair sampler, optimizer, and fixed-support entropy definition as the transitive protocol. The only change is the orientation of the fixed binary labels.

For each prompt, the four responses are ordered by decreasing mean HelpSteer score and denoted y_0, y_1, y_2, y_3. We construct the full six-edge tournament

$$\begin{aligned} &y_0 \succ y_1, \quad y_1 \succ y_2, \quad y_2 \succ y_3, \quad y_3 \succ y_0, \\ &y_0 \succ y_2, \quad y_1 \succ y_3. \end{aligned}$$

The first four edges form the directed cycle, while the last two complete the tournament so that cyclic and transitive prompts contribute the same number of offline pairs. The builder stores $\delta = 1$ for each synthetic comparison; the trainers consume the resulting chosen/rejected orientation. The six orientations remain fixed throughout training.

Pair sampling and weighting use the same $q_t(p \mid x)$ as in the transitive setup, now over the six cyclic-tournament edges. In particular, $\tau = 1$, $\varepsilon = 0.05$, two pairs are drawn with replacement per selected prompt, normalized weights are clipped to $[0.1, 10]$, IPO uses 500 prompts (1,000 pair draws) per outer iteration, and DPO uses 250 prompts (500 pair draws).

The common completed configurations plotted for both objectives are

$$(\alpha, \lambda, \beta) \in \{(0.1, 0.1, 0.01),\ (0.5, 0.5, 5),\ (0.99, 0.8, 1)\}.$$

They use seed 0 and are compared over $t = 0, \ldots, 99$. The IPO-only run at $(0.99, 0.8, 10)$ is excluded from the matched comparison. The shared $(0, 0, 0.01)$ baseline is also excluded because the available runs end at $t = 49$. For evaluation, 500 prompts are selected with seed 123 from the cyclic data file, and the same four ordered dataset responses are held fixed as $\mathcal{Y}(x)$ for the entire run.

H.4 Additional entropy results

Transitive terminal entropy. Table 1 reports the iteration-99 values corresponding to Figure 2. For IPO, terminal entropy declines steadily as α increases and approaches zero at the most self-referential settings. DPO enters the low-entropy regime at substantially smaller α, although its terminal values are not strictly ordered at the most aggressive settings.

Across nearly the entire grid, IPO retains more fixed-support diversity than DPO at the same (α, λ, β). Together with the trajectories, the endpoint table shows that the conservative IPO target delays feedback-driven concentration, whereas DPO reaches the low-entropy regime under weaker self-reference.

Cyclic terminal entropy. Table 2 gives the matched iteration-99 values for the three completed cyclic configurations common to IPO and DPO. The middle configuration remains diffuse for both objectives, while the other two configurations are concentrated by the endpoint.

	MRS-IPO, $t = 99$		MRS-DPO, $t = 99$	
α	$\lambda = 0.5$	$\lambda = 0.9$	$\lambda = 0.5$	$\lambda = 0.9$
0.50	1.187	1.174	0.484	0.320
0.60	1.145	1.134	0.377	0.265
0.70	1.075	1.067	0.285	0.196
0.80	0.946	0.922	0.188	0.174
0.90	0.621	0.605	0.123	0.098
0.95	0.357	0.329	0.084	0.093
0.97	0.217	0.216	0.119	0.134
0.99	0.106	0.091	0.189	0.210

Table 1: Mean fixed-support entropy at iteration 99 for the complete transitive Qwen2.5-1.5B sweep at $\beta = 1$. Each entry is one seed-0 run averaged over the fixed 500-prompt, four-response evaluation panel.

(α, λ, β)	MRS-IPO, $t = 99$	MRS-DPO, $t = 99$
$(0.1, 0.1, 0.01)$	0.150	0.219
$(0.5, 0.5, 5)$	1.282	1.234
$(0.99, 0.8, 1)$	0.096	0.177

Table 2: Mean fixed-support entropy at iteration 99 for the completed cyclic configurations shared by IPO and DPO. Each entry is one seed-0 run averaged over the fixed 500-prompt, four-response evaluation panel.