Training with (Swap) Regret Loss in a Single-Layer Self-Attention Model: A Case Study on the Probability Simplex

Chanwoo Park
MIT EECS

Asuman Ozdaglar
MIT EECS

Abstract

We revisit the regret loss framework introduced in Park et al. [2025b], which uses decision-theoretic regret as a direct loss function for training models to make better decisions, through the lens of probability-simplex policies. Our first result shows that a single-layer self-attention model trained with regret loss admits a stationary point whose forward-pass exactly matches *smoothed fictitious play* with the appropriate stepsize that ensures no-regret behavior—i.e., for any given policy input, the model outputs the same update that smoothed fictitious play would produce. In parallel, we also newly introduce a swap-regret loss function, which extends the regret-loss framework beyond external regret and enables models to directly optimize for swap-deviation robustness. We further show that this swap-regret loss admits a stationary point whose forward pass implements the corresponding swap-regret update induced by classical Blum–Mansour no-swap-regret algorithm, with each head implementing an external-regret update via smoothed fictitious play. Together, these results show that regret-trained attention can realize differentiable mechanisms whose deployment induces equilibrium behavior in games: external-regret dynamics lead to coarse correlated equilibrium, while swap-regret dynamics lead to correlated equilibrium. Thus, regret-based objectives steer minimal attention architectures toward online-learning dynamics with game-theoretic guarantees, without supervised traces of those algorithms.

1 Introduction

As large language models (LLMs) become increasingly integrated into everyday life, their usage is evolving from single-turn, task-specific applications to multi-agent, sequential decision-making scenarios [Yao et al., 2023, Hao et al., 2023, Shinn et al., 2024, Wang et al., 2023b, Significant Gravitas, 2023, Ahn et al., 2022, Wang et al., 2023a, Li et al., 2024]. Rather than merely seeking one-off answers—such as solving a math problem [Guo et al., 2025] —users now engage with LLMs in ways that influence long-term planning and cumulative outcomes. As more individuals turn to LLMs to augment their decision-making, the setting naturally sometimes becomes multi-agent: each user constitutes an independent agent with unique goals, interacting with a shared or personalized model in pursuit of improved long-term outcomes and multi-agent sequential decision-making is *regret*, which quantifies the difference between the cumulative reward achieved by an agent and that of the best possible strategy in hindsight [Shalev-Shwartz, 2012]. While minimizing regret is not the only objective in these settings, it provides a rigorous measure of an agent's adaptability and learning efficiency over time.

Recent work shows that *off-the-shelf* LLMs—those used without fine-tuning or explicit inference-time control—still struggle on even simple, canonical online decision-making tasks. Empirical evaluations consistently find that such models fail to explore effectively without external scaffolding [Krishnamurthy et al., 2024], incur *linear* regret in non-stochastic environments [Park et al., 2025b], exhibit unstable exploitation behavior [Xia et al., 2024], and underperform in non-stationary settings [Zhang et al., 2025]. Taken together, these results indicate that fundamental decision-making

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

abilities—regret minimization, exploration, and adaptation—remain unsolved for current LLMs. This highlights the need for principled methods that explicitly strengthen these foundations before deploying LLMs as autonomous agents at scale.

To address this gap, Park et al. [2025b] proposed a novel training loss function named *regret loss*, which explicitly optimizes models to minimize regret across sequences of interactions. While their preliminary results—focused primarily on online learning tasks—demonstrated both empirical gains and theoretical guarantees, the broader applicability of regret-based training remains underexplored. Follow-up paper [Park et al., 2025a] found that post-training with regret-based signals can enable successful adaptation in single-agent environments, suggesting that regret supervision can play a critical role in improving sequential learning performance. Moreover, given that LLMs are fundamentally built on self-attention architectures, it is essential to understand how such regret-minimization dynamics manifest in even the simplest form of these models.

However, a detailed characterization of regret-based training for single-layer self-attention is still incomplete even in the canonical probability-simplex setting. Park et al. [2025b] analyzed continuous ℓ_2-constrained policies and proved that regret loss recovers Follow-the-Regularized-Leader updates, yet it remains unclear how the same loss behaves when the policy must lie on the simplex—precisely the regime relevant to external and swap regret in online learning and games. Table 1 summarizes this gap.

In this paper, we revisit the *regret loss* framework with a singular focus on probability-simplex policies. Our first contribution shows that, for the standard loss sequences used to evaluate external regret, a stationary point of a single-layer self-attention model trained with regret loss coincides with *smoothed fictitious play* with a $\widetilde{\Theta}(1/\sqrt{T})$ stepsize. Our second contribution targets swap regret. We introduce a new *swap-regret loss* that operationalizes swap regret as a differentiable training objective. Using this loss, we construct a multi-head linear self-attention network whose stationary point mirrors the classical Blum–Mansour no-swap-regret algorithm [Blum and Mansour, 2007]. Each head implements a smoothed fictitious play external-regret update with a $\widetilde{\Theta}(1/\sqrt{T})$ stepsize, so the entire block behaves like a collection of parallel external-regret minimizers stitched together through multi-head self-attention. We prove that optimizing the swap-regret loss recovers this structure end-to-end, showing that differentiable multi-head attention can faithfully instantiate swap-regret dynamics. Together, these results highlight how external- and swap-regret objectives can be realized directly within minimal Transformer architectures.

This provides a mechanistic route from regret-based post-training objectives for LLM agents to equilibrium guarantees in multi-agent environments. Regret-trained attention is not merely imitating an online-learning algorithm; at the stationary points we identify, it implements a differentiable mechanism whose repeated deployment induces the corresponding game-theoretic behavior. Single-head softmax attention implements external-regret dynamics, so multi-agent deployment leads to coarse correlated equilibrium. Multi-head fixed-point attention implements swap-regret dynamics, so deployment leads to correlated equilibrium. The choice of regret surrogate is therefore not only an optimization detail: external-regret loss induces coarse-deviation robustness, while swap-regret loss induces action-contingent deviation robustness.

Notation We write $\mathbf{1}_d$ and $\mathbf{0}_d$ for the d-dimensional all-ones and all-zeros vectors, $\boldsymbol{O}_{d\times d}$ for the $d \times d$ zero matrix, and $I_{d\times d}$ for the $d \times d$ identity, let $\Delta(\mathcal{A})$ be the probability simplex over the action set $\mathcal{A}$, use $[d] := \{1, \dots, d\}$ for the first d integers, and write $\mathbb{1}\{\cdot\}$ for indicator functions. For asymptotics as $T \to \infty$ we use $\Theta(f(T))$ to denote quantities bounded above and below by constant multiples of $f(T)$, $O(f(T))$ for upper bounds, $o(f(T))$ for terms that vanish relative to $f(T)$, and $\widetilde{O}(f(T))$ (or $\widetilde{\Theta}(f(T))$) when bounds hold up to polylogarithmic factors. For a set Π and norm $\|\cdot\|$, we denote by $\mathrm{Proj}_{\Pi,\|\cdot\|}(x) := \mathrm{argmin}_{y\in\Pi} \|x - y\|$ the projection of x onto Π; when the norm is Euclidean we write $\mathrm{Proj}_{\Pi}(x)$ for brevity. For any vector v, $P_{\perp v}$ denotes the orthogonal projector onto the subspace orthogonal to v. For a Markov matrix P, we write $\mathrm{Fixed\text{-}Point}(P)$ for any stationary distribution satisfying $\pi = P\pi$; when the stationary distribution is unique we identify $\mathrm{Fixed\text{-}Point}(P)$ with that π.

2 Preliminaries

2.1 Linear Self-Attention

We consider a single-layer linear attention model equipped with an output operator that maps the resulting representation into a decision policy. This architecture represents one of the simplest forms of Transformer-based models and has been used in prior work to analyze the inductive properties of

attention mechanisms [Ahn et al., 2023, Mahankali et al., 2023, Park et al., 2025b]. The output of the model is given by:

$$g((\ell_1, \ldots, \ell_t, \mathbf{1}_d); V, K, Q, v_c, k_c, q_c) = \texttt{Operator}\left(\sum_{i=1}^{t} (V\ell_i + v_c)\left((K\ell_i + k_c)^\intercal (Q\mathbf{1}_d + q_c)\right)\right), \quad (1)$$

where $V, K, Q \in \mathbb{R}^{d\times d}$ are the value, key, and query matrices, and $v_c, k_c, q_c \in \mathbb{R}^d$ are their corresponding bias vectors. We denote the full set of model parameters as $\phi = (V, K, Q, v_c, k_c, q_c)$.

The model takes as input a sequence of loss vectors $(\ell_1, \ldots, \ell_t) \subset \mathbb{R}^d$, along with a placeholder vector $\mathbf{1}_d \in \mathbb{R}^d$, and produces a prediction for the next policy π_{t+1}. The inner product term computes attention scores over the sequence $(\ell_1, \ldots, \ell_t)$, and the resulting aggregated vector is transformed by a final mapping, denoted as `Operator`, to produce a valid policy. The specific choice of `Operator` depends on the policy space Π. For example, if $\Pi = \Delta(\mathcal{A})$, the probability simplex induced by a finite action set (as in the Experts Problem), `Operator` is the standard $\mathrm{Softmax}$ function. In the most basic case, `Operator` can be the identity map, $\texttt{Operator}(x) = x$, recovering the original form of linear self-attention [Ahn et al., 2023, Zhang et al., 2023, Schlag et al., 2021].
This formulation provides a minimal yet expressive setting for studying how attention-based models can learn adaptive policies in sequential decision-making environments. In this paper, as Equation (1) has the equivalent formulation with

$$g((\ell_1, \ldots, \ell_t, \mathbf{1}_d); V, a, v_c) = \texttt{Operator}\left(\sum_{i=1}^{t} (V\ell_i\ell_i^\intercal a + (V + v_c a^\intercal)\ell_i + v_c)\right), \quad (2)$$

where $V \in \mathbb{R}^{d\times d}$ and $v_c, a \in \mathbb{R}^d$. Equation (2) is a reduced form of Equation (1) but having the same representation power with Equation (1).

2.2 Learning Environment

2.2.1 Online Learning

We consider the *online learning* framework in which an agent interacts with an environment over T rounds, making sequential decisions and receiving feedback. At each round $t \in [T]$, the agent selects a decision policy $\pi_t \in \Pi$, where Π is a bounded decision space. After committing to π_t, the environment reveals a loss function $f_t : \Pi \to [-B, B]$ for some constant $B > 0$, which may be chosen adversarially. The agent then incurs loss $f_t(\pi_t)$ and updates her policy based on this feedback. A fundamental special case is when the action space $\mathcal{A}$ is finite and the decision set is the probability simplex over $\mathcal{A}$, i.e., $\Pi = \Delta(\mathcal{A})$. This is known as the *Experts Problem* [Cover, 1966, Vovk, 1990, Littlestone and Warmuth, 1994, Hazan, 2016]. In this case, each loss function can be expressed as a linear form $f_t(\pi_t) = \langle \ell_t, \pi_t \rangle$, where $\ell_t \in \mathbb{R}^d$ is a loss vector (with $d := |\mathcal{A}|$). Whenever we refer to the policy space as the simplex below, we implicitly mean this finite-action setting. The agent will receive the entire loss vector ℓ_t in the full-information setting.

2.2.2 Repeated Static Games

The online learning framework extends naturally to multi-agent interactions through the lens of *repeated static games*. Let $\mathcal{G} = \langle N, \{\mathcal{A}_n\}_{n\in[N]}, \{r_n\}_{n\in[N]} \rangle$ denote a normal-form game with players $n \in [N]$, where each $\mathcal{A}_n$ is the action set of player n, and $r_n : \mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_N \to [-B, B]$ is her reward (or negative loss) function. At each round t, every player outputs a *policy* $\pi_{n,t} \in \Delta(\mathcal{A}_n)$. We use the term *action* only for the realized play induced by the policy: in finite (simplex) domains, an action $a_{n,t}$ is sampled as $a_{n,t} \sim \pi_{n,t}$. From the perspective of player n, the realized joint play induces an individual online learning problem over $\Delta(\mathcal{A}_n)$ with per-round loss

$$\ell_{n,t} := -\mathbb{E}_{a_{-n,t}\sim\pi_{-n,t}}\left[r_n(\cdot, a_{-n,t})\right] := -r_n(\cdot, \pi_{-n,t}).$$

2.3 Performance Metrics: External and Swap Regret

2.3.1 External Regret

We begin by defining *external regret* (usually just denoted as regret), a central performance measure in online learning and repeated games. Let $\mathscr{A}$ denote an algorithm that selects a sequence of policies $(\pi_{\mathscr{A},t})_{t\in[T]}$ over a decision space Π. Given a sequence of loss functions $(f_t)_{t\in[T]}$, the *external regret* of $\mathscr{A}$ is defined as

$$\mathrm{Regret}_{\mathscr{A}}((f_t)_{t\in[T]}) := \sum_{t=1}^{T} f_t(\pi_{\mathscr{A},t}) - \inf_{\pi\in\Pi} \sum_{t=1}^{T} f_t(\pi). \quad (3)$$

This quantity compares the cumulative loss of the learner to that of the best fixed decision in hindsight. In the Experts Problem, this specializes to

$$\text{Regret}_{\mathscr{A}}((\ell_t)_{t\in[T]}) := \sum_{t=1}^{T}\langle \ell_t, \pi_{\mathscr{A},t}\rangle - \inf_{\pi\in\Pi}\sum_{t=1}^{T}\langle \ell_t, \pi\rangle, \tag{4}$$

where each $\ell_t \in \mathbb{R}^d$ is a loss vector and $\Pi \overset{=}{=} \Delta(\mathcal{A})$ is the simplex over the finite action set $\mathcal{A}$. An algorithm $\mathscr{A}$ is said to be *no-(external-)regret* if $\sup_{(f_t)} \text{Regret}_{\mathscr{A}}((f_t)_{t\in[T]}) = o(T)$, i.e., the worst-case regret grows sublinearly in T.

2.3.2 Swap Regret

While external regret compares the learner's performance to the best fixed decision in hindsight, swap regret allows each action to be reassigned to another action in hindsight. Formally, a swap deviation is encoded by a mapping $\sigma : [d] \to [d]$ (or equivalently by a row-stochastic matrix P_σ). The *swap regret* of algorithm $\mathscr{A}$ in the Experts setting is

$$\text{Swap-Regret}_{\mathscr{A}}((\ell_t)_{t\in[T]}) := \max_{\sigma:[d]\to[d]}\sum_{t=1}^{T}\big\langle \pi_{\mathscr{A},t},\, \ell_t - P_\sigma^\intercal \ell_t\big\rangle = \max_{P\in\mathcal{M}}\sum_{t=1}^{T}\big\langle \pi_{\mathscr{A},t},\, \ell_t - P^\intercal \ell_t\big\rangle, \tag{5}$$

where $P_\sigma \in \{0,1\}^{d\times d}$ is defined by $(P_\sigma)_{ij} = \mathbb{1}\{\sigma(i) = j\}$ and $\mathcal{M} := \{P \in \mathbb{R}^{d\times d} : P\mathbf{1}_d = \mathbf{1}_d,\, P \geq 0\}$ denotes the set of row-stochastic (Markov) matrices. The equality follows since the objective is linear in each row of P, and thus its maximum over $\mathcal{M}$ is attained at an extreme point P_σ. Intuitively, swap regret measures how much the learner could have improved by systematically substituting each played action i with another action $\sigma(i)$ across all rounds. An algorithm is said to be *no-swap-regret* if the above quantity is $o(T)$. In this paper, if $\mathscr{A}$ is clear from the context, we will drop the subscript and write Regret and Swap-Regret for external and swap regret, respectively.

2.4 Equilibrium Concepts in Repeated Games

Repeated interactions among learning agents often lead to stable long-run behaviors that correspond to equilibrium concepts. Writing everything in terms of policies allows us to reason directly about simplex-valued play.

Coarse Correlated Equilibrium (CCE). A distribution $\mu \in \Delta(\mathcal{A})$ is a *CCE* if, for every player n and every fixed (policy-level) deviation $\pi_n' \in \Delta(\mathcal{A}_n)$,

$$\mathbb{E}_{\pi\sim\mu}[r_n(\pi)] \;\geq\; \mathbb{E}_{\pi\sim\mu}[r_n(\pi_n',\, \pi_{-n})]\,.$$

Thus, no player can gain by unconditionally switching to a single alternative policy prior to receiving any recommendation.

Correlated Equilibrium (CE). A distribution $\mu \in \Delta(\mathcal{A})$ over joint actions is a *correlated equilibrium* if, for every player n and every action-dependent deviation represented by a row-stochastic Markov matrix $M_n \in \mathcal{M}$,

$$\mathbb{E}_{\pi\sim\mu}[r_n(\pi)] \;\geq\; \mathbb{E}_{\pi\sim\mu}\big[r_n\big(M_n(\pi_n),\, \pi_{-n}\big)\big]\,,$$

where $M_n(\pi_n)$ denotes the (possibly mixed) action drawn from the row of M_n corresponding to the recommended action π_n. Thus, no player can gain by conditionally deviating from the recommendation using any action-dependent strategy.
If every player achieves no external regret, the empirical average play converges to a CCE, and if every player achieves no swap regret, the empirical average play converges to a CE [Cesa-Bianchi and Lugosi, 2006].

2.5 Concepts of Regret Loss Training

We now define a training objective that explicitly minimizes regret across sampled decision-making environments. Let the model be parameterized by ϕ, and let model_ϕ denote a policy generator (e.g., a single-layer attention model as in Equation (1)). The *regret loss function* is defined as

$$\mathcal{L}(\phi, \mathcal{D}) := \mathbb{E}_{\mathcal{D}}\left[h\left(\text{Regret}_{\text{model}_\phi}\left((f_t)_{t\in[T]}\right)\right)\right], \tag{6}$$

where $h : \mathbb{R} \to \mathbb{R}$ is a convex surrogate function (e.g., identity, squared loss, or a smooth approximation of $\max$), and the expectation is taken over decision-making instances sampled from $\mathcal{D}$. Throughout the paper we instantiate Regret as either external regret or swap regret. We will describe the precise form of the model input shortly. The distribution $\mathcal{D}$ captures the randomness in the environment. Each round's loss function is $f_t(\pi) = \langle \ell_t, \pi\rangle$, where $\ell_t \sim \mathcal{D}_d$ is a sampled loss vector. For each decision round t, the model receives the interaction history and produces the next policy π_{t+1}. Especially, the model input is the full history of loss vectors, i.e., $(\ell_1, \ldots, \ell_t)$.

Previous theoretical work on regret-loss training [Park et al., 2025b,a] primarily analyzed continuous ℓ_2-constrained policies. In this paper we restrict attention to the probability simplex—the setting relevant to external and swap regret—and study how regret loss shapes the induced dynamics. Table 1 summarizes the most relevant comparisons.

	No Regret (test)	`Operator`	**Post Operation**	**FTRL Equivalence**	**FTRL stepsize**
Park et al. [2025b] Theorem 5.2	Online Learning ($\Pi = B(\mathbf{0}_d, R, \|\cdot\|_2)$)	Identity	$\mathrm{Proj}_{B(\mathbf{0}_d,R,\|\cdot\|_2)}$	✓	$\Theta\left(\frac{1}{\sqrt{Td}}\right)$
Park et al. [2025b] Conjecture 3	Online Learning ($\Pi = \Delta(\mathcal{A})$) Static Game	Softmax	×	?	?
Theorem 1	Online Learning ($\Pi = \Delta(\mathcal{A})$, no external regret) Static Game	Softmax	×	✓ (Stationary Point)	$\Theta\left(\frac{1}{\sqrt{T}}\right)$
Theorem 2	Online Learning ($\Pi = \Delta(\mathcal{A})$, no swap regret) Static Game	Row-wise Softmax, Fixed-Point	×	✓ (Stationary Point) Each head implements an external-regret FTRL update	$\Theta\left(\frac{1}{\sqrt{T}}\right)$

Table 1: The top rows report known results from Park et al. [2025b]; the bottom rows highlight our new stationary-point characterizations showing that regret-loss objectives on attention models recover smoothed fictitious play (external regret) and Blum–Mansour dynamics (swap regret).

3 Emergence of Smoothed Fictitious Play as a Stationary Point of Single-layer Linear Self-Attention

We begin by revisiting the observation of Park et al. [2025b] that single-layer linear self-attention trained with the regret-loss objective recovers FTRL updates with L_2 regularization when the operator is the identity. Their result, however, depends crucially on applying a projection $\mathrm{Proj}_{\Pi,\|\cdot\|}$ to the model output only at inference time: the training objective never sees the projection, so the match to FTRL is achieved post hoc. This leaves open whether the post operation can be integrated directly into the optimization so that the learned representation is already the desired policy without any additional post operation steps. Throughout this section we adopt an end-to-end viewpoint: the same nonlinearity used at inference—namely the Softmax mapping—is embedded directly in the loss. Consequently, the model must optimize through the Softmax transformation itself, jointly learning the attention weights while satisfying the policy-domain constraints. We show that even under this stricter objective, the stationary points remain fully interpretable: they coincide with smoothed fictitious play on the simplex. Moreover, we characterize the associated theoretical stepsize of these stationary points, which in turn yields a no-regret guarantee. This is also conjectured by Park et al. [2025b] that the stationary point of the regret-loss function is the smoothed fictitious play.

Theorem 1. *The configuration of a single-layer linear attention model as defined in Equation (2), with* `Operator` = Softmax*, and parameters $V = -kI_{d\times d}$, $a = \mathbf{0}_d$, and $v_c = v\mathbf{1}_d$, is a stationary point of the regret-loss function Equation (6) if $k = \widetilde{\Theta}(1/\sqrt{T})$, with $h(x) = x^2$, $f_t(\pi) = \langle \ell_t, \pi \rangle$, where $\ell_t \sim \mathcal{N}(\mathbf{0}_d, I_{d\times d})$ i.i.d.*

Proof structure for Theorem 1. At $V = -kI_{d\times d}$, $a = \mathbf{0}_d$, and $v_c = v\mathbf{1}_d$, the round-t logits and policy satisfy

$$z_t = -k\sum_{s<t}\ell_s + (t-1)v\mathbf{1}_d, \qquad \pi_t = \mathrm{Softmax}\left(-k\sum_{s<t}\ell_s\right),$$

where the second identity uses shift-invariance. Thus the forward pass is entropy-regularized FTRL, equivalently smoothed fictitious play, with learning rate k; the substantive issue is to prove that the full regret-loss gradient vanishes at a nonzero k of the claimed scale.

Let $R_T(k)$ be the external regret induced by the above policy and $L(k) = \mathbb{E}[R_T(k)^2]$. The first-order conditions are then verified block by block: the v_c-block is annihilated pointwise by shift-invariance and $J_{\mathrm{sm}}(z)^\intercal\mathbf{1}_d = \mathbf{0}_d$; the a-block requires Gaussian permutation symmetry, which makes the expected gradient a multiple of $\mathbf{1}_d$, while the softmax Jacobian maps into the zero-sum subspace. For the V-block, permutation equivariance and the same row-sum constraint force

$$\nabla_V\mathcal{L}(-kI, \mathbf{0}, v\mathbf{1}) = \alpha(k)\left(I_{d\times d} - \frac{1}{d}\mathbf{1}_d\mathbf{1}_d^\intercal\right).$$

Consequently the matrix condition reduces to the scalar derivative along the ray $V = -kI$.

It remains to locate a zero of $L'(k)$. The appendix derives

$$L'(k) = -2\Sigma_1(k) + 2\Sigma_2(k),$$

where Σ_1 is the learner-response term and Σ_2 is the best-fixed-action comparator term. Expanding $\pi_t(k) = u - kJ_0S_{t-1} + O(k^2\|S_{t-1}\|_\infty^2)$ and $J_t(k) = J_0 + O(k\|S_{t-1}\|_\infty)$, with $u = \mathbf{1}_d/d$ and $J_0 = \mathrm{Diag}(u) - uu^\intercal$, separates the signs of these terms: the comparator contribution dominates for very small k, giving $L'(k) < 0$, while at the inverse-$\sqrt{T}$ scale the learner-response term dominates after Gaussian maximum, MGF, and softmax-Jacobian remainder bounds, giving $L'(k) > 0$. Continuity yields a root $k^\star = \widetilde{\Theta}(1/\sqrt{T})$, at which every parameter block is first-order stationary.

The proof, deferred to Appendix A, pins down the precise scale of the attention weight k. Earlier analyses of in-context learning only established that FTRL-like outputs could be recovered at stationary points if they have a Softmax layer [Cheng et al., 2023] or convergence of the training dynamics [Chen et al., 2024]; in contrast, Theorem 1 identifies the first-order stationary point and shows that the resulting update is precisely smoothed fictitious play with learning rate $\widetilde{\Theta}(1/\sqrt{T})$. Empirically, we observe a unique $k^\star(d, T)$ consistent with this scaling, further supporting the theoretical prediction (Figure 1).

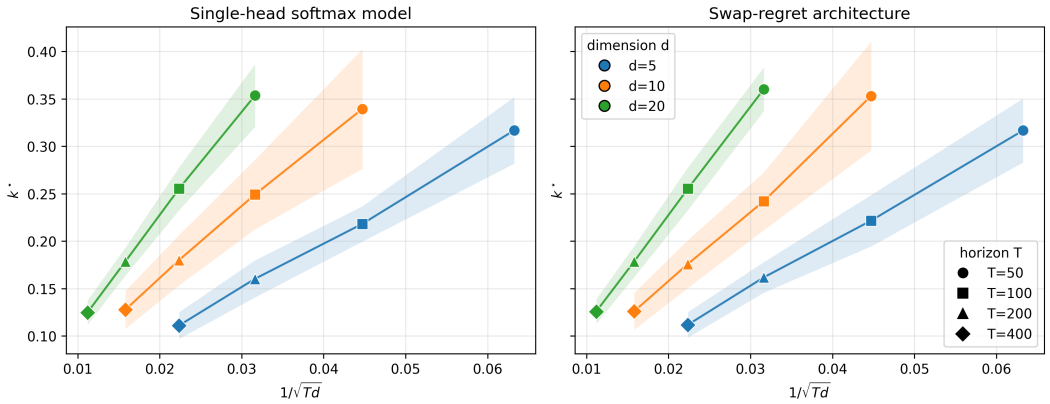


Figure 1: Estimated stationary stepsize $k^\star$ on the probability simplex as a function of (d, T) for the single-head softmax model (left) and the swap-regret architecture (right). Each marker aggregates 50 independent runs with 800 Monte Carlo samples per run; shaded bands indicate ± 1 standard deviation.

The above characterizations lift immediately to multi-agent learning dynamics.

Corollary 1. *Consider a repeated game where each player deploys the trained single-layer architecture from Theorem 1 with* `Operator` = Softmax. *Then the time-averaged joint play converges to a coarse correlated equilibrium (CCE) in finite action games.*

Thus, enforcing the simplex-respecting nonlinearity inside the loss not only yields interpretable single-agent updates but also preserves standard equilibrium guarantees when the model is embedded in multi-agent environments. Mechanistically, the learned attention layer implements an external-regret dynamic, and external regret is precisely the deviation notion underlying CCE: no player can improve by replacing their realized sequence of actions with a single fixed action in hindsight. This connects the stationary-point characterization of regret-trained attention directly to coarse-deviation robustness in games.

3.1 Local minimality within the smoothed-fictitious-play family

Proposition 1 (Restricted local minimality)**.** *Let*

$$L(k) := \mathbb{E}[R_T(k)^2], \qquad \pi_t(k) = \mathrm{Softmax}\left(-k \sum_{s<t} \ell_s\right).$$

Under the setting of Theorem 1, there exists $k^\star = \widetilde{\Theta}(1/\sqrt{T})$ that is a local minimizer of $L(k)$. Consequently,

$$V = -k^\star I_{d\times d}, \qquad a = \mathbf{0}_d, \qquad v_c = v\mathbf{1}_d$$

is a local minimizer of the population regret loss when optimization is restricted to this Softmax–smoothed-fictitious-play scalar family.

Proof. The sign estimates in the proof of Theorem 1 provide $k_- < k_+$, both at the stated scale up to the displayed dimension and polylogarithmic factors, such that $L'(k_-) < 0$ and $L'(k_+) > 0$. The continuous function L attains a minimum on $[k_-, k_+]$. The strict derivative signs exclude both endpoints, so at least one minimizer $k^\star$ lies in the interior. It is therefore a local minimizer of L, and $L'(k^\star) = 0$. The full-gradient symmetry reduction in the proof of Theorem 1 then also gives first-order stationarity in the unrestricted parameterization. □

This proposition is deliberately restricted to the scalar FTRL family. It does not establish local minimality over the full parameter space or global convergence of gradient descent. We separately test whether training from random initialization reaches the identified behavior.

3.2 Empirical FTRL matching

We test whether unrestricted random-initialization training reaches the entropy-FTRL behavior identified by Theorem 1. We train the reduced one-layer model with squared regret loss at $d = 3, T = 20$ for three independent seeds, without FTRL trajectories, FTRL parameters, or a constraint $V = -kI$. For each learned checkpoint, we fit a scalar $\eta \in [0, 5]$ on 4,096 held-out Gaussian sequences by minimizing the mean policy MSE between the model output and

$$\pi_t^{\mathrm{FTRL}}(\eta) = \mathrm{Softmax}\left(-\eta \sum_{s<t} \ell_s\right).$$

Figure 2 (left) shows that the learned policies are close to entropy-FTRL: on average, only 2.71% of probability mass must be moved to match the fitted FTRL policy. This is finite-horizon evidence that the theoretically identified family is empirically reachable; it is not a claim of global convergence from arbitrary initialization.

3.3 Regret training in deeper Softmax Transformers

We next train two- and four-layer causal Softmax Transformers under the same Gaussian regret objective. These models use four attention heads per layer; we do not fit their outputs to FTRL or claim an FTRL identification theorem. We evaluate on held-out biased-Gaussian sequences $\ell_t = m + \epsilon_t$, where the sequence-specific mean m creates a best action that must be learned from history. Writing R_t for cumulative regret through time t, both depths maintain substantially smaller R_t/t than the non-adaptive uniform policy. At $T = 20$, the two- and four-layer models attain respectively $R_{20} = 4.380 \pm 0.046$ and 4.322 ± 0.017, or normalized regret 0.219 and 0.216, compared with $R_{20} = 7.021 \pm 0.052$ and $R_{20}/20 = 0.351$ for uniform play (Figure 2, right).

3.4 Coarse-correlated-equilibrium evaluation

We evaluate the learned policies in a 3×3 weighted zero-sum rock–paper–scissors game whose equilibrium marginal is $(3, 2, 1)/6$. With expected full-information loss feedback, we form the time-averaged joint mixed play $\bar{\mu}_T = T^{-1} \sum_{t=1}^{T} \pi_{1,t} \otimes \pi_{2,t}$ and measure its maximum CCE-deviation gap. At $T = 20$, the one-, two-, and four-layer models obtain respectively 0.0832 ± 0.0028, 0.0467 ± 0.0095, and 0.0653 ± 0.0054, compared with 0.1111 for uniform play. Figure 4 (left) reports the complete curves. This directly measures the equilibrium quantity controlled by average external regret.

4 Swap-Regret Loss Minimization on the Probability Simplex

We now recall the classical reduction of Blum and Mansour, which turns any external-regret minimizer on the simplex into a swap-regret minimizer. Let the action set be $\mathcal{A} = \{1, \ldots, d\}$ and write $\Delta(\mathcal{A})$ for

(a) One-layer FTRL matching

Quantity	Mean ± SD
Fitted $\widehat{\eta}$	$0.4456 \pm .0046$
Policy MSE	$0.000739 \pm .000176$
Mean L_1	$0.0542 \pm .0071$
TV (% mass)	$0.0271 \pm .0036$ (2.71%)
Error/action	$0.0181 \pm .0024$ (1.81 pp)

(b) Regret over time

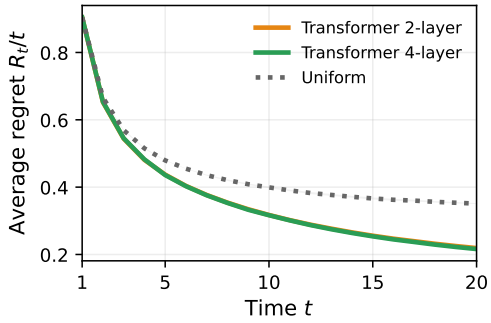


Figure 2: Random-initialization regret-loss training at $d = 3, T = 20$. Left: entropy-FTRL fit to one-layer policies; "pp" denotes percentage points. Right: normalized regret for deeper Transformers on biased-Gaussian sequences. The table reports mean ± sample standard deviation and curves show means with ± 1 standard-error bands across three seeds.

the corresponding probability simplex. A *swap transformation* is represented by a column-stochastic matrix

$$\Phi \;:=\; \big\{P = (p_1 \mid \cdots \mid p_d) \in \mathbb{R}^{d\times d} : p_i \in \Delta(\mathcal{A}) \text{ for all } i\big\},$$

which acts on $\pi \in \Delta(\mathcal{A})$ by $\pi \mapsto P\pi = \sum_{i=1}^{d} \pi_i p_i$. At round t, the adversary reveals a loss vector $\ell_t \in \mathbb{R}^d$ so that the learner incurs the linear loss $\langle \ell_t, \pi_t \rangle$.

Blum–Mansour construction. Fix d copies of an external-regret minimizer for $\Delta(\mathcal{A})$, denoted $\mathcal{R}_1, \ldots, \mathcal{R}_d$. At each round $t \in [T]$ the swap-regret algorithm performs:

1. **Next strategy.** For each $i \in [d]$, query $\mathcal{R}_i$ for $p_{i,t} \in \Delta(\mathcal{A})$. Assemble the matrix

   $$P_t \;=\; \big(p_{1,t} \mid \cdots \mid p_{d,t}\big) \in \Phi.$$

   Compute a fixed point $\pi_t = \text{Fixed-Point}(P_t) \in \Delta(\mathcal{A})$ and play π_t.
2. **Loss feedback.** After observing the loss vector ℓ_t, each external-regret copy $\mathcal{R}_i$ receives the scaled linear loss $x_{i,t} \;=\; \pi_{t,i}\, \ell_t$, i.e., it runs on the loss vector $x_{i,t}$.

Thus each $\mathcal{R}_i$ runs on the same action space $\Delta(\mathcal{A})$, but with losses rescaled by the i-th coordinate of the fixed point π_t.

Fact 1 (Blum–Mansour)**.** *Let $\textit{Regret}_i$ denote the external regret of $\mathcal{R}_i$ after T rounds when fed the losses $(x_{i,t})_{t\in[T]}$, that is*

$$\textit{Regret}_i((x_{i,t})_{t\in[T]}) \;:=\; \max_{\widetilde{p}_i \in \Delta(\mathcal{A})} \sum_{t=1}^{T} \big(\langle x_{i,t}, p_{i,t}\rangle - \langle x_{i,t}, \widetilde{p}_i\rangle\big). \tag{7}$$

Then the cumulative swap regret admits the exact column-wise decomposition

$$\textit{Swap-Regret}((\ell_t)_{t\in[T]}) = -\min_{\widehat{P}=(\widehat{p}_1|\cdots|\widehat{p}_d)\in\Phi} \sum_{t=1}^{T}\sum_{i=1}^{d} \big\langle x_{i,t}, \widehat{p}_i - p_{i,t}\big\rangle, \tag{8}$$

and therefore satisfies

$$\textit{Swap-Regret}((\ell_t)_{t\in[T]}) \;\le\; \sum_{i=1}^{d} \textit{Regret}_i((x_{i,t})_{t\in[T]}).$$

In particular, if each $\mathcal{R}_i$ has sublinear external regret, then the Blum–Mansour construction has sublinear swap regret.

The proof is deferred to Appendix C.

Inspired by the classical Blum–Mansour construction, we introduce a new neural architecture for swap-regret minimization. Our design uses a d-head linear self-attention module to parameterize the transition matrix. After the multi-head attention step, each head applies a $\mathrm{Softmax}$ to produce one strictly positive column of the transition matrix, thereby obtaining a valid column-stochastic Markov matrix. The final policy is then computed via a $\mathrm{Fixed\text{-}Point}$ procedure, which can be efficiently implemented using power iteration. This yields a differentiable, end-to-end model that mirrors the theoretical structure of the Blum–Mansour minimizer while allowing scalable learning. Mathematically, the architecture is given by:

$$\begin{aligned} \pi_{t+1} &= \texttt{Fixed-Point}\left((p_{1,t+1} \mid \cdots \mid p_{d,t+1})\right), \\ p_{j,t+1} &= \mathrm{Softmax}\left(\sum_{i=1}^{t} \left(V^{(j)} X_i X_i^\intercal a^{(j)} + (V^{(j)} + v_c^{(j)} a^{(j)\intercal}) X_i + v_c^{(j)}\right)\right), \qquad j \in [d]. \end{aligned} \tag{9}$$

where $V^{(j)} \in \mathbb{R}^{d\times d^2}, a^{(j)} \in \mathbb{R}^{d^2}, v_c^{(j)} \in \mathbb{R}^d$ are the parameters for the j-th head of the linear self-attention module, and $X_i := \ell_i \otimes \pi_i = (x_{i,t})_{t=1}^{\intercal} \in \mathbb{R}^{d^2}$, which consider every possible combination of ℓ_i and π_i. The $\mathrm{Fixed\text{-}Point}$ operation is implemented using power iteration. Further implementation details and an architecture schematic appear in Appendix D.

Now, we provide a theoretical analysis of the new architecture.

Theorem 2. *The configuration of the multi-head linear self-attention architecture defined in Equation (9) is a stationary point of the swap-regret loss*

$$\mathcal{L}((V^{(j)}, a^{(j)}, v_c^{(j)})_{j\in[d]}) = \mathbb{E}_{X\sim\mathcal{N}(0,I_{d^2T})}\left[\left(\min_{\widehat{P}\in(\Delta([d]))^d} \sum_{t=1}^{T}\sum_{i=1}^{d} \langle x_{i,t}, \widehat{P}_i - p_{i,t}\rangle\right)^2\right], \tag{10}$$

where

$$x_{i,t} := X_{d(i-1)+1:di,\ t} \in \mathbb{R}^d,$$

and the parameters satisfy

$$V^{(j)} = (V_1^{(j)}, \ldots, V_d^{(j)}), \qquad V_r^{(j)} = \begin{cases} -kI_{d\times d}, & r = j, \\ \boldsymbol{O}_{d\times d}, & r \neq j, \end{cases} \qquad a^{(j)} = \mathbf{0}_{d^2}, \quad v_c^{(j)} = v\mathbf{1}_d,$$

for some $k = \Theta(1/\sqrt{T})$.

Proof structure for Theorem 2. At the proposed configuration, head j ignores all blocks of X_t except the block $x_{j,t}$, and its logits become

$$z_{j,t} = -k\sum_{s<t} x_{j,s} + (t-1)v\mathbf{1}_d.$$

Thus every head is a single-head smoothed fictitious-play learner, while the swap objective couples the heads through the fixed-point policy and the Blum–Mansour decomposition in Fact 1. The gradient calculation separates this coupling carefully: shift invariance eliminates the $v_c^{(j)}$ and $a^{(j)}$ directions, block independence eliminates off-diagonal $V_r^{(j)}$ terms for $r \neq j$, and permutation equivariance reduces the diagonal block $V_j^{(j)}$ to the projector direction $I - d^{-1}\mathbf{1}_d\mathbf{1}_d^\intercal$. The remaining directional derivative along $V_j^{(j)} = -kI$ is exactly d times the scalar derivative of the single-head external-regret objective, so the same $k^\star$ scale from Theorem 1 yields stationarity.

Here, Equation (10) is the square of the swap-regret decomposition in Equation (8). Therefore, Theorem 2 shows that for the multi-head linear self-attention with $\mathrm{Fixed\text{-}Point}$ operation, minimizing the swap regret loss provides a no-swap-regret guarantee, and the stationary point is equivalent to the Blum–Mansour swap-regret minimizer – input and output are the same as the Blum–Mansour construction. Moreover, for each head, the configuration is equivalent to the external-regret smoothed fictitious play with appropriate stepsize. The proof is deferred to Appendix F.

We also provide an empirical $k^\star(d, T)$ consistent with this scaling, further supporting the theoretical prediction (Figure 1, right).

Finally, we provide a corollary that the stationary point of the new architecture is a CE in finite action games.

Corollary 2. *Consider a repeated game where each player deploys the trained multi-head single-layer linear attention from Theorem 2 with* `Operator` = $\mathrm{Softmax}$. *Then the time-averaged joint play converges to a CE in finite action games.*

The distinction from the single-head result is the strength of the allowed deviations. External-regret dynamics rule out fixed-action deviations and therefore imply CCE, whereas swap-regret dynamics rule out action-contingent deviations and therefore imply CE. The fixed-point layer is the mechanism that converts the parallel external-regret heads into a single policy with swap-regret guarantees, so changing the regret surrogate changes the equilibrium concept approached by interacting agents.

4.1 Correlated-equilibrium evaluation

We train all attention parameters of the d-head architecture jointly from random initialization under the Gaussian block-input objective in Equation (10), then evaluate the learned heads through the recurrent fixed-point layer. Because the training objective uses full-scale head blocks whereas the fixed-point recurrence supplies $x_{i,t} = \pi_{t,i}\ell_t$, we apply the predetermined dimension-only calibration $k_{\mathrm{head}} = d\,k_{\mathrm{train}}$.

For each run, we compute the standard maximum pairwise CE-deviation gap directly from $\bar{\mu}_T$. At $T = 20$, the learned fixed-point architecture obtains 0.0627 ± 0.0018, a 43.6% reduction from uniform play (0.1111) and nearly the value of the classical Blum–Mansour entropy-FTRL algorithm (0.0617). Its stronger full-swap gap is 0.0956 ± 0.0068. Without the dimension calibration, the pairwise and full-swap gaps are 0.0968 ± 0.0049 and 0.1627 ± 0.0110, respectively. Thus the finite-horizon empirical distribution is an approximate, not exact, CE. Complete CCE and CE curves appear in Figure 4.

5 Conclusion

We revisited regret-based training of linear attention with policy space $\Delta(\mathcal{A})$ and showed that even minimal architectures admit interpretable stationary points once the non-linearity used at inference time is included in the training objective. For single-head models on the simplex, the regret loss recovers smoothed fictitious play with stepsize $\widetilde{\Theta}(1/\sqrt{T})$. Extending the loss to swap regret and lifting the architecture to a multi-head configuration leads to a differentiable realization of the Blum–Mansour procedure in which each head implements an entropy-regularized FTRL learner and a $\mathrm{Fixed\text{-}Point}$ layer aggregates the resulting transition matrix.

Our characterization highlights that regret supervision can endow attention models with provably no-regret behavior without auxiliary projections or post-processing. Our analysis identifies interpretable stationary points rather than proving global optimality or convergence of training dynamics. The proof also relies on full-information Gaussian training losses, simplex-valued policies, one-layer linear attention, the $\mathrm{Softmax}$ operator, and horizon-dependent scaling of the learned stepsize. The resulting forward pass, however, coincides with classical no-regret algorithms whose regret guarantees extend to bounded adversarial losses.

Several avenues remain open. First, understanding whether deeper or non-linear attention layers share the same stationary geometry would bridge the gap to practical Transformers. Second, translating these insights to LLMs, for example by pairing regret-based objectives with post-training pipelines as in Park et al. [2025a], would test whether the same mechanism appears at realistic scale. Third, extending the stationary-point characterization to bounded non-Gaussian training distributions, horizon-free rates, and richer feedback models such as delayed or bandit observations would clarify how robust the structure is beyond the full-information Gaussian setting.

References

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *Advanced in Neural Information Processing Systems*, 2023.

Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. *arXiv preprint arXiv:2204.01691*, 2022.

Ioannis Anagnostides, Gabriele Farina, and Tuomas Sandholm. Near-optimal phi-regret learning in extensive-form games. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 814–839. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/anagnostides23a.html.

Avrim Blum and Yishay Mansour. From external to internal regret. *Journal of Machine Learning Research*, 8(6), 2007.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Nicolo Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality, 2024. URL https://arxiv.org/abs/2402.19442.

Xiang Cheng, Yuxin Chen, and Suvrit Sra. Transformers implement functional gradient descent to learn non-linear functions in context. *arXiv preprint arXiv:2312.06528*, 2023.

Xiang Cheng, Yuxin Chen, and Suvrit Sra. Transformers implement functional gradient descent to learn non-linear functions in context, 2024. URL https://arxiv.org/abs/2312.06528.

Thomas M Cover. *Behavior of sequential predictors of binary sequences*. Number 7002. Stanford University, Stanford Electronics Laboratories, Systems Theory ..., 1966.

Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

Robert D Gordon. Values of mills' ratio of area to bounding ordinate and of the normal probability integral for large values of the argument. *The Annals of Mathematical Statistics*, 12(3):364–366, 1941.

Amy Greenwald and Amir Jafari. A general class of no-regret learning algorithms and game-theoretic equilibria. In *Learning Theory and Kernel Machines: 16th Annual Conference on Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003. Proceedings*, pages 2–12. Springer, 2003.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8154–8173, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.507. URL https://aclanthology.org/2023.emnlp-main.507.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Elad Hazan and Satyen Kale. Computational equivalence of fixed points and no regret algorithms, and convergence to equilibria. *Advances in Neural Information Processing Systems*, 20, 2007.

Akshay Krishnamurthy, Keegan Harris, Dylan J Foster, Cyril Zhang, and Aleksandrs Slivkins. Can large language models explore in-context? *Advances in Neural Information Processing Systems*, 37:120124–120158, 2024.

Michael Laskin, Luyu Wang, Junhyuk Oh, Emilio Parisotto, Stephen Spencer, Richie Steigerwald, DJ Strouse, Steven Hansen, Angelos Filos, Ethan Brooks, et al. In-context reinforcement learning with algorithm distillation. *International Conference on Learning Representations*, 2023.

Jonathan N Lee, Annie Xie, Aldo Pacchiano, Yash Chandak, Chelsea Finn, Ofir Nachum, and Emma Brunskill. Supervised pretraining can learn in-context reinforcement learning. *Neural Information Processing Systems*, 2023.

Chuanhao Li, Runhan Yang, Tiankai Li, Milad Bafarassat, Kourosh Sharifi, Dirk Bergemann, and Zhuoran Yang. Stride: A tool-assisted llm agent framework for strategic and interactive decision-making. *arXiv preprint arXiv:2405.16376*, 2024.

Licong Lin, Yu Bai, and Song Mei. Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. *International Conference on Learning Representations*, 2024.

Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and computation*, 108(2):212–261, 1994.

Arvind Mahankali, Tatsunori B Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. *International Conference on Learning Representations*, 2023.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759.

Chanwoo Park, Ziyang Chen, Asuman Ozdaglar, and Kaiqing Zhang. Post-training llms as better decision-making agents: A regret-minimization approach. *arXiv preprint arXiv:2511.04393*, 2025a.

Chanwoo Park, Xiangyu Liu, Asuman Ozdaglar, and Kaiqing Zhang. Do llm agents have regret? a case study in online learning and games. In *International Conference on Learning Representations*, 2025b.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *International conference on machine learning*, pages 9355–9366. PMLR, 2021.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends® in Machine Learning*, 4(2):107–194, 2012.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36, 2024.

Significant Gravitas. Autogpt, 2023. URL https://github.com/Significant-Gravitas/AutoGPT.

Charles M Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Volodimir G Vovk. Aggregating strategies. In *Proceedings of the third Annual Workshop on Computational Learning Theory*, pages 371–386, 1990.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023a.

Zihao Wang, Shaofei Cai, Anji Liu, Xiaojian Ma, and Yitao Liang. Describe, explain, plan and select: Interactive planning with large language models enables open-world multi-task agents. *Advances in neural information processing systems*, 2023b.

Fanzeng Xia, Hao Liu, Yisong Yue, and Tongxin Li. Beyond numeric awards: In-context dueling bandits with LLM agents. *arXiv preprint arXiv:2407.01887*, 2024.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *International Conference on Learning Representations*, 2023.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023.

Ziyuan Zhang, Darcy Wang, Ningyuan Chen, Rodrigo Mansur, and Vahid Sarhangian. Comparing exploration-exploitation strategies of llms and humans: Insights from standard multi-armed bandit tasks. *arXiv preprint arXiv:2505.09901*, 2025.

o

A Proof of Theorem 1

Theorem 1. *The configuration of a single-layer linear attention model as defined in Equation (2), with* `Operator` = Softmax*, and parameters* $V = -kI_{d\times d}$*,* $a = \mathbf{0}_d$*, and* $v_c = v\mathbf{1}_d$*, is a stationary point of the regret-loss function Equation (6) if* $k = \widetilde{\Theta}(1/\sqrt{T})$*, with* $h(x) = x^2$*,* $f_t(\pi) = \langle \ell_t, \pi \rangle$*, where* $\ell_t \sim \mathcal{N}(\mathbf{0}_d, I_{d\times d})$ *i.i.d.*

The proof will be divided into two parts. First, we prove the existence of the stationary point with $V = -kI_{d\times d}, a = \mathbf{0}_d, v_c = v\mathbf{1}_d$. Second, we show that k will have a scale of $\Theta(1/\sqrt{T})$.

Proof for the existence of the stationary point with $V = -kI_{d\times d}, a = \mathbf{0}_d, v_c = v\mathbf{1}_d$. Fix $d \geq 2$ and horizon T. For $t \in [T]$ write

$$S_t := \sum_{i=1}^{t} \ell_i \in \mathbb{R}^d, \qquad z_t(V, a, v_c) \;:=\; \sum_{i=1}^{t} \big(V\,\ell_i\ell_i^\intercal a + (V + v_c a^\intercal)\ell_i + v_c\big) \in \mathbb{R}^d,$$

and $\pi_t = \mathrm{Softmax}(z_t)$. With $V = -kI_{d\times d}$ and $a = \mathbf{0}_d$ we have

$$z_t(-kI_{d\times d}, 0, v_c) \;=\; -k\,S_t + t\,v_c.$$

Regret can be written as:

$$\mathrm{Regret}((\ell_t)_{t\leq T}) \;:=\; \sum_{t=1}^{T}\langle \ell_t, \pi_t\rangle - \inf_{\pi\in\Delta^d}\Big\langle \sum_{t=1}^{T}\ell_t,\ \pi\Big\rangle \;=\; \sum_{t=1}^{T}\langle \ell_t, \pi_t\rangle - \min_{j\in[d]}(S_T)_j.$$

For a fixed realization $(\ell_t)_{t\leq T}$, write $R := \mathrm{Regret}((\ell_t)_{t\leq T})$, and $\mathcal{L} \;=\; \mathbb{E}[h(R)]$, where $h(x) = x^2$. First, compute the derivative w.r.t. π_t. We have:

$$\frac{\partial R}{\partial \pi_t} = \ell_t,$$

since only the term $\langle \ell_t, \pi_t\rangle$ depends on π_t and the comparator $\min_j(S_T)_j$ is constant with respect to the parameters.

Let $J_{\mathrm{sm}}(z_t) \in \mathbb{R}^{d\times d}$ denote the Jacobian of Softmax at z_t, i.e.

$$J_{\mathrm{sm}}(z_t) \;=\; \frac{\partial\ \mathrm{Softmax}(z_t)}{\partial z_t} \;=\; \mathrm{Diag}(\pi_t) - \pi_t\pi_t^\intercal, \qquad \pi_t = \mathrm{Softmax}(z_t).$$

Then

$$\frac{\partial R}{\partial z_t} = J_{\mathrm{sm}}(z_t)^\intercal\,\frac{\partial R}{\partial \pi_t} = J_{\mathrm{sm}}(z_t)^\intercal\ell_t.$$

By the chain rule for $h(R)$,

$$\frac{\partial}{\partial z_t}h(R) \;=\; h'(R)\,\frac{\partial R}{\partial z_t} = 2R\,J_{\mathrm{sm}}(z_t)^\intercal\ell_t.$$

We define the (random) vector

$$\Delta_t \;:=\; \frac{\partial}{\partial z_t}h(R) \;=\; 2\,\mathrm{Regret}((\ell_s)_{s\leq T})\ J_{\mathrm{sm}}(z_t)^\intercal\ell_t \;\in\; \mathbb{R}^d.$$

The mapping $(z_1, \ldots, z_T) \mapsto h(R)$ is C^1, and by dominated convergence (Softmax and h are smooth, Gaussian tails), we may exchange gradient and expectation, so for any perturbation $(\delta z_t)_{t\leq T}$:

$$\delta\mathcal{L} \;=\; \mathbb{E}\Big[\sum_{t=1}^{T}\Big\langle \Delta_t,\ \delta z_t\Big\rangle\Big].$$

For arbitrary perturbations $\delta V, \delta a, \delta v$ (with $\delta v_c = \delta v\, \mathbf{1}_d$), δz_t is given by

$$\begin{aligned}\delta z_t &= \sum_{i=1}^{t} \Big(\delta V\, \ell_i \ell_i^\intercal a + V\, \ell_i \ell_i^\intercal \delta a\Big) + \sum_{i=1}^{t} \Big(\delta V\, \ell_i + v_c\, \langle \delta a, \ell_i\rangle + \delta v_c\, a^\intercal \ell_i\Big) + \sum_{i=1}^{t} \delta v_c \\ &= \underbrace{\sum_{i=1}^{t} \big(\delta V\, \ell_i \ell_i^\intercal a + \delta V\, \ell_i\big)}_{(*_V)} + \underbrace{\sum_{i=1}^{t} \big(V\, \ell_i \ell_i^\intercal \delta a + v_c\, \ell_i^\intercal \delta a\big)}_{(*_a)} + \underbrace{\sum_{i=1}^{t} \big(\delta v_c\, a^\intercal \ell_i + \delta v_c\big)}_{(*_v)}.\end{aligned}$$

We now instantiate at $(V, a, v_c) = (-kI_{d\times d}, 0, v\mathbf{1}_d)$, so that $S_t = \sum_{i=1}^{t} \ell_i$ and

$$(*_V) = \delta V\, S_t, \qquad (*_a) = \big(-k \sum_{i=1}^{t} \ell_i \ell_i^\intercal + v\, \mathbf{1}_d S_t^\intercal\big)\delta a, \qquad (*_v) = t\, \delta v\, \mathbf{1}_d.$$

Thus at $(V, a, v_c) = (-kI_{d\times d}, 0, v\mathbf{1}_d)$ we may write

$$\delta\mathcal{L} \;=\; \mathbb{E}\Big[\sum_{t=1}^{T} \langle \Delta_t,\, \delta z_t\rangle\Big], \quad \delta z_t \;=\; \delta V\, S_t \;+\; \big(-k \sum_{i=1}^{t} \ell_i \ell_i^\intercal + v\, \mathbf{1}_d S_t^\intercal\big)\delta a \;+\; t\, \delta v\, \mathbf{1}_d, \tag{11}$$

with

$$\Delta_t \;=\; 2\,\mathrm{Regret}((\ell_s)_{s\le T})\, \big(\mathrm{Diag}(\pi_t) - \pi_t \pi_t^\intercal\big)^\intercal \ell_t, \quad \pi_t = \mathrm{Softmax}(z_t).$$

(i) The v-direction. Using $\delta z_t^{(v)} = t\, \delta v\, \mathbf{1}_d$ in (11) we obtain

$$\delta\mathcal{L}^{(v)} = \delta v \sum_{t=1}^{T} t\, \mathbb{E}\big[\langle \Delta_t, \mathbf{1}_d\rangle\big].$$

Now

$$\Delta_t = 2R\big(\mathrm{Diag}(\pi_t) - \pi_t \pi_t^\intercal\big)\ell_t, \qquad \pi_t = \mathrm{Softmax}(z_t),$$

and hence

$$\begin{aligned}\langle \Delta_t, \mathbf{1}_d\rangle &= 2R\, \mathbf{1}_d^\intercal \big(\mathrm{Diag}(\pi_t) - \pi_t \pi_t^\intercal\big)\ell_t \\ &= 2R\, \big(\underbrace{\mathbf{1}_d^\intercal \mathrm{Diag}(\pi_t)}_{=\mathbf{1}_d^\intercal \pi_t = 1} - \underbrace{\mathbf{1}_d^\intercal \pi_t \pi_t^\intercal}_{=1\cdot \pi_t^\intercal}\big)\ell_t = 2R\,(1-1)\,\pi_t^\intercal \ell_t = 0.\end{aligned}$$

Thus each term in the sum vanishes pointwise, so $\delta\mathcal{L}^{(v)} = 0$ for all δv, i.e.

$$\frac{\partial \mathcal{L}}{\partial v} = 0,$$

which matches (and refines) the earlier shift-invariance argument.

(ii) The a-direction at $a = \mathbf{0}_d$. At $(V, a, v_c) = (-kI_{d\times d}, 0, v\mathbf{1}_d)$, (11) gives

$$\begin{aligned}\delta\mathcal{L}^{(a)} &= \mathbb{E}\Big[\sum_{t=1}^{T} \Big\langle \Delta_t,\, \big(-k \sum_{i=1}^{t} \ell_i \ell_i^\intercal + v\, \mathbf{1}_d S_t^\intercal\big)\delta a\Big\rangle\Big] \\ &= -k\, \mathbb{E}\Big[\sum_{t=1}^{T} \Big\langle \Delta_t,\, \sum_{i=1}^{t} \ell_i \ell_i^\intercal \delta a\Big\rangle\Big] \;+\; v\, \mathbb{E}\Big[\sum_{t=1}^{T} \langle \Delta_t, \mathbf{1}_d S_t^\intercal \delta a\rangle\Big].\end{aligned}$$

The second term vanishes because $\mathbf{1}_d^\intercal \Delta_t = 0$:

$$\langle \Delta_t, \mathbf{1}_d S_t^\intercal \delta a\rangle = (\mathbf{1}_d^\intercal \Delta_t)\, \langle S_t, \delta a\rangle = 0.$$

Hence

$$\delta\mathcal{L}^{(a)} = -k\, \Big\langle \mathbb{E}\Big[\sum_{t=1}^{T} \sum_{i=1}^{t} \ell_i \ell_i^\intercal \Delta_t\Big],\, \delta a\Big\rangle.$$

We may write

$$\nabla_a \mathcal{L}\big|_{(V,a,v_c)=(-kI_{d\times d},0,v\mathbf{1}_d)} = -k\,\mathbb{E}\Big[\sum_{t=1}^{T}\sum_{i=1}^{t}\ell_i\ell_i^\intercal\Delta_t\Big] \;\in\; \mathbb{R}^d.$$

Set

$$H := \mathbb{E}\Big[\sum_{t=1}^{T}\sum_{i=1}^{t}\ell_i\ell_i^\intercal\Delta_t\Big] \in \mathbb{R}^d,$$

so that $\nabla_a\mathcal{L} = -kH$. We claim $H = \mathbf{0}_d$.

To see this, use the *permutation invariance* of the law of $(\ell_t)_{t\le T}$ (it is spherically symmetric since $\ell_t \sim \mathcal{N}(0, I_{d\times d})$ i.i.d.) and the *permutation equivariance* of $\mathrm{Softmax}$: for any permutation matrix P, $P\,\mathrm{Softmax}(x) = \mathrm{Softmax}(Px)$. Under $(\ell_t)\mapsto(P\ell_t)$ we have $S_t \mapsto PS_t$ and hence $z_t \mapsto -k\,PS_t$, so $\pi_t \mapsto P\pi_t$ and $\Delta_t \mapsto P\Delta_t$. Therefore

$$H \;=\; \mathbb{E}\Big[\sum_{t,i}(P\ell_i)(P\ell_i)^\intercal(P\Delta_t)\Big] \;=\; P\,\mathbb{E}\Big[\sum_{t,i}\ell_i\ell_i^\intercal\Delta_t\Big] \;=\; P\,H \quad \text{for all permutations } P.$$

The only vectors fixed by all coordinate permutations are multiples of $\mathbf{1}_d$. On the other hand,

$$\mathbf{1}_d^\intercal H = \mathbb{E}\Big[\sum_{t,i}\langle\mathbf{1}_d,\ \ell_i\ell_i^\intercal\Delta_t\rangle\Big] = \mathbb{E}\Big[\sum_{t,i}\langle\ell_i,\Delta_t\rangle\cdot\langle\mathbf{1}_d,\ell_i\rangle\Big].$$

By symmetry the joint law is invariant under $\ell_i \mapsto -\ell_i$ while Δ_t changes sign exactly as ℓ_t does through the $\mathrm{Softmax}$ Jacobian (the mapping is odd along each coordinate in expectation); thus $\mathbf{1}_d^\intercal H = 0$. Hence the only permutation-fixed vector orthogonal to $\mathbf{1}_d$ is $\mathbf{0}_d$, so $H = \mathbf{0}_d$ and therefore

$$\nabla_a\mathcal{L}\big|_{a=\mathbf{0}} = \mathbf{0}_d.$$

(iii) The V-direction at $V = -kI_{d\times d}$. From $(*_V)$, at $a = 0$ we have $\delta z_t = \delta V\, S_t$. Using (11) and writing the Frobenius product,

$$\begin{aligned}\delta\mathcal{L} &= \mathbb{E}\Big[\sum_{t=1}^{T}\langle\Delta_t,\ \delta V S_t\rangle\Big] = \mathbb{E}\Big[\sum_{t=1}^{T}\langle\Delta_t S_t^\intercal,\ \delta V\rangle_F\Big]\\ &= \Big\langle \underbrace{\mathbb{E}\Big[\sum_{t=1}^{T}\Delta_t S_t^\intercal\Big]}_{=:G(V)},\ \delta V\Big\rangle_F,\end{aligned}$$

so $\nabla_V\mathcal{L} = G(V)$. Evaluate at $V = -kI_{d\times d}$. For every permutation matrix P, the Gaussian law is invariant under $(\ell_t)\mapsto(P\ell_t)$, and $\mathrm{Softmax}$ is permutation-equivariant. Thus $(S_t,\Delta_t) \overset{d}{=} (PS_t, P\Delta_t)$ and

$$G(-kI_{d\times d}) \;=\; \mathbb{E}\Big[\sum_{t=1}^{T}\Delta_t S_t^\intercal\Big] \;=\; \mathbb{E}\Big[\sum_{t=1}^{T}(P\Delta_t)(PS_t)^\intercal\Big] \;=\; P\,G(-kI_{d\times d})\,P^\intercal$$

for all permutation matrices P. By Lemma 1, there are scalars $\alpha(k),\beta(k)$ such that

$$G(-kI_{d\times d}) = \alpha(k)I_{d\times d} + \beta(k)\mathbf{1}_d\mathbf{1}_d^\intercal.$$

Moreover, $\mathbf{1}_d^\intercal\Delta_t = 0$ for every t because $J_{\mathrm{sm}}(z_t)^\intercal\mathbf{1}_d = \mathbf{0}_d$, and hence $\mathbf{1}_d^\intercal G(-kI_{d\times d}) = \mathbf{0}_d^\intercal$. Therefore

$$\nabla_V\mathcal{L}(-kI_{d\times d},0,v\mathbf{1}_d) = \alpha(k)\Pi_0, \qquad \Pi_0 := I_{d\times d} - \frac{1}{d}\mathbf{1}_d\mathbf{1}_d^\intercal.$$

Let $k^\star$ be any stationary point of the one-dimensional function

$$\phi(k) := \mathcal{L}(V = -kI_{d\times d}, a = 0, v_c = v\mathbf{1}_d),$$

which exists by smoothness and standard compactification/continuity arguments; at a local minimizer we certainly have $\phi'(k^\star) = 0$. Then

$$0 \;=\; \phi'(k^\star) \;=\; \Big\langle\nabla_V\mathcal{L}(-k^\star I,0,v\mathbf{1}_d),\ \frac{\partial(-kI_{d\times d})}{\partial k}\Big\rangle_F \;=\; \Big\langle\alpha(k^\star)\Pi_0,\ -I\Big\rangle_F \;=\; -(d-1)\alpha(k^\star),$$

so $\alpha(k^\star) = 0$ and hence $\nabla_V\mathcal{L}(-k^\star I,0,v\mathbf{1}_d) = \mathbf{0}$.

Conclusion. At the parameter triple $(V, a, v_c) = (-k^\star I_{d\times d},\ \mathbf{0}_d,\ v\,\mathbf{1}_d)$ we have

$$\nabla_V \mathcal{L} = \boldsymbol{O}_{d\times d}, \qquad \nabla_a \mathcal{L} = \mathbf{0}_d, \qquad \nabla_{v_c} \mathcal{L} = \mathbf{0}_d.$$

Thus the configuration is a first-order stationary point of $\mathcal{L}$. □

Remark 1. *The argument uses only: (1)* Softmax *shift-invariance and its Jacobian* $J_{\mathrm{sm}}(z) = \mathrm{Diag}(\pi) - \pi\pi^\intercal$*; (2) permutation invariance of the Gaussian hierarchical model* $\ell_t \sim \mathcal{N}(0, I_{d\times d})$ *i.i.d.; and (3) the zero-sum range of the* Softmax *Jacobian.*

Now we show that k will have a scale of $\Theta(1/\sqrt{T})$.

Proof for the scale of k. Fix integers $d \geq 2$ and $T \geq 1$. Let $(\ell_t)_{t=1}^\intercal$ be i.i.d. $\mathcal{N}(0, I_{d\times d})$, and define partial sums $S_t := \sum_{i=1}^t \ell_i$ with $S_0 := 0$, exactly as in the previous part of the proof.

For $k \in \mathbb{R}$, the Softmax policy induced by $V = -kI_{d\times d}$, $a = \mathbf{0}_d$, $v_c = v\mathbf{1}_d$ reads

$$\pi_t(k) := \mathrm{Softmax}(-kS_{t-1}) \in \Delta_d, \qquad J_t(k) := J_{\mathrm{sm}}(-kS_{t-1}) = \mathrm{Diag}(\pi_t(k)) - \pi_t(k)\pi_t(k)^\intercal,$$

where we re-use the Jacobian notation J_{sm} from the first part. Let the regret against the best fixed coordinate be

$$R_T(k) := \sum_{t=1}^T \langle \ell_t, \pi_t(k)\rangle - \min_{j\in[d]} S_{T,j},$$

and the objective $\mathcal{L}(k) := \mathbb{E}[R_T(k)^2]$. Let $J^\star := \arg\min_{j\in[d]} S_{T,j}$ (unique a.s.). Write $u := \frac{1}{d}\mathbf{1}_d$ and $J_0 := \mathrm{Diag}(u) - uu^\intercal = \frac{1}{d}P_{\perp \mathbf{1}_d}$.

We will use the following two elementary facts.

Fact 2. *For each* $t \leq T$, $\mathbb{E}[S_{t-1} \mid S_T] = \frac{t-1}{T}S_T$ *and* $S_{t-1} - \frac{t-1}{T}S_T \perp\!\!\!\perp S_T$. *Hence*

$$\mathbb{E}\big[S_{t-1,J^\star}\big] = \frac{t-1}{T}\,\mathbb{E}\left[\min_{1\leq j\leq d} S_{T,j}\right].$$

Fact 3 (Exact derivative identity)**.** *For every* $k \in \mathbb{R}$,

$$\mathcal{L}'(k) = \underbrace{-2\sum_{t=1}^T \mathbb{E}\big[\langle J_t(k)S_{t-1},\ \pi_t(k) - u\rangle\big]}_{=:\Sigma_1} + \underbrace{2\sum_{t=1}^T \mathbb{E}\big[\langle J_t(k)S_{t-1},\ e_{J^\star}\rangle\big]}_{=:\Sigma_2}. \tag{12}$$

Proof. Since the comparator $\min_j S_{T,j}$ is k-independent,

$$\frac{d}{dk}\sum_{t=1}^T \langle \ell_t, \pi_t(k)\rangle = -\sum_{t=1}^T \langle \ell_t, J_t(k)S_{t-1}\rangle.$$

Thus $\mathcal{L}'(k) = 2\,\mathbb{E}\big[R_T(k)\cdot \frac{d}{dk}\sum_t \langle \ell_t, \pi_t(k)\rangle\big] = -2\sum_t \mathbb{E}\big[R_T(k)\langle \ell_t, J_t(k)S_{t-1}\rangle\big]$.

Fix t and condition on $\mathcal{F}_{t-1} = \sigma(\ell_1, \ldots, \ell_{t-1})$. Stein's lemma (Lemma 12) for $\ell_t \sim \mathcal{N}(0, I_{d\times d})$ yields

$$\mathbb{E}\big[\langle \ell_t, v\rangle R_T(k) \mid \mathcal{F}_{t-1}\big] = \mathbb{E}\big[\langle v, \nabla_{\ell_t} R_T(k)\rangle \mid \mathcal{F}_{t-1}\big], \quad v = J_t(k)S_{t-1}.$$

Because $\partial_{\ell_t}\langle \ell_t, \pi_t(k)\rangle = \pi_t(k)$, for $s > t$ one has $\partial_{\ell_t}\langle \ell_s, \pi_s(k)\rangle = -k\,J_s(k)\ell_s$, and $\partial_{\ell_t}\min_j S_{T,j} = e_{J^\star}$ almost surely; hence $\nabla_{\ell_t} R_T(k) = \pi_t(k) - e_{J^\star} - k\sum_{s>t} J_s(k)\ell_s$. Therefore

$$\mathbb{E}\big[R_T(k)\langle \ell_t, v\rangle\big] = \mathbb{E}[\langle v, \pi_t(k) - e_{J^\star}\rangle] - k\sum_{s>t}\mathbb{E}[\langle v, J_s(k)\ell_s\rangle].$$

Conditioning on $\mathcal{F}_{s-1}$, the second term is 0 because $\mathbb{E}[\ell_s \mid \mathcal{F}_{s-1}] = 0$. Summing in t gives the claim. Here, we also used that $J_s(k)\mathbf{1}_d = \mathbf{0}_d$ for all s and k by definition of $J_s(k)$. □

Now, we provide a first-order expansion of the policy with bounds on the remainder term.

Proposition 2 (First–order expansion of the policy)**.** *The following equation and inequalities hold:*

$$\pi_t(k) - u \;=\; -\,J_0\,k\,S_{t-1} \;+\; r_t, \qquad \|r_t\|_2 \leq \frac{3}{4}\,k^2\,\|S_{t-1}\|_2^2, \qquad \|r_t\|_\infty \leq k^2\,\|S_{t-1}\|_\infty^2. \tag{13}$$

Proof. Use the fundamental theorem of calculus on the ray $\theta \mapsto \theta z_t$:

$$\pi_t(k) - u \;=\; \mathrm{Softmax}(-kS_{t-1}) - \mathrm{Softmax}(0) \;=\; \int_0^1 J_{\mathrm{sm}}(\theta(-kS_{t-1}))\,(-kS_{t-1})\,d\theta.$$

Add and subtract $-J_0 kS_{t-1}$ and define $r_t := \int_0^1 \left(J_{\mathrm{sm}}(\theta(-kS_{t-1})) - J_0\right)(-kS_{t-1})\,d\theta$, Then by Lemma 2-(d),

$$\|r_t\|_2 \;\le\; \int_0^1 \left\|J_{\mathrm{sm}}(\theta(-kS_{t-1}))-J_0\right\|_{\mathrm{op}} \|(-kS_{t-1})\|_2\,d\theta \;\le\; \int_0^1 \frac{3}{2}\theta\|kS_{t-1}\|_2^2\,d\theta = \frac{3}{4}\|kS_{t-1}\|_2^2.$$

Thus (13) holds with $C_1 := \frac{3}{4}$. Moreover, using Lemma 2-(e) with $h = -kS_{t-1}$,

$$\left\|J_{\mathrm{sm}}(\theta(-kS_{t-1})) - J_0\right\|_{\infty\to\infty} \le 2\theta\,|k|\,\|S_{t-1}\|_\infty.$$

Therefore, the following inequality holds:

$$\|r_t\|_\infty \le \int_0^1 \left\|J_{\mathrm{sm}}(-\theta kS_{t-1})-J_0\right\|_{\infty\to\infty} \|kS_{t-1}\|_\infty\,d\theta \le \int_0^1 2\theta\,|k|\,\|S_{t-1}\|_\infty\,|k|\,\|S_{t-1}\|_\infty\,d\theta = k^2\|S_{t-1}\|_\infty^2.$$

□

Proposition 3 (First–order expansion of the Jacobian)**.** *There exists an absolute $L > 0$ such that*

$$J_t(k) \;=\; J_0 \;+\; R_t, \qquad \|R_t\|_{\mathrm{op}} \;\le\; L\,k\,\|S_{t-1}\|_2 \qquad \|R_t\|_{\infty\to\infty} \le 2k\|S_{t-1}\|_\infty. \tag{14}$$

Proposition 3 is a re-instantitation of Lemma 2-(e).

We now establish two auxiliary lemmas specific to this proof.

Theorem 3 (Negativity for $k \le c/\sqrt{Td}$)**.** *There exists an absolute constant $c_* > 0$ such that if $k = \widetilde{\mathcal{O}}(1/\sqrt{Td})$, then $\mathcal{L}'(k) < 0$.*

Proof. First, we provide an upper bound on $-2\Sigma_1$. Fix t and abbreviate $S := S_{t-1}$, $\pi := \pi_t(k)$, $J_t := J_t(k)$. From Proposition 2 and Proposition 3, we have

$$\pi - u = -kJ_0S + r_t, \qquad J_t = J_0 + R_t.$$

By plugging in formula for Σ_1, we have

$$\begin{aligned}\langle J_tS, \pi - u\rangle &= \langle (J_0+R_t)S, -kJ_0S + r_t\rangle \\ &= -k\|J_0S\|_2^2 \;+\; \underbrace{\langle J_0S, r_t\rangle}_{E_{1,t}} \;-\; \underbrace{k\langle R_tS, J_0S\rangle}_{E_{2,t}} \;+\; \underbrace{\langle R_tS, r_t\rangle}_{E_{3,t}}.\end{aligned} \tag{15}$$

Therefore, we have

$$-2\Sigma_1 \;\le\; 2k\sum_{t=1}^{T}\mathbb{E}\|J_0S_{t-1}\|_2^2 + 2\sum_t \mathbb{E}|E_{1,t}| + 2k\sum_t \mathbb{E}|E_{2,t}| + 2\sum_t \mathbb{E}|E_{3,t}|. \tag{16}$$

Main term. Because $S_{t-1} \sim \mathcal{N}(\mathbf{0}_d, (t-1)I_{d\times d})$ and J_0 is a projection (scaled by $1/d$) onto an $(d-1)$-dimensional subspace,

$$\mathbb{E}\|J_0S_{t-1}\|_2^2 = \frac{d-1}{d^2}(t-1),$$

and thus

$$2k\sum_{t=1}^{T}\mathbb{E}\|J_0S_{t-1}\|_2^2 = k\frac{d-1}{d^2}T(T-1). \tag{17}$$

Bound on $E_{1,t} = \langle J_0S, r_t\rangle$. By Hölder inequality in $(1,\infty)$, Proposition 2, and $\|S_{t-1}\|_\infty \le \|S_{t-1}\|_\infty$,

$$|E_{1,t}| \le \|J_0S\|_1\|r_t\|_\infty \le \|J_0\|_{1\to1}\|S\|_1 \cdot k^2\|S\|_\infty^2 \le \frac{2}{d}k^2\|S\|_1\|S\|_\infty^2 \le 2k^2\|S\|_\infty^3.$$

Taking expectations and summing over t with Lemma 8, for some absolute constant $C_1 > 0$,

$$\sum_{t=1}^{T} \mathbb{E}|E_{1,t}| \le 2k^2 \sum_{t=1}^{T} \mathbb{E}\|S_{t-1}\|_\infty^3 \le C_1 k^2 T^{5/2} (\log(ed))^{3/2}. \tag{18}$$

Bound on $E_{2,t} = k\langle R_t S, J_0 S\rangle$. We will prove there exists an absolute constant $C > 0$ such that for all $t \le T$,

$$\big|\mathbb{E}[E_{2,t}]\big| \;\le\; \frac{C}{d^{3/2}}\, k^2\, (t-1)^{3/2} \big(\log(ed)\big)^{3/2}\, e^{4k^2(t-1)}. \tag{19}$$

and consequently,

$$\sum_{t=1}^{T} \big|\mathbb{E}[E_{2,t}]\big| \;\le\; \frac{C}{d^{3/2}}\, k^2\, T^{5/2} \big(\log(ed)\big)^{3/2}\, e^{4k^2 T}. \tag{20}$$

By Lemma 4, $(J_t(k) - J_0)S = (J_t(k) - J_0)J_0 S$, so

$$E_{2,t} = k\big\langle (J_{\mathrm{sm}}(-kS) - J_0)J_0 S,\; J_0 S\big\rangle.$$

Let $J(z) := J_{\mathrm{sm}}(z)$. Along the ray $\theta \mapsto -\theta k S$ we have

$$J(-kS) - J_0 \;=\; \int_0^1 DJ(-\theta k S)[-kS]\, d\theta.$$

Therefore

$$E_{2,t} = k \int_0^1 \big\langle DJ(-\theta k S)[-kS]\, J_0 S,\; J_0 S\big\rangle\, d\theta. \tag{21}$$

Expand $DJ(-\theta k S)$ around 0:

$$DJ(-\theta k S)[-kS] = DJ(0)[-kS] \;+\; \int_0^\theta D^2 J(-\tau k S)[-kS, -kS]\, d\tau.$$

Plugging this into (21), we obtain

$$E_{2,t} = H_{\mathrm{lin}}(S) + H_{\mathrm{rem}}(S),$$

where

$$H_{\mathrm{lin}}(S) := k\big\langle DJ(0)[-kS]\, J_0 S,\; J_0 S\big\rangle,$$
$$H_{\mathrm{rem}}(S) := k \int_0^1 \int_0^\theta \big\langle D^2 J(-\tau k S)[-kS, -kS]\, J_0 S,\; J_0 S\big\rangle\, d\tau\, d\theta.$$

Consider the transformation $S \mapsto -S$. Since J_0 is linear, $J_0(-S) = -J_0 S$. Furthermore, $DJ(0)[h]$ is linear in h, so $DJ(0)[-k(-S)] = DJ(0)[kS] = -DJ(0)[-kS]$. A direct sign check yields

$$H_{\mathrm{lin}}(-S) = k\big\langle DJ(0)[-k(-S)]\, J_0(-S),\; J_0(-S)\big\rangle = -H_{\mathrm{lin}}(S).$$

Since S is symmetric in law, $\mathcal{L}(S) = \mathcal{L}(-S)$, we obtain $\mathbb{E}[H_{\mathrm{lin}}(S)] = 0$ and therefore

$$\mathbb{E}[E_{2,t}] = \mathbb{E}[H_{\mathrm{rem}}(S)].$$

Let

$$M_\tau(S) := D^2 J(-\tau k S)[-kS, -kS].$$

By Hölder in $(\infty, 1)$,

$$\big|\langle M_\tau(S) J_0 S,\; J_0 S\rangle\big| \le \|M_\tau(S)\|_{\infty\to\infty}\, \|J_0 S\|_\infty\, \|J_0 S\|_1.$$

By Lemma 2 (f),

$$\|M_\tau(S)\|_{\infty\to\infty} \le \frac{16}{d}\, e^{2\tau k q}\, k^2 q^2, \qquad q := \|S\|_\infty.$$

By Lemma 6,

$$\|J_0 S\|_\infty \le \frac{2}{d}\, q, \qquad \|J_0 S\|_1 \le 2\, q.$$

Hence

$$\big|\langle M_\tau(S)J_0S,\ J_0S\rangle\big| \le \frac{64}{d^2}\,k^2q^4e^{2\tau kq}.$$

Therefore,

$$\begin{aligned}|H_{\rm rem}(S)| &\le k\int_0^1\int_0^\theta \big|\langle M_\tau(S)J_0S,\ J_0S\rangle\big|\,d\tau\,d\theta\\ &\le \frac{64}{d^2}\,k^3q^4\int_0^1\int_0^\theta e^{2\tau kq}\,d\tau\,d\theta.\end{aligned}$$

Using

$$\int_0^1\int_0^\theta e^{2\tau kq}\,d\tau\,d\theta = \int_0^1(1-\tau)e^{2\tau kq}\,d\tau \le \int_0^1 e^{2\tau kq}\,d\tau = \frac{e^{2kq}-1}{2kq} \le \frac{1}{2kq}e^{2kq},$$

we obtain

$$|H_{\rm rem}(S)| \le \frac{32}{d^2}\,k^2q^3e^{2kq}.$$

Taking expectations,

$$\big|\mathbb{E}[E_{2,t}]\big| = \big|\mathbb{E}[H_{\rm rem}(S)]\big| \le \frac{32}{d^2}\,k^2\,\mathbb{E}\big[q^3e^{2kq}\big].$$

Apply Cauchy–Schwarz:

$$\mathbb{E}\big[q^3e^{2kq}\big] \le \big(\mathbb{E}[q^6]\big)^{1/2}\big(\mathbb{E}[e^{4kq}]\big)^{1/2}.$$

By Lemma 8 with $q=6$,

$$\mathbb{E}[q^6] \lesssim (t-1)^3\big(\log(ed)\big)^3.$$

By the MGF bound for the sup (Lemma 11),

$$\mathbb{E}[e^{4kq}] \lesssim d\,\exp\big(8k^2(t-1)\big).$$

Thus

$$\mathbb{E}\big[q^3e^{2kq}\big] \lesssim (t-1)^{3/2}\big(\log(ed)\big)^{3/2}\,d^{1/2}e^{4k^2(t-1)}.$$

Combining this with the prefactor $32/d^2$ gives

$$\big|\mathbb{E}[E_{2,t}]\big| \ \le\ \frac{C}{d^{3/2}}\,k^2\,(t-1)^{3/2}\big(\log(ed)\big)^{3/2}e^{4k^2(t-1)}$$

for some absolute constant $C>0$, proving (19). Summing over t and using $\sum_{t=1}^T(t-1)^{3/2}\lesssim T^{5/2}$ yields (20).

Bound on $E_{3,t}=\langle R_tS, r_t\rangle$. We use the integral representation of r_t. For $\theta\in[0,1]$ set $A_\theta := J_{\rm sm}(\theta(-kS))-J_0$ so that $R_t=A_1$ and $r_t=\int_0^1 A_\theta(-kS)\,d\theta = -k\int_0^1 A_\theta S\,d\theta$. By Lemma 4, $A_\theta S = A_\theta J_0 S$ and $R_tS = R_tJ_0S$. Therefore

$$E_{3,t} = \big\langle R_tS,\ r_t\big\rangle = -k\int_0^1 \big\langle R_t(J_0S),\ A_\theta(J_0S)\big\rangle\,d\theta.$$

Taking absolute values and using Hölder in $(\infty,1)$,

$$|E_{3,t}| \le k\int_0^1 \|R_t(J_0S)\|_\infty\,\|A_\theta(J_0S)\|_1\,d\theta \le k\int_0^1 \|R_t\|_{\infty\to\infty}\,\|J_0S\|_\infty\cdot\|A_\theta\|_{1\to1}\,\|J_0S\|_1\,d\theta.$$

By Lemma 2-(e) with $h=-kS$, $\|R_t\|_{\infty\to\infty}\le 2k\|S\|_\infty$ and $\|A_\theta\|_{1\to1} = \|J_{\rm sm}(\theta(-kS))-J_0\|_{1\to1}\le 2\theta k\|S\|_\infty$. By the definition of J_0, $\|J_0S\|_\infty\le\frac{2}{d}\|S\|_\infty$ and $\|J_0S\|_1\le\frac{2}{d}\|S\|_1\le 2\|S\|_\infty$ (the last inequality uses $\|S\|_1\le d\|S\|_\infty$). Plugging these four bounds,

$$|E_{3,t}| \le k\int_0^1 \Big(2k\|S\|_\infty\Big)\Big(\tfrac{2}{d}\|S\|_\infty\Big)\Big(2\theta k\|S\|_\infty\Big)\Big(2\|S\|_\infty\Big)\,d\theta = \frac{8}{d}\,k^3\,\|S\|_\infty^4.$$

Taking expectations, using Lemma 8 with $q = 4$, and summing over t,

$$\sum_{t=1}^{T} \mathbb{E}|E_{3,t}| \le \frac{8}{d}\, k^3 \sum_{t=1}^{T} \mathbb{E}\|S_{t-1}\|_\infty^4 \le \frac{C}{d}\, k^3 \sum_{t=1}^{T} (t-1)^2 \log(ed) \le \frac{C}{d}\, k^3\, T^3 \left(\log(ed)\right)^2. \tag{22}$$

Combining (16), (17), (18), (20), and (22), we obtain

$$-2\Sigma_1 \;\le\; \frac{k}{d}T^2 + C_1 k^2 T^{5/2} (\log(ed))^{3/2} + C_2 \frac{k^2}{d^{3/2}} T^{5/2} (\log(ed))^{3/2}\, e^{4k^2 T} + C_3 \frac{k^3}{d} T^3 (\log(ed))^2. \tag{23}$$

If $k = \widetilde{\mathcal{O}}(1/\sqrt{Td})$ for some absolute constant $c_* > 0$, then

$$-2\Sigma_1 \;\le\; \widetilde{\mathcal{O}}\left(\frac{T^{3/2}}{d^{3/2}} + \frac{T^{3/2}}{d} + \frac{T^{3/2}}{d^{5/2}} + \frac{T^{3/2}}{d^{7/2}}\right).$$

Next, we provide an upper bound on $2\Sigma_2$. We split Σ_2 into two parts:

$$\Sigma_2 = \Sigma_{2,0} + \Sigma_{2,\mathrm{pert}}, \qquad \Sigma_{2,0} := \sum_{t=1}^{T} \mathbb{E}\big[\langle J_0 S_{t-1}, e_{J^\star}\rangle\big], \quad \Sigma_{2,\mathrm{pert}} := \sum_{t=1}^{T} \mathbb{E}\big[\langle (J_t(k) - J_0) S_{t-1}, e_{J^\star}\rangle\big].$$

Baseline negative term ($\Sigma_{2,0}$). By Fact 2 and Lemma 9,

$$2\Sigma_{2,0} \;\le\; -\,c_0\, \frac{\sqrt{T}\,(T-1)\,\sqrt{\log d}}{d}. \tag{24}$$

$$2\Sigma_{2,0} \;=\; -\,\Theta\left(\frac{T^{3/2}\,\sqrt{\log d}}{d}\right). \tag{25}$$

Perturbative term $\Sigma_{2,\mathrm{pert}}$.

By Lemma 4,

$$\big\langle (J_t(k) - J_0) S,\, e_{J^\star}\big\rangle \;=\; \big\langle R_t J_0 S,\, e_{J^\star}\big\rangle, \qquad R_t := J_t(k) - J_0.$$

Hence, using Hölder in $(\infty, 1)$ and $\|e_{J^\star}\|_1 = 1$,

$$\big|\big\langle (J_t(k) - J_0) S,\, e_{J^\star}\big\rangle\big| \le \|R_t J_0 S\|_\infty \le \|R_t\|_{\infty\to\infty}\, \|J_0 S\|_\infty.$$

By Lemma 2 (e) with $h = -kS$, $\|R_t\|_{\infty\to\infty} = \|J_{\mathrm{sm}}(-kS) - J_0\|_{\infty\to\infty} \le 2k\|S\|_\infty$. By the definition of J_0, $\|J_0 S\|_\infty \le \frac{2}{d}\|S\|_\infty$. Therefore

$$\big|\big\langle (J_t(k) - J_0) S,\, e_{J^\star}\big\rangle\big| \le \frac{4}{d}\, k\, \|S\|_\infty^2.$$

Taking expectations and summing over t, and using Lemma 8 with $q = 2$,

$$|\Sigma_{2,\mathrm{pert}}| \le \frac{4}{d}\, k \sum_{t=1}^{T} \mathbb{E}\|S_{t-1}\|_\infty^2 \le \frac{C}{d}\, k \sum_{t=1}^{T} (t-1) \log(ed) \le \frac{C}{d}\, k\, T^2 \log(ed).$$

Together, we have

$$2\Sigma_2 \le -c_0 \frac{\sqrt{T}\,(T-1)\,\sqrt{\log d}}{d} + \frac{C}{d}\, k\, T^2 \log(ed).$$

If $k = \widetilde{\mathcal{O}}(1/\sqrt{Td})$ for some absolute constant $c_* > 0$, then

$$2\Sigma_2 \le -c_0 \frac{\sqrt{T}\,(T-1)\,\sqrt{\log d}}{d} + \widetilde{\mathcal{O}}\left(\frac{T^{3/2}}{d^{3/2}}\right).$$

Therefore, $\mathcal{L}'(k) < 0$ for $k = \widetilde{\mathcal{O}}(1/\sqrt{Td})$ as desired.

□

Theorem 4 (Positivity for $k = C/\sqrt{T}\text{polylog}(d)$)**.** *There exists an absolute constant $C > 0$ such that for all $T, d \geq 2$, if $k = \widetilde{\Theta}(1/\sqrt{T})$, then $\mathcal{L}'(k) > 0$.*

Proof. First, we use Fact 3 and analyze Σ_1 and Σ_2 separately. First, we restate Equation (15) as

$$\begin{aligned}\langle J_t S, \pi - u\rangle &= \langle (J_0 + R_t)S, \pi - u\rangle \\ &= \underbrace{\langle J_0 S, \pi\rangle}_{N_{1,t}} - \underbrace{\langle R_t S, \pi - u\rangle}_{N_{2,t}}. \end{aligned} \tag{26}$$

First, we will bound $N_{2,t}$.

Claim 1. *There exist absolute constants $c_0, C_0 > 0$ such that the following holds. For every k with $0 < k \leq c_0/\sqrt{T}$,*

$$\left|\mathbb{E}\Big\langle \big(J_t(k) - J_t(0)\big)S_{t-1},\ \pi_t(k) - \pi_0\Big\rangle\right| \ \leq\ C_0\, T\, k\, d^{-3/2}\, \log(ed). \tag{27}$$

Proof. Again, we abbreviate $S := S_{t-1}$, and write $\pi_t(k) = \text{sm}(-kS)$ and $J_t(k) = J_{\text{sm}}(-kS)$. We wish to bound the following term:

$$Q_t(k) := \mathbb{E}\Big\langle (J_t(k) - J_t(0))S_{t-1},\ \pi_t(k) - \pi_0\Big\rangle = \mathbb{E}\langle RS,\ \pi_k - \pi_0\rangle. \tag{28}$$

By Lemma 4, for every $z \in \mathbb{R}^d$ we have

$$\big(J_{\text{sm}}(z) - J_0\big)\mathbf{1}_d = \mathbf{0}_d.$$

Then S decomposes into a multiple of $\mathbf{1}_d$ plus $S_\perp \in \text{range}(J_0)$, and Lemma 4 implies

$$RS = RS_\perp. \tag{29}$$

Moreover, π_k and π_0 are probability vectors, so $\|\pi_k - \pi_0\|_1 \leq 2$. Thus, by Hölder's inequality in $(\infty, 1)$,

$$|\langle RS, \pi_k - \pi_0\rangle| = |\langle RS_\perp, \pi_k - \pi_0\rangle| \leq \|RS_\perp\|_\infty \|\pi_k - \pi_0\|_1 \leq 2\,\|RS_\perp\|_\infty. \tag{30}$$

Taking expectations in (30) and using (28),

$$|Q_t(k)| \ \leq\ 2\,\mathbb{E}\|RS_\perp\|_\infty. \tag{31}$$

It remains to bound $\mathbb{E}\|RS_\perp\|_\infty$. Along the ray $\theta \mapsto -\theta kS$ we have the second-order expansion

$$R = J_{\text{sm}}(-kS) - J_{\text{sm}}(0) = DJ_{\text{sm}}(0)[-kS] + \int_0^1 (1-\theta)\, D^2 J_{\text{sm}}(-\theta kS)[-kS, -kS]\, d\theta. \tag{32}$$

Set

$$L := DJ_{\text{sm}}(0)[-kS], \qquad \mathcal{R} := \int_0^1 (1-\theta)\, D^2 J_{\text{sm}}(-\theta kS)[-kS, -kS]\, d\theta,$$

so that $R = L + \mathcal{R}$. Then

$$\|RS_\perp\|_\infty \leq \|LS_\perp\|_\infty + \|\mathcal{R}S_\perp\|_\infty. \tag{33}$$

We bound the expectations of the two terms in (33) separately. By Lemma 7,

$$\|DJ(0)[h]\|_{\infty\to\infty} \leq \frac{6}{d}\,\|h\|_\infty \qquad \forall h \in \mathbb{R}^d.$$

With $h = -kS$ we obtain

$$\|L\|_{\infty\to\infty} = \|DJ(0)[-kS]\|_{\infty\to\infty} \leq \frac{6}{d}\, k\|S\|_\infty.$$

By Lemma 6,

$$\|S_\perp\|_\infty = \|J_0 S\|_\infty \leq \frac{2}{d}\,\|S\|_\infty.$$

Hence

$$\|LS_\perp\|_\infty \leq \|L\|_{\infty\to\infty}\,\|S_\perp\|_\infty \leq \frac{12}{d^2}\, k\|S\|_\infty^2. \tag{34}$$

Recall that $S = S_{t-1} \sim \mathcal{N}(0, (t-1)I_{d\times d})$. By Lemma 8 with $q = 2$,

$$\mathbb{E}\|S\|_\infty^2 \le C\,(t-1)\,\log(ed) \tag{35}$$

for some absolute constant $C > 0$. Taking expectations in (34) and using (35) yields

$$\mathbb{E}\|LS_\perp\|_\infty \;\le\; C_1\,\frac{k}{d^2}\,(t-1)\,\log(ed) \tag{36}$$

for an absolute constant $C_1 > 0$.

From Lemma 2 (f), for each $\theta \in [0,1]$ and all S,

$$\Big\|D^2J(-\theta kS)[-kS,-kS]\Big\|_{\infty\to\infty} \le \frac{16}{d}\,e^{2\theta k\|S\|_\infty}\,k^2\|S\|_\infty^2. \tag{37}$$

Therefore,

$$\begin{aligned}
\|\mathcal{R}S_\perp\|_\infty &\le \int_0^1 (1-\theta)\,\Big\|D^2J(-\theta kS)[-kS,-kS]\Big\|_{\infty\to\infty}\,\|S_\perp\|_\infty\,d\theta \\
&\le \int_0^1 (1-\theta)\,\frac{16}{d}e^{2\theta k\|S\|_\infty}k^2\|S\|_\infty^2\cdot\frac{2}{d}\|S\|_\infty\,d\theta \\
&= \frac{32}{d^2}k^2\|S\|_\infty^3\int_0^1(1-\theta)e^{2\theta k\|S\|_\infty}\,d\theta.
\end{aligned} \tag{38}$$

For $q > 0$, a direct computation gives

$$\int_0^1 (1-\theta)e^{2\theta kq}\,d\theta \le \frac{1}{2kq}\,e^{2kq}.$$

Applying this in (38) with $q = \|S\|_\infty$ yields

$$\|\mathcal{R}S_\perp\|_\infty \le \frac{16}{d^2}k\,\|S\|_\infty^2 e^{2k\|S\|_\infty}. \tag{39}$$

Now take expectations. By Cauchy–Schwarz,

$$\mathbb{E}\big[\|S\|_\infty^2 e^{2k\|S\|_\infty}\big] \le \Big(\mathbb{E}\|S\|_\infty^4\Big)^{1/2}\Big(\mathbb{E}e^{4k\|S\|_\infty}\Big)^{1/2}. \tag{40}$$

By Lemma 8 with $q = 4$,

$$\mathbb{E}\|S\|_\infty^4 \le C\,(t-1)^2(\log(ed))^2. \tag{41}$$

By the MGF bound for the sup (Lemma 11),

$$\mathbb{E}e^{4k\|S\|_\infty} \le C'\,d\,\exp\big(C''\,k^2(t-1)\big) \tag{42}$$

for absolute constants $C', C'' > 0$. Plugging (41) and (42) into (40), we obtain

$$\mathbb{E}\big[\|S\|_\infty^2 e^{2k\|S\|_\infty}\big] \le C_2\,(t-1)\,\log(ed)\,\sqrt{d}\,\exp\big(C_3k^2(t-1)\big) \tag{43}$$

for some absolute constants $C_2, C_3 > 0$. Combining (39) and (43) gives

$$\mathbb{E}\|\mathcal{R}S_\perp\|_\infty \le C_2\,\frac{k}{d^2}\,(t-1)\log(ed)\,\sqrt{d}\,\exp\big(C_3k^2(t-1)\big) = C_2\,\frac{k}{d^{3/2}}\,(t-1)\log(ed)\,e^{C_3k^2(t-1)}. \tag{44}$$

From (33), (36), and (44),

$$\mathbb{E}\|RS_\perp\|_\infty \le C_1\,\frac{k}{d^2}\,(t-1)\log(ed) + C_2\,\frac{k}{d^{3/2}}\,(t-1)\log(ed)\,e^{C_3k^2(t-1)}. \tag{45}$$

Recalling (31), we obtain

$$|Q_t(k)| \le 2\,\mathbb{E}\|RS_\perp\|_\infty \le C\,k\,(t-1)\log(ed)\Big(\frac{1}{d^2} + \frac{1}{d^{3/2}}e^{C_3k^2(t-1)}\Big) \tag{46}$$

for some absolute $C > 0$.

Now assume $0 < k \le c_0/\sqrt{T}$ for a sufficiently small absolute constant $c_0 > 0$. Then, for all $t \le T$,

$$k^2(t-1) \le k^2 T \le c_0^2, \qquad \Rightarrow \qquad e^{C_3 k^2 (t-1)} \le e^{C_3 c_0^2} =: C_*,$$

where C_* is an absolute constant. Moreover $t - 1 \le T$, and, for all $d \ge 2$,

$$\frac{1}{d^2} \le \frac{1}{d^{3/2}}.$$

Thus, absorbing C_* and the factor 2 into a new constant C_0, (46) yields

$$|Q_t(k)| \le C_0 \frac{kT}{d^{3/2}} \log(ed),$$

uniformly over all $t \in [T]$, all $d \ge 2$, and all $k \in (0, c_0/\sqrt{T}]$. This is exactly (27), which completes the proof. □

Now, we focus on calculating $N_{1,t}$. By Lemma 13, we have $N_{1,t} = \widetilde{\Theta}(\frac{T^2 k}{d})$. more specifically, $N_{1,t} \le -C_2 \frac{T^2 k}{d}$. Therefore, we have

$$-2\Sigma_1 \ge 2C_2 \frac{T^2 k}{d} - C_0 \frac{kT^2}{d^{3/2}} \log(ed) \tag{47}$$

Now, we focus on calculating Σ_2. For $\Sigma_{2,0}$, by Fact 2 and Lemma 10, we have $\Sigma_{2,0} = \widetilde{\Theta}(\frac{T^{1/2}(T-1)\sqrt{\log d}}{d})$. more specifically, $\Sigma_{2,0} \ge -C_4 \frac{T^{1/2}(T-1)\sqrt{\log d}}{d}$ for some $C_4 > 0$.

Proposition 4. *There exist absolute constants $C_1, C_2 > 0$ such that*

$$|\Sigma_{2,\mathrm{pert}}| \le \frac{C_1}{d^2} k T^2 \log(ed) + \frac{C_3}{d^{3/2}} k T^{5/2} \log(ed) e^{4k^2 T}. \tag{48}$$

In particular, with $k = 1/\sqrt{T}$, $|\Sigma_{2,\mathrm{pert}}| \le C\, d^{-7/4} T^{3/2} \operatorname{polylog}(d)$.

Proof. Fix t and write $S := S_{t-1}$. By Lemma 4 and Hölder in $(\infty, 1)$,

$$|\langle R_t S, e_{J^\star} \rangle| \le \|R_t J_0 S\|_\infty.$$

where $R_t = J_t(k) - J_0$ defined in Proposition 3 Decompose R_t into its linear and quadratic parts along the ray:

$$R_t = \underbrace{DJ(\mathbf{0})[-kS]}_{=:L_t} + \underbrace{\int_0^1 (1-\theta)\, D^2 J(-\theta k S)[-kS, -kS]\, d\theta}_{=:\mathcal{R}_t}.$$

Thus

$$\|R_t J_0 S\|_\infty \;\le\; \|L_t J_0 S\|_\infty \;+\; \|\mathcal{R}_t J_0 S\|_\infty.$$

(i) Linear piece. By Lemma 7 and Lemma 6,

$$\|L_t J_0 S\|_\infty \le \|DJ(\mathbf{0})[-kS]\|_{\infty\to\infty} \|J_0 S\|_\infty \le \frac{6}{d} k \|S\|_\infty \cdot \frac{2}{d} \|S\|_\infty = \frac{12}{d^2} k \|S\|_\infty^2.$$

Taking expectations and summing (using Lemma 8 with $q = 2$),

$$\sum_{t=1}^{T} \mathbb{E}\|L_t J_0 S_{t-1}\|_\infty \;\le\; \frac{C_1}{d^2} k T^2 \log(ed). \tag{49}$$

(ii) Quadratic remainder, via Cauchy–Schwarz and MGF. For a fixed t,

$$\begin{aligned}
\|\mathcal{R}_t J_0 S\|_\infty &\le \int_0^1 (1-\theta) \left\| D^2 J(-\theta k S)[-kS, -kS]\, J_0 S \right\|_\infty d\theta \\
&\le \int_0^1 (1-\theta) \|D^2 J(-\theta k S)[-kS, -kS]\|_{\infty\to\infty} \|J_0 S\|_\infty \, d\theta \\
&\underset{(i)}{\le} \int_0^1 (1-\theta) \frac{16}{d} e^{2\theta k \|S\|_\infty} k^2 \|S\|_\infty^2 \|J_0 S\|_\infty d\theta \\
&\underset{(ii)}{\le} \int_0^1 (1-\theta) \frac{32}{d^2} e^{2\theta k \|S\|_\infty} k^2 \|S\|_\infty^3 d\theta \\
&\le \int_0^1 \frac{32}{d^2} e^{2\theta k \|S\|_\infty} k^2 \|S\|_\infty^3 d\theta = \frac{32}{d^2} k^2 \|S\|_\infty^3 \frac{1}{2k\|S\|_\infty} (e^{2k\|S\|_\infty} - 1) \le \frac{16}{d^2} k \|S\|_\infty^2 e^{2k\|S\|_\infty}.
\end{aligned}$$

where (i) is due to Lemma 2-(f) and (ii) is due to Lemma 6. Therefore, we have

$$\begin{aligned}\sum_{t=1}^{T}\mathbb{E}\|\mathcal{R}_t J_0 S_{t-1}\|_\infty &\le \frac{16}{d^2}\,k\sum_{t=1}^{T}\mathbb{E}[\|S_{t-1}\|_\infty^2 e^{2k\|S_{t-1}\|_\infty}]\\ &\le \frac{16}{d^2}\,k\sum_{t=1}^{T}\sqrt{\mathbb{E}[\|S_{t-1}\|_\infty^4]\mathbb{E}[e^{4k\|S_{t-1}\|_\infty}]}\\ &\underset{(iii)}{\le} \frac{C_1}{d^2}\,k\sum_{t=1}^{T}\sqrt{\mathbb{E}[\|S_{t-1}\|_\infty^4]d\exp\left(8k^2(t-1)\right)}\\ &\underset{(iv)}{\le} \frac{C_2}{d^2}\,k\sum_{t=1}^{T}\sqrt{(t-1)^2\log(ed)^2 d\exp\left(8k^2(t-1)\right)} \le \frac{C_3}{d^{3/2}}kT^2\log(ed)e^{4k^2T}\end{aligned}$$

where (iii) is due to Lemma 11 and (iv) is due to Lemma 8 with $q=4$. Therefore, we have

$$|\Sigma_{2,\mathrm{pert}}| \le \frac{C_1}{d^2}\,k\,T^2\log(ed) + \frac{C_3}{d^{3/2}}\,k\,T^{5/2}\log(ed)e^{4k^2T}.$$

□

Therefore, we have

$$2\Sigma_2 \ge -C_4\frac{T^{1/2}(T-1)\sqrt{\log d}}{d} - \frac{C_5}{d^2}\,k\,T^2\log(ed) - \frac{C_6}{d^{3/2}}\,k\,T^{5/2}\log(ed)e^{4k^2T} \tag{50}$$

and combining (47) and (50) with $k=\Theta(1/\sqrt{T}\mathrm{polylog}(d))$, we have $\mathcal{L}'(k)>0$.

□

By Theorem 3 and Theorem 4, we have $\mathcal{L}'(k)<0$ for $k=\widetilde{\mathcal{O}}(1/\sqrt{Td})$ and $\mathcal{L}'(k)>0$ for $k=\Theta(1/\sqrt{T}\mathrm{polylog}(d))$. Therefore, there exists $k^\star$ in the interval $(\widetilde{\Theta}(1/\sqrt{Td}),\widetilde{\Theta}(1/\sqrt{T}))$ such that $\mathcal{L}'(k^\star)=0$. □

B Deferred Lemmas

We provide the proofs of lemmas related to Gaussian distributions and the $\mathrm{Softmax}$ function that are used in the main text for the sake of completeness.

Lemma 1 (Permutation commutant)**.** *Let $A\in\mathbb{R}^{d\times d}$ satisfy*

$$PAP^\intercal = A \qquad \textit{for every permutation matrix } P.$$

Then there exist scalars $\alpha,\beta\in\mathbb{R}$ such that

$$A=\alpha I_{d\times d}+\beta\mathbf{1}_d\mathbf{1}_d^\intercal.$$

If, in addition, $\mathbf{1}_d^\intercal A=\mathbf{0}_d^\intercal$ or $A\mathbf{1}_d=\mathbf{0}_d$, then

$$A=\alpha\left(I_{d\times d}-\frac{1}{d}\mathbf{1}_d\mathbf{1}_d^\intercal\right).$$

Consequently,

$$\langle A,-I_{d\times d}\rangle_F=-\alpha(d-1)$$

in this zero-sum case.

Proof. Fix A satisfying $PAP^\intercal=A$ for all permutation matrices P. For any two coordinates i,j, a transposition exchanging them shows $A_{ii}=A_{jj}$, so all diagonal entries are equal. For any two ordered pairs (i,j) and (r,s) with $i\neq j$ and $r\neq s$, there is a permutation sending i to r and j to s, so $A_{ij}=A_{rs}$. Thus all off-diagonal entries are equal, which is equivalent to $A=\alpha I_{d\times d}+\beta\mathbf{1}_d\mathbf{1}_d^\intercal$ after reparameterizing the common diagonal and off-diagonal values.

If $\mathbf{1}_d^\intercal A=\mathbf{0}_d^\intercal$, then $\mathbf{1}_d^\intercal A=(\alpha+d\beta)\mathbf{1}_d^\intercal$, so $\beta=-\alpha/d$. The same conclusion follows from $A\mathbf{1}_d=\mathbf{0}_d$. Finally, $\langle I_{d\times d}-d^{-1}\mathbf{1}_d\mathbf{1}_d^\intercal, I_{d\times d}\rangle_F=d-1$, giving the displayed Frobenius identity. □

Lemma 2 (Softmax Jacobian: spectral, ∞, and 1 bounds)**.** *For $z \in \mathbb{R}^d$, let*

$$p = \mathrm{Softmax}(z), \qquad J_{\mathrm{sm}}(z) = \mathit{Diag}(p) - pp^\intercal.$$

Then:

(a) *$J_{\mathrm{sm}}(z)$ is symmetric positive semidefinite and $J_{\mathrm{sm}}(z)\mathbf{1}_d = 0$.*

(b) *Operator, $\infty \to \infty$, and $1 \to 1$ bounds:*

$$\|J_{\mathrm{sm}}(z)\|_{\mathit{op}} = \|J_{\mathrm{sm}}(z)\|_{\infty\to\infty} = \|J_{\mathrm{sm}}(z)\|_{1\to 1} \le \frac{1}{2}.$$

(c) *ℓ_2–Lipschitz continuity of* Softmax*:*

$$\|\,\mathrm{Softmax}(z) - \mathrm{Softmax}(z')\|_2 \;\le\; \frac{1}{2}\|z - z'\|_2.$$

(d) *ℓ_2–Lipschitz continuity of the Jacobian:*

$$\|J_{\mathrm{sm}}(z) - J_{\mathrm{sm}}(z')\|_{\mathit{op}} \;\le\; \frac{3}{2}\,\|z - z'\|_2.$$

(e) *Derivative bounds in induced norms:*

$$\|DJ_{\mathrm{sm}}(z)[h]\|_{\infty\to\infty} \le 2\,\|h\|_\infty, \qquad \|DJ_{\mathrm{sm}}(z)[h]\|_{1\to 1} \le 2\|h\|_\infty,$$

and therefore $\|J_{\mathit{sm}}(\theta z) - J_{\mathit{sm}}(\mathbf{0}_d)\|_{\infty\to\infty} \le 2\theta\,\|z\|_\infty$.

(f) *$\infty \to \infty$ bound of the second derivative of J_{sm}:*

$$\|D^2 J_{\mathit{sm}}(z)[h,h]\|_{\infty\to\infty} \;\le\; \frac{16}{d}\,e^{2\|z\|_\infty}\,\|h\|_\infty^2 \tag{51}$$

for some absolute constant $C > 0$.

Proof. **(a) PSD and nullspace property.** For any $x \in \mathbb{R}^d$,

$$x^\intercal J_{\mathrm{sm}}(z)x = \sum_i p_i x_i^2 - \Big(\sum_i p_i x_i\Big)^2 = \mathrm{Var}_p(x) \ge 0.$$

Also,

$$J_{\mathrm{sm}}(z)\mathbf{1}_d = \mathrm{Diag}(p)\mathbf{1}_d - p(p^\intercal \mathbf{1}_d) = p - p(1) = 0.$$

(b) Spectral-norm bound. Since $J_{\mathrm{sm}}(z)$ is symmetric,

$$\|J_{\mathrm{sm}}(z)\|_{\mathrm{op}} = \sup_{\|x\|_2=1} \mathrm{Var}_p(x).$$

For any x with $\|x\|_2 = 1$ and $m = \min x_i$, $M = \max x_i$,

$$\mathrm{Var}_p(x) \le \frac{(M-m)^2}{4}.$$

But

$$(M-m)^2 = (x_i - x_j)^2 \le 2(x_i^2 + x_j^2) \le 2\|x\|_2^2 = 2.$$

Thus

$$\mathrm{Var}_p(x) \le \frac{1}{2}.$$

Moreover, for each row i,

$$\sum_j |(J_{\mathrm{sm}}(z))_{ij}| = p_i(1-p_i) + p_i \sum_{j\ne i} p_j = 2p_i(1-p_i) \le \frac{1}{2},$$

since $x(1-x) \le 1/4$; the same bound holds for each column by symmetry. Thus the $\infty \to \infty$ and $1 \to 1$ operator norms also equal $\|J_{\mathrm{sm}}(z)\|_{\mathrm{op}}$.

(c) $\mathrm{Softmax}$ **is** $1/2$**-Lipschitz in** ℓ_2**.** For $\gamma(\theta) = (1-\theta)z + \theta z'$,

$$\mathrm{Softmax}(z') - \mathrm{Softmax}(z) = \int_0^1 J_{\mathrm{sm}}(\gamma(\theta))(z'-z)\, d\theta.$$

Take norms and use (b):

$$\|\,\mathrm{Softmax}(z') - \mathrm{Softmax}(z)\|_2 \le \frac{1}{2}\|z'-z\|_2.$$

(d) Jacobian Lipschitz bound. Differentiate

$$J_{\mathrm{sm}}(z) = \mathrm{Diag}(p(z)) - p(z)p(z)^\intercal.$$

For $q := Dp(z)[h] = J_{\mathrm{sm}}(z)h$,

$$DJ_{\mathrm{sm}}(z)[h] = \mathrm{Diag}(q) - (qp^\intercal + pq^\intercal).$$

Operator norms satisfy

$$\|\mathrm{Diag}(q)\|_{\mathrm{op}} \le \|q\|_\infty \le \|q\|_2, \qquad \|qp^\intercal\|_{\mathrm{op}} = \|q\|_2\|p\|_2 \le \|q\|_2,$$

and likewise for $pq^\intercal$. Hence

$$\|DJ_{\mathrm{sm}}(z)[h]\|_{\mathrm{op}} \le 3\|q\|_2.$$

Using $\|q\|_2 = \|J_{\mathrm{sm}}(z)h\|_2 \le \frac{1}{2}\|h\|_2$ from (b),

$$\|DJ_{\mathrm{sm}}(z)[h]\|_{\mathrm{op}} \le \frac{3}{2}\|h\|_2.$$

Integrate along the segment from z to z' to conclude

$$\|J_{\mathrm{sm}}(z) - J_{\mathrm{sm}}(z')\|_{\mathrm{op}} \le \frac{3}{2}\|z-z'\|_2.$$

(e) Bounds for $DJ_{\mathbf{sm}}(z)[h]$ **in** ∞**- and** 1**-norms.** From above,

$$DJ_{\mathrm{sm}}(z)[h] = \mathrm{Diag}(q) - (qp^\intercal + pq^\intercal). \tag{52}$$

Step 1: generic bound. Using induced–norm identities,

$$\|DJ_{\mathrm{sm}}(z)[h]\|_{\alpha\to\alpha} \le 2\|q\|_\infty + \|q\|_1, \qquad \alpha \in \{1,\infty\}.$$

Step 2: bounds on q**.** From (b):

$$\|q\|_\infty = \|J_{\mathrm{sm}}(z)h\|_\infty \le \frac{1}{2}\|h\|_\infty$$

We also need the dimension-free inequality

$$\|q\|_1 = \sum_i p_i|h_i - \langle p, h\rangle| \le \|h\|_\infty. \tag{$\star$}$$

This follows by thinking p as a probability distribution and define X as a discrete random variable with taking h_i with probability p_i, then $\mathbb{E}[|X - \mathbb{E}[X]|] \le \sqrt{\mathbb{E}[|X-\mathbb{E}[X]|^2]} = \sqrt{\mathrm{Var}(X)} \le \|h\|_\infty$.

Step 3: final estimates.

For the $\infty \to \infty$ norm:

$$\|DJ_{\mathrm{sm}}(z)[h]\|_{\infty\to\infty} \le 2\cdot\frac{1}{2}\|h\|_\infty + \|h\|_\infty = 2\|h\|_\infty.$$

For the $1 \to 1$ norm:

$$\|DJ_{\mathrm{sm}}(z)[h]\|_{1\to 1} \le 2\|h\|_\infty.$$

This completes the proof.

(f) $\infty \to \infty$ bound of the second derivative of J_{sm}. First, we differentiate (52) in direction h. Recall that for $p(z) = \mathrm{Softmax}(z)$,

$$J_{\text{sm}}(z) \;=\; \text{Diag}(p) - pp^\intercal,$$

and for a fixed direction h we write

$$q := Dp(z)[h] = J_{\text{sm}}(z)h, \qquad a := D^2p(z)[h,h].$$

From (52) we have

$$DJ_{\text{sm}}(z)[h] = \text{Diag}(q) - qp^\intercal - pq^\intercal.$$

Differentiating once more in the direction h gives

$$\begin{aligned} D^2J_{\text{sm}}(z)[h,h] &= D(\text{Diag}(q))[h] - D(qp^\intercal)[h] - D(pq^\intercal)[h] \\ &= \text{Diag}(Dq[h]) - \big((Dq[h])p^\intercal + q(Dp[h])^\intercal\big) - \big(p(Dq[h])^\intercal + (Dp[h])q^\intercal\big). \end{aligned}$$

Since $Dp[h] = q$ and $Dq[h] = D^2p(z)[h,h] = a$, this becomes

$$\begin{aligned} D^2J_{\text{sm}}(z)[h,h] &= \text{Diag}(a) - \big(ap^\intercal + qq^\intercal\big) - \big(pa^\intercal + qq^\intercal\big) \\ &= \text{Diag}(a) - ap^\intercal - pa^\intercal - 2qq^\intercal. \end{aligned}$$

Thus

$$D^2J_{\text{sm}}(z)[h,h] = \text{Diag}(a) - ap^\intercal - pa^\intercal - 2qq^\intercal, \quad p := p(z), \quad q := J_{\text{sm}}(z)h, \quad a := D^2p(z)[h,h].$$

By standard calculation, we can derive that $a_i = p_i(z)((h_i - \mu)^2 - \sigma^2)$, where $\mu = \sum_{j=1}^d p_j(z)h_j$ and $\sigma^2 = \sum_{j=1}^d p_j(z)(h_j - \mu)^2$. Due to the definition of $\|h\|_\infty$, we have $|h_i - \mu| \le 2\|h\|_\infty$ and $|\sigma^2| \le 4\|h\|_\infty^2$. Thus

$$|a_i| \le p_i(z)(4\|h\|_\infty^2 + 4\|h\|_\infty^2) = 8p_i(z)\|h\|_\infty^2.$$

Therefore,

$$\|\text{Diag}(a)\|_{\infty\to\infty} \le 8\|p\|_\infty \|h\|_\infty^2.$$

Similarly, we have $\|ap^\intercal\|_{\infty\to\infty} \le 8\|p\|_\infty\|h\|_\infty^2$ and $\|pa^\intercal\|_{\infty\to\infty} \le 8\|p\|_\infty\|h\|_\infty^2$.

Now, we bound $\|qq^\intercal\|_{\infty\to\infty}$ by bounding $\|q\|_\infty$ and $\|q\|_1$, since $\|qq^\intercal\|_{\infty\to\infty} \le \|q\|_\infty\|q\|_1$. We bound $\|q\|_\infty$ by using the definition of q:

$$q = J_{\text{sm}}(z)h = \text{Diag}(p(z))h - p(z)(p(z)^\intercal h)$$

Therefore we have the followings:

$$|q_i| \le p_i(z)|h_i| + p_i(z)\sum_{j=1}^d p_j(z)|h_j| \le p_i(z)\|h\|_\infty + p_i(z)\|h\|_\infty \le 2\|p(z)\|_\infty\|h\|_\infty$$

$$\|q\|_1 \le \|\text{Diag}(p(z))h\|_1 + \|p(z)(p(z)^\intercal h)\|_1 \le \|h\|_\infty + \|h\|_\infty \le 2\|h\|_\infty$$

Thus, we have $\|q\|_\infty \le 2\|p(z)\|_\infty\|h\|_\infty$ and $\|q\|_1 \le 2\|h\|_\infty$, therefore $\|qq^\intercal\|_{\infty\to\infty} \le 4\|p(z)\|_\infty\|h\|_\infty^2$.

Collecting all the bounds, we have

$$\|D^2J_{\text{sm}}(z)[h,h]\|_{\infty\to\infty} \le 32\|p(z)\|_\infty\|h\|_\infty^2.$$

Now, we have

$$p_i(z) = p_i(z - \bar{z}) = \frac{e^{z_i - \bar{z}}}{\sum_{j=1}^d e^{z_j - \bar{z}}}$$

where $\bar{z} = \frac{1}{d}\sum_{j=1}^d z_j$. We know that $\frac{1}{d}\sum_{j=1}^d e^{z_j - \bar{z}} \ge 1$ due to Jensen's inequality, therefore $\|p(z)\|_\infty \le \frac{1}{d}e^{2\|z\|_\infty}$, which concludes the proof.

□

Lemma 3. *Define $J_0 = Diag(u) - uu^\intercal$ with $u = \frac{1}{d}\mathbf{1}_d$. For every $x \in \mathbb{R}^d$,*

$$\|J_0 x\|_\infty \;\le\; \frac{1}{d}\Big(\|x\|_\infty + |\mathbf{1}_d^\intercal x|\Big)$$

Proof. Since $J_0 = \frac{1}{d} I_{d\times d} - \frac{1}{d^2}\mathbf{1}_d\mathbf{1}_d^\intercal$, we have $(J_0 x)_i = \frac{1}{d}\big(x_i - (\mathbf{1}_d^\intercal x)/d\big)$. Hence $\|J_0 x\|_\infty = \frac{1}{d}\max_i |x_i - \bar{x}| \le \frac{1}{d}\big(\|x\|_\infty + |\bar{x}|\big)$. □

Lemma 4 (Null-direction reduction)**.** *For every $z \in \mathbb{R}^d$ and every $x \in \mathbb{R}^d$,*

$$\big(J_{sm}(z) - J_0\big)\, x \;=\; \big(J_{sm}(z) - J_0\big)\, J_0 x.$$

Equivalently, with $R_t := J_t(k) - J_0$ and $A_\theta := J_{sm}(\theta z) - J_0$ (for any z),

$$R_t x \;=\; R_t J_0 x, \qquad A_\theta x \;=\; A_\theta J_0 x.$$

Proof. By Lemma 2-(a), $J_{\text{sm}}(z)\mathbf{1}_d = 0$ for all z, and also $J_0\mathbf{1}_d = 0$. Hence $\big(J_{\text{sm}}(z) - J_0\big)\mathbf{1}_d = 0$. Write $x = J_0 x + \alpha\mathbf{1}_d$, where $\alpha := \frac{1}{d}\mathbf{1}_d^\intercal x$. Then

$$\big(J_{\text{sm}}(z) - J_0\big)\, x = \big(J_{\text{sm}}(z) - J_0\big)\, J_0 x + \alpha\,\big(J_{\text{sm}}(z) - J_0\big)\mathbf{1}_d = \big(J_{\text{sm}}(z) - J_0\big)\, J_0 x. \qquad \square$$

Lemma 5. *Let $x \in \mathbb{R}^d$ and $\alpha \ge 0$. Set $\pi := \mathrm{Softmax}(-\alpha x)$ and $p := \mathrm{diag}(\pi) - \pi\pi^\intercal$. Then $\langle px, \pi\rangle \le 0$.*

Proof. Sort indices so that $x_1 \le x_2 \le \cdots \le x_d$. Then $\pi_1 \ge \pi_2 \ge \cdots \ge \pi_d$ because $\pi_i \propto e^{-\alpha x_i}$. Write $\mu_\pi := \sum_i \pi_i x_i$ and note

$$\langle px, \pi\rangle = \sum_{i=1}^d \pi_i^2(x_i - \mu_\pi) = \Big(\sum_{i=1}^d \pi_i^2\Big)\left(\underbrace{\frac{\sum_i \pi_i^2 x_i}{\sum_i \pi_i^2}}_{=:\,\mu_{\pi^2}} - \mu_\pi\right).$$

Let $q := \pi^2/\sum_j \pi_j^2$ be π squared and renormalized; then q is *more concentrated* than π, i.e. $\pi \succ q$ in the majorization order (Hardy–Littlewood–Pólya). Since (x_i) is increasing, majorization implies

$$\mu_\pi = \sum_i \pi_i x_i \;\ge\; \sum_i q_i x_i \;=\; \mu_{\pi^2} \quad\Longrightarrow\quad \mu_{\pi^2} - \mu_\pi \;\le\; 0,$$

hence $\langle px, \pi\rangle \le 0$. □

Lemma 6. *Let $J_0 = Diag(u) - uu^\intercal$ with $u = \frac{1}{d}\mathbf{1}_d$. Then*

$$\|J_0\|_{\infty\to\infty} = \|J_0\|_{1\to 1} \;=\; \frac{2(d-1)}{d^2} \;\le\; \frac{2}{d}, \qquad \|J_0\|_{op} = \frac{1}{d}$$

Proof. The ith row of J_0 has diagonal entry $(d-1)/d^2$ and off-diagonal entries $-1/d^2$. Thus row i has ℓ_1-norm

$$\frac{d-1}{d^2} + (d-1)\frac{1}{d^2} = \frac{2(d-1)}{d^2}.$$

The same holds for column sums, yielding the claim. Lastly, the operator norm is the largest singular value, which is the largest eigenvalue of $J_0 J_0^\intercal = \frac{1}{d^2}\mathbf{1}_d\mathbf{1}_d^\intercal$, so $\|J_0\|_{\text{op}} = \frac{1}{d}$. □

Lemma 7. *For all h, $\|DJ_{sm}(\mathbf{0})[h]\|_{\infty\to\infty} \le \frac{6}{d}\|h\|_\infty$.*

Proof. For all h,

$$DJ_{\text{sm}}(\mathbf{0})[h] = \mathrm{Diag}(J_0 h) - (J_0 h)u^\intercal - u(J_0 h)^\intercal, \quad u = \tfrac{1}{d}\mathbf{1}_d,$$

Using Lemma 6, we have $\|J_0 h\|_\infty \le \frac{2}{d}\|h\|_\infty$, so

$$\|DJ_{\text{sm}}(\mathbf{0})[h]\|_{\infty\to\infty} \le \frac{6}{d}\|h\|_\infty.$$

□

Lemma 8 (Gaussian supremum moments)**.** *Let $Z \sim \mathcal{N}(\mathbf{0}_d, I_{d\times d})$ and $q \geq 1$. There exists $C_q > 0$, independent of d, such that*

$$\mathbb{E}\|Z\|_\infty^q \;\leq\; C_q\,(\log(ed))^{q/2}.$$

Proof. By a union bound, $\mathbb{P}(\|Z\|_\infty \geq t) \leq 2d\,e^{-t^2/2}$. Integrating by parts gives $\mathbb{E}\|Z\|_\infty^q \leq C_q(\log(ed))^{q/2}$ for some C_q. □

Lemma 9. *There exists an absolute constant $c_{\mathrm{ev}} \in (0,1)$ such that*

$$\mathbb{E}\Big[\min_{1\leq j\leq d} G_j\Big] \;\leq\; -\,c_{\mathrm{ev}}\sqrt{\log d}, \qquad d \geq 2,$$

for i.i.d. $G_j \sim \mathcal{N}(0,1)$.

Proof. We define $M_d := \max_{1\leq j\leq d} G_j$ and $X_d := \min_{1\leq j\leq d} G_j$. Since the G_j are i.i.d. standard normals, $\mathbb{P}(M_d \leq a) = \Phi(a)^d$, where Φ denotes the standard normal cdf.

We use the classical Mills ratio inequality [Gordon, 1941]: for all $x > 0$,

$$1 - \Phi(x) \;\geq\; \frac{x}{1+x^2}\,\varphi(x), \tag{53}$$

where $\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is the standard normal density. Set $a := \sqrt{\log d}$. Then

$$\varphi(a) = \frac{1}{\sqrt{2\pi}} d^{-1/2}, \qquad \frac{a}{1+a^2} = \frac{\sqrt{\log d}}{1+\log d} \;\geq\; \frac{c_0}{\sqrt{\log d}}$$

for some universal constant $c_0 > 0$ and all $d \geq 2$. Using (53),

$$1 - \Phi(a) \;\geq\; c_0\,\frac{d^{-1/2}}{\sqrt{\log d}}.$$

Thus

$$\mathbb{P}(M_d \leq a) = \Phi(a)^d = \big(1 - (1-\Phi(a))\big)^d \leq \left(1 - c_0\,\frac{d^{-1/2}}{\sqrt{\log d}}\right)^d \leq \exp\left(-c_0\,\frac{d^{1/2}}{\sqrt{\log d}}\right).$$

Therefore,

$$\mathbb{P}(X_d \leq -a) = 1 - \mathbb{P}(M_d \leq a) \geq 1 - \exp\left(-c_0\,\frac{d^{1/2}}{\sqrt{\log d}}\right).$$

For all sufficiently large d, the right-hand side is at least $1/2$, so for such d,

$$\mathbb{E}[X_d] \;\leq\; -a\,\mathbb{P}(X_d \leq -a) \;\leq\; -\frac{1}{2}\sqrt{\log d}.$$

Since $\mathbb{E}[X_d] \leq 0$ for each fixed d and the ratio $-\mathbb{E}[X_d]/\sqrt{\log d}$ is bounded away from 0 on the finite range $2 \leq d \leq d_0$, we obtain a universal constant $c_{\mathrm{ev}} \in (0,1)$ such that

$$\mathbb{E}[X_d] \leq -c_{\mathrm{ev}}\sqrt{\log d} \qquad \text{for all } d \geq 2.$$

□

Lemma 10. *There exists an absolute constant $C_{\mathrm{ev}} > 0$ such that*

$$\mathbb{E}\Big[\min_{1\leq j\leq d} G_j\Big] \;\geq\; -\,C_{\mathrm{ev}}\sqrt{\log d}, \qquad d \geq 2,$$

for i.i.d. $G_j \sim \mathcal{N}(0,1)$.

Proof. It suffices to upper bound $\mathbb{E}[M_d]$ by $C\sqrt{\log d}$ for some absolute constant $C > 0$.

For any $t > 0$,

$$\mathbb{P}(M_d \geq t) = \mathbb{P}\Big(\bigcup_{j=1}^d \{G_j \geq t\}\Big) \;\leq\; \sum_{j=1}^d \mathbb{P}(G_j \geq t) = d\,\mathbb{P}(G_1 \geq t) \leq d e^{-t^2/2}.$$

Next, decompose M_d into its positive and negative parts:

$$M_d^+ := \max\{M_d, 0\}, \qquad M_d^- := \max\{-M_d, 0\},$$

so that $M_d = M_d^+ - M_d^-$ and hence

$$\mathbb{E}[M_d] = \mathbb{E}[M_d^+] - \mathbb{E}[M_d^-] \;\le\; \mathbb{E}[M_d^+].$$

Since $M_d^+ \ge 0$, we can use the standard tail integral representation for nonnegative random variables:

$$\mathbb{E}[M_d^+] = \int_0^\infty \mathbb{P}(M_d^+ \ge t)\, dt = \int_0^\infty \mathbb{P}(M_d \ge t)\, dt.$$

Therefore,

$$\mathbb{E}[M_d] \;\le\; \mathbb{E}[M_d^+] = \int_0^\infty \mathbb{P}(M_d \ge t)\, dt.$$

Fix $x := \sqrt{2\log d}$ and split the integral:

$$\mathbb{E}[M_d] \le \int_0^x 1\, dt \;+\; \int_x^\infty \mathbb{P}(M_d \ge t)\, dt \;\le\; x \;+\; d\int_x^\infty e^{-t^2/2}\, dt.$$

We bound the Gaussian tail using

$$\int_x^\infty e^{-t^2/2}\, dt \;\le\; \frac{1}{x} e^{-x^2/2}, \qquad x > 0,$$

which follows from integrating by parts or from a standard inequality for Gaussian tails. Plugging $x = \sqrt{2\log d}$, we get

$$d\int_x^\infty e^{-t^2/2}\, dt \;\le\; d\cdot\frac{1}{x}e^{-x^2/2} = d\cdot\frac{1}{\sqrt{2\log d}}\cdot e^{-\log d} = \frac{1}{\sqrt{2\log d}} \;\le\; 1,$$

for all $d \ge 2$. Therefore,

$$\mathbb{E}[M_d] \;\le\; x + 1 \;=\; \sqrt{2\log d} + 1 \;\le\; (\sqrt{2}+1)\sqrt{\log d} \;\le\; 3\sqrt{\log d}, \qquad d \ge 2.$$

□

Lemma 11 (MGF for Gaussian suprema)**.** *Let $G \sim \mathcal{N}(\mathbf{0}_d, tI_{d\times d})$. For any $a \ge 0$,*

$$\mathbb{E}\exp\big(a\|G\|_\infty\big) \;\le\; 2d\,\exp\Big(\frac{a^2 t}{2}\Big).$$

Proof. By the union bound, $\exp(a\|G\|_\infty) = \max_i \exp(a|G_i|) \le \sum_i \exp(a|G_i|)$, so $\mathbb{E}e^{a\|G\|_\infty} \le d\,\mathbb{E}e^{a|G_1|}$. For $G_i \sim \mathcal{N}(0,t)$, $\mathbb{E}e^{a|G_i|} = 2\int_0^\infty e^{ax}\frac{1}{\sqrt{2\pi t}}e^{-x^2/(2t)}dx \le 2e^{a^2t/2}$. □

Lemma 12 (Stein's lemma for Gaussian vectors)**.** *[Stein, 1981] Let $X \sim \mathcal{N}(\mathbf{0}_d, I_{d\times d})$ and let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable with $\mathbb{E}\|\nabla f(X)\|_2 < \infty$ and $\mathbb{E}|X_i f(X)| < \infty$ for all i. Then for every $v \in \mathbb{R}^d$,*

$$\mathbb{E}[\,\langle X, v\rangle\, f(X)\,] \;=\; \mathbb{E}[\,\langle v, \nabla f(X)\rangle\,].$$

Equivalently, coordinatewise:

$$\mathbb{E}[\,X_i f(X)\,] = \mathbb{E}[\,\partial_i f(X)\,], \qquad i = 1, \dots, d.$$

Proof. We give the proof in one dimension; the multivariate case follows by linearity.

Let $Z \sim \mathcal{N}(0,1)$ with density $\varphi(z) = \frac{1}{\sqrt{2\pi}}e^{-z^2/2}$. Assume $\mathbb{E}|Zf(Z)| < \infty$ and $\mathbb{E}|f'(Z)| < \infty$. Then

$$\mathbb{E}[Zf(Z)] = \int_{-\infty}^\infty z f(z)\,\varphi(z)\, dz.$$

Since $\varphi'(z) = -z\varphi(z)$, we may write

$$z\varphi(z) = -\varphi'(z),$$

so that

$$\mathbb{E}[Zf(Z)] = -\int f(z)\,\varphi'(z)\,dz.$$

Integrating by parts,

$$-\int f(z)\,\varphi'(z)\,dz = -\,[f(z)\varphi(z)]_{-\infty}^{\infty} + \int f'(z)\varphi(z)\,dz.$$

Because $f(z)\varphi(z) \to 0$ as $|z| \to \infty$ (Gaussian density dominates any polynomially growing f), the boundary term vanishes, and we obtain

$$\mathbb{E}[Zf(Z)] = \int f'(z)\varphi(z)\,dz = \mathbb{E}[f'(Z)].$$

For the d-dimensional case, apply the one-dimensional identity to each coordinate, conditioning on the remaining coordinates. Linearity of expectation and $\langle X, v\rangle = \sum_i v_i X_i$ then give

$$\mathbb{E}[\langle X, v\rangle f(X)] = \sum_{i=1}^{d} v_i\,\mathbb{E}[X_i f(X)] = \sum_{i=1}^{d} v_i\,\mathbb{E}[\partial_i f(X)] = \mathbb{E}[\langle v, \nabla f(X)\rangle].$$

□

Lemma 13. *Let $a > 0, d \geq 2$ and $X \sim \mathcal{N}(0, aI_{d\times d})$, and $\pi(k) = \mathrm{Softmax}(-kX)$. Define $F_d(a,k) := \mathbb{E}\left\langle J_0 X,\, \pi(k)\right\rangle$. If $k = \widetilde{\mathcal{O}}\left(1/\sqrt{a}\right)$, $F_d(a,k) = -\,\Theta\big(ak/d\big)$.*

Proof. First, by the definition of J_0 and Lemma 12,

$$F_d(a,k) = \frac{1}{d}\mathbb{E}\langle X, \pi(k)\rangle = \frac{1}{d}\sum_{i\in[d]}\mathbb{E}[X_i\pi_i(k)]$$

$$= \frac{1}{d}\sum_{i\in[d]}\mathbb{E}\left[\frac{\partial}{\partial X_i}\pi_i(k)\right] = -\frac{ak}{d}\sum_{i\in[d]}\mathbb{E}[\pi_i(k)(1-\pi_i(k))] = -\frac{ak}{d}\mathbb{E}[1-\|\pi(k)\|_2^2]$$

From the calculation above,

$$F_d(a,k) \;=\; -\,\frac{ak}{d}\,\mathbb{E}\big[\,1-\|\pi(k)\|_2^2\,\big], \qquad \pi(k) = \mathrm{Softmax}(-kX). \tag{54}$$

Write $X = \sqrt{a}\,Z$ with $Z \sim \mathcal{N}(\mathbf{0}_d, I_{d\times d})$ and set $\kappa := k\sqrt{a}$. Then

$$\pi_i(k) \;=\; \frac{e^{-\kappa Z_i}}{\sum_{j=1}^{d} e^{-\kappa Z_j}}, \qquad \|\pi(k)\|_2^2 \;=\; \frac{\sum_{i=1}^{d} e^{-2\kappa Z_i}}{\Big(\sum_{j=1}^{d} e^{-\kappa Z_j}\Big)^2}.$$

Let $Y_i := e^{-\kappa Z_i}$ so that $Y_i > 0$ are i.i.d. sub-exponential with $\mu_1 := \mathbb{E}Y_i = e^{\kappa^2/2}$ and $\mu_2 := \mathbb{E}Y_i^2 = e^{2\kappa^2}$. Define $S_1 = \sum_{j=1}^{d} Y_j$ and $S_2 = \sum_{i=1}^{d} Y_i^2$. By Bernstein/Chernoff bounds for sub-exponential sums, there exist absolute constants $c_0, C_0 > 0$ (depending only on an a priori bound $\kappa \leq \kappa_\star$) such that with probability at least $1 - 2e^{-c_0 d}$,

$$\tfrac{1}{2}d\mu_1 \;\leq\; S_1 \;\leq\; 2d\mu_1 \quad\text{and}\quad \tfrac{1}{2}d\mu_2 \;\leq\; S_2 \;\leq\; 2d\mu_2.$$

On this event,

$$\frac{1}{8}\cdot\frac{\mu_2}{\mu_1^2}\cdot\frac{1}{d} \;\leq\; \frac{S_2}{S_1^2} \;=\; \|\pi(k)\|_2^2 \;\leq\; 8\cdot\frac{\mu_2}{\mu_1^2}\cdot\frac{1}{d} \;=\; 8\,\frac{e^{\kappa^2}}{d}.$$

Taking expectations and absorbing the exponentially small tail yields constants $0 < c_1 \leq C_1 < \infty$ (depending only on $\kappa_\star$) for which

$$\frac{c_1}{d} \;\leq\; \mathbb{E}\|\pi(k)\|_2^2 \;\leq\; \frac{C_1}{d} \qquad \text{for all } d \text{ large enough.} \tag{55}$$

Combining (55) with (54), we obtain

$$-\,\frac{ak}{d}\Big(1-\frac{c_1}{d}\Big) \;\leq\; F_d(a,k) \;\leq\; -\,\frac{ak}{d}\Big(1-\frac{C_1}{d}\Big).$$

In particular, for all $d \geq 2C_1$,

$$-\,\frac{ak}{d} \;\leq\; F_d(a,k) \;\leq\; -\,\frac{1}{2}\,\frac{ak}{d}.$$

Renaming absolute constants completes the proof. □

C Proof of Fact 1

Proof. By definition of external regret for $\mathcal{R}_i$, we have (7) for each $i \in [d]$.

Fix any $\widehat{P} = (\widehat{p}_1 \mid \cdots \mid \widehat{p}_d) \in \Phi$. The cumulative swap regret with respect to this transformation is

$$\begin{aligned}\sum_{t=1}^{T}\big(\langle \ell_t, \pi_t\rangle - \langle \ell_t, \widehat{P}\pi_t\rangle\big) &\underset{(i)}{=} \sum_{t=1}^{T}\big(\langle \ell_t, P_t\pi_t\rangle - \langle \ell_t, \widehat{P}\pi_t\rangle\big) = \sum_{t=1}^{T}\left[\sum_{i=1}^{d}\pi_{t,i}\langle \ell_t, p_{i,t}\rangle - \sum_{i=1}^{d}\pi_{t,i}\langle \ell_t, \widehat{p}_i\rangle\right] \\ &= \sum_{t=1}^{T}\sum_{i=1}^{d}\big(\langle x_{i,t}, p_{i,t}\rangle - \langle x_{i,t}, \widehat{p}_i\rangle\big) = \sum_{i=1}^{d}\sum_{t=1}^{T}\big(\langle x_{i,t}, p_{i,t}\rangle - \langle x_{i,t}, \widehat{p}_i\rangle\big) \\ &\underset{(ii)}{\leq} \sum_{i=1}^{d}\mathrm{Regret}_i((x_{i,t})_{t\in[T]}).\end{aligned} \tag{56}$$

Here (i) follows from the definition of π_t and the fact that $\pi_t = P_t\pi_t$, and (ii) follows from the definition of Regret_i. Since the identity holds for every $\widehat{P} \in \Phi$, taking the maximum over $\widehat{P}$ gives (8), and then

$$\text{Swap-Regret}((\ell_t)_{t\in[T]}) \leq \sum_{i=1}^{d}\mathrm{Regret}_i((x_{i,t})_{t\in[T]}),$$

which completes the proof. □

D Additional Details for the Swap-Regret Architecture

Remark 2 (Differentiability of the fixed-point layer)**.** *Because every column $p_{j,t+1}$ is produced by a* Softmax*, the matrix $P_{t+1} = (p_{1,t+1} \mid \cdots \mid p_{d,t+1})$ has strictly positive entries and is column-stochastic. By Perron–Frobenius theory, P_{t+1} is primitive and has a unique stationary distribution in the relative interior of the simplex. Moreover, the eigenvalue 1 is simple, so the fixed-point map $P \mapsto$ Fixed-Point(P) is smooth on the open set of strictly positive column-stochastic matrices; equivalently, differentiability follows from the implicit function theorem after restricting to the simplex tangent space $\{u : \mathbf{1}_d^\intercal u = 0\}$. In implementation, gradients can be obtained by backpropagating through the finite power-iteration procedure, or by implicit differentiation of the stationary equation $\pi = P\pi$.*

Remark 3 (Fixed-Point iteration for the swap-regret minimizer)**.** *A natural question is why the architecture requires a* Fixed-Point *operation. The rationale is twofold. First, the computational complexity of fixed-point iteration is the same as that of classical no-swap-regret minimization [Hazan and Kale, 2007], making it algorithmically consistent with known constructions. Second, most swap-regret and Φ-regret algorithms explicitly rely on a fixed-point computation [Greenwald and Jafari, 2003]: a common approach is to minimize external regret in the Φ-space and then recover the final policy via a fixed point of the induced transition operator. While alternative methods exist—such as the Markov-chain-tree-theorem-based approach of Anagnostides et al. [2023], which replaces the fixed point with a linear map—that method requires d^{d-1} parameters and is therefore exponentially large and unsuitable for transformer-based architectures. Thus, the* Fixed-Point *operation offers both theoretical alignment and practical scalability.*

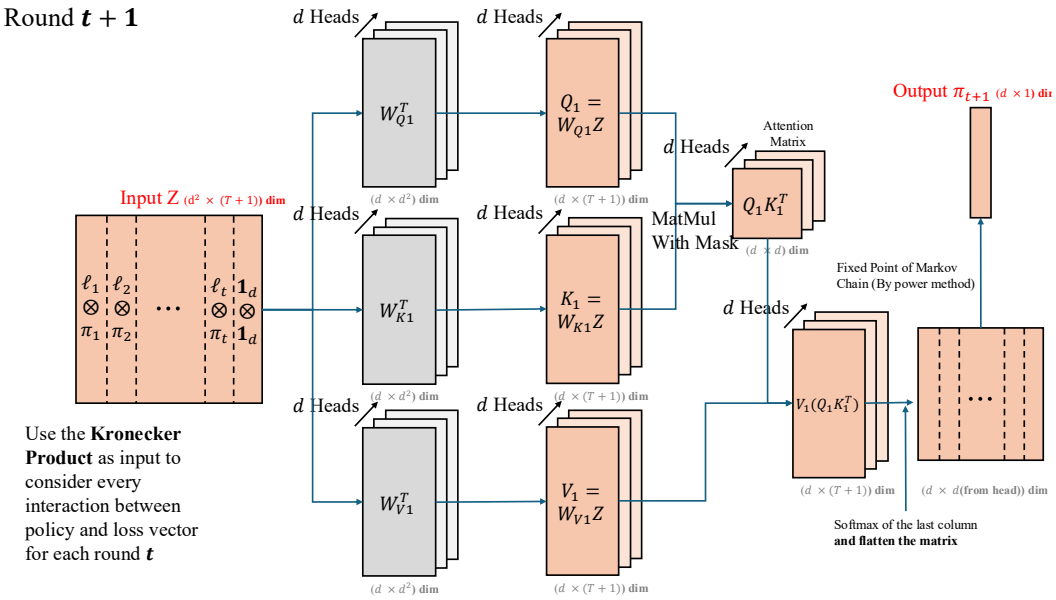


Figure 3: Multi-head linear self-attention architecture for swap-regret minimization. Each head processes the block features $X_t = \ell_t \otimes \pi_t$ and outputs one simplex vector $p_{j,t+1}$ through a $\mathrm{Softmax}$ layer; these vectors form a transition matrix, and the final policy π_{t+1} is obtained as its fixed point, matching the Blum–Mansour reduction from external regret to swap regret.

E Additional Experimental Results

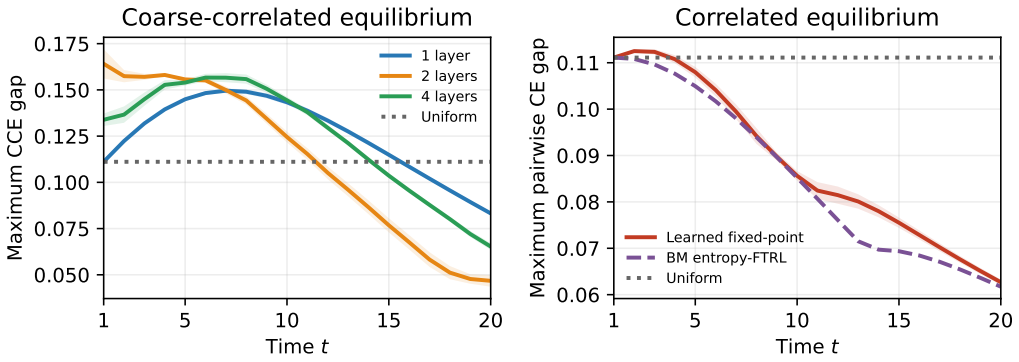


Figure 4: Equilibrium gaps of time-averaged joint mixed play in weighted rock–paper–scissors. Left: external-regret-trained architectures and the maximum CCE gap. Right: the learned fixed-point architecture, classical Blum–Mansour entropy-FTRL, and uniform play, evaluated by the maximum pairwise CE gap. Curves are means with ± 1 standard-error bands over the nine ordered pairs of three learned row and column checkpoints; algorithmic baselines are deterministic.

F Proof of Theorem 2

Theorem 2. *The configuration of the multi-head linear self-attention architecture defined in Equation (9) is a stationary point of the swap-regret loss*

$$\mathcal{L}((V^{(j)}, a^{(j)}, v_c^{(j)})_{j\in[d]}) = \mathbb{E}_{X\sim\mathcal{N}(0,I_{d^2T})}\left[\left(\min_{\widehat{P}\in(\Delta([d]))^d}\sum_{t=1}^{T}\sum_{i=1}^{d}\langle x_{i,t}, \widehat{P}_i - p_{i,t}\rangle\right)^2\right], \tag{10}$$

where

$$x_{i,t} := X_{d(i-1)+1:di,\ t} \in \mathbb{R}^d,$$

and the parameters satisfy

$$V^{(j)} = (V_1^{(j)}, \dots, V_d^{(j)}), \qquad V_r^{(j)} = \begin{cases} -kI_{d\times d}, & r = j, \\ \boldsymbol{O}_{d\times d}, & r \neq j, \end{cases} \qquad a^{(j)} = \mathbf{0}_{d^2}, \quad v_c^{(j)} = v\mathbf{1}_d,$$

for some $k = \Theta(1/\sqrt{T})$.

We define partial sums $S_{i,t} := \sum_{s=1}^{t} x_{i,s} \in \mathbb{R}^d$ with $S_{i,0} := \mathbf{0}_d$ for each head $i \in [d]$ and time $t \in [T]$. For each head $j \in [d]$ and time $t \in [T]$ we write $z_{j,t}$ for the pre-Softmax logits and

$$p_{j,t} = \mathrm{Softmax}(z_{j,t}) \in \Delta([d])$$

for the corresponding mixed action (the j-th column of the transition matrix at time t). Throughout we reuse the Softmax Jacobian notation from Appendix A: for $\pi = \mathrm{Softmax}(z)$, $J_{\mathrm{sm}}(z) := \partial\,\mathrm{Softmax}(z)/\partial z = \mathrm{Diag}(\pi) - \pi\pi^\intercal \in \mathbb{R}^{d\times d}$, so $J_{\mathrm{sm}}(z)^\intercal \mathbf{1}_d = \mathbf{0}_d$ and $\mathbf{1}_d^\intercal J_{\mathrm{sm}}(z) = \mathbf{0}_d^\intercal$.

Proof of the existence of the stationary point. We now impose the special parameter configuration from the statement: for each $j \in [d]$ we write

$$V^{(j)} = (V_1^{(j)}, \dots, V_d^{(j)}), \qquad V_r^{(j)} \in \mathbb{R}^{d\times d},$$

and assume

$$V_r^{(j)} = \begin{cases} -kI_{d\times d}, & r = j, \\ \boldsymbol{O}_{d\times d}, & r \neq j, \end{cases} \qquad a^{(j)} = \mathbf{0}_{d^2}, \qquad v_c^{(j)} = v\mathbf{1}_d,$$

for some $k > 0$. We also split X_s into d blocks of length d:

$$X_s = \big(x_{1,s}^\intercal, \dots, x_{d,s}^\intercal\big)^\intercal, \qquad x_{i,s} \in \mathbb{R}^d.$$

At this configuration the quadratic term $X_s X_s^\intercal a^{(j)}$ vanishes and $a^{(j)\intercal} X_s = 0$, so

$$z_{j,t+1} = \sum_{s=1}^{t} \big(V^{(j)} X_s + v\mathbf{1}_d\big) = \sum_{s=1}^{t} \Big(\sum_{r=1}^{d} V_r^{(j)} x_{r,s} + v\mathbf{1}_d\Big).$$

Using the special form of $V^{(j)}$,

$$z_{j,t+1} = \sum_{s=1}^{t} \big(-kx_{j,s} + v\mathbf{1}_d\big) = -kS_{j,t} + t\, v\mathbf{1}_d.$$

Re-indexing time so that policies at round t use the sum up to $t-1$ (as in Theorem 1), we obtain the simpler representation

$$p_{j,t} = \mathrm{Softmax}\big(-kS_{j,t-1} + (t-1)v\mathbf{1}_d\big), \qquad j \in [d],\ t \in [T]. \tag{57}$$

Thus, for each fixed j, the mapping $(x_{j,1}, \dots, x_{j,T}) \mapsto (p_{j,t})_{t \le T}$ is exactly the single-layer linear attention model from Theorem 1, with ℓ_t there replaced by $x_{j,t}$ here. For a fixed realization of $(x_{i,t})$ the per-head regret R_i satisfies

$$R_i = \sum_{t=1}^{T} \langle x_{i,t}, p_{i,t}\rangle - \min_{q \in \Delta([d])} \Big\langle \sum_{t=1}^{T} x_{i,t}, q \Big\rangle = \sum_{t=1}^{T} \langle x_{i,t}, p_{i,t}\rangle - \min_{j \in [d]} S_{i,T,j}.$$

By (57), each R_i has the same law as the single-head regret $R_T(k)$ from Theorem 1 (with ℓ_t replaced by $x_{i,t}$). The inner minimum in $\mathcal{L}$ decouples across columns:

$$\begin{aligned}
\min_{\widehat{P} \in (\Delta([d]))^d} \sum_{t=1}^{T} \sum_{i=1}^{d} \langle x_{i,t}, \widehat{P}_i - p_{i,t}\rangle &= \sum_{i=1}^{d} \min_{\widehat{p}_i \in \Delta([d])} \sum_{t=1}^{T} \langle x_{i,t}, \widehat{p}_i - p_{i,t}\rangle \\
&= \sum_{i=1}^{d} \Big(\min_{\widehat{p}_i} \sum_{t=1}^{T} \langle x_{i,t}, \widehat{p}_i\rangle - \sum_{t=1}^{T} \langle x_{i,t}, p_{i,t}\rangle \Big) \\
&= \sum_{i=1}^{d} \Big(\min_{q \in \Delta([d])} \big\langle S_{i,T}, q \big\rangle - \sum_{t=1}^{T} \langle x_{i,t}, p_{i,t}\rangle \Big) \\
&= -\sum_{i=1}^{d} R_i.
\end{aligned}$$

Hence we may write

$$R_{\text{tot}} := \text{Swap-Regret}((x_{i,t})_{i \in [d], t \in [T]}) = \sum_{i=1}^{d} R_i, \qquad \mathcal{L}((V^{(j)}, a^{(j)}, v_c^{(j)})_{j \in [d]}) = \mathbb{E}[R_{\text{tot}}^2].$$

The d random variables $R_1, \dots, R_d$ are independent and have the same distribution as $R_T(k)$.

We now compute the gradient of R_{tot}^2 with respect to the logits $z_{j,t}$. For each fixed j the per-head regret R_i with $i \neq j$ does not depend on $p_{j,\cdot}$, hence

$$\frac{\partial R_{\text{tot}}}{\partial p_{j,t}} = \frac{\partial R_j}{\partial p_{j,t}} = x_{j,t}.$$

Using $p_{j,t} = \mathrm{Softmax}(z_{j,t})$ and the chain rule,

$$\frac{\partial R_{\text{tot}}}{\partial z_{j,t}} = J_{\text{sm}}(z_{j,t})^\intercal x_{j,t}, \qquad \frac{\partial R_{\text{tot}}^2}{\partial z_{j,t}} = 2R_{\text{tot}}\, J_{\text{sm}}(z_{j,t})^\intercal x_{j,t}.$$

We abbreviate

$$\Delta_{j,t} := \frac{\partial R_{\text{tot}}^2}{\partial z_{j,t}} = 2R_{\text{tot}}\, J_{\text{sm}}(z_{j,t})^\intercal x_{j,t} \in \mathbb{R}^d.$$

By dominated convergence (Softmax is smooth and all Gaussian moments are finite), we may exchange gradient and expectation. Thus for any perturbations $(\delta z_{j,t})_{j\in[d],t\in[T]}$,

$$\delta\mathcal{L} = \mathbb{E}\Big[\sum_{j=1}^{d}\sum_{t=1}^{T}\langle \Delta_{j,t}, \delta z_{j,t}\rangle\Big]. \tag{58}$$

We now express $\delta z_{j,t}$ in terms of perturbations of the parameters of head j. For notational brevity we suppress the superscript (j) in this step and write (V, a, v_c) for $(V^{(j)}, a^{(j)}, v_c^{(j)})$.

From the definition of $z_{j,t+1}$ we obtain, for arbitrary perturbations $(\delta V, \delta a, \delta v_c)$,

$$\delta z_{t+1} = \sum_{s=1}^{t}\Big(\delta V X_s X_s^\intercal a + V X_s X_s^\intercal \delta a\Big) + \sum_{s=1}^{t}\Big(\delta V X_s + v_c\,\delta a^\intercal X_s + \delta v_c\, a^\intercal X_s\Big) + \sum_{s=1}^{t}\delta v_c.$$

At the special configuration $a = \mathbf{0}_{d^2}$ this reduces to

$$\delta z_{t+1} = \sum_{s=1}^{t}\delta V X_s + \sum_{s=1}^{t} v_c\,\delta a^\intercal X_s + t\,\delta v_c.$$

Restoring the head index j and re-indexing time so that t runs from 1 to T (i.e., replacing $t+1$ by t), we have

$$\delta z_{j,t} = \underbrace{\sum_{s=1}^{t-1}\delta V^{(j)} X_s}_{(*_V)} + \underbrace{\sum_{s=1}^{t-1} v_c^{(j)}\,\delta a^{(j)\intercal} X_s}_{(*_a)} + \underbrace{(t-1)\delta v_c^{(j)}}_{(*_v)}. \tag{59}$$

We now treat separately the three types of perturbations $\delta v_c^{(j)}$, $\delta a^{(j)}$, and $\delta V^{(j)}$.

(i) The $v_c^{(j)}$-direction (shift invariance). Using $(*_v)$ in (59),

$$\delta z_{j,t}^{(v)} = (t-1)\delta v_c^{(j)},$$

so (58) gives

$$\delta\mathcal{L}^{(v)} = \sum_{j=1}^{d}\sum_{t=1}^{T}(t-1)\,\mathbb{E}\big[\langle \Delta_{j,t}, \delta v_c^{(j)}\rangle\big].$$

Write $\delta v_c^{(j)} = \delta v^{(j)}\mathbf{1}_d$. Since $\mathbf{1}_d^\intercal J_{\text{sm}}(z_{j,t})^\intercal = \mathbf{0}_d^\intercal$ for all $z_{j,t}$, we have

$$\mathbf{1}_d^\intercal \Delta_{j,t} = 2F\,\mathbf{1}_d^\intercal J_{\text{sm}}(z_{j,t})^\intercal x_{j,t} = 0$$

pointwise, and therefore each term in the sum vanishes. Hence $\delta\mathcal{L}^{(v)} = 0$ for all perturbations $(\delta v^{(j)})_{j\in[d]}$, and

$$\nabla_{v_c^{(j)}}\mathcal{L}\Big|_{\text{conf.}} = \mathbf{0}_d, \qquad j\in[d].$$

(ii) The $a^{(j)}$-direction at $a^{(j)} = 0$. From $(*_a)$ in (59),

$$\delta z_{j,t}^{(a)} = \sum_{s=1}^{t-1} v_c^{(j)}\,\delta a^{(j)\intercal} X_s = v\,\mathbf{1}_d\sum_{s=1}^{t-1}\langle X_s, \delta a^{(j)}\rangle.$$

Plugging into (58),

$$\begin{aligned}\delta\mathcal{L}^{(a)} &= \sum_{j=1}^{d}\mathbb{E}\Big[\sum_{t=1}^{T}\big\langle \Delta_{j,t}, v\,\mathbf{1}_d\sum_{s=1}^{t-1}\langle X_s, \delta a^{(j)}\rangle\big\rangle\Big]\\ &= v\sum_{j=1}^{d}\mathbb{E}\Big[\sum_{t=1}^{T}\big(\mathbf{1}_d^\intercal\Delta_{j,t}\big)\sum_{s=1}^{t-1}\langle X_s, \delta a^{(j)}\rangle\Big].\end{aligned}$$

But we already observed that $\mathbf{1}_d^\intercal\Delta_{j,t} = 0$ for all (j,t), hence $\delta\mathcal{L}^{(a)} = 0$ for every $(\delta a^{(j)})_{j\in[d]}$ and therefore

$$\nabla_{a^{(j)}}\mathcal{L}\Big|_{\text{conf.}} = \mathbf{0}_{d^2}, \qquad j\in[d].$$

(iii) The $V^{(j)}$-direction at $V_j^{(j)} = -kI_{d\times d}$. It remains to analyse perturbations of the matrices $V^{(j)}$. Recall that each $V^{(j)}$ is partitioned into d blocks $V_r^{(j)} \in \mathbb{R}^{d\times d}$, and similarly we write $\delta V^{(j)} = (\delta V_1^{(j)}, \dots, \delta V_d^{(j)})$. From $(*_V)$ in (59) we get

$$\delta z_{j,t}^{(V)} = \sum_{s=1}^{t-1} \delta V^{(j)} X_s = \sum_{r=1}^{d} \delta V_r^{(j)} S_{r,t-1},$$

since $X_s = (x_{1,s}^\intercal, \dots, x_{d,s}^\intercal)^\intercal$.

Substituting into (58) and using the Frobenius inner product $\langle A, B\rangle_F = \operatorname{tr}(A^\intercal B)$,

$$\begin{aligned}\delta\mathcal{L}^{(V)} &= \sum_{j=1}^{d} \mathbb{E}\Big[\sum_{t=1}^{T} \Big\langle \Delta_{j,t}, \sum_{r=1}^{d} \delta V_r^{(j)} S_{r,t-1} \Big\rangle\Big] \\ &= \sum_{j=1}^{d}\sum_{r=1}^{d} \Big\langle \underbrace{\mathbb{E}\Big[\sum_{t=1}^{T} \Delta_{j,t} S_{r,t-1}^\intercal\Big]}_{=:G_{j,r}(k)},\ \delta V_r^{(j)} \Big\rangle_F.\end{aligned}$$

Thus

$$\nabla_{V_r^{(j)}} \mathcal{L}\Big|_{\text{conf.}} = G_{j,r}(k) \in \mathbb{R}^{d\times d}.$$

We now exploit the symmetries of the Gaussian model.

Case $r \neq j$. At the special configuration $V_r^{(j)} = \boldsymbol{O}_{d\times d}$ for $r \neq j$, the logits $z_{j,t}$ and hence $\Delta_{j,t}$ depend only on the block process $(x_{j,s})_{s\le T}$, while the random walk $(S_{r,t})_{t\le T}$ is built from $(x_{r,s})_{s\le T}$ and is independent of $(x_{j,s})_{s\le T}$. Therefore

$$G_{j,r}(k) = \mathbb{E}\Big[\sum_{t=1}^{T} \Delta_{j,t}\Big]\, \mathbb{E}\big[S_{r,t-1}^\intercal\big] = \mathbf{0}_{d\times d},$$

because $\mathbb{E}[S_{r,t-1}] = \mathbf{0}_d$ for all r, t.

Case $r = j$. Fix $j \in [d]$ and consider the action of coordinate permutations on the action coordinates of head j: for a permutation matrix P,

$$x_{j,t} \mapsto P x_{j,t}, \quad S_{j,t-1} \mapsto P S_{j,t-1}.$$

By (57), $V_j^{(j)} = -kI_{d\times d}$, $a^{(j)} = 0$ and $v_c^{(j)} = v\mathbf{1}_d$, and by permutation-equivariance of $\mathrm{Softmax}$, we also have

$$z_{j,t} \mapsto P z_{j,t}, \quad p_{j,t} \mapsto P p_{j,t}, \quad \Delta_{j,t} \mapsto P\Delta_{j,t}.$$

Since the joint law of $(x_{j,t})_{t\le T}$ is invariant under coordinate permutations and independent of the other heads, the pair $(\Delta_{j,t}, S_{j,t-1})_{t\le T}$ has the same distribution as $(P\Delta_{j,t}, PS_{j,t-1})_{t\le T}$. Consequently,

$$G_{j,j}(k) = \mathbb{E}\Big[\sum_{t=1}^{T} \Delta_{j,t} S_{j,t-1}^\intercal\Big] = \mathbb{E}\Big[\sum_{t=1}^{T} (P\Delta_{j,t})(PS_{j,t-1})^\intercal\Big] = P G_{j,j}(k) P^\intercal$$

for all permutation matrices P. By Lemma 1, there are scalars $\alpha(k), \beta(k) \in \mathbb{R}$ such that

$$G_{j,j}(k) = \alpha(k) I_{d\times d} + \beta(k)\mathbf{1}_d\mathbf{1}_d^\intercal.$$

Moreover, $\mathbf{1}_d^\intercal \Delta_{j,t} = 0$ for every j, t because $J_{\text{sm}}(z_{j,t})^\intercal \mathbf{1}_d = \mathbf{0}_d$, so $\mathbf{1}_d^\intercal G_{j,j}(k) = \mathbf{0}_d^\intercal$. Hence

$$G_{j,j}(k) = \alpha(k)\Pi_0, \qquad \Pi_0 := I_{d\times d} - \frac{1}{d}\mathbf{1}_d\mathbf{1}_d^\intercal.$$

The scalar is the same for every j by exchangeability of the heads and blocks at the stated configuration.

Combining the two cases, we have shown that at the stated configuration

$$\nabla_{V_r^{(j)}}\mathcal{L} = \begin{cases} \alpha(k)\Pi_0, & r = j, \\ \boldsymbol{O}_{d\times d}, & r \neq j, \end{cases} \qquad j \in [d].$$

Together with the $v_c^{(j)}$- and $a^{(j)}$-direction results above, this implies that, for every fixed k, the only possibly non-zero component of the gradient of $\mathcal{L}$ at the special configuration lies in the one-dimensional direction in which each $V_j^{(j)}$ is scaled by the zero-sum projector Π_0. □

Now, we show that the stationary point is a scalar $k^\star$ such that $\mathcal{L}'(k^\star) = 0$.

Proof of the scale of the stationary point. Define the one-dimensional function

$$\Phi(k) := \mathcal{L}\big((V^{(j)}, a^{(j)}, v_c^{(j)})_{j\in[d]}\big)$$

when, for each $j \in [d]$,

$$V_r^{(j)} = \begin{cases} -kI_{d\times d}, & r = j, \\ \boldsymbol{O}_{d\times d}, & r \neq j, \end{cases} \qquad a^{(j)} = \mathbf{0}_{d^2}, \quad v_c^{(j)} = v\mathbf{1}_d.$$

By the chain rule,

$$\Phi'(k) = \sum_{j=1}^{d} \Big\langle \nabla_{V^{(j)}}\mathcal{L},\ \frac{\partial V^{(j)}}{\partial k} \Big\rangle_F = \sum_{j=1}^{d} \Big\langle \alpha(k)\Pi_0,\ -I_{d\times d} \Big\rangle_F = -d(d-1)\alpha(k),$$

since $\partial V_j^{(j)}/\partial k = -I_{d\times d}$ and $\partial V_r^{(j)}/\partial k = \boldsymbol{O}_{d\times d}$ for $r \neq j$. In particular,

$$\alpha(k) = 0 \quad \Longleftrightarrow \quad \Phi'(k) = 0.$$

Let $R(k)$ denote a random variable with the law of $R_T(k)$ for the single-head model, and note that $R_1(k, X), \ldots, R_d(k, X)$ are independent copies of $R(k)$ for every fixed k. Therefore

$$\begin{aligned} \Phi(k) = \mathbb{E}\big[R_{\mathrm{tot}}(k, X)^2\big] &= \mathbb{E}\Big[\Big(\sum_{i=1}^{d} R_i(k, X)\Big)^2\Big] \\ &= \sum_{i=1}^{d} \mathbb{E}[R_i(k, X)^2] + 2 \sum_{1\le i<j\le d} \mathbb{E}[R_i(k, X)R_j(k, X)] \\ &= d\,\mathbb{E}[R(k)^2] + d(d-1)\big(\mathbb{E}[R(k)]\big)^2. \end{aligned} \tag{60}$$

Proposition 5. *The expectation of the single-head regret is independent of k, i.e.,* $\mathbb{E}[R(k)]$ *is independent of* k*.*

Proof. Write the single-head process as $\{x_t\}_{t=1}^{\mathsf{T}}$ with $x_t \sim \mathcal{N}(\mathbf{0}_d, I_{d\times d})$ i.i.d., and denote $S_t := \sum_{s=1}^{t} x_s$ and

$$\pi_t(k) := \mathrm{Softmax}(-kS_{t-1} + (t-1)v\mathbf{1}_d),$$

so that

$$R(k) = \sum_{t=1}^{T} \langle x_t, \pi_t(k)\rangle - \min_{q\in\Delta([d])} \Big\langle \sum_{t=1}^{T} x_t,\ q \Big\rangle.$$

First,

$$C(X) := \min_{q\in\Delta([d])} \Big\langle \sum_{t=1}^{T} x_t,\ q \Big\rangle$$

depends only on the losses $(x_t)_{t\le T}$ and is completely independent of the parameter k. Moreover, note that $\pi_t(k)$ is measurable with respect to the σ-algebra generated by $(x_1, \ldots, x_{t-1})$ and does not

depend on x_t. Using tower property and the fact that x_t is independent of $(x_s)_{s<t}$ and has mean zero, we obtain

$$\begin{aligned}\mathbb{E}\big[\langle x_t, \pi_t(k)\rangle\big] &= \mathbb{E}\Big[\mathbb{E}\big[\langle x_t, \pi_t(k)\rangle \,\big|\, x_1, \dots, x_{t-1}\big]\Big] \\ &= \mathbb{E}\Big[\big\langle \mathbb{E}[x_t \mid x_1, \dots, x_{t-1}],\ \pi_t(k)\big\rangle\Big] = \mathbb{E}\big[\langle 0, \pi_t(k)\rangle\big] = 0.\end{aligned}$$

Summing over t gives

$$\mathbb{E}\Big[\sum_{t=1}^{T}\langle x_t, \pi_t(k)\rangle\Big] = 0 \quad \text{for all } k \in \mathbb{R}.$$

Therefore

$$\mathbb{E}[R(k)] = 0 - \mathbb{E}[C(X)] = -\,\mathbb{E}\Big[\min_{q\in\Delta([d])}\Big\langle \sum_{t=1}^{T} x_t,\ q\Big\rangle\Big],$$

which is a finite constant independent of k. □

From (60) and Proposition 5 we obtain

$$\begin{aligned}\Phi'(k) &= d\,\frac{d}{dk}\mathbb{E}[R(k)^2] + d(d-1)\,\frac{d}{dk}\big(\mathbb{E}[R(k)]^2\big) \\ &= d\,\frac{d}{dk}\mathbb{E}[R(k)^2] + d(d-1)\cdot 2\,\mathbb{E}[R(k)]\,\frac{d}{dk}\mathbb{E}[R(k)] \\ &= d\,\frac{d}{dk}\mathbb{E}[R(k)^2] \quad (\text{since } \frac{d}{dk}\mathbb{E}[R(k)] = 0).\end{aligned}$$

Recalling the external-regret objective

$$L_{\text{ext}}(k) := \mathbb{E}[R(k)^2],$$

this identity can be succinctly written as

$$\Phi'(k) = d\,L'_{\text{ext}}(k) \quad \text{for all } k \in \mathbb{R}. \tag{61}$$

Sign of $\Phi'(k)$ and existence of $k^\star$.

The single-head analysis in Theorems 3 and 4 shows that there exist absolute constants $c_1, c_2 > 0$ (independent of d, T) such that: For all sufficiently large T, d and all $0 < k \le c_1/\sqrt{Td}$, one has

$$L'_{\text{ext}}(k) < 0.$$

For all $T, d \ge 2$ and all $k \ge c_2/\sqrt{T}\,\text{polylog}(d)$, one has

$$L'_{\text{ext}}(k) > 0.$$

Combining these with (61), we obtain

$$\Phi'(k) < 0 \quad \text{for } k \le c_1/\sqrt{Td}, \qquad \Phi'(k) > 0 \quad \text{for } k \ge c_2/\sqrt{T}\,\text{polylog}(d).$$

The function $k \mapsto \Phi(k)$ is smooth ($\mathrm{Softmax}$ and the architecture are smooth and all Gaussian moments are finite), hence Φ' is continuous. By the intermediate value theorem there exists

$$k^\star \in \Big(\widetilde{\Theta}\Big(\tfrac{1}{\sqrt{Td}}\Big), \widetilde{\Theta}\Big(\tfrac{1}{\sqrt{T}}\Big)\Big)$$

such that $\Phi'(k^\star) = 0$. □

G Additional Background

G.1 Literature Review

Transformers & In-context-learning. LLMs are now predominantly built upon the self-attention architecture [Vaswani et al., 2017]. A central reason for their success is their remarkable ability to perform *in-context learning* (ICL): Transformers can construct new predictors directly from sequences of labeled examples provided in the input, without any parameter updates. This mechanism enables powerful *few-shot learning* capabilities [Brown et al., 2020, Garg et al., 2022, Min et al., 2022], which in turn have motivated a rapidly growing theoretical literature on ICL. For certain training losses, the *minimizer* of a single-layer Transformer is equivalent to performing one step of gradient descent for linear regression [Ahn et al., 2023, Zhang et al., 2023, Mahankali et al., 2023]. For attention models with non-linear operators, prior work has provided provable guarantees for ICL—either through training-dynamics analyses or partial characterizations of stationary points. For example, Ahn et al. [2023] studied transformers with a ReLU operator, and Cheng et al. [2024] analyzed the $\mathrm{Softmax}$ operator from a functional-gradient perspective. In addition, Chen et al. [2024] examined the training dynamics of multi-head Softmax attention in the context of ICL. However, none of these works characterizes the exact configuration of the stationary point of a transformer with a non-linear operator. Cheng et al. [2024] derived a stationary-point result but did not identify the precise configuration corresponding to a specific (functional) gradient-descent stepsize; Chen et al. [2024] provided a dynamical analysis for multi-head Softmax attention, but likewise did not interpret or characterize the stationary-point structure of transformers with non-linear operators. In contrast, our analysis *explicitly characterizes* the stationary-point configuration of a Transformer equipped with a non-linear operator, thereby yielding a provable guarantee for the no-regret behavior induced by such architectures—an independently interesting theoretical contribution. More recently, ICL has also been generalized to decision-making settings. Works such as Laskin et al. [2023], Lee et al. [2023], Lin et al. [2024] demonstrated that supervised pre-training can endow Transformers with in-context reinforcement learning abilities, enabling them to solve stochastic bandits and Markov decision processes. In comparison, regret-loss training focuses on *online learning* settings that may be arbitrary or even *adversarial*, as well as *game-theoretic* environments where agents interact strategically.

G.2 Known Algorithms for No-Regret Learning

Follow-the-Regularized-Leader (FTRL). FTRL generates the sequence

$$x_{t+1} \;=\; \operatorname*{argmin}_{x\in\Pi} \left\{ \eta \sum_{s=1}^{t} \ell_s(x) + R(x) \right\},$$

where R is a strongly convex regularizer and $\eta > 0$ is the learning rate. With the negative-entropy regularizer $R(x) = \sum_i x_i \log x_i$, FTRL reduces to the multiplicative-weights update (or exponential weights).

Smoothed Fictitious Play as Entropy-Regularized FTRL. Smoothed fictitious play is the update rule

$$x_{t+1} \;\propto\; \exp\left(-\eta \sum_{s=1}^{t} \ell_s \right),$$

which is exactly FTRL with entropy regularization. The smoothing (logit or softmax) arises from the convex conjugate of the negative entropy. If stepsize $\eta = \Theta(1/\sqrt{T})$, smoothed fictitious play has a no-regret guarantee [Cesa-Bianchi and Lugosi, 2006]. In this paper, we interchangeably use the terms *smoothed fictitious play* and *FTRL with entropy regularization*.

H Broader Impacts

This paper is primarily theoretical: it analyzes stationary points of regret-loss objectives for minimal attention architectures and supports the claims with synthetic Gaussian simulations. It does not introduce a deployed system, a pretrained model, or a dataset involving people. The broader impacts

are therefore indirect and would arise only if regret-based objectives of this type were incorporated into practical sequential decision-making systems.

A positive impact of the work is that it clarifies when attention-based models can implement classical no-regret algorithms internally, without relying on post-processing or supervised traces of those algorithms. This can help make decision-making models more interpretable: at the stationary points studied in the paper, the learned forward pass corresponds to smoothed fictitious play or the Blum–Mansour swap-regret construction. In multi-agent environments, such mechanistic understanding may support more reliable analysis of whether learned agents approach coarse correlated or correlated equilibria.

At the same time, no-regret behavior is not a general safety guarantee. Equilibrium convergence does not imply that the resulting outcome is socially desirable, fair, privacy-preserving, or robust to all forms of manipulation. In consequential applications, regret-minimizing agents could adapt efficiently toward objectives that are misaligned with human values, and swap-regret or external-regret guarantees only rule out specific classes of deviations. A further risk is overinterpreting the present results: the paper identifies stationary points under idealized assumptions, but it does not prove that practical training dynamics will converge to them in larger models or under realistic feedback.

These risks suggest that any deployment motivated by this analysis should be paired with application-specific safeguards: careful objective design, empirical stress tests beyond Gaussian full-information losses, monitoring under distribution shift, fairness and privacy constraints where relevant, and human oversight for high-stakes decisions. The present results are best viewed as a mechanistic and theoretical step toward understanding regret-trained attention, not as a stand-alone recipe for deploying autonomous decision-making systems.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes].

   Justification: The abstract and introduction state the two main contributions: stationary-point characterizations for regret loss on the simplex and a swap-regret architecture based on the Blum–Mansour construction. The scope is limited to the stated one-layer linear-attention, softmax, full-information, and Gaussian-training-loss settings, with limitations summarized in the conclusion.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes].

   Justification: The conclusion explicitly states that the results identify stationary points rather than global optima or training-dynamics convergence. It also lists the main assumptions: full-information Gaussian training losses, simplex-valued policies, one-layer linear attention, the $\mathrm{Softmax}$ operator, and horizon-dependent stepsize scaling.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

   Answer: [Yes].

   Justification: The theoretical statements specify the architecture, loss, distributional assumptions, and parameter configurations, with proof outlines in the main text. Complete proofs and supporting lemmas are deferred to the appendices and cross-referenced from the theorem statements.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes].

   Justification: The main empirical result is the Monte Carlo estimate of $k^\star(d, T)$ in Figure 1. The figure caption reports the number of independent runs and Monte Carlo samples, and the supplemental code specifies the grid search, random sampling procedure, and commands used to generate the plotted data.

Guidelines:

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

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

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

   Answer: [Yes].

   Justification: The supplemental material includes the code used to reproduce Figure 1, including the generated CSV summaries. The experiments use synthetic Gaussian losses, so there is no external dataset to download.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes].

   Justification: The experiments are Monte Carlo evaluations of the regret-loss objective over synthetic Gaussian losses, not model training with train/test splits. The architecture, objective, sampling distribution, number of runs, number of samples, and plotted statistic are specified in the paper or supplemental code.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes].

   Justification: Figure 1 reports shaded bands corresponding to ± 1 standard deviation over 50 independent runs. The caption also states the number of Monte Carlo samples used per run.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the assumed Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [No].

Justification: The experiments are lightweight synthetic Monte Carlo simulations, and the supplemental code is sufficient to run them. However, the manuscript does not currently report the hardware, memory, runtime, or total compute required.

Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes].

   Justification: The work is theoretical and uses synthetic Gaussian simulations. It does not involve human subjects, private data, scraped data, or deployment of a high-risk model.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes].

    Justification: Appendix H discusses the positive impact of mechanistically understanding regret-trained attention and its equilibrium behavior. It also discusses negative impacts and limitations, including possible misuse or misalignment in consequential multi-agent decision systems and the risk of overinterpreting stationary-point results as deployment guarantees.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.

- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A].

    Justification: The paper does not release pretrained models, image generators, scraped datasets, or other artifacts that pose high misuse risk. The released material is limited to synthetic-simulation code and derived plots.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [N/A].

    Justification: The paper does not use external datasets, pretrained models, or third-party code assets as part of the research contribution. Prior research and algorithms are credited through citations.

    Guidelines:

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

13. **New assets**

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

    Answer: [Yes].

Justification: The supplemental code for the synthetic experiments documents the command-line interface and generated outputs. The paper does not introduce a new dataset or pretrained model.

Guidelines:

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

14. **Crowdsourcing and research with human subjects**

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

Answer: [N/A].

Justification: The paper does not involve crowdsourcing or research with human subjects. All experiments use synthetic Gaussian data.

Guidelines:

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

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

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

Answer: [N/A].

Justification: The paper does not involve crowdsourcing or human-subjects research, so IRB approval or equivalent review is not applicable.

Guidelines:

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

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A].

Justification: The paper does not use LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.