# CAN IN-CONTEXT LEARNING SURVIVE RECURSIVE PRETRAINING?

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Can an in-context learner generate data from which its successors can relearn the same abilities? We study recursive pretraining of attention models on regression tasks, with each generation trained from scratch and evaluated on the original distribution. We distinguish optimization failure caused by small labels from structural degradation that survives energy normalization. For linear–quadratic tasks under population-context gradient flow, arbitrarily accurate single-generation learning coexists with a nonvanishing recursive clean-risk gap at every fixed training time. We derive a cubic threshold for preserving quadratic structure under polynomial training schedules. Alternatively, a calibrated signed combination of two attention-temperature views, followed by normalization, exactly preserves the task law and every generation's clean risk at fixed training time in the same population setting, without real-data refresh or parameter inheritance. We provide fixed-horizon SGD approximations and demonstrate improved retention in dense-attention experiments.

## 1 INTRODUCTION

In-context learning (ICL) allows a pretrained model to infer a task from demonstrations without updating its parameters (Brown et al., 2020). Theory shows that transformers can implement learning algorithms and attain small error on new regression tasks (Bai et al., 2023; Zhang et al., 2024; Wu et al., 2024; Cheng et al., 2023). These results primarily assume a fixed task distribution. What happens when a previously trained model generates that distribution?

Repeatedly replacing training data with model-generated data can erase information about the original distribution, a phenomenon known as *model collapse* (Shumailov et al., 2024). Here this replacement affects a distribution of regression tasks. We ask:

*When does training on a predecessor's predictions preserve in-context learning?*

At generation $r$, a model is initialized from scratch and trained on the task law $P_{r-1}$, then frozen to generate the next task law $P_r$. Each generated function uses its own sampled parent task and seed context, both held fixed across queries. Every model is evaluated on fresh episodes from the original task law $P_0$. This protocol provides a stylized abstraction of LLM pretraining on real-world corpora that may contain outputs from earlier models, including erroneous or hallucinated content. It isolates how model-generated data reshape the training distribution across generations. By keeping the evaluation law fixed at $P_0$, we examine whether models trained on these evolving data retain the ability to learn the original, ground-truth tasks in context.

In this paper, we investigate the gap between *learning a task family* and *generating data from which successive models can relearn it*. We trace how learned attention reshapes the next generation's training tasks and how these changes affect ICL risk on the original law $P_0$. Our analysis distinguishes task *scale*, measured by total label energy, from task *composition*, measured by the relative energies of its components. Shrinking labels can leave the optimal predictor unchanged yet obstruct finite-budget learning. Conversely, restoring total energy does not correct unequal component attenuation, which can progressively weaken the training signal needed to learn the original tasks. We therefore examine whether this degradation can be prevented by increasing training time or by changing generation. For linear–quadratic tasks in the population-context setting, we derive a training-time



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

threshold for component preservation under polynomial schedules and construct a calibrated signed two-temperature correction that preserves the task law and each cold-start learner's risk at fixed training time, without real-data refresh or parameter inheritance. We summarize our contributions as follows:

- **Learning does not ensure preservation.** Extending single-generation ICL analysis, we separate amplitude-induced SGD failure from structural degradation. For normalized linear–quadratic tasks under population-context gradient flow, we prove a nonvanishing recursive risk gap despite arbitrarily accurate single-generation learning (Theorems 1–3).
- **A training-time barrier.** Rather than fixing the generation operator, we derive component attenuation from query/key dynamics. In the same population setting, a cubic boundary governs quadratic-energy preservation under polynomial schedules; sufficiently large super-cubic schedules yield vanishing clean risk (Theorem 4).
- **Generation-only preservation.** Complementing real-data mixing, calibrated signed attention-temperature views and normalization exactly preserve the task law and fixed-time cold-start risk in this population setting, without real-data refresh or parameter inheritance (Theorem 5). Dense-attention experiments demonstrate improved finite-context retention.

## 2 Related work

**In-context learning.** The in-context learning (ICL) ability has motivated empirical and theoretical studies of simple task families. Garg et al. (2022) show that transformers pretrained on regression tasks can learn unseen functions in context, with linear-regression performance close to least squares. Xie et al. (2021) interpret ICL as inference of a shared latent concept in a probabilistic sequence model. Attention layers can also implement gradient-based learning algorithms (Von Oswald et al., 2023; Ahn et al., 2023), and transformers can implement and select among statistical procedures (Bai et al., 2023). A separate question is whether pretraining learns such predictors. Zhang et al. (2024) prove gradient-flow convergence for a linear-attention model under suitable initialization. For a restricted linear-attention parameterization, Wu et al. (2024) bound pretraining task complexity and compare the learned predictor with optimally tuned ridge regression at matched context lengths. For softmax attention, Chen et al. (2024) analyze multi-head training dynamics for linear regression, while He et al. (2025) relate learned attention patterns to debiased gradient descent. Beyond linear tasks, Cheng et al. (2023) connect attention to functional gradient descent, and Collins et al. (2024) show how softmax attention adapts to the smoothness of the pretraining functions.

**Model collapse.** Repeated training on model-generated data raises the question of whether information about the original distribution survives across generations. Shumailov et al. (2024) identify model collapse in recursive training, including the loss of low-probability events and eventual distortion of the learned distribution. Alemohammad et al. (2024) study related self-consuming loops for generative image models and document degradation in sample quality or diversity. For regression, Dohmatob et al. (2024) characterize the effects of recursive generation on downstream test error, including the roles of sample size, noise, and regularization. A complementary line asks how access to real data changes this behavior. Gerstgrasser et al. (2024) distinguish replacement from accumulation: retaining the original data and successive synthetic datasets can prevent collapse in their linear-model analysis and experiments. Fu et al. (2025) establish stability and generalization guarantees for recursive learning with real and synthetic data, including transformer ICL when the real-data proportion is sufficiently large.

## 3 Preliminaries

**ICL tasks and episodes.** Let $d \geq 2$ and let $\nu$ be uniform on $\mathbb{S}^{d-1} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\| = 1\}$. For a task law $P$ over functions $f : \mathbb{S}^{d-1} \to \mathbb{R}$, an *in-context episode* samples $f \sim P$ and independent inputs $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, \boldsymbol{x}_q \sim \nu$. A predictor $F_{\boldsymbol{\theta}}(\mathcal{C}_n, \boldsymbol{x}_q)$ uses demonstrations $\mathcal{C}_n = \{(\boldsymbol{x}_i, f(\boldsymbol{x}_i))\}_{i=1}^n$ to estimate $y_q = f(\boldsymbol{x}_q)$ without changing $\boldsymbol{\theta}$. Each new episode independently resamples the task and inputs. We define the *task energy* as $\mathcal{E}(P) := \mathbb{E}_{f\sim P, \boldsymbol{x}\sim\nu}[f(\boldsymbol{x})^2]$, which tracks the overall scale of generated labels.

**Attention models.** We represent an in-context episode in the matrix form

$$\boldsymbol{H}_n = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_n, \boldsymbol{h}_q] = \begin{bmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_n & \boldsymbol{x}_q \\ y_1 & \cdots & y_n & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times(n+1)}, \qquad y_i = f(\boldsymbol{x}_i). \tag{1}$$

An attention layer with parameters $\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V \in \mathbb{R}^{(d+1)\times(d+1)}$ updates the query token as

$$\widetilde{\boldsymbol{h}}_q = \boldsymbol{h}_q + \frac{\sum_{i=1}^n (\boldsymbol{W}_V \boldsymbol{h}_i) e^{(\boldsymbol{W}_K \boldsymbol{h}_i)^\top (\boldsymbol{W}_Q \boldsymbol{h}_q)}}{\sum_{j=1}^n e^{(\boldsymbol{W}_K \boldsymbol{h}_j)^\top (\boldsymbol{W}_Q \boldsymbol{h}_q)}}. \tag{2}$$

The prediction is the last coordinate $F_{\boldsymbol{\theta}}(\mathcal{C}_n, \boldsymbol{x}_q) = (\widetilde{\boldsymbol{h}}_q)_{d+1}$. We study the following block parameterization, analogous to that of Collins et al. (2024):

$$\boldsymbol{W}_Q = \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{0} \\ \boldsymbol{0}^\top & 0 \end{bmatrix}, \boldsymbol{W}_K = \begin{bmatrix} \boldsymbol{K} & \boldsymbol{u} \\ \boldsymbol{0}^\top & 0 \end{bmatrix}, \boldsymbol{W}_V = \begin{bmatrix} \boldsymbol{0}_{d\times d} & \boldsymbol{0} \\ \boldsymbol{0}^\top & c \end{bmatrix}, \tag{3}$$

We train $\boldsymbol{\theta} = (\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{u}, c)$, where $\boldsymbol{Q}, \boldsymbol{K} \in \mathbb{R}^{d\times d}$, $\boldsymbol{u} \in \mathbb{R}^d$, and $c \in \mathbb{R}$. The resulting predictor is

$$F_{\boldsymbol{\theta}}(\mathcal{C}_n, \boldsymbol{x}_q) = c \frac{\sum_{i=1}^n y_i e^{(\boldsymbol{Q}\boldsymbol{x}_q)^\top (\boldsymbol{K}\boldsymbol{x}_i + \boldsymbol{u} y_i)}}{\sum_{j=1}^n e^{(\boldsymbol{Q}\boldsymbol{x}_q)^\top (\boldsymbol{K}\boldsymbol{x}_j + \boldsymbol{u} y_j)}}. \tag{4}$$

We write $s_i = (\boldsymbol{Q}\boldsymbol{x}_q)^\top (\boldsymbol{K}\boldsymbol{x}_i + \boldsymbol{u} y_i)$ for the score of the $i$-th demonstration and $\pi_i = e^{s_i} / \sum_{j=1}^n e^{s_j}$ for its attention weight, with $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_n]^\top$. The weights depend on labels through $\boldsymbol{u}$; setting $\boldsymbol{u} = \boldsymbol{0}$ and $c = 1$ recovers Collins et al. (2024). Appendix G details the parameterization.

For the comparison in Theorem 1, we also use the linear-attention predictor of Wu et al. (2024):

$$F_{\boldsymbol{\Gamma}}(\mathcal{C}_n, \boldsymbol{x}_q) = \boldsymbol{x}_q^\top \boldsymbol{\Gamma} \left( \frac{d}{n} \sum_{i=1}^n \boldsymbol{x}_i y_i \right), \qquad \boldsymbol{\Gamma} \in \mathbb{R}^{d\times d}, \tag{5}$$

where only $\boldsymbol{\Gamma}$ is trained.

**Loss and evaluation.** Write $\xi = (\mathcal{C}_n, \boldsymbol{x}_q, y_q)$ for a complete episode and define its squared loss by

$$\ell(\boldsymbol{\theta}; \xi) := \frac{1}{2} \big[ F_{\boldsymbol{\theta}}(\mathcal{C}_n, \boldsymbol{x}_q) - y_q \big]^2. \tag{6}$$

For an original task law $P_0$ with $\mathcal{E}(P_0) > 0$, define the *population loss* and *clean ICL risk* by

$$L_P^{(n)}(\boldsymbol{\theta}) := \mathbb{E}_P[\ell(\boldsymbol{\theta}; \xi)], \tag{7}$$

$$R_0^{(n)}(\boldsymbol{\theta}) := \frac{2 L_{P_0}^{(n)}(\boldsymbol{\theta})}{\mathcal{E}(P_0)}. \tag{8}$$

The expectation covers the entire episode. Training uses $P$; clean risk uses $P_0$. The zero predictor has risk one and is optimal among query-only predictors when $\mathbb{E}_{f\sim P_0}[f(\boldsymbol{x})] = 0$, as in the task families below.

**Recursive pretraining.** At generation $r \geq 1$, train $\boldsymbol{\theta}_r$ on $P_{r-1}$, then generate $P_r$ (Figure 1). Each generation starts from $\boldsymbol{\theta} = (\boldsymbol{I}_d, \boldsymbol{I}_d, \boldsymbol{0}, 0)$ or $\boldsymbol{\Gamma} = \boldsymbol{0}$, unless stated otherwise.

Suppressing $r$, SGD draws $B$ fresh independent length-$n$ episodes $\xi_k^{(1)}, \ldots, \xi_k^{(B)}$ from $P_{r-1}$ at step $k$ and updates

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \frac{\eta_k}{B} \sum_{b=1}^B \nabla_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}_k; \xi_k^{(b)}), \tag{9}$$

with step size $\eta_k$ and budget $T_r = \sum_k \eta_{r,k}$. For linear attention, replace $\boldsymbol{\theta}$ by $\boldsymbol{\Gamma}$. For generation, fix $\boldsymbol{\theta}_r$, draw $f \sim P_{r-1}$ and independent $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m \sim \nu$, and set $\mathcal{C}_m = \{(\boldsymbol{z}_i, f(\boldsymbol{z}_i))\}_{i=1}^m$. Define

$$\widetilde{f}_r(\boldsymbol{x}) = F_{\boldsymbol{\theta}_r}(\mathcal{C}_m, \boldsymbol{x}), \qquad \widetilde{P}_r = \mathrm{Law}(\widetilde{f}_r). \tag{10}$$

Here Law denotes the distribution over generated functions conditional on $\boldsymbol{\theta}_r$ and $P_{r-1}$. Each function keeps $f$ and $\mathcal{C}_m$ fixed across queries; independent draws define independent functions. We use $m = n$ unless stated otherwise. *Unnormalized generation* sets $P_r = \widetilde{P}_r$; *unit-energy generation* sets $P_r = \mathrm{Law}(\widetilde{f}_r / \sqrt{\mathcal{E}(\widetilde{P}_r)})$ when $\mathcal{E}(\widetilde{P}_r) > 0$, using one population scale for all functions.

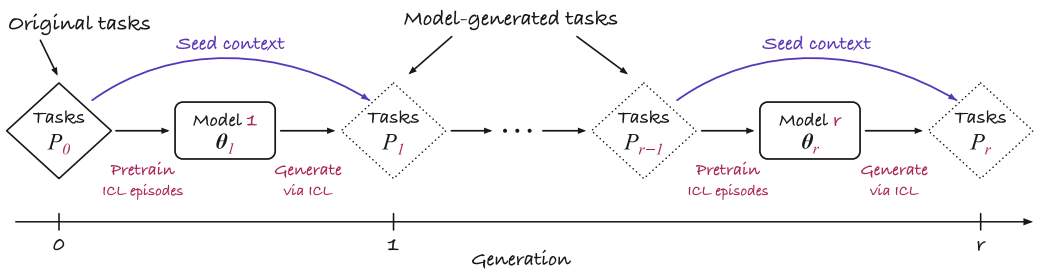


Figure 1: **Recursive pretraining for ICL.** Train $F_{\boldsymbol{\theta}_r}$ on $P_{r-1}$, then generate $P_r$ using a fixed context from $P_{r-1}$ for each new function (purple arcs).

**Population contexts and gradient flow.** With *population contexts*, replace context averages by integrals under $\nu$ for each task and query in training, generation, and evaluation. Compute squared error after this replacement, obtaining $L_P^{(\infty)}$ and $R_0^{(\infty)}$. Full-parameter gradient flow follows

$$\frac{d\boldsymbol{\theta}_r(t)}{dt} = -\nabla_{\boldsymbol{\theta}} L^{(\infty)}_{P_{r-1}}(\boldsymbol{\theta}_r(t)), \qquad \boldsymbol{\theta}_r(0) = (\boldsymbol{I}_d, \boldsymbol{I}_d, \mathbf{0}, 0). \tag{11}$$

Here $T_r$ is the flow duration and $\boldsymbol{\theta}_r = \boldsymbol{\theta}_r(T_r)$ is fixed during generation.

## 4 Main Results

### 4.1 Task Energy and ICL Risk

Previous works on model collapse have identified the phenomenon of "covariance collapse" during recursive pretraining of models (Shumailov et al., 2024), where the data covariance collapses to 0. However, does covariance collapse leads to ICL ability collapse? We show in the following theorem that a collapse in task energy (under standard Gaussian assumption the task energy equals covariance) doesn't necessarily lead to a loss of ICL ability.

**Theorem 1** (Population task energy and ICL risk)**.** *Let $P_0$ be the law of $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$ with $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$. Use the recursive protocol in Section 3 with linear attention* (5)*, training and generation context lengths $m = n$, and unnormalized generation $P_r = \widetilde{P}_r$. At each generation, minimize the population loss $L^{(n)}_{P_{r-1}}$ in* (7)*. Then for every $r \geq 1$:*

*1. The population minimizer is unique and given by*

$$\boldsymbol{\Gamma}_r = \frac{n}{n+d-1}\boldsymbol{I}_d.$$

*Moreover, at each fixed generation the SGD process* (9) *converges to this minimizer with decreasing step sizes satisfying $\sum_{k=0}^{\infty} \eta_k = \infty, \sum_{k=0}^{\infty} \eta_k^2 < \infty$.*

*2. The task energy is given by*

$$\mathcal{E}(P_r) = \left(\frac{n}{n+d-1}\right)^r \longrightarrow 0 \qquad \text{as } r \to \infty.$$

*3. The clean ICL risk defined in* (8) *is given by*

$$R_0^{(n)}(\boldsymbol{\Gamma}_r) = \frac{d-1}{n+d-1} < 1. \tag{12}$$

The population optimizer is independent of task scale. Thus all generations make the same predictions on fresh $P_0$ episodes, and the risk in (12) is the minimum attainable by this model with $n$ demonstrations. The theorem assumes complete optimization at each generation, and the SGD statement gives convergence at each fixed generation.

Realistically a model is almost never sufficiently trained for infinite time to the optimum. So we next study softmax attention trained by SGD under finite training time.

**Assumption 4.1** (Small task amplitude and bounded training budgets)**.** The initial task law satisfies $\mathcal{E}(P_0) > 0$ and $\mathbb{E}_{f\sim P_0}\left[\|f\|_\infty^2\right] < \infty$. There exist a deterministic generation index $r_0 \in \{0, 1, \dots\}$ and deterministic constants $M \geq 0$ and $T < \infty$ such that the task law at generation $r_0$ satisfies $P_{r_0}(\{f : \|f\|_\infty \leq M\}) = 1$ almost surely. The subsequent training budgets satisfy $\sup_{r>r_0} T_r \leq T, TM^2 < \ln 2$.

We show that when labels are sufficiently small, a bounded training budget limits the learned prediction magnitude. Those predictions supply still smaller labels at the next generation.

**Theorem 2** (Finite time SGD task energy and ICL risk)**.** *Suppose Assumption 4.1 holds. Use the recursive protocol in Section 3 with softmax attention* (4)*, full-parameter SGD* (9)*, nonnegative step sizes, and unnormalized generation $P_r = \widetilde{P}_r$. At every generation, initialize $c = 0$ and arbitrary $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{u}$. Define $q_0 = e^{TM^2} - 1 \in [0, 1)$ and $\kappa = \left(\frac{\mathbb{E}_{f\sim P_0}[\|f\|_\infty^2]}{\mathcal{E}(P_0)}\right)^{1/2}$. Then, almost surely over the training data, for every integer $j \geq 1$:*

*1. The task energy satisfies*

$$\mathcal{E}(P_{r_0+j}) \leq M^2 q_0^{3^j-1}, \qquad \lim_{r\to\infty} \mathcal{E}(P_r) = 0.$$

*2. The clean ICL risk defined in* (8) *satisfies*

$$\left|R_0^{(n)}(\boldsymbol{\theta}_{r_0+j}) - 1\right| \leq 2\kappa q_0^{3^{j-1}} + \kappa^2 q_0^{2\cdot 3^{j-1}}, \qquad \lim_{r\to\infty} R_0^{(n)}(\boldsymbol{\theta}_r) = 1. \tag{13}$$

We see that once the task has small enough $M$ s.t. $TM^2 < \ln 2$ holds, task energy decays doubly exponentially in the number of subsequent generations, and clean risk approaches that of the zero predictor.

### 4.2 Task Composition under Fixed Training Time

Under the the assumption of Theorem 2, task energy vanishes across generations and clean ICL risk approaches that of the zero predictor. We next ask whether keeping task energy fixed is sufficient to preserve ICL performance. To study this question, we use the unit-energy protocol of Section 3, with population contexts and gradient flow (11). Global normalization fixes the overall scale of each generated task law, allowing us to examine changes in the relative energies of its linear and quadratic components. We consider tasks of the form

$$f(\boldsymbol{x}) = f_1(\boldsymbol{x}) + f_2(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} + \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}, \qquad \boldsymbol{A} = \boldsymbol{A}^\top, \quad \operatorname{tr}(\boldsymbol{A}) = 0. \tag{14}$$

For a normalized task law $P_r$ in this family, define

$$p_r := \mathbb{E}_{P_r}[\|f_2\|_{L^2(\nu)}^2] = \frac{2\mathbb{E}_{P_r}[\|\boldsymbol{A}\|_{\mathrm{F}}^2]}{d(d+2)}, \qquad \mathcal{E}(P_r) = 1. \tag{15}$$

Thus $p_r$ is the fraction of task energy in quadratic interactions, and the linear fraction is $1 - p_r$. We use the following original task distribution throughout Subsections 4.2–4.4.

**Assumption 4.2** (Isotropic linear–quadratic tasks)**.** Let $p_0 \in (0, 1)$. Under $P_0$, the coefficients $\boldsymbol{w}$ and $\boldsymbol{A}$ in (14) are independent and uniformly distributed on spheres of fixed radii in $\mathbb{R}^d$ and the space of symmetric trace-free matrices, respectively, with $\|\boldsymbol{w}\|^2 = d(1 - p_0)$, $\|\boldsymbol{A}\|_{\mathrm{F}}^2 = \frac{1}{2}d(d+2)p_0$.

This assumption gives a symmetric setting in which the two components can be followed exactly. Uniform coefficient directions match the rotational symmetry of the inputs, while independence separates the sources of linear and quadratic variation. The trace-free condition removes the constant part of a quadratic form on the sphere. The fixed radii give every task linear energy $1 - p_0$, quadratic energy $p_0$, and bounded labels. Changes in the component ratio therefore arise from training and generation, rather than from variation in the initial component energies.

**Theorem 3** (ICL risk and quadratic proportion)**.** *Let $d \geq 2$ and suppose Assumption 4.2 holds. Use softmax attention* (4)*, population-context gradient flow* (11)*, and global unit-energy normalization. Set $T_r = T > 0$ at every generation. Then we have*

*1. The clean ICL risk satisfies*

$$\lim_{T\to\infty} R_0^{(\infty)}(\boldsymbol{\theta}_1(T)) = 0 \quad \textit{and} \quad \lim_{r\to\infty} R_0^{(\infty)}(\boldsymbol{\theta}_r(T)) \geq \frac{p_0}{8}. \tag{16}$$

*2. For every fixed $T > 0$, there are constants $C < \infty$ and $q_T \in (0, 1)$, depending only on $d, p_0, T$, such that*

$$p_r \leq C q_T^r \quad \textit{for all } r \geq 1.$$

*As $r \to \infty$, $\boldsymbol{\theta}_r(T)$ converges to the parameters obtained by training for time $T$ on the unit-energy purely linear task law.*

The two risk limits compare increasing the training time of the first model with increasing the number of generations at fixed training time, showing that accurately learning in one generation does not guarantee preservation through recursive generation. The limiting clean risk is at least $p_0/8$ uniformly over $d \geq 2$ and $T > 0$, despite unit task energy.

**Remark 4.1.** Proposition C.1 compares these trajectories with full-matrix SGD over a prescribed finite generation horizon. Sufficiently long contexts, small steps, large batches and small independent initialization perturbations make the parameters, generated laws and clean errors arbitrarily close with prescribed high probability.

### 4.3 Training Time Needed to Preserve Task Composition

Now we keep the model, task family, and normalization of Theorem 3, and now allow training time to increase with the generation index. How fast must it grow to preserve the quadratic component? The next theorem answers this question for polynomial schedules.

**Theorem 4** (Polynomial training-time threshold)**.** *Suppose Assumption 4.2 holds. Use softmax attention* (4)*, population-context gradient flow* (11)*, and global unit-energy normalization. Fix $A > 0$ and $\gamma \geq 0$, and set $T_r = Ar^\gamma$. Then the quadratic fraction $p_r$ in* (15) *and clean risk $R_0^{(\infty)}$ satisfy:*

*1. If $0 \leq \gamma \leq 3$, then for every finite $A > 0$, the quadratic fraction satisfies*

$$\lim_{r\to\infty} p_r = 0.$$

*2. If $\gamma > 3$, there exist constants $A_0, C \in (0, \infty)$, depending only on $d, p_0, \gamma$, such that for every $A \geq A_0$ and $r \geq 1$,*

$$p_r \geq p_0/2, \qquad R_0^{(\infty)}(\boldsymbol{\theta}_r(T_r)) \leq C(1 + T_r)^{-2/3}. \tag{17}$$

*In particular, $R_0^{(\infty)}(\boldsymbol{\theta}_r(T_r)) \to 0$ as $r \to \infty$.*

**Remark 4.2.** Appendix D shows that attention selectivity grows at most as $(1+T)^{1/3}$. Consequently, each generation decreases the logarithm of the quadratic-to-linear energy ratio $\ln \frac{p_r}{1-p_r}$ by at least a constant times $(1 + T_r)^{-1/3}$. A matching upper bound holds while $p_{r-1} \in [p_0/2, p_0]$. The threshold therefore comes from the series $\sum_{r=1}^{\infty} r^{-\gamma/3}$. At $\gamma = 3$, this is the divergent harmonic series so the accumulated decrease is unbounded for every finite $A$, forcing $p_r \to 0$. For $\gamma > 3$, the series converges, and a sufficiently large $A$ makes the total decrease small enough to preserve $p_r \geq p_0/2$.

**Finite generation contexts.** Theorem 4 uses population contexts. With finitely many seed examples, the sampling-error bound grows with attention selectivity. For one frozen population teacher, Proposition D.1 gives the sufficient condition

$$m \geq C_d \, \|f\|_\infty^2 \, \beta^{(d-1)/2} / \epsilon^2 \tag{18}$$

for mean-squared generation sampling error at most $\epsilon^2$. This one-generation upper bound describes an additional context-length cost; it does not establish a necessary sample complexity or control error over infinitely many generations.

### 4.4 Preserving Tasks with Two-Temperature Generation

Instead of increasing training time as in Theorem 4, we keep it fixed and modify generation: combine predictions at two attention temperatures to give both components the same multiplier, then normalize the generated law.

**Attention at temperature $\tau$.** For $\tau > 0$, divide the attention scores of (4) by $\tau$. Given a context $\mathcal{C}_m = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^m$, the resulting prediction is

$$F_{\boldsymbol{\theta}}(\mathcal{C}_m, \boldsymbol{x}; \tau) = c\frac{\sum_{i=1}^m y_i e^{s_i/\tau}}{\sum_{j=1}^m e^{s_j/\tau}}, \qquad s_i = (\boldsymbol{Q}\boldsymbol{x})^\top(\boldsymbol{K}\boldsymbol{x}_i + \boldsymbol{u}y_i). \tag{19}$$

At $\tau = 1$ this is the original predictor. Smaller temperatures concentrate the attention weights, while larger temperatures make them more uniform. Temperature changes affect only generation and the trained parameters $\boldsymbol{\theta}$ remain fixed.

**Two-temperature generation.** Algorithm 1 combines predictions from the same context, with temperatures and weights shared across tasks. We allow signed weights whose sum need not be one and normalize the resulting law.

**Algorithm 1** Two-temperature task generation

**Require:** Trained parameters $\boldsymbol{\theta}$; task law $P$; seed length $m$; distinct temperatures $\tau_1, \tau_2 > 0$; weights $\lambda_1, \lambda_2 \in \mathbb{R}$.

1: Draw $f \sim P$ and $\boldsymbol{x}_1, \dots, \boldsymbol{x}_m \overset{\text{i.i.d.}}{\sim} \nu$; set $\mathcal{C}_m = \{(\boldsymbol{x}_i, f(\boldsymbol{x}_i))\}_{i=1}^m$.

2: Keep $\mathcal{C}_m$ fixed and define $\widetilde{f}$ at every query $\boldsymbol{x}$ by
$\widetilde{f}(\boldsymbol{x}) = \lambda_1 F_{\boldsymbol{\theta}}(\mathcal{C}_m, \boldsymbol{x}; \tau_1) + \lambda_2 F_{\boldsymbol{\theta}}(\mathcal{C}_m, \boldsymbol{x}; \tau_2)$.

3: Let $\widetilde{P} = \mathrm{Law}(\widetilde{f})$ over the draws of $f$ and $\boldsymbol{x}_1, \dots, \boldsymbol{x}_m$.

4: **return** $\mathrm{Law}\left(\widetilde{f}/\sqrt{\mathcal{E}(\widetilde{P})}\right)$, provided $0 < \mathcal{E}(\widetilde{P}) < \infty$.

At generation $r$, the algorithm takes $\boldsymbol{\theta}_r, P_{r-1}$ and returns $P_r$. Training and clean risk use (4). Its population version replaces the context averages in (19) by integrals (Section 3).

**Lemma 4.1** (Population response coefficients)**.** *Let $\beta > 0$ and $\boldsymbol{\theta} = (\sqrt{\beta}\boldsymbol{I}_d, \sqrt{\beta}\boldsymbol{I}_d, \boldsymbol{0}, c)$. For any task $f = f_1 + f_2$ in* (14)*, the population-context prediction is*

$$F_{\boldsymbol{\theta},\infty}(f, \boldsymbol{x}) = c\big[a(\beta)f_1(\boldsymbol{x}) + b(\beta)f_2(\boldsymbol{x})\big],$$

*where the coefficients depend only on $d$ and $\beta$ and are given by*

$$a(\beta) = \frac{\int (\boldsymbol{x}^\top \boldsymbol{z}) e^{\beta \boldsymbol{x}^\top \boldsymbol{z}}\, d\nu(\boldsymbol{z})}{\int e^{\beta \boldsymbol{x}^\top \boldsymbol{z}}\, d\nu(\boldsymbol{z})}, \qquad b(\beta) = 1 - \frac{d\, a(\beta)}{\beta}. \tag{20}$$

*They satisfy $0 < b(\beta) < a(\beta) < 1$, and $b(\beta)/a(\beta)$ is strictly increasing on $(0, \infty)$.*

Lemma C.1 in Appendix C.2 proves these properties and provides the derivative bounds and asymptotic expansions used in Theorems 3 and 4.

**Choosing the temperatures and weights.** Under Assumption 4.2, population gradient flow gives $\boldsymbol{Q}_r = \boldsymbol{K}_r = \sqrt{\beta_r}\boldsymbol{I}_d$ and $\boldsymbol{u}_r = \boldsymbol{0}$ at the end of training (Lemma C.2). Thus $\beta_r \geq 1$ is the learned scale of the scores $\beta_r \boldsymbol{x}^\top \boldsymbol{z}$. Choose any two fixed target scales $0 < \gamma_1 < \gamma_2$, for example $\gamma_1 = 1$ and $\gamma_2 = 2$. Setting $\tau_{r,j} = \beta_r/\gamma_j$ makes the scores equal to $\gamma_j \boldsymbol{x}^\top \boldsymbol{z}$.

By Lemma 4.1, the prediction at target scale $\gamma_j$ is $c_r[a(\gamma_j)f_1 + b(\gamma_j)f_2]$, with $a, b$ given by (20). To make the weighted prediction equal to $c_r(f_1 + f_2)$, we require a multiplier of one for each component:

$$\tau_{r,j} = \frac{\beta_r}{\gamma_j} \quad (j = 1, 2), \qquad \begin{cases} \lambda_1 a(\gamma_1) + \lambda_2 a(\gamma_2) = 1, \\ \lambda_1 b(\gamma_1) + \lambda_2 b(\gamma_2) = 1. \end{cases} \tag{21}$$

Writing $a_j = a(\gamma_j)$ and $b_j = b(\gamma_j)$, the solution is

$$\lambda_1 = \frac{b_2 - a_2}{a_1 b_2 - a_2 b_1}, \qquad \lambda_2 = \frac{a_1 - b_1}{a_1 b_2 - a_2 b_1}. \tag{22}$$

The denominator is positive and $\lambda_1 < 0 < \lambda_2$ (Lemma E.1). We adopt this choice and derive the following theorem.

**Theorem 5** (Exact preservation by two-temperature generation)**.** *Suppose Assumption 4.2 holds. Use softmax attention* (4) *and population-context gradient flow* (11)*. Fix $T > 0$ and $0 < \gamma_1 < \gamma_2 < \infty$. At every generation, set $T_r = T$ and generate $P_r$ by the population version of Algorithm 1, with* (21)*. Then:*

*1. For each sampled $f$, the normalized generated function equals $f$. Consequently,*

$$P_r = P_0, \qquad p_r = p_0, \qquad r \geq 1. \tag{23}$$

*2. The training trajectories are identical across generations:*

$$\boldsymbol{\theta}_r(t) = \boldsymbol{\theta}_1(t), \qquad r \geq 1, \quad 0 \leq t \leq T. \tag{24}$$

*3. The clean ICL risk of the original predictor* (4) *satisfies*

$$\sup_{r \geq 1} R_0^{(\infty)}(\boldsymbol{\theta}_r(T)) = R_0^{(\infty)}(\boldsymbol{\theta}_1(T)) \leq C_{d,p_0}(1+T)^{-2/3}, \tag{25}$$

*where $C_{d,p_0} < \infty$ depends only on $d, p_0$.*

The exact identity depends on population contexts and the response matching in (21). Since each attention evaluation attenuates the quadratic component more strongly, a nonnegative combination cannot remove the response gap (Proposition E.1), which explains the signed weights. Task preservation concerns the generation rule, whereas (25) evaluates the original trained predictor, whose error still depends on $T$.

**Remark 4.3.** For finite seed contexts, Proposition E.2 bounds the mean-square error before normalization by a quantity proportional to $1/m$, with explicit dependence on the temperatures and weights. The bound allows both evaluations to share the same seed. Proposition E.3 establishes stability under normalization and approximation by full-matrix SGD over any fixed generation horizon.

## 5 Experiments

We test whether the distinction between task scale and composition, and the benefit of response matching, persist under finite-context SGD with dense attention.

**Experimental setup.** Inputs follow the distribution $\nu$. We set linear tasks with $d = 8$ and $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_8)$. We then train linear models (5) from zero with 500 SGD updates, $B = 512$, $\eta = 1$, and $n = m \in \{16, 64, 128\}$. Softmax tasks follow Assumption 4.2 with $d = 3$, initially $p_0 = 1/2$ unless varied. The model has one layer and head, dense query, key, value, and output matrices, and no MLP. Scores are $(\boldsymbol{W}_Q \boldsymbol{h}_q)^\top \boldsymbol{W}_K \boldsymbol{h}_i / 2$. SGD minimizes half-MSE with $B = 128$, $\eta = 0.03$, and no momentum or weight decay. We use $n = 128$, $m = 256$ (512 for two-temperature comparisons), and fresh Xavier initialization unless specified below. Recursion retains complete functions; normalization uses one estimated population scale.

We evaluate $P_0$ MSE, equal to $R_0^{(n)}$, before generation correction: exactly for linear attention and on independent episodes for softmax. Means and 95% pointwise Student-$t$ intervals include all 100 linear or 20 softmax runs. Appendix F details numerical integration and sampling.

**Energy decay and optimization.** Linear attention initially maintains near-optimal MSE despite declining energy (Figure 2(a,b)): energy decay alone does not determine ICL performance. Later degradation at fixed budgets highlights the complete-optimization assumption in Theorem 1. Softmax attention likewise degrades as energy falls (c,d), qualitatively matching the failure mechanism studied in Theorem 2, without verifying its amplitude condition or asymptotic rate.

**Normalization does not preserve task composition.** Even after normalization, quadratic-to-linear energy declines and recursive MSE rises above fresh-task controls (Figure 4). Thus preserving scale is insufficient, consistent with Theorem 3. Constant components also grow, taking these dense-model trajectories outside the theoretical family. Increasing training budgets slows component loss (Figure 5), supporting the role of training time in Theorem 4. All schedules $K_r = 8{,}000 r^\gamma$, $\gamma \in \{0, 2, 3, 4\}$, still lose quadratic fraction over five generations; this comparison does not identify the cubic threshold.

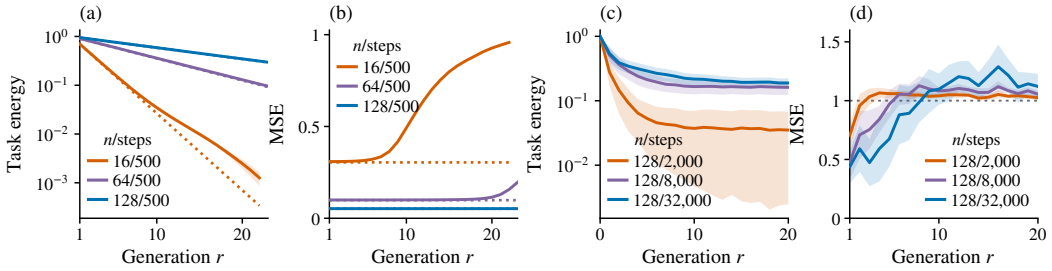


Figure 2: **Task energy and original-task MSE.** (a,b) Linear attention at different context lengths. (c,d) Softmax attention at different per-generation training budgets.

**Response matching improves task retention.** Figure 3 compares response matching, ordinary normalized generation, and $10\%$ original-task refresh in training and generation, each with $8{,}000$ updates. We see that response matching retains substantially more quadratic energy and improves endpoint MSE over both controls in every run. This supports Theorem 5's principle: correcting unequal component responses helps successive models relearn the original tasks at fixed training budgets. Preservation remains approximate, including one retained run with an anisotropic response failure and the extra attention view also increases generation cost.

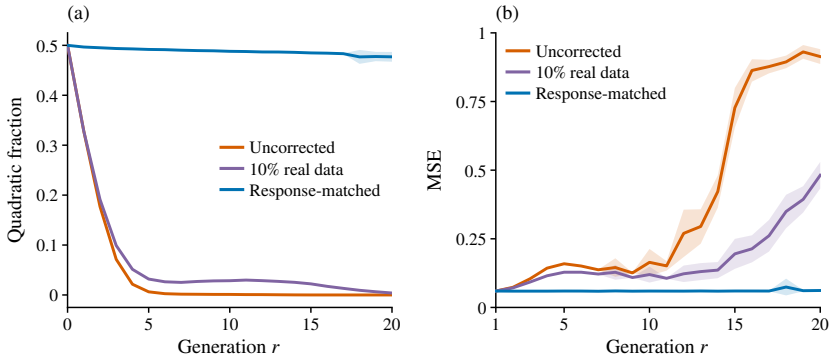


Figure 3: **Two-temperature generation with dense attention.** Quadratic-energy fraction (left) and original-task MSE (right) over 20 generations.

## 6 Discussion and Conclusion

Learning a task family does not guarantee that a model can generate data from which its successors will relearn it. Shrinking labels can leave the optimal linear-attention predictor unchanged yet impede finite-budget softmax training. Even after energy normalization, unequal component attenuation can weaken the signal needed for subsequent learning. For linear–quadratic tasks under population-context gradient flow, this produces a persistent clean-risk gap at every fixed per-generation training time. Sufficiently increasing training time can preserve task structure, but generation correction offers another route: calibrated signed two-temperature views, followed by normalization, preserve the task law and each cold-start learner's clean risk at fixed training time in this setting, without real-data refresh or parameter inheritance.

Our experiments show improved retention with finite contexts, but not universal preservation. Anisotropy can invalidate response matching, signed weights can amplify sampling errors, and the additional attention evaluation increases generation cost. Extending correction to more general learned operators while controlling accumulated generation error remains an important direction. Ultimately, recursive generation should be judged by what subsequent models can learn from its outputs, not by label energy alone.

### AI USE STATEMENT

In this work, we used generative AI tools to assist with experimental methodology and the interpretation and presentation of experimental results. Additionally, we used these tools to identify and summarize relevant literature for the related work section, improve the writing and organization of the manuscript, and prepare scientific figures and LaTeX source. The research direction and core ideas originated with the authors. AI assistance also included revising the presentation of theorem statements and proofs, and organizing and checking mathematical arguments. The reported experimental measurements come from recorded GPU training runs, with the complete seed cohorts specified in Appendix F. We take responsibility for the final content of this work, including all text, claims, citations, and artifacts prepared with AI assistance.

### REPRODUCIBILITY STATEMENT

The model, training procedures, and generation rules are defined in Section 3. Assumptions accompany the main results, whose proofs appear in Appendices A–E. Appendix F specifies the task sampling, initialization, optimization budgets, seed cohorts, normalization, and evaluation procedures for the experiments.

### ETHICS STATEMENT

This study uses synthetic regression tasks and no human-subject or personal data. Its guarantees concern the stated models and training regimes; they do not establish the safety, fairness, or reliability of language-model outputs. Applications to real data require separate evaluation.

## REFERENCES

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36:45614–45650, 2023.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2022.

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. Self-consuming generative models go mad. In *International Conference on Learning Representations*, volume 2024, pp. 53581–53608, 2024.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *Advances in neural information processing systems*, 36:57125–57211, 2023.

Etienne Boursier and Claire Boyer. Softmax as linear attention in the large-prompt regime: a measure-based perspective. *arXiv preprint arXiv:2512.11784*, 2025.

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.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. *arXiv preprint arXiv:2402.19442*, 2024.

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.

Liam Collins, Advait Parulekar, Aryan Mokhtari, Sujay Sanghavi, and Sanjay Shakkottai. In-context learning with transformers: Softmax attention adapts to function lipschitzness. *Advances in Neural Information Processing Systems*, 37:92638–92696, 2024.

Elvis Dohmatob, Yunzhen Feng, and Julia Kempe. Model collapse demystified: The case of regression. *Advances in Neural Information Processing Systems*, 37:46979–47013, 2024.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

Shi Fu, Yingjie Wang, Yuzhu Chen, Xinmei Tian, and Dacheng Tao. A theoretical perspective: How to prevent model collapse in self-consuming training loops. In *International Conference on Learning Representations*, volume 2025, pp. 93254–93287, 2025.

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.

Khashayar Gatmiry, Nikunj Saunshi, Sashank J Reddi, Stefanie Jegelka, and Sanjiv Kumar. Can looped transformers learn to implement multi-step gradient descent for in-context learning? *arXiv preprint arXiv:2410.08292*, 2024.

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

Jianliang He, Xintian Pan, Siyu Chen, and Zhuoran Yang. In-context linear regression demystified: Training dynamics and mechanistic interpretability of multi-head softmax attention. *arXiv preprint arXiv:2503.12734*, 2025.

Yinpeng Liu, Jiawei Liu, Xiang Shi, Qikai Cheng, Yong Huang, and Wei Lu. Let's learn step by step: Enhancing in-context learning ability with curriculum learning. *arXiv preprint arXiv:2402.10738*, 2024.

Hossein Mobahi, Mehrdad Farajtabar, and Peter Bartlett. Self-distillation amplifies regularization in hilbert space. *Advances in Neural Information Processing Systems*, 33:3351–3361, 2020.

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

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.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

Zhijie Wang, Bo Jiang, and Shuai Li. In-context learning on function classes unveiled for transformers. In *Forty-first International Conference on Machine Learning*, 2024.

Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter Bartlett. How many pretraining tasks are needed for in-context learning of linear regression? In *International Conference on Learning Representations*, volume 2024, pp. 56587–56641, 2024.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

# Contents

**1 Introduction** **1**

**2 Related work** **2**

**3 Preliminaries** **2**

**4 Main Results** **4**
4.1 Task Energy and ICL Risk . . . . . 4
4.2 Task Composition under Fixed Training Time . . . . . 5
4.3 Training Time Needed to Preserve Task Composition . . . . . 6
4.4 Preserving Tasks with Two-Temperature Generation . . . . . 6

**5 Experiments** **8**

**6 Discussion and conclusion** **9**

**A Proof of Theorem 1** **13**
A.1 Proof of the theorem . . . . . 13
A.2 Context moments and the population optimizer . . . . . 14
A.3 An auxiliary SGD convergence result . . . . . 16

**B Proof of Theorem 2** **19**
B.1 Proof of the theorem . . . . . 19
B.2 Prediction bounds for one generation . . . . . 20

**C Proof of Theorem 3** **22**
C.1 Proof of the theorem . . . . . 22
C.2 Full response lemma and supporting results . . . . . 23
C.3 Finite-horizon approximation by SGD . . . . . 32

**D Proof of Theorem 4** **41**
D.1 Proof of Theorem 4 . . . . . 41
D.2 Proofs of the supporting lemmas . . . . . 43
D.3 Extension to a finite generation context . . . . . 47

**E Proof of Theorem 5** **49**
E.1 Proof of Theorem 5 . . . . . 49
E.2 Why response matching requires signed weights . . . . . 50
E.3 Finite-context sampling . . . . . 50
E.4 A fixed-horizon extension to full-matrix SGD . . . . . 51

**F Experimental Details** **54**
F.1 Models and evaluation . . . . . 54
F.2 Generation and training comparisons . . . . . 55

**G Attention parameterization** **56**

# A PROOF OF THEOREM 1

The argument has two parts. Lemma A.1 computes the population optimizer from the coefficient covariance. Applying this formula across generations gives the energy and risk identities in Theorem 1. Proposition A.1 then supplies the stated SGD convergence guarantee.

## A.1 PROOF OF THE THEOREM

*Proof of Theorem 1.* Throughout, $\nu$ is uniform on $\mathbb{S}^{d-1}$, and generation uses $n$ seed inputs, as does training. We use the context-moment matrix $\boldsymbol{H}_n(\boldsymbol{C})$ in Lemma A.1. The lemma applies to every centered linear-task law with nonzero coefficient covariance, so the generated laws need not remain Gaussian.

**1. Covariance of the generated tasks.** We first determine how generation transforms the coefficient covariance. Write $f_{r-1}(\boldsymbol{x}) = \boldsymbol{w}_{r-1}^\top \boldsymbol{x}$, where $\mathbb{E}[\boldsymbol{w}_{r-1}] = \boldsymbol{0}$ and $\boldsymbol{C}_{r-1} = \mathbb{E}[\boldsymbol{w}_{r-1}\boldsymbol{w}_{r-1}^\top]$. Draw the $n$ seed inputs $\boldsymbol{z}_{r,1}, \dots, \boldsymbol{z}_{r,n}$ independently of $\boldsymbol{w}_{r-1}$ and put

$$\boldsymbol{S}_r := \frac{d}{n} \sum_{i=1}^{n} \boldsymbol{z}_{r,i} \boldsymbol{z}_{r,i}^\top. \tag{A.1}$$

The generation rule (10) and the predictor (5) give

$$\begin{aligned} f_r(\boldsymbol{x}) &= F_{\boldsymbol{\Gamma}_r}\left(\{(\boldsymbol{z}_{r,i}, f_{r-1}(\boldsymbol{z}_{r,i}))\}_{i=1}^n, \boldsymbol{x}\right) \\ &= \boldsymbol{x}^\top \boldsymbol{\Gamma}_r \left( \frac{d}{n} \sum_{i=1}^{n} \boldsymbol{z}_{r,i} \boldsymbol{z}_{r,i}^\top \boldsymbol{w}_{r-1} \right) = \boldsymbol{x}^\top \boldsymbol{\Gamma}_r \boldsymbol{S}_r \boldsymbol{w}_{r-1}. \end{aligned} \tag{A.2}$$

Thus $\boldsymbol{w}_r = \boldsymbol{\Gamma}_r \boldsymbol{S}_r \boldsymbol{w}_{r-1}$. Independence of $\boldsymbol{S}_r$ and $\boldsymbol{w}_{r-1}$ preserves the zero mean:

$$\mathbb{E}[\boldsymbol{w}_r] = \boldsymbol{\Gamma}_r \mathbb{E}[\boldsymbol{S}_r] \mathbb{E}[\boldsymbol{w}_{r-1}] = \boldsymbol{0}. \tag{A.3}$$

Lemma A.1, applied to this seed statistic, gives the covariance recursion

$$\begin{aligned} \boldsymbol{C}_r := \mathbb{E}[\boldsymbol{w}_r \boldsymbol{w}_r^\top] &= \boldsymbol{\Gamma}_r \mathbb{E}\left[ \boldsymbol{S}_r \mathbb{E}[\boldsymbol{w}_{r-1}\boldsymbol{w}_{r-1}^\top \mid \boldsymbol{S}_r] \boldsymbol{S}_r \right] \boldsymbol{\Gamma}_r^\top \\ &= \boldsymbol{\Gamma}_r \mathbb{E}[\boldsymbol{S}_r \boldsymbol{C}_{r-1} \boldsymbol{S}_r] \boldsymbol{\Gamma}_r^\top = \boldsymbol{\Gamma}_r \boldsymbol{H}_n(\boldsymbol{C}_{r-1}) \boldsymbol{\Gamma}_r^\top. \end{aligned} \tag{A.4}$$

This covariance also determines task energy. For a linear-task law with coefficient covariance $\boldsymbol{C}$,

$$\mathcal{E}(P) = \mathbb{E}[(\boldsymbol{w}^\top \boldsymbol{x}_q)^2] = \mathbb{E}[\boldsymbol{w}^\top \mathbb{E}[\boldsymbol{x}_q \boldsymbol{x}_q^\top \mid \boldsymbol{w}] \boldsymbol{w}] = \frac{1}{d} \mathbb{E}[\boldsymbol{w}^\top \boldsymbol{w}] = \frac{\operatorname{tr}(\boldsymbol{C})}{d}. \tag{A.5}$$

**2. The isotropic recursion and clean risk.** In the isotropic setting of Theorem 1, the covariance recursion reduces to a scalar recursion. For a law $P$ with covariance $\boldsymbol{C} = \mathcal{E}\boldsymbol{I}_d$, where $\mathcal{E} = \mathcal{E}(P) > 0$, the context-moment identity (A.15) becomes

$$\begin{aligned} \boldsymbol{H}_n(\mathcal{E}\boldsymbol{I}_d) &= \left(1 + \frac{d-2}{n(d+2)}\right) \mathcal{E}\boldsymbol{I}_d + \frac{d}{n(d+2)} \operatorname{tr}(\mathcal{E}\boldsymbol{I}_d) \boldsymbol{I}_d \\ &= \mathcal{E}\left(1 + \frac{d-2}{n(d+2)} + \frac{d^2}{n(d+2)}\right) \boldsymbol{I}_d = \mathcal{E}\frac{n+d-1}{n} \boldsymbol{I}_d. \end{aligned} \tag{A.6}$$

Substituting this expression into the population optimum (A.18) gives

$$\boldsymbol{\Gamma}^*(\mathcal{E}\boldsymbol{I}_d) = (\mathcal{E}\boldsymbol{I}_d) \boldsymbol{H}_n(\mathcal{E}\boldsymbol{I}_d)^{-1} = (\mathcal{E}\boldsymbol{I}_d) \left[ \mathcal{E}\frac{n+d-1}{n} \boldsymbol{I}_d \right]^{-1} = \frac{n}{n+d-1} \boldsymbol{I}_d. \tag{A.7}$$

The common task scale $\mathcal{E}$ cancels from the optimizer. Set $q = n/(n+d-1) \in (0,1)$. We claim that $\boldsymbol{C}_r = q^r \boldsymbol{I}_d$ and $\boldsymbol{\Gamma}_r = q\boldsymbol{I}_d$. Starting from $\boldsymbol{C}_0 = \boldsymbol{I}_d$, suppose that $\boldsymbol{C}_{r-1} = q^{r-1}\boldsymbol{I}_d$. The optimizer is then $\boldsymbol{\Gamma}_r = q\boldsymbol{I}_d$, and (A.4) gives the induction step:

$$\boldsymbol{C}_r = \boldsymbol{\Gamma}_r \boldsymbol{H}_n(\boldsymbol{C}_{r-1}) \boldsymbol{\Gamma}_r^\top = (q\boldsymbol{I}_d) \left[ q^{r-1} \frac{n+d-1}{n} \boldsymbol{I}_d \right] (q\boldsymbol{I}_d) = q^r \boldsymbol{I}_d. \tag{A.8}$$

Since $q^{r-1} > 0$ at each finite generation, Lemma A.1 guarantees that this population optimum is unique. Applying (A.5) to $P_r$ gives

$$\mathcal{E}(P_r) = \frac{\operatorname{tr} \boldsymbol{C}_r}{d} = \frac{\operatorname{tr}(q^r \boldsymbol{I}_d)}{d} = q^r \xrightarrow[r\to\infty]{} 0. \tag{A.9}$$

Clean risk is evaluated on the original law, whose covariance is $\boldsymbol{C}_0 = \boldsymbol{I}_d$. Thus $\mathcal{E}(P_0) = 1$ and $\boldsymbol{H}_n(\boldsymbol{I}_d) = q^{-1}\boldsymbol{I}_d$, and the loss formula (A.17) gives

$$\begin{aligned} R_0^{(n)}(\boldsymbol{\Gamma}_r) &= \frac{2L_{\boldsymbol{I}_d}^{(n)}(q\boldsymbol{I}_d)}{\mathcal{E}(P_0)} = \frac{\operatorname{tr} \boldsymbol{I}_d - 2\operatorname{tr}(q\boldsymbol{I}_d) + \operatorname{tr}((q\boldsymbol{I}_d)(q^{-1}\boldsymbol{I}_d)(q\boldsymbol{I}_d)^\top)}{\operatorname{tr} \boldsymbol{I}_d} \\ &= \frac{d - 2qd + q^2 q^{-1} d}{d} = 1 - q = \frac{d-1}{n+d-1} < 1. \end{aligned} \tag{A.10}$$

**3. Verification of the SGD assumptions.** To apply Proposition A.1 at generation $r$, we need $\mathbb{E}[\boldsymbol{w}_{r-1}] = \boldsymbol{0}$, $\operatorname{tr} \boldsymbol{C}_{r-1} > 0$, and $\mathbb{E}[\|\boldsymbol{w}_{r-1}\|^4] < \infty$. The first two properties follow from $\boldsymbol{C}_{r-1} = q^{r-1}\boldsymbol{I}_d$ and the mean calculation above. For the third, use $\|\boldsymbol{S}_r\|_{\mathrm{op}} \le d$ from (A.16) and $\boldsymbol{\Gamma}_r = q\boldsymbol{I}_d$ to obtain

$$\begin{aligned} \mathbb{E}[\|\boldsymbol{w}_r\|^4] &= \mathbb{E}[\|q\boldsymbol{S}_r\boldsymbol{w}_{r-1}\|^4] \le q^4 \mathbb{E}[\|\boldsymbol{S}_r\|_{\mathrm{op}}^4 \|\boldsymbol{w}_{r-1}\|^4] \\ &\le (qd)^4 \mathbb{E}[\|\boldsymbol{w}_{r-1}\|^4] \le (qd)^{4r} \mathbb{E}[\|\boldsymbol{w}_0\|^4] = (qd)^{4r} d(d+2) < \infty. \end{aligned} \tag{A.11}$$

The last equality uses $\boldsymbol{w}_0 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$. Thus the fourth moment is finite at every fixed generation. The step-size assumptions imply $\eta_k \to 0$, so the upper bound in Proposition A.1 holds after finitely many updates. The same proof applies from that point, since a finite prefix preserves a finite second moment of the iterate. The proposition therefore gives mean-square and almost-sure convergence of SGD to the unique optimizer $\boldsymbol{\Gamma}_r = q\boldsymbol{I}_d$ at each fixed generation, under the stated step-size conditions. Together with the energy and risk identities above, this proves Theorem 1. □

### A.2 Context moments and the population optimizer

The following calculation is used both in the theorem and in the SGD analysis. It depends on the coefficient covariance, without imposing a Gaussian task law.

**Lemma A.1** (Population loss for linear attention)**.** *Let $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$, where $\mathbb{E}[\boldsymbol{w}] = \boldsymbol{0}$, $\boldsymbol{C} = \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top]$, and $0 < \operatorname{tr} \boldsymbol{C} < \infty$. Draw the query and $n$ context inputs independently from the uniform distribution on $\mathbb{S}^{d-1}$, independently of $\boldsymbol{w}$. Define*

$$\boldsymbol{S}_n = \frac{d}{n}\sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\top, \qquad \boldsymbol{s} = \boldsymbol{S}_n \boldsymbol{w}, \qquad \boldsymbol{H}_n(\boldsymbol{C}) = \mathbb{E}[\boldsymbol{s}\boldsymbol{s}^\top].$$

*Write $L_{\boldsymbol{C}}^{(n)}$ for the population loss of the linear-attention model* (5) *under this task law. The context moments are given by* (A.14) *and* (A.15)*. The loss has the unique minimizer*

$$\boldsymbol{\Gamma}^* = \boldsymbol{C}\boldsymbol{H}_n(\boldsymbol{C})^{-1},$$

*and its excess loss is given by* (A.19)*.*

*Proof.* We first compute the context moments, then substitute them into the squared loss.

**Context moments.** The second and fourth moments of the inputs determine the population loss (7).

For the coordinate calculations below, $\boldsymbol{x}_i$ denotes coordinate $i$ of a single input vector. Sign symmetry makes $\mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_j] = 0$ for $i \ne j$. Equal coordinate distributions and $\sum_i \boldsymbol{x}_i^2 = 1$ then give $\mathbb{E}[\boldsymbol{x}_i^2] = \frac{1}{d}$ and $\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \frac{1}{d}\boldsymbol{I}_d$. For fourth moments, let $a = \mathbb{E}[\boldsymbol{x}_1^4]$ and $b = \mathbb{E}[\boldsymbol{x}_1^2 \boldsymbol{x}_2^2]$. Rotation in the first two coordinates implies

$$a = \mathbb{E}[\left(\frac{\boldsymbol{x}_1 + \boldsymbol{x}_2}{\sqrt{2}}\right)^4] = \frac{1}{4}\big(\mathbb{E}[\boldsymbol{x}_1^4] + 6\mathbb{E}[\boldsymbol{x}_1^2\boldsymbol{x}_2^2] + \mathbb{E}[\boldsymbol{x}_2^4]\big) = \frac{1}{2}a + \frac{3}{2}b.$$

The odd terms vanish by sign symmetry, so $a = 3b$. The unit-norm constraint also gives

$$1 = \mathbb{E}[\left(\sum_{i=1}^{d} \boldsymbol{x}_i^2\right)^2] = \sum_{i=1}^{d} \mathbb{E}[\boldsymbol{x}_i^4] + 2 \sum_{1 \le i < j \le d} \mathbb{E}[\boldsymbol{x}_i^2 \boldsymbol{x}_j^2] = da + d(d-1)b = d(d+2)b.$$

Consequently, $b = 1/[d(d+2)]$ and $a = 3/[d(d+2)]$. Together with sign symmetry, these identities give the fourth-moment formula, where $\delta_{ij} = 1$ if $i = j$ and 0 otherwise:

$$\mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_j \boldsymbol{x}_k \boldsymbol{x}_\ell] = \frac{\delta_{ij}\delta_{k\ell} + \delta_{ik}\delta_{j\ell} + \delta_{i\ell}\delta_{jk}}{d(d+2)}. \tag{A.12}$$

For any symmetric matrix $\boldsymbol{C}$, contraction with the fourth-moment identity (A.12) yields

$$\begin{aligned}\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{C} \boldsymbol{x}\boldsymbol{x}^\top]_{ij} &= \sum_{k,\ell} \boldsymbol{C}_{k\ell} \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_k \boldsymbol{x}_\ell \boldsymbol{x}_j] = \frac{\sum_{k,\ell} \boldsymbol{C}_{k\ell}(\delta_{ik}\delta_{\ell j} + \delta_{i\ell}\delta_{kj} + \delta_{ij}\delta_{k\ell})}{d(d+2)} \\ &= \frac{\boldsymbol{C}_{ij} + \boldsymbol{C}_{ji} + \delta_{ij}\sum_k \boldsymbol{C}_{kk}}{d(d+2)} = \frac{2\boldsymbol{C}_{ij} + \delta_{ij}\operatorname{tr}(\boldsymbol{C})}{d(d+2)}.\end{aligned} \tag{A.13}$$

We now return to context inputs, with $\boldsymbol{x}_i$ denoting input $i$ and $\boldsymbol{s} = \boldsymbol{S}_n \boldsymbol{w}$. Since $\mathbb{E}[\boldsymbol{S}_n] = \frac{d}{n}\sum_{i=1}^{n} \frac{\boldsymbol{I}_d}{d} = \boldsymbol{I}_d$, independence of the coefficient and the inputs gives

$$\mathbb{E}[\boldsymbol{s}\boldsymbol{w}^\top] = \mathbb{E}\left[\boldsymbol{S}_n \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top \mid \boldsymbol{S}_n]\right] = \mathbb{E}[\boldsymbol{S}_n \boldsymbol{C}] = \boldsymbol{C}. \tag{A.14}$$

For the second moment, separate the diagonal terms from the independent pairs and apply (A.13):

$$\begin{aligned}\mathbb{E}[\boldsymbol{s}\boldsymbol{s}^\top] = \mathbb{E}[\boldsymbol{S}_n \boldsymbol{C} \boldsymbol{S}_n] &= \frac{d^2}{n^2} \sum_{i=1}^{n}\sum_{j=1}^{n} \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{C} \boldsymbol{x}_j \boldsymbol{x}_j^\top] \\ &= \frac{d^2}{n^2}\left\{\sum_{i=1}^{n} \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{C} \boldsymbol{x}_i \boldsymbol{x}_i^\top] + \sum_{i=1}^{n}\sum_{\substack{j=1 \\ j\neq i}}^{n} \mathbb{E}[\boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{C} \boldsymbol{x}_j \boldsymbol{x}_j^\top]\right\} \\ &= \frac{d^2}{n^2}\left\{n\frac{2\boldsymbol{C} + \operatorname{tr}(\boldsymbol{C})\boldsymbol{I}_d}{d(d+2)} + n(n-1)\frac{\boldsymbol{C}}{d^2}\right\} \\ &= \left(1 + \frac{d-2}{n(d+2)}\right)\boldsymbol{C} + \frac{d}{n(d+2)}\operatorname{tr}(\boldsymbol{C})\boldsymbol{I}_d =: \boldsymbol{H}_n(\boldsymbol{C}).\end{aligned} \tag{A.15}$$

Here the $i = j$ terms use the fourth-moment identity, while the $i \neq j$ terms factor by input independence. We will also use the following deterministic bound, which follows from the unit norm of each input:

$$\|\boldsymbol{S}_n\|_{\mathrm{op}} \le \frac{d}{n}\sum_{i=1}^{n} \left\|\boldsymbol{x}_i \boldsymbol{x}_i^\top\right\|_{\mathrm{op}} = d. \tag{A.16}$$

**Population loss.** Averaging over the independent query converts prediction error into coefficient-estimation error. The two context moments (A.14) and (A.15) then express the population loss (7) as a quadratic function of the trainable matrix:

$$\begin{aligned}L_{\boldsymbol{C}}^{(n)}(\boldsymbol{\Gamma}) &= \frac{1}{2}\mathbb{E}\left[(\boldsymbol{x}_q^\top(\boldsymbol{\Gamma}\boldsymbol{s} - \boldsymbol{w}))^2\right] = \frac{1}{2d}\mathbb{E}[\|\boldsymbol{\Gamma}\boldsymbol{s} - \boldsymbol{w}\|^2] \\ &= \frac{1}{2d}\mathbb{E}\left[\boldsymbol{s}^\top \boldsymbol{\Gamma}^\top \boldsymbol{\Gamma}\boldsymbol{s} - 2\boldsymbol{w}^\top \boldsymbol{\Gamma}\boldsymbol{s} + \boldsymbol{w}^\top \boldsymbol{w}\right] \\ &= \frac{1}{2d}\left\{\operatorname{tr}(\boldsymbol{\Gamma}\boldsymbol{H}_n(\boldsymbol{C})\boldsymbol{\Gamma}^\top) - 2\operatorname{tr}(\boldsymbol{\Gamma}\boldsymbol{C}) + \operatorname{tr}\boldsymbol{C}\right\}.\end{aligned} \tag{A.17}$$

The matrix $\boldsymbol{H}_n(\boldsymbol{C})$ is positive definite whenever $\operatorname{tr}\boldsymbol{C} > 0$, since $\boldsymbol{H}_n(\boldsymbol{C}) \succeq d\operatorname{tr}(\boldsymbol{C})\boldsymbol{I}_d/[n(d+2)] \succ 0$. The stationary point of this quadratic loss is therefore

$$\nabla_{\boldsymbol{\Gamma}} L_{\boldsymbol{C}}^{(n)}(\boldsymbol{\Gamma}) = \frac{1}{d}\big(\boldsymbol{\Gamma}\boldsymbol{H}_n(\boldsymbol{C}) - \boldsymbol{C}\big), \qquad \boldsymbol{\Gamma}^* = \boldsymbol{C}\boldsymbol{H}_n(\boldsymbol{C})^{-1}. \tag{A.18}$$

For $\Delta\boldsymbol{\Gamma} = \boldsymbol{\Gamma} - \boldsymbol{\Gamma}^*$, the identity $\boldsymbol{H}_n(\boldsymbol{C})(\boldsymbol{\Gamma}^*)^\top = \boldsymbol{C}$ cancels the mixed terms:

$$\begin{aligned} 2d\{L_{\boldsymbol{C}}^{(n)}(\boldsymbol{\Gamma}) - L_{\boldsymbol{C}}^{(n)}(\boldsymbol{\Gamma}^*)\} &= \operatorname{tr}\big(\Delta\boldsymbol{\Gamma}\boldsymbol{H}_n(\boldsymbol{C})\Delta\boldsymbol{\Gamma}^\top\big) \\ &\quad + 2\operatorname{tr}\big(\Delta\boldsymbol{\Gamma}\boldsymbol{H}_n(\boldsymbol{C})(\boldsymbol{\Gamma}^*)^\top\big) - 2\operatorname{tr}(\Delta\boldsymbol{\Gamma}\boldsymbol{C}) \\ &= \operatorname{tr}\big(\Delta\boldsymbol{\Gamma}\boldsymbol{H}_n(\boldsymbol{C})\Delta\boldsymbol{\Gamma}^\top\big). \end{aligned} \tag{A.19}$$

Positive definiteness makes the loss difference strictly positive unless $\Delta\boldsymbol{\Gamma} = \mathbf{0}$. Thus the stationary point is the unique population optimizer.

□

### A.3 An auxiliary SGD convergence result

We next verify that decreasing-step-size SGD reaches the population optimizer characterized in Lemma A.1. The task law is fixed throughout this proposition; its application across generations was checked in the theorem proof above.

**Proposition A.1** (SGD convergence for linear attention)**.** *Fix $d \geq 2$ and positive integers $n, B$. Let $P$ be the law of $f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$, where*

$$\mathbb{E}[\boldsymbol{w}] = \mathbf{0}, \qquad \boldsymbol{C} = \mathbb{E}[\boldsymbol{w}\boldsymbol{w}^\top], \qquad \operatorname{tr}\boldsymbol{C} > 0, \qquad \mathbb{E}[\|\boldsymbol{w}\|^4] < \infty.$$

*In each episode, draw $n$ context inputs and one query independently from the uniform distribution on $\mathbb{S}^{d-1}$, independently of $\boldsymbol{w}$. Train only $\boldsymbol{\Gamma}$ in* (5) *by SGD* (9)*, using $B$ fresh independent episodes per update and a deterministic initialization $\boldsymbol{\Gamma}_0$. There exists $\eta_{\max} > 0$, depending only on $P, d, n, B$, such that any deterministic step-size sequence satisfying*

$$0 < \eta_k \leq \eta_{\max}, \qquad \sum_{k=0}^{\infty} \eta_k = \infty, \qquad \sum_{k=0}^{\infty} \eta_k^2 < \infty$$

*gives convergence to the unique population optimizer $\boldsymbol{\Gamma}^* = \boldsymbol{C}\boldsymbol{H}_n(\boldsymbol{C})^{-1}$:*

$$\mathbb{E}[\|\boldsymbol{\Gamma}_k - \boldsymbol{\Gamma}^*\|_{\mathrm{F}}^2] \xrightarrow[k\to\infty]{} 0, \qquad \boldsymbol{\Gamma}_k \xrightarrow[k\to\infty]{\text{a.s.}} \boldsymbol{\Gamma}^*. \tag{A.20}$$

*Here $\boldsymbol{H}_n(\boldsymbol{C})$ is the context second-moment matrix in* (A.15)*.*

*Proof.* We express SGD as a linear recursion around the population optimizer, bound its conditional second moment, and use the step-size assumptions to prove both modes of convergence.

**1. Error recursion.** Lemma A.1 gives the two context moments and the unique population optimizer $\boldsymbol{\Gamma}^*$. For one episode, the sample gradient is

$$\nabla_{\boldsymbol{\Gamma}} \tfrac{1}{2}(\boldsymbol{x}_q^\top \boldsymbol{\Gamma}\boldsymbol{s} - y_q)^2 = (\boldsymbol{x}_q^\top \boldsymbol{\Gamma}\boldsymbol{s} - y_q)\boldsymbol{x}_q\boldsymbol{s}^\top. \tag{A.21}$$

Column-vectorize $\boldsymbol{\Gamma}$ as $\boldsymbol{a}$, and let $\boldsymbol{z} = \boldsymbol{s} \otimes \boldsymbol{x}_q$, so $\boldsymbol{z}^\top \boldsymbol{a} = \boldsymbol{x}_q^\top \boldsymbol{\Gamma}\boldsymbol{s}$. At update $k$, let $\boldsymbol{z}^{(b)}$ and $y^{(b)}$ denote the feature vector and query label of episode $\xi_k^{(b)}$. We omit $k$ from individual batch entries. Define the empirical second-moment matrix and the batch gradient at the optimum by $\widehat{\boldsymbol{G}}_k = \frac{1}{B}\sum_{b=1}^{B} \boldsymbol{z}^{(b)}(\boldsymbol{z}^{(b)})^\top$ and $\widehat{\boldsymbol{g}}_k^* = \frac{1}{B}\sum_{b=1}^{B}((\boldsymbol{z}^{(b)})^\top \boldsymbol{a}^* - y^{(b)})\boldsymbol{z}^{(b)}$. Subtracting the optimum from the SGD update gives

$$\begin{aligned} \boldsymbol{e}_{k+1} := \boldsymbol{a}_{k+1} - \boldsymbol{a}^* &= \boldsymbol{e}_k - \frac{\eta_k}{B}\sum_{b=1}^{B}((\boldsymbol{z}^{(b)})^\top \boldsymbol{a}_k - y^{(b)})\boldsymbol{z}^{(b)} \\ &= \boldsymbol{e}_k - \frac{\eta_k}{B}\sum_{b=1}^{B}\Big[\boldsymbol{z}^{(b)}(\boldsymbol{z}^{(b)})^\top \boldsymbol{e}_k + ((\boldsymbol{z}^{(b)})^\top \boldsymbol{a}^* - y^{(b)})\boldsymbol{z}^{(b)}\Big] \\ &= \boldsymbol{e}_k - \eta_k(\widehat{\boldsymbol{G}}_k \boldsymbol{e}_k + \widehat{\boldsymbol{g}}_k^*), \qquad \boldsymbol{e}_k = \boldsymbol{a}_k - \boldsymbol{a}^*. \end{aligned} \tag{A.22}$$

**2. Conditional moments of the batch gradient.** We now compute the conditional moments of the two batch quantities. Let $\mathcal{F}_k$ denote the history before update $k$, including the initialization and all previous data and parameter updates, and write $\mathbb{E}_k[\cdot] := \mathbb{E}[\cdot|\mathcal{F}_k]$. Since the current batch is independent of this history,

$$\begin{aligned}\boldsymbol{G} := \mathbb{E}_k[\widehat{\boldsymbol{G}}_k] &= \frac{1}{B}\sum_{b=1}^{B}\mathbb{E}_k[\boldsymbol{z}^{(b)}(\boldsymbol{z}^{(b)})^\top] = \mathbb{E}[\boldsymbol{z}\boldsymbol{z}^\top] \\ &= \mathbb{E}[(\boldsymbol{s}\boldsymbol{s}^\top)\otimes(\boldsymbol{x}_q\boldsymbol{x}_q^\top)] = \boldsymbol{H}_n(\boldsymbol{C})\otimes\frac{\boldsymbol{I}_d}{d}. \end{aligned} \tag{A.23}$$

For each episode, write $\boldsymbol{g}_*^{(b)} := ((\boldsymbol{z}^{(b)})^\top\boldsymbol{a}^* - y^{(b)})\boldsymbol{z}^{(b)}$, so that $\widehat{\boldsymbol{g}}_k^* = \frac{1}{B}\sum_{b=1}^{B}\boldsymbol{g}_*^{(b)}$. The first-order condition at the population optimizer gives

$$\begin{aligned}\mathbb{E}_k[\widehat{\boldsymbol{g}}_k^*] &= \frac{1}{B}\sum_{b=1}^{B}\mathbb{E}_k[\boldsymbol{g}_*^{(b)}] = \mathbb{E}[(\boldsymbol{z}^\top\boldsymbol{a}^* - y_q)\boldsymbol{z}] \\ &= \mathrm{vec}\ \mathbb{E}[(\boldsymbol{x}_q^\top\boldsymbol{\Gamma}^*\boldsymbol{s} - y_q)\boldsymbol{x}_q\boldsymbol{s}^\top] = \mathrm{vec}\ \nabla L(\boldsymbol{\Gamma}^*) = \boldsymbol{0}. \end{aligned} \tag{A.24}$$

For the batch variance, define $\sigma^2 := \mathbb{E}[\left\|(\boldsymbol{z}^\top\boldsymbol{a}^* - y_q)\boldsymbol{z}\right\|^2] = \mathbb{E}_k[\left\|\boldsymbol{g}_*^{(b)}\right\|^2]$. Independence of the episodes yields

$$\begin{aligned}\mathbb{E}_k[\|\widehat{\boldsymbol{g}}_k^*\|^2] &= \frac{1}{B^2}\sum_{b=1}^{B}\sum_{b'=1}^{B}\mathbb{E}_k[\left\langle\boldsymbol{g}_*^{(b)},\boldsymbol{g}_*^{(b')}\right\rangle] \\ &= \frac{1}{B^2}\left\{\sum_{b=1}^{B}\mathbb{E}_k[\left\|\boldsymbol{g}_*^{(b)}\right\|^2] + \sum_{b=1}^{B}\sum_{\substack{b'=1\\b'\neq b}}^{B}\left\langle\mathbb{E}_k[\boldsymbol{g}_*^{(b)}],\mathbb{E}_k[\boldsymbol{g}_*^{(b')}]\right\rangle\right\} = \frac{B\sigma^2}{B^2} = \frac{\sigma^2}{B}. \end{aligned} \tag{A.25}$$

The cross terms vanish because the episode gradients are independent and centered at the optimum. The fourth-moment assumption ensures that the moments used here are finite. Indeed, (A.16) gives $\|\boldsymbol{S}_n\|_{\mathrm{op}} \le d$, and hence $\|\boldsymbol{z}\| \le d\,\|\boldsymbol{w}\|$ and $|y_q| \le \|\boldsymbol{w}\|$. Thus $\mathbb{E}[\|\boldsymbol{z}\|^4] \le d^4\mathbb{E}[\|\boldsymbol{w}\|^4] < \infty$, and

$$\sigma^2 \le d^2(d\,\|\boldsymbol{\Gamma}^*\|_{\mathrm{op}} + 1)^2\mathbb{E}[\|\boldsymbol{w}\|^4] < \infty. \tag{A.26}$$

**3. Mean-square and almost-sure convergence.** The preceding bounds give $\mu = \lambda_{\min}(\boldsymbol{G}) > 0$ and $M_B = \left\|\mathbb{E}[\widehat{\boldsymbol{G}}_k^2]\right\|_{\mathrm{op}} < \infty$. Set $\eta_{\max} = \mu/(2M_B)$. Since $M_B \ge \|\boldsymbol{G}\|_{\mathrm{op}}^2 \ge \mu^2$, the step-size bound ensures $0 < \mu\eta_k \le 1/2$. Expanding the squared error recursion and applying the conditional-moment bounds gives

$$\begin{aligned}\mathbb{E}_k[\|\boldsymbol{e}_{k+1}\|^2] &= \|\boldsymbol{e}_k\|^2 - 2\eta_k\boldsymbol{e}_k^\top\boldsymbol{G}\boldsymbol{e}_k - 2\eta_k\boldsymbol{e}_k^\top\mathbb{E}_k[\widehat{\boldsymbol{g}}_k^*] + \eta_k^2\mathbb{E}_k[\left\|\widehat{\boldsymbol{G}}_k\boldsymbol{e}_k + \widehat{\boldsymbol{g}}_k^*\right\|^2] \\ &\le \|\boldsymbol{e}_k\|^2 - 2\mu\eta_k\,\|\boldsymbol{e}_k\|^2 + 2\eta_k^2\mathbb{E}_k[\left\|\widehat{\boldsymbol{G}}_k\boldsymbol{e}_k\right\|^2] + 2\eta_k^2\mathbb{E}_k[\|\widehat{\boldsymbol{g}}_k^*\|^2] \\ &\le (1 - 2\mu\eta_k + 2M_B\eta_k^2)\,\|\boldsymbol{e}_k\|^2 + \frac{2\eta_k^2\sigma^2}{B} \\ &\le (1 - \mu\eta_k)\,\|\boldsymbol{e}_k\|^2 + \frac{2\eta_k^2\sigma^2}{B}, \qquad 0 < \eta_k \le \frac{\mu}{2M_B}. \end{aligned} \tag{A.27}$$

For mean-square convergence, let $a_K = \mathbb{E}[\|\boldsymbol{e}_K\|^2]$ and fix a tail index $N < K$. Taking expectations and iterating (A.27) gives

$$\begin{aligned}a_K &\le a_N\prod_{j=N}^{K-1}(1-\mu\eta_j) + \frac{2\sigma^2}{B}\sum_{i=N}^{K-1}\eta_i^2\prod_{j=i+1}^{K-1}(1-\mu\eta_j) \\ &\le a_N e^{-\mu\sum_{j=N}^{K-1}\eta_j} + \frac{2\sigma^2}{B}\sum_{i=N}^{\infty}\eta_i^2. \end{aligned} \tag{A.28}$$

The sum of the step sizes diverges, so the first term vanishes as $K$ grows. Square summability then gives

$$\limsup_{K\to\infty} a_K \le \frac{2\sigma^2}{B} \sum_{i=N}^{\infty} \eta_i^2 \xrightarrow[N\to\infty]{} 0. \tag{A.29}$$

This proves mean-square convergence. For almost-sure convergence, add the deterministic tail $(2\sigma^2/B)\sum_{j\ge k}\eta_j^2$ to $\|\boldsymbol{e}_k\|^2$. By (A.27), the resulting process is a nonnegative supermartingale whose conditional decrement is at least $\mu\eta_k \|\boldsymbol{e}_k\|^2$. The supermartingale converges, and telescoping expectations gives $\sum_k \eta_k \|\boldsymbol{e}_k\|^2 < \infty$ almost surely. Since $\sum_k \eta_k = \infty$ and $\|\boldsymbol{e}_k\|^2$ has a limit, that limit must be zero. Column vectorization preserves the Frobenius norm, so these two limits are precisely the convergence claims in (A.20). □

# B PROOF OF THEOREM 2

The proof uses one property of softmax attention: its weights form a convex combination of the context labels. Lemma B.1 translates this property into a bound on the prediction magnitude after one generation of full-parameter SGD. We iterate that bound to obtain the energy decay, then apply it to fresh $P_0$ episodes to control clean risk.

## B.1 PROOF OF THE THEOREM

*Proof of Theorem 2.* Work on the probability-one event in Assumption 4.1 on which the task-amplitude and training-budget bounds hold. Lemma B.1 applies to every subsequent training realization whose sampled tasks satisfy their almost-sure amplitude bounds. Since there are countably many sampled episodes, these bounds hold simultaneously with probability one.

**1. Recursion for task amplitude and energy.** We now apply the one-generation estimate recursively, starting from the amplitude bound assumed at generation $r_0$. If $TM^2 = 0$, then either all subsequent step sizes are zero or every task under $P_{r_0}$ is zero. In either case the gains remain zero from generation $r_0 + 1$ onward, so $\mathcal{E}(P_{r_0+j}) = 0$ and $R_0^{(n)}(\boldsymbol{\theta}_{r_0+j}) = 1$ for all $j \geq 1$.

For the remaining case, assume $a := TM^2 > 0$. The threshold assumption gives $q_0 = e^a - 1 \in (0, 1)$. Starting from $M_{r_0} = M$, define

$$M_{r+1} = M_r(e^{TM_r^2} - 1), \qquad r \geq r_0. \tag{B.1}$$

Since $T_{r+1} \leq T$, Lemma B.1 ensures that $M_r$ bounds the task values under $P_r$. To show that this sequence remains below its initial threshold, suppose $M_r \leq M$ and use the integral form of the exponential increment:

$$e^{TM_r^2} - 1 = TM_r^2 \int_0^1 e^{sTM_r^2}\, ds \leq TM_r^2 \int_0^1 e^{sa}\, ds = q_0 \frac{M_r^2}{M^2}. \tag{B.2}$$

Consequently, $M_{r+1} \leq q_0 M_r^3/M^2 \leq q_0 M_r \leq M$. This proves that every subsequent generation remains below the threshold. Iterating the cubic bound gives the stated rate: the base case $j = 0$ is immediate, and the induction step is

$$M_{r_0+j+1} \leq \frac{q_0}{M^2} M_{r_0+j}^3 \leq \frac{q_0}{M^2} \left(M q_0^{(3^j-1)/2}\right)^3 = M q_0^{(3^{j+1}-1)/2}. \tag{B.3}$$

Thus $M_{r_0+j} \leq M q_0^{(3^j-1)/2}$ for every $j \geq 0$. Integrating the pointwise task bound yields

$$\mathcal{E}(P_{r_0+j}) = \mathbb{E}_{f\sim P_{r_0+j}} \left[\int_{\mathbb{S}^{d-1}} f(\boldsymbol{x})^2\, d\nu(\boldsymbol{x})\right] \leq M_{r_0+j}^2 \leq M^2 q_0^{3^j-1}. \tag{B.4}$$

**2. Clean risk on the original task law.** To evaluate the trained models on real tasks, first bound their final gains using the amplitude recursion. Let $c_r$ now denote the final gain of the generation-$r$ model. Combining (B.12) with (B.2) gives, for $j \geq 1$,

$$|c_{r_0+j}| \leq e^{T_{r_0+j} M_{r_0+j-1}^2} - 1 \leq e^{TM_{r_0+j-1}^2} - 1 \leq q_0 \frac{M_{r_0+j-1}^2}{M^2} \leq q_0^{3^{j-1}}. \tag{B.5}$$

For a fresh real episode, write $F_r = F_{\boldsymbol{\theta}_r}(\mathcal{C}_n, \boldsymbol{x}_q)$ and $\mathcal{E}_0 = \mathcal{E}(P_0) = \mathbb{E}[y_q^2]$. Holding the trained parameters fixed, the prediction bound (B.9) gives

$$\|F_r\|_{L^2} = \left(\mathbb{E}[F_r^2]\right)^{1/2} \leq |c_r| \left(\mathbb{E}_{f\sim P_0}[\|f\|_\infty^2]\right)^{1/2} = |c_r| \kappa \sqrt{\mathcal{E}_0}. \tag{B.6}$$

Expanding the normalized squared risk and applying Cauchy–Schwarz now bounds its distance from the risk of the zero predictor:

$$\begin{aligned} \left|R_0^{(n)}(\boldsymbol{\theta}_{r_0+j}) - 1\right| &= \frac{\left|\mathbb{E}[F_{r_0+j}^2] - 2\mathbb{E}[F_{r_0+j} y_q]\right|}{\mathcal{E}_0} \leq \frac{\|F_{r_0+j}\|_{L^2}^2 + 2\,\|F_{r_0+j}\|_{L^2}\sqrt{\mathcal{E}_0}}{\mathcal{E}_0} \\ &\leq \kappa^2 |c_{r_0+j}|^2 + 2\kappa |c_{r_0+j}| \leq 2\kappa q_0^{3^{j-1}} + \kappa^2 q_0^{2\cdot 3^{j-1}}. \end{aligned} \tag{B.7}$$

Because $q_0 < 1$, the task-energy bound and the deviation of clean risk from one both tend to zero. This proves Theorem 2. □

### B.2 PREDICTION BOUNDS FOR ONE GENERATION

**Lemma B.1** (Amplitude bound under full-parameter SGD)**.** *Suppose $\|f\|_\infty \le M$ almost surely under the training law. Train the softmax predictor* (4) *using full-parameter SGD* (9)*, with $c_0 = 0$, arbitrary initial $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{u}$, and nonnegative step sizes of total budget $T_r$. Then the final gain satisfies*

$$|c_r| \le e^{T_r M^2} - 1$$

*almost surely over the training data. For every bounded task $f$ and every context $\mathcal{C}_m = \{(\boldsymbol{z}_i, f(\boldsymbol{z}_i))\}_{i=1}^m$, the resulting predictor satisfies*

$$\sup_{\boldsymbol{x}\in\mathbb{S}^{d-1}} |F_{\boldsymbol{\theta}_r}(\mathcal{C}_m, \boldsymbol{x})| \le (e^{T_r M^2} - 1)\,\|f\|_\infty .$$

*In particular, unnormalized generation from the training law produces tasks of amplitude at most $M(e^{T_r M^2} - 1)$.*

*Proof.* Fix a generation $r$ and write $\eta_k = \eta_{r,k}$. Let $c_k$ denote the gain after $k$ SGD updates, where $c_0 = 0$ and $T_r = \sum_{k=0}^{K-1} \eta_{r,k}$. Work on the probability-one event on which every sampled task satisfies $\|f\|_\infty \le M$.

**1. Softmax averaging and the gain update.** The softmax predictor averages context labels before multiplying by the learned gain. For one episode, write $t = \sum_{i=1}^n \pi_i y_i$ and $F = ct$. Nonnegativity and unit sum of the softmax weights give

$$|t| = \left|\sum_{i=1}^n \pi_i f(\boldsymbol{x}_i)\right| \le \sum_{i=1}^n \pi_i |f(\boldsymbol{x}_i)| \le \|f\|_\infty \sum_{i=1}^n \pi_i = \|f\|_\infty \tag{B.8}$$

and

$$|F| = |c|\,|t| \le |c|\,\|f\|_\infty . \tag{B.9}$$

These prediction bounds control the gain update directly. At a fixed update $k$, let $t^{(b)}$ and $y^{(b)}$ denote the weighted context label and query label of episode $\xi_k^{(b)}$, suppressing $k$ on these batch entries. The training-task bound gives $|t^{(b)}| \le M$ by (B.8), and $|y^{(b)}| \le M$. With the other parameter coordinates held fixed, the attention scores do not depend on $c$, so $\partial_c F = t$. Thus the gain coordinate of full-parameter SGD satisfies

$$c_{k+1} = c_k - \frac{\eta_k}{B}\sum_{b=1}^B \partial_c \tfrac{1}{2}(ct^{(b)} - y^{(b)})^2\Big|_{c=c_k} = c_k - \frac{\eta_k}{B}\sum_{b=1}^B (c_k t^{(b)} - y^{(b)})t^{(b)}. \tag{B.10}$$

Using $\eta_k \ge 0$ and the bounds on the weighted labels and query labels,

$$\begin{aligned} |c_{k+1}| &\le |c_k| + \frac{\eta_k}{B}\sum_{b=1}^B |c_k t^{(b)} - y^{(b)}|\,|t^{(b)}| \le |c_k| + \frac{\eta_k}{B}\sum_{b=1}^B (|c_k||t^{(b)}| + |y^{(b)}|)|t^{(b)}| \\ &\le |c_k| + \frac{\eta_k}{B}\sum_{b=1}^B (|c_k|M + M)M = |c_k| + \eta_k M^2(1 + |c_k|). \end{aligned} \tag{B.11}$$

Adding 1 to both sides and iterating the resulting inequality yields

$$1 + |c_K| \le (1 + |c_0|)\prod_{k=0}^{K-1}(1 + \eta_k M^2) \le \prod_{k=0}^{K-1} e^{\eta_k M^2} = e^{M^2\sum_{k=0}^{K-1}\eta_k} = e^{T_r M^2}. \tag{B.12}$$

The product bound uses $c_0 = 0$ and $1+x \le e^x$ for $x \ge 0$. Hence the final gain obeys $|c_K| \le e^{T_r M^2} - 1$ on every training path under consideration.

**2. Prediction and generation.** The same gain bound controls prediction for every task and context, regardless of their use in training. Indeed, (B.9) gives $|F| \le (e^{T_r M^2} - 1)\,\|f\|_\infty$. For a task from the training law, with $\|f\|_\infty \le M$, apply (B.9) to any seed of length $m$ in the generation rule. For every $\boldsymbol{x} \in \mathbb{S}^{d-1}$,

$$|f_r(\boldsymbol{x})| = \left|F_{\boldsymbol{\theta}_r}\left(\{(\boldsymbol{z}_i, f(\boldsymbol{z}_i))\}_{i=1}^m, \boldsymbol{x}\right)\right| \le |c_K|\,\|f\|_\infty \le M(e^{T_r M^2} - 1). \tag{B.13}$$

This gives a uniform amplitude bound for the generated task law.

□

# C Proof of Theorem 3

The proof has three parts. We first establish learnability of the original mixture, then show that every finite training duration reduces the relative quadratic energy, and finally evaluate the limiting model. Appendix C.2 contains the response and gradient-flow calculations used in this argument. Appendix C.3 treats finite-context SGD separately.

## C.1 Proof of the theorem

*Proof of Theorem 3.* We use three facts proved in Appendix C.2: the population response properties from Lemma C.1, the full-parameter training dynamics from Lemma C.2, and the pure-linear risk bound from Lemma C.3. Write $\rho = b/a$, and let $(c_T(p), \beta_T(p))$ denote the parameters reached after training for time $T$ on a unit-energy law with quadratic fraction $p$.

**Step 1: learning the original task law.** Assumption 4.2 gives positive energies $1 - p_0$ and $p_0$ in the two components. Lemma C.2 therefore gives $\beta_T(p_0) \to \infty$ and $c_T(p_0) \to 1$ as $T \to \infty$. Since $a(\beta), b(\beta) \to 1$ by Lemma C.1, the risk identity (C.8) implies the first limit in (16).

**Step 2: task composition and parameter convergence.** Lemma C.2 gives a bound uniform over all unit-energy mixtures, including the pure-component endpoints. Writing $E_1 = 1 - p$, $E_2 = p$, and $C = 1/b(1)$ gives

$$\dot{\beta}(t) \le \frac{2C(1+C)}{d}\{E_1 + 2E_2\} \le \frac{4C(1+C)}{d},$$
$$\beta(T) \le 1 + \frac{4C(1+C)}{d}T =: B_T < \infty. \tag{C.1}$$

Each generation multiplies the two task components by positive deterministic scalars and then applies a common normalization factor. The coefficient directions remain independent and uniform, so the population-flow lemma applies at every generation.

For a unit-energy input law with quadratic fraction $p$, let $p^+$ denote the fraction after generation and normalization. Before normalization, the component energies are $(1-p)c^2a^2$ and $pc^2b^2$. Dividing the quadratic energy by their sum yields

$$p^+ = \frac{pc^2b^2}{(1-p)c^2a^2 + pc^2b^2} = \frac{pb^2}{(1-p)a^2 + pb^2}. \tag{C.2}$$

Lemma C.2 gives $c > 0$ after every positive training time. Together with $a, b > 0$, this ensures that normalization is defined. The complementary linear fraction is

$$1 - p^+ = \frac{(1-p)a^2 + pb^2 - pb^2}{(1-p)a^2 + pb^2} = \frac{(1-p)a^2}{(1-p)a^2 + pb^2}. \tag{C.3}$$

Taking the ratio cancels the common normalization:

$$\frac{p^+}{1-p^+} = \frac{pb^2}{(1-p)a^2} = \frac{p}{1-p}\left(\frac{b}{a}\right)^2 = \frac{p}{1-p}\rho(\beta)^2. \tag{C.4}$$

Iterating this identity from $p_0 \in (0,1)$ gives

$$\frac{p_r}{1-p_r} = \frac{p_{r-1}}{1-p_{r-1}}\rho(\beta_r)^2 = \frac{p_0}{1-p_0}\prod_{j=1}^{r}\rho(\beta_j)^2. \tag{C.5}$$

Set $q_T := \rho(B_T)^2 \in (0,1)$. Combining the odds recursion (C.5) with the uniform bound (C.1) gives

$$\frac{p_r}{1-p_r} \le \frac{p_0}{1-p_0}q_T^r,$$
$$p_r \le \frac{p_0}{1-p_0}q_T^r \longrightarrow 0. \tag{C.6}$$

Since $B_T < \infty$, the factor $q_T$ is strictly below one. This establishes geometric decay of the quadratic fraction for every finite $T$.

For parameter convergence, use smooth dependence of the reduced vector field $V_p$ in (C.45)–(C.46) on $(c, \beta, p)$. On the common compact region reached during $[0, T]$, choose finite constants $L_T \geq 1$ and $C_T$ bounding its parameter Lipschitz constant and its derivative in $p$. For $\boldsymbol{z}_p(t) = (c(t;p), \beta(t;p))$, the common initialization and Gronwall's inequality give

$$\begin{aligned}\|\boldsymbol{z}_p(t) - \boldsymbol{z}_0(t)\| &\leq C_T p t + L_T \int_0^t \|\boldsymbol{z}_p(s) - \boldsymbol{z}_0(s)\|\, ds \\ &\leq C_T p \int_0^t e^{L_T(t-s)}\, ds = \frac{C_T p}{L_T}(e^{L_T t} - 1). \end{aligned} \tag{C.7}$$

This continuity estimate shows that the generation-$r$ parameters converge to those trained for time $T$ on purely linear tasks as $p_{r-1} \to 0$. To evaluate their performance, apply (C.38) to the fixed unit-energy real law. For any model in the symmetric family,

$$R_0^{(\infty)}(\beta, c) = 2L_{P_0}^{(\infty)}(\beta, c) = (1 - p_0)(ca(\beta) - 1)^2 + p_0(cb(\beta) - 1)^2. \tag{C.8}$$

Substituting the limiting parameters yields

$$\begin{aligned}\lim_{r\to\infty} R_0^{(\infty)}(\boldsymbol{\theta}_r(T)) = (1 - p_0)\{c_T(0)a(\beta_T(0)) - 1\}^2 \\ + p_0\{c_T(0)b(\beta_T(0)) - 1\}^2. \end{aligned} \tag{C.9}$$

The weights $1 - p_0, p_0$ come from the original evaluation law.

**Step 3: a positive limiting clean risk.** For every $d \geq 2$, apply Lemma C.3 at time $T$ to the limiting parameters from Step 2:

$$\lim_{r\to\infty} R_0^{(\infty)}(\boldsymbol{\theta}_r(T)) \geq \frac{16p_0}{121} > \frac{p_0}{8}.$$

This proves (16). □

### C.2 Full response lemma and supporting results

We first state and prove the full version of Lemma 4.1, including the derivative bounds and asymptotic expansions needed for the training analysis. We then establish the gradient-flow and pure-linear risk lemmas used in the theorem proof. These facts also support Theorems 4 and 5.

**Lemma C.1** (Population response coefficients: full statement)**.** *Let $d \geq 2$, let $\nu$ be uniform on $\mathbb{S}^{d-1}$, and let $\beta > 0$. For any fixed $\boldsymbol{x} \in \mathbb{S}^{d-1}$, define*

$$Z_d(\beta) = \int e^{\beta \boldsymbol{x}^\top \boldsymbol{z}}\, d\nu(\boldsymbol{z}), \qquad a(\beta) = \frac{Z_d'(\beta)}{Z_d(\beta)}, \qquad b(\beta) = 1 - \frac{da(\beta)}{\beta}, \qquad \rho(\beta) = \frac{b(\beta)}{a(\beta)}. \tag{C.10}$$

*For $f_1(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x}$ and $f_2(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}$ with $\boldsymbol{A}$ symmetric and trace free, the population-context prediction $F_{\boldsymbol{\theta},\infty}(f, \boldsymbol{x})$ at $(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{u}, c) = (\sqrt{\beta}\boldsymbol{I}_d, \sqrt{\beta}\boldsymbol{I}_d, \boldsymbol{0}, c)$ is*

$$F_{\boldsymbol{\theta},\infty}(f_1 + f_2, \boldsymbol{x}) = ca(\beta)f_1(\boldsymbol{x}) + cb(\beta)f_2(\boldsymbol{x}).$$

*The responses satisfy $0 < b < a < 1$, $a', b', \rho' > 0$, $a'/a \leq 1/\beta$, and $b'/b \leq 2/\beta$. As $\beta \to \infty$,*

$$Z_d(\beta) = \frac{2^{(d-3)/2}\Gamma(d/2)}{\sqrt{\pi}} e^\beta \beta^{-(d-1)/2}\big(1 + O(\beta^{-1})\big),$$

*and the response coefficients satisfy*

$$\begin{aligned} a(\beta) &= 1 - \frac{d-1}{2\beta} + O(\beta^{-2}), & a'(\beta) &= \frac{d-1}{2\beta^2} + O(\beta^{-3}), \\ b(\beta) &= 1 - \frac{d}{\beta} + O(\beta^{-2}), & b'(\beta) &= \frac{d}{\beta^2} + O(\beta^{-3}), \\ \rho(\beta) &= 1 - \frac{d+1}{2\beta} + O(\beta^{-2}). \end{aligned}$$

*Proof of Lemmas 4.1 and C.1.* **Response coefficients.** For $\boldsymbol{Q} = \boldsymbol{K} = \sqrt{\beta}\,\boldsymbol{I}_d$ and $\boldsymbol{u} = \boldsymbol{0}$, write the predictor in terms of the averaging operator

$$(K_\beta f)(\boldsymbol{x}) := \frac{\int e^{\beta \boldsymbol{x}^\top \boldsymbol{z}} f(\boldsymbol{z})\, d\nu(\boldsymbol{z})}{Z_d(\beta)},$$
$$F_{\boldsymbol{\theta},\infty}(f, \boldsymbol{x}) = c(K_\beta f)(\boldsymbol{x}). \tag{C.11}$$

Rotation invariance makes $Z_d$ independent of the unit query $\boldsymbol{x}$.

To obtain identities that also compare different dimensions, first consider an arbitrary dimension $s \geq 2$. In spherical coordinates, the angle $\theta$ from a fixed query has surface element proportional to $\sin^{s-2}\theta\, d\theta$. Setting $t = \cos\theta$ gives $|dt| = \sin\theta\, d\theta$, and hence an inner-product density proportional to $(1-t^2)^{(s-3)/2}$. After normalization, this density is

$$C_s(1-t^2)^{(s-3)/2}\,\mathbf{1}_{\{-1<t<1\}}, \qquad C_s = \frac{\Gamma(s/2)}{\sqrt{\pi}\,\Gamma((s-1)/2)}. \tag{C.12}$$

The substitution $u = t^2$ determines both the normalizing constant and the even moments:

$$\int_{-1}^{1} t^{2j}(1-t^2)^{(s-3)/2}\, dt = 2\int_0^1 t^{2j}(1-t^2)^{(s-3)/2}\, dt = \int_0^1 u^{j-1/2}(1-u)^{(s-3)/2}\, du$$
$$= \frac{\Gamma(j+1/2)\Gamma((s-1)/2)}{\Gamma(j+s/2)}. \tag{C.13}$$

Consequently,

$$\mathbb{E}[t^{2j}] = C_s\int_{-1}^{1} t^{2j}(1-t^2)^{(s-3)/2}\, dt = \frac{\Gamma(s/2)}{\sqrt{\pi}\,\Gamma((s-1)/2)}\,\frac{\Gamma(j+1/2)\Gamma((s-1)/2)}{\Gamma(j+s/2)}$$
$$= \frac{\Gamma(s/2)\Gamma(j+1/2)}{\sqrt{\pi}\,\Gamma(j+s/2)}. \tag{C.14}$$

Symmetry makes the odd moments zero. We may expand the exponential under the integral because its absolute series is bounded by $e^{|\beta|}$ on the integration domain. Using $\Gamma(j+1/2) = (2j)!\sqrt{\pi}/(4^j j!)$ then gives

$$Z_s(\beta) = \sum_{j=0}^{\infty}\frac{\beta^{2j}}{(2j)!}\mathbb{E}[t^{2j}] = \Gamma(s/2)\sum_{j=0}^{\infty}\frac{(\beta^2/4)^j}{j!\,\Gamma(j+s/2)}. \tag{C.15}$$

Let $\mathbb{E}_{s,\beta}[\cdot]$ denote expectation under this scalar density tilted by $e^{\beta t}/Z_s(\beta)$, and define $a_s = Z_s'/Z_s$. Differentiating the normalizer gives

$$a_s(\beta) = \frac{Z_s'(\beta)}{Z_s(\beta)} = \frac{\int_{-1}^{1} t e^{\beta t} C_s(1-t^2)^{(s-3)/2}\, dt}{Z_s(\beta)} = \mathbb{E}_{s,\beta}[t]. \tag{C.16}$$

The derivative of the tilted mean is its variance:

$$a_s'(\beta) = \frac{Z_s''(\beta)Z_s(\beta) - Z_s'(\beta)^2}{Z_s(\beta)^2} = \frac{Z_s''(\beta)}{Z_s(\beta)} - \left(\frac{Z_s'(\beta)}{Z_s(\beta)}\right)^2 = \mathbb{E}_{s,\beta}[t^2] - \big(\mathbb{E}_{s,\beta}[t]\big)^2$$
$$= \mathrm{Var}_{s,\beta}(t) > 0. \tag{C.17}$$

To locate the tilted mean, pair $t$ and $-t$ in the numerator: the resulting integrand $t(e^{\beta t} - e^{-\beta t})$ is positive on $(0,1)$, which proves positivity of $a_s$ for $\beta > 0$. Moreover, $a_s < 1$ because $t < 1$ almost surely. In dimension $d$, write $a = a_d$. Rotations fixing $\boldsymbol{x}$ force the tilted vector mean to be parallel to $\boldsymbol{x}$, with component $a$ along $\boldsymbol{x}$. Therefore

$$\mathbb{E}_{d,\beta}[\boldsymbol{z}] = \boldsymbol{x}\,\boldsymbol{x}^\top\mathbb{E}_{d,\beta}[\boldsymbol{z}] = \boldsymbol{x}\,\mathbb{E}_{d,\beta}[\boldsymbol{x}^\top\boldsymbol{z}] = a(\beta)\boldsymbol{x}. \tag{C.18}$$

For the linear component,

$$(K_\beta f_1)(\boldsymbol{x}) = \mathbb{E}_{d,\beta}[\boldsymbol{w}^\top\boldsymbol{z}] = \boldsymbol{w}^\top\mathbb{E}_{d,\beta}[\boldsymbol{z}] = a(\beta)\boldsymbol{w}^\top\boldsymbol{x}. \tag{C.19}$$

For the quadratic component, put $s_{\|} = \mathbb{E}_{d,\beta}[t^2]$. Integrating the derivative of a function that vanishes at both endpoints gives

$$\begin{aligned} 0 &= \int_{-1}^{1} \frac{d}{dt}\left[e^{\beta t}(1-t^2)^{(d-1)/2}\right] dt = \int_{-1}^{1} e^{\beta t}(1-t^2)^{(d-3)/2}\{\beta(1-t^2) - (d-1)t\}\, dt, \\ 0 &= \beta\{1 - \mathbb{E}_{d,\beta}[t^2]\} - (d-1)a, \\ s_{\|} &= 1 - \frac{d-1}{\beta}a. \end{aligned} \tag{C.20}$$

The integration by parts remains valid for $d = 2$, since the endpoint values vanish and the derivative has integrable endpoint singularities. To recover the full second-moment matrix, use rotations fixing $\boldsymbol{x}$: all $d-1$ perpendicular variances are equal, and their sum is $1 - s_{\|}$. Thus

$$\begin{aligned} s_{\perp} &= \frac{1 - s_{\|}}{d-1} = \frac{a}{\beta}, \\ \mathbb{E}_{d,\beta}[\boldsymbol{z}\boldsymbol{z}^{\top}] &= s_{\perp}(\boldsymbol{I}_d - \boldsymbol{x}\boldsymbol{x}^{\top}) + s_{\|}\boldsymbol{x}\boldsymbol{x}^{\top} = \frac{a}{\beta}\boldsymbol{I}_d + \left(1 - \frac{da}{\beta}\right)\boldsymbol{x}\boldsymbol{x}^{\top}. \end{aligned} \tag{C.21}$$

Contracting the second-moment matrix with the trace-free matrix $\boldsymbol{A}$ gives the quadratic response:

$$\begin{aligned} (K_\beta f_2)(\boldsymbol{x}) &= \mathbb{E}_{d,\beta}[\mathrm{tr}(\boldsymbol{A}\boldsymbol{z}\boldsymbol{z}^{\top})] = \mathrm{tr}\big(\boldsymbol{A}\,\mathbb{E}_{d,\beta}[\boldsymbol{z}\boldsymbol{z}^{\top}]\big) = \frac{a}{\beta}\,\mathrm{tr}\,\boldsymbol{A} + \left(1 - \frac{da}{\beta}\right)\mathrm{tr}(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^{\top}) \\ &= b(\beta)\boldsymbol{x}^{\top}\boldsymbol{A}\boldsymbol{x}, \qquad b(\beta) = 1 - \frac{da(\beta)}{\beta}. \end{aligned} \tag{C.22}$$

The trace-free condition removes the constant term.

**Ordering and derivative bounds.** Termwise differentiation of the normalizing series gives

$$Z_s'(\beta) = \Gamma(s/2)\sum_{j=1}^{\infty} \frac{(j\beta/2)(\beta^2/4)^{j-1}}{j!\,\Gamma(j+s/2)} = \frac{\beta}{2}\Gamma(s/2)\sum_{j=0}^{\infty} \frac{(\beta^2/4)^j}{j!\,\Gamma(j+s/2+1)} = \frac{\beta}{s}Z_{s+2}(\beta). \tag{C.23}$$

Differentiating this dimension recurrence once more yields

$$\frac{Z_s''}{Z_s} = \frac{Z_{s+2} + \beta Z_{s+2}'}{sZ_s} = \frac{Z_s'}{\beta Z_s} + \frac{\beta Z_{s+2}}{sZ_s}\frac{Z_{s+2}'}{Z_{s+2}} = \frac{a_s}{\beta} + a_s a_{s+2}. \tag{C.24}$$

The integration-by-parts identity in dimension $s$ also gives $Z_s''/Z_s = 1 - (s-1)a_s/\beta$. Equating the two formulas yields

$$\begin{aligned} 1 - \frac{sa_s}{\beta} &= a_s a_{s+2}, \\ b = a_d a_{d+2}, \qquad \rho &:= \frac{b}{a} = a_{d+2}, \\ 0 < b < a < 1, \qquad a' > 0, \quad \rho' > 0, \quad b' &= a'\rho + a\rho' > 0. \end{aligned} \tag{C.25}$$

To bound the logarithmic derivatives, put $z = \beta^2/4$ and assign the nonnegative integer $k$ probabilities proportional to the weights $w_j = z^j/[j!\Gamma(j+s/2)]$. The normalizing series then gives

$$\frac{a_s(\beta)}{\beta} = \frac{1}{2}\frac{\sum_{j\geq 0} w_j/(j+s/2)}{\sum_{j\geq 0} w_j} = \frac{1}{2}\mathbb{E}[\frac{1}{k+s/2}]. \tag{C.26}$$

The monotonicity of this ratio follows by differentiating the normalized probabilities: $d\Pr(k = j)/d\ln z = (j - \mathbb{E}[k])\Pr(k = j)$. Consequently,

$$\begin{aligned}
\frac{d}{d\ln z}\mathbb{E}[\frac{1}{k+s/2}] &= \mathrm{Cov}\left(k, \frac{1}{k+s/2}\right) \\
&= \frac{1}{2}\sum_{i,j}\Pr(k=i)\Pr(k=j)(i-j)\left(\frac{1}{i+s/2} - \frac{1}{j+s/2}\right) \\
&= -\frac{1}{2}\sum_{i,j}\Pr(k=i)\Pr(k=j)\frac{(i-j)^2}{(i+s/2)(j+s/2)} \le 0, \\
\frac{d}{d\beta}\left(\frac{a_s(\beta)}{\beta}\right) &= \frac{\beta a_s' - a_s}{\beta^2} \le 0, \\
\frac{a_s'}{a_s} \le \frac{1}{\beta}, &\qquad \frac{b'}{b} = \frac{a_d'}{a_d} + \frac{a_{d+2}'}{a_{d+2}} \le \frac{2}{\beta}.
\end{aligned} \tag{C.27}$$

Absolute convergence on bounded $\beta$ intervals justifies these differentiations.

**Large-selectivity response bounds.** Fix a dimension $s \ge 2$, put $k = (s-1)/2 > 0$, and change variables $u = \beta(1-t)$ in the scalar sphere integral. For $j = 0, 1, 2$, define the auxiliary integrals

$$I_j(\beta) := \int_0^{2\beta} e^{-u}u^{k+j-1}(2 - u/\beta)^{k-1}\,du. \tag{C.28}$$

Under this substitution, $1 - t^2 = (u/\beta)(2 - u/\beta)$ and $|dt| = du/\beta$. Thus

$$Z_s(\beta) = C_s\int_0^{2\beta} e^{\beta-u}(u/\beta)^{k-1}(2-u/\beta)^{k-1}\frac{du}{\beta} = C_s e^{\beta}\beta^{-k}I_0(\beta). \tag{C.29}$$

The same change of variables expresses the first two rescaled moments as ratios of these integrals:

$$\mathbb{E}_{s,\beta}\left[(\beta(1-t))^j\right] = \frac{C_s e^{\beta}\beta^{-k}I_j(\beta)}{Z_s(\beta)} = \frac{I_j(\beta)}{I_0(\beta)}. \tag{C.30}$$

On $0 \le u \le \beta$, Taylor's formula on the fixed interval $u/(2\beta) \in [0, 1/2]$ yields

$$\begin{aligned}
(2 - u/\beta)^{k-1} &= 2^{k-1}\left[1 - \frac{(k-1)u}{2\beta} + R_k(u,\beta)\right], \\
|R_k(u,\beta)| &\le C_k u^2/\beta^2.
\end{aligned} \tag{C.31}$$

The remainder is integrable against $e^{-u}u^{k+j-1}$. The other endpoint contributes only an exponentially small term, including when $s = 2$:

$$\int_{\beta}^{2\beta} e^{-u}u^{k+j-1}(2-u/\beta)^{k-1}du \le e^{-\beta}\beta^{k+j}\int_1^2 v^{k+j-1}(2-v)^{k-1}dv = O(e^{-\beta}\beta^{k+j}). \tag{C.32}$$

The last integral is finite because $k > 0$. Extending the integrals from the Taylor expansion to infinity contributes only exponential tails. Combining the two estimates gives

$$\begin{aligned}
I_j(\beta) &= 2^{k-1}\left[\Gamma(k+j) - \frac{k-1}{2\beta}\Gamma(k+j+1) + O(\beta^{-2})\right] \\
&= 2^{k-1}\Gamma(k+j)\left[1 - \frac{(k-1)(k+j)}{2\beta} + O(\beta^{-2})\right].
\end{aligned} \tag{C.33}$$

The ratio identity $(1+x/\beta+O(\beta^{-2}))/(1+y/\beta+O(\beta^{-2}))=1+(x-y)/\beta+O(\beta^{-2})$ now gives

$$\begin{aligned}
\mathbb{E}_{s,\beta}[\beta(1-t)] &= k-\frac{k(k-1)}{2\beta}+O(\beta^{-2}),\\
\mathbb{E}_{s,\beta}\big[(\beta(1-t))^2\big] &= k(k+1)\left[1-\frac{k-1}{\beta}+O(\beta^{-2})\right],\\
\mathrm{Var}_{s,\beta}[\beta(1-t)] &= k+O(\beta^{-1}),\\
a_s(\beta) &= 1-\frac{k}{\beta}+O(\beta^{-2}),\\
a_s'(\beta) &= \mathrm{Var}_{s,\beta}(t)=\frac{k}{\beta^2}+O(\beta^{-3}).
\end{aligned}\tag{C.34}$$

Finally, specialize to dimension $d$. The identity $Z_d(\beta)=C_d e^{\beta}\beta^{-(d-1)/2}I_0(\beta)$ and the expansion of $I_0$ yield the normalizer asymptotics. Combining the moment expansions with the response identities gives all the claimed estimates:

$$\begin{aligned}
Z_d(\beta) &= \frac{2^{(d-3)/2}\Gamma(d/2)}{\sqrt{\pi}}e^{\beta}\beta^{-(d-1)/2}\big(1+O(\beta^{-1})\big),\\
a(\beta) &= 1-\frac{d-1}{2\beta}+O(\beta^{-2}), \qquad a'(\beta)=\frac{d-1}{2\beta^2}+O(\beta^{-3}),\\
b(\beta) &= 1-\frac{d}{\beta}+O(\beta^{-2}),\\
b'(\beta) &= \frac{da(\beta)}{\beta^2}-\frac{da'(\beta)}{\beta}=\frac{d}{\beta^2}+O(\beta^{-3}),\\
\rho(\beta) &= a_{d+2}(\beta)=1-\frac{d+1}{2\beta}+O(\beta^{-2}).
\end{aligned}\tag{C.35}$$

□

**Lemma C.2** (Population gradient flow)**.** *Let $p\in[0,1]$, and use the coefficient construction of Assumption 4.2 with component energies $E_1=1-p$ and $E_2=p$; a zero-energy component is identically zero. Full-parameter population gradient flow from $(\boldsymbol{I}_d,\boldsymbol{I}_d,\mathbf{0},0)$ exists for every $t\ge 0$ and preserves*

$$\boldsymbol{Q}(t)=\boldsymbol{K}(t)=\sqrt{\beta(t)}\boldsymbol{I}_d, \qquad \boldsymbol{u}(t)=\mathbf{0}.$$

*The loss and scalar dynamics are given by* (C.38)*,* (C.45)*, and* (C.46)*, with $\beta(0)=1$ and $c(0)=0$. Put $C=1/b(1)$. For $t>0$,*

$$0<c(t)\le C, \qquad 1\le\beta(t)\le 1+\frac{2C(1+C)}{d}(E_1+2E_2)t, \qquad \dot\beta(t)\ge 0.$$

*If $0<p<1$, then $\beta(t)\to\infty$, $c(t)\to 1$, and the prediction risk on this task law tends to zero as $t\to\infty$.*

*Proof.* **Loss on the symmetric parameter family.** The sphere moment identities in Appendix A show that linear and trace-free quadratic functions are orthogonal, because all third moments vanish:

$$\int(\boldsymbol{w}^\top\boldsymbol{x})(\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x})\,d\nu(\boldsymbol{x})=\sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d}\boldsymbol{w}_i\boldsymbol{A}_{jk}\mathbb{E}[\boldsymbol{x}_i\boldsymbol{x}_j\boldsymbol{x}_k]=0.\tag{C.36}$$

To identify the component energies, the query second-moment calculation in (A.5) gives linear energy $\|\boldsymbol{w}\|^2/d$ for a fixed coefficient $\boldsymbol{w}$. For the quadratic energy, apply the symmetric-matrix moment identity (A.13) to $\boldsymbol{A}$ and use $\mathrm{tr}\,\boldsymbol{A}=0$:

$$\int(\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x})^2\,d\nu(\boldsymbol{x})=\mathrm{tr}\big(\boldsymbol{A}\,\mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^\top]\big)=\frac{\mathrm{tr}(\boldsymbol{A}\{2\boldsymbol{A}+\mathrm{tr}(\boldsymbol{A})\boldsymbol{I}_d\})}{d(d+2)}=\frac{2\,\|\boldsymbol{A}\|_{\mathrm{F}}^2}{d(d+2)}.\tag{C.37}$$

The component energies are therefore $E_1 = \mathbb{E}[\|\boldsymbol{w}\|^2]/d$ and $E_2 = 2\mathbb{E}[\|\boldsymbol{A}\|_{\mathrm{F}}^2]/[d(d+2)]$. Expanding the prediction error $(ca-1)f_1 + (cb-1)f_2$ and using orthogonality for each task gives

$$\begin{aligned} L(\beta, c) &= \frac{1}{2}\mathbb{E}\left[\int \{(ca-1)f_1 + (cb-1)f_2\}^2 \, d\nu\right] \\ &= \frac{1}{2}(ca-1)^2\mathbb{E}\left[\int f_1^2 \, d\nu\right] + (ca-1)(cb-1)\mathbb{E}\left[\int f_1 f_2 \, d\nu\right] \\ &\quad + \frac{1}{2}(cb-1)^2\mathbb{E}\left[\int f_2^2 \, d\nu\right] = \frac{1}{2}E_1(ca-1)^2 + \frac{1}{2}E_2(cb-1)^2. \end{aligned} \tag{C.38}$$

**Invariance under full-parameter gradient flow.** We first show that the label-dependent key remains zero under the population gradient. At $\boldsymbol{u} = \boldsymbol{0}$, write $t = \sum_i \pi_i y_i$. Since $\nabla_{\boldsymbol{u}} s_i = y_i \boldsymbol{Q}\boldsymbol{x}_q$, differentiating the softmax weights and their weighted label gives

$$\begin{aligned} \nabla_{\boldsymbol{u}} \pi_i &= \pi_i \left(y_i - \sum_j \pi_j y_j\right) \boldsymbol{Q}\boldsymbol{x}_q = \pi_i(y_i - t)\boldsymbol{Q}\boldsymbol{x}_q, \\ \nabla_{\boldsymbol{u}} F &= c\sum_i y_i \nabla_{\boldsymbol{u}} \pi_i = c\left(\sum_i \pi_i y_i^2 - t^2\right)\boldsymbol{Q}\boldsymbol{x}_q. \end{aligned} \tag{C.39}$$

Task-sign symmetry now cancels the expected label-key gradient. The task distribution is invariant under $f \mapsto -f$; at $\boldsymbol{u} = \boldsymbol{0}$, this transformation preserves the weights and prediction derivative and changes the residual $ct - y_q$ to its negative. Thus

$$\begin{aligned} \nabla_{\boldsymbol{u}} L\big|_{\boldsymbol{u}=\boldsymbol{0}} &= \mathbb{E}\left[(ct - y_q)c\left(\sum_i \pi_i y_i^2 - t^2\right)\boldsymbol{Q}\boldsymbol{x}_q\right] \\ &= -\mathbb{E}\left[(ct - y_q)c\left(\sum_i \pi_i y_i^2 - t^2\right)\boldsymbol{Q}\boldsymbol{x}_q\right] = -\nabla_{\boldsymbol{u}} L\big|_{\boldsymbol{u}=\boldsymbol{0}} = \boldsymbol{0}. \end{aligned} \tag{C.40}$$

Replacing the finite sums by tilted integrals gives the same cancellation for population contexts. The cancellation uses expectation over tasks and therefore applies to the population gradient; an individual stochastic batch may have a nonzero label-key gradient.

For the query and key gradients, use rotational symmetry. When $\boldsymbol{u} = \boldsymbol{0}$, the score depends on $\boldsymbol{B} = \boldsymbol{K}^\top \boldsymbol{Q}$; denote its population loss by $\mathcal{L}(\boldsymbol{B}, c)$. Applying the same orthogonal transformation to the input and task gives $\mathcal{L}(\boldsymbol{O}^\top \boldsymbol{B}\boldsymbol{O}, c) = \mathcal{L}(\boldsymbol{B}, c)$. At $\boldsymbol{B} = \beta \boldsymbol{I}_d$, the gradient consequently commutes with every orthogonal $\boldsymbol{O}$. Diagonal sign matrices force the off-diagonal entries to vanish, and permutation matrices force the diagonal entries to agree. The gradient is therefore a scalar multiple of the identity, with the scalar determined by its trace:

$$\begin{aligned} \nabla_{\boldsymbol{B}} \mathcal{L}(\beta \boldsymbol{I}_d, c) &= g \boldsymbol{I}_d, \\ \partial_\beta L(\beta, c) &= \langle \nabla_{\boldsymbol{B}} \mathcal{L}, \boldsymbol{I}_d \rangle_{\mathrm{F}} = dg, \\ \nabla_{\boldsymbol{B}} \mathcal{L}(\beta \boldsymbol{I}_d, c) &= \frac{1}{d}\partial_\beta L(\beta, c)\boldsymbol{I}_d. \end{aligned} \tag{C.41}$$

To recover the gradients in the trained matrix coordinates, substitute the differential $d\boldsymbol{B} = (d\boldsymbol{K})^\top \boldsymbol{Q} + \boldsymbol{K}^\top d\boldsymbol{Q}$ into the loss differential:

$$\begin{aligned} dL &= \mathrm{tr}\big[(\nabla_{\boldsymbol{B}} \mathcal{L})^\top \{(d\boldsymbol{K})^\top \boldsymbol{Q} + \boldsymbol{K}^\top d\boldsymbol{Q}\}\big] \\ &= \mathrm{tr}\big[\boldsymbol{Q}(\nabla_{\boldsymbol{B}} \mathcal{L})^\top (d\boldsymbol{K})^\top\big] + \mathrm{tr}\big[(\nabla_{\boldsymbol{B}} \mathcal{L})^\top \boldsymbol{K}^\top d\boldsymbol{Q}\big] \\ &= \big\langle \boldsymbol{Q}(\nabla_{\boldsymbol{B}} \mathcal{L})^\top, d\boldsymbol{K} \big\rangle_{\mathrm{F}} + \langle \boldsymbol{K}\nabla_{\boldsymbol{B}} \mathcal{L}, d\boldsymbol{Q} \rangle_{\mathrm{F}}. \end{aligned} \tag{C.42}$$

Identifying the coefficients of the two differentials yields

$$\begin{aligned} \nabla_{\boldsymbol{Q}} L &= \boldsymbol{K}\nabla_{\boldsymbol{B}} \mathcal{L}, \\ \nabla_{\boldsymbol{K}} L &= \boldsymbol{Q}(\nabla_{\boldsymbol{B}} \mathcal{L})^\top. \end{aligned} \tag{C.43}$$

At $\boldsymbol{Q} = \boldsymbol{K} = q\boldsymbol{I}_d$ with $q = \sqrt{\beta} > 0$, gradient flow hence satisfies

$$\dot{\boldsymbol{Q}} = \dot{\boldsymbol{K}} = -\frac{q}{d}\partial_\beta L\, \boldsymbol{I}_d,$$
$$\dot{q} = -\frac{q}{d}\partial_\beta L,$$
$$\dot{\beta} = 2q\dot{q} = -\frac{2\beta}{d}\partial_\beta L, \qquad \dot{c} = -\partial_c L. \tag{C.44}$$

These derivatives show that the symmetric parameter family is invariant under full-parameter population gradient flow. The label-key derivative vanishes by (C.40), and the remaining derivatives are tangent to the family. Local uniqueness of the smooth full ODE then implies that the trajectory from the stated initialization stays in this family under the unprojected optimizer.

Differentiating (C.38) gives the scalar gain dynamics:

$$\dot{c} = -\partial_c L = -E_1 a(ca-1) - E_2 b(cb-1) = E_1 a(1-ca) + E_2 b(1-cb). \tag{C.45}$$

The selectivity dynamics are

$$\dot{\beta} = -\frac{2\beta}{d}\partial_\beta L = -\frac{2\beta}{d}\{E_1 ca'(ca-1) + E_2 cb'(cb-1)\}$$
$$= \frac{2\beta c}{d}\{E_1 a'(1-ca) + E_2 b'(1-cb)\}. \tag{C.46}$$

All response coefficients are evaluated at the current $\beta$. The factor $2\beta/d$ comes from training the two matrices separately.

**Global existence and learning.** For $0 < p < 1$, it remains to show that $\beta(t) \to \infty$ and $c(t) \to 1$. We first obtain bounds valid for every $p \in [0, 1]$. The response formulas (C.35) already give $a, b, \rho \to 1$ as $\beta \to \infty$.

First take $E_1, E_2 > 0$. Define

$$A = E_1 a + E_2 b, \quad B = E_1 a^2 + E_2 b^2,$$
$$A_1 = E_1 a' + E_2 b', \quad B_1 = E_1 aa' + E_2 bb',$$
$$c_* = A/B, \qquad c_\dagger = A_1/B_1. \tag{C.47}$$

The two velocities are $\dot{c} = A - Bc = B(c_* - c)$ and $\dot{\beta} = (2\beta c/d)B_1(c_\dagger - c)$. Since $c(0) = 0$, an integrating factor gives

$$\frac{d}{dt}\left[e^{\int_0^t B(u)\,du} c(t)\right] = e^{\int_0^t B(u)\,du} A(t),$$
$$c(t) = e^{-\int_0^t B(u)\,du}\int_0^t e^{\int_0^s B(u)\,du} A(s)\,ds = \int_0^t A(s) e^{-\int_s^t B(u)\,du}\,ds > 0, \qquad t > 0. \tag{C.48}$$

The same formula gives positive gain at the pure-component endpoints whenever $E_1 + E_2 > 0$. To order the two gain thresholds, expand the numerator of their difference:

$$A_1 B - A B_1 = (E_1 a' + E_2 b')(E_1 a^2 + E_2 b^2) - (E_1 a + E_2 b)(E_1 aa' + E_2 bb')$$
$$= E_1 E_2 (a' b^2 + b' a^2 - abb' - baa')$$
$$= E_1 E_2 (a-b)(ab' - ba') = E_1 E_2 (a-b) a^2 \rho' > 0,$$
$$c_\dagger - c_* = \frac{E_1 E_2 (a-b) a^2 \rho'}{B B_1} > 0. \tag{C.49}$$

We check the boundaries of $0 \le c \le c_\dagger(\beta)$, $\beta \ge 1$. At the upper gain boundary, $\dot{\beta} = 0$, so

$$\frac{d}{dt}\{c - c_\dagger(\beta)\}\Big|_{c=c_\dagger(\beta)} = \{\dot{c} - c_\dagger'(\beta)\dot{\beta}\}\Big|_{c=c_\dagger(\beta)} = B(c_* - c_\dagger) < 0. \tag{C.50}$$

At the lower gain boundary, $\dot{c}|_{c=0} = A > 0$. At $\beta = 1$ within the gain boundaries,

$$\dot{\beta}\Big|_{\beta=1} = \frac{2c}{d} B_1 (c_\dagger - c) \ge 0. \tag{C.51}$$

The boundary derivatives prove invariance and imply $\dot{\beta} \geq 0$ along the trajectory. To obtain a uniform upper bound on the gain, set $C := 1/b(1)$ and use monotonicity of $a$ and $b$:

$$\dot{c}|_{c=C} = E_1 a\left(1 - \frac{a}{b(1)}\right) + E_2 b\left(1 - \frac{b}{b(1)}\right) \leq 0. \tag{C.52}$$

Combining the boundary calculations gives

$$\begin{aligned} \beta(t) \geq 1, \qquad \dot{\beta}(t) \geq 0, \\ 0 \leq c(t) \leq \min\{c_\dagger(\beta(t)), 1/b(1)\}. \end{aligned} \tag{C.53}$$

At the pure-component endpoints, the same boundary argument establishes these bounds with non-strict inequalities.

To establish global existence, combine the logarithmic derivative bounds (C.27) with $a, b < 1$ and the gain bound. These estimates give a uniform bound on the selectivity velocity:

$$\begin{aligned} \dot{\beta} \leq \frac{2C(1+C)}{d}\{E_1\beta a' + E_2\beta b'\} \leq \frac{2C(1+C)}{d}(E_1 + 2E_2) =: C', \\ \beta(t) \leq 1 + C't. \end{aligned} \tag{C.54}$$

Thus the full trajectory exists globally and remains bounded on every finite interval.

We now rule out convergence to finite selectivity for a mixture with both components present. Suppose for contradiction that $\beta(t)$ had a finite limit. The full parameter trajectory would then remain in a compact set. Along full gradient flow,

$$\begin{aligned} \frac{d}{dt}L(\boldsymbol{\theta}(t)) = \langle \nabla L(\boldsymbol{\theta}(t)), -\nabla L(\boldsymbol{\theta}(t))\rangle = -\left\|\nabla L(\boldsymbol{\theta}(t))\right\|^2, \\ \int_0^T \left\|\nabla L(\boldsymbol{\theta}(t))\right\|^2 dt = L(\boldsymbol{\theta}(0)) - L(\boldsymbol{\theta}(T)) \leq L(\boldsymbol{\theta}(0)). \end{aligned} \tag{C.55}$$

The energy identity implies that accumulation points must be stationary. Indeed, boundedness of the gradient and Hessian on the compact set bounds the time derivative of the squared gradient, $-2\nabla L^\top \nabla^2 L \nabla L$, and makes the squared gradient uniformly continuous. If it exceeded some positive $\epsilon$ at arbitrarily late times, there would be disjoint intervals on which it exceeds $\epsilon/2$, contradicting its finite integral. A stationary accumulation point is impossible: at $c = 0$, we have $\dot{c} > 0$, while at any positive stationary gain $c = c_*$, the strict ordering $c_* < c_\dagger$ implies $\dot{\beta} > 0$. Hence $\beta(t) \to \infty$.

Finally, since $a, b \to 1$, we have $B \to E_1 + E_2 > 0$ and $c_* \to 1$. Solving $\dot{c} + Bc = Bc_*$ with an integrating factor gives

$$\begin{aligned} c(t) - 1 = e^{-\int_{t_0}^t B(s)ds}\{c(t_0) - 1\} \\ + \int_{t_0}^t e^{-\int_s^t B(u)du} B(s)\{c_*(s) - 1\}\, ds, \\ \int_{t_0}^t e^{-\int_s^t B(u)du} B(s) ds = 1 - e^{-\int_{t_0}^t B(s)ds} \leq 1. \end{aligned} \tag{C.56}$$

Given $\epsilon > 0$, choose $t_0$ so $|c_*(s) - 1| \leq \epsilon$ thereafter. Then

$$\begin{aligned} |c(t) - 1| &\leq e^{-\int_{t_0}^t B(s)\, ds}|c(t_0) - 1| + \epsilon \int_{t_0}^t e^{-\int_s^t B(u)\, du} B(s)\, ds \\ &\leq e^{-\int_{t_0}^t B(s)\, ds}|c(t_0) - 1| + \epsilon. \end{aligned} \tag{C.57}$$

Because $B(s) \to E_1 + E_2 > 0$, the first term tends to zero. Letting $\epsilon \downarrow 0$ proves $c \to 1$, and therefore $ca, cb \to 1$. The normalized prediction risk on the training law consequently satisfies

$$\frac{2L(\beta(t), c(t))}{E_1 + E_2} = \frac{E_1(ca - 1)^2 + E_2(cb - 1)^2}{E_1 + E_2} \longrightarrow 0. \tag{C.58}$$

This proves convergence of the prediction risk for $0 < p < 1$. □

**Lemma C.3** (Risk after training on purely linear tasks)**.** *Let $d \geq 2$. Train by full-parameter population gradient flow from $(\boldsymbol{I}_d, \boldsymbol{I}_d, \mathbf{0}, 0)$ on the unit-energy purely linear law, with $\boldsymbol{w}$ uniform on the sphere of radius $\sqrt{d}$. For every finite $t \geq 0$,*

$$c(t)a(\beta(t)) < 1, \qquad \beta(t) \leq 1 + \frac{1}{da(1)^2}.$$

*Evaluation on the original law from Assumption 4.2 satisfies the dimension-independent bound*

$$R_0^{(\infty)}(\boldsymbol{\theta}(t)) \geq \frac{16p_0}{121} > \frac{p_0}{8}.$$

*Proof.* Set $E_1 = 1$ and $E_2 = 0$ in (C.45)–(C.46). Write $e = 1 - ca$ and $a_0 = a(1) > 0$. The dynamics induced by the query and key matrix factors give

$$\begin{aligned}
&\dot{c} = ae, \qquad \dot{\beta} = \frac{2\beta c}{d} a' e, \\
&\dot{e} = -a\dot{c} - ca'\dot{\beta} = -a^2 e - \frac{2\beta c^2}{d}(a')^2 e = -\left[a^2 + \frac{2\beta c^2}{d}(a')^2\right] e, \\
&e(t) = e^{-\int_0^t \left(a(\beta(s))^2 + \frac{2\beta(s)c(s)^2}{d} a'(\beta(s))^2\right) ds} > 0.
\end{aligned} \tag{C.59}$$

The positive residual implies that $c$ and $\beta$ are nondecreasing, $ca < 1$, and $c \leq 1/a_0$. We can now bound selectivity in terms of this bounded gain. For $t > 0$, we have $\dot{c} > 0$, so (C.27) gives

$$\frac{d\beta}{dc} = \frac{(2\beta c/d)a'e}{ae} = \frac{2\beta c}{d}\frac{a'}{a} \leq \frac{2c}{d}. \tag{C.60}$$

Integrating from the initialization $c(0) = 0$ gives

$$\beta(t) - 1 = \int_0^{c(t)} \frac{d\beta}{dc}(u)\, du \leq \int_0^{c(t)} \frac{2u}{d}\, du = \frac{c(t)^2}{d} \leq \frac{1}{da_0^2}. \tag{C.61}$$

We next bound the quadratic-to-linear response ratio uniformly over dimensions. Let $a_s = Z_s'/Z_s$ denote the linear response in dimension $s \geq 2$, as in Lemma C.1. Thus $a = a_d$ and $\rho = a_{d+2}$. Rearranging (C.25) and using $a_{s+2} > 0$ gives, for $\beta > 0$,

$$\begin{aligned}
a_s(\beta) &= \frac{\beta}{s + \beta a_{s+2}(\beta)} \leq \frac{\beta}{s}, \\
a_d(1) &= \frac{1}{d + a_{d+2}(1)} \geq \frac{1}{d + 1/(d+2)}.
\end{aligned} \tag{C.62}$$

Substituting this lower bound into (C.61) yields

$$\begin{aligned}
\beta(t) &\leq 1 + \frac{1}{da_d(1)^2} \leq 1 + \frac{(d + 1/(d+2))^2}{d} = d + 1 + \frac{2}{d+2} + \frac{1}{d(d+2)^2} \\
&\leq d + 1 + \frac{1}{2} + \frac{1}{32} = d + \frac{49}{32} < d + \frac{8}{5} =: \bar{\beta}_d.
\end{aligned} \tag{C.63}$$

Here both denominator bounds use $d \geq 2$.

To control $\rho$ at this value of $\beta$, combine the variance and integration-by-parts identities from the proof of Lemma C.1 with (C.27):

$$\begin{aligned}
a_s'(\beta) = \frac{Z_s''(\beta)}{Z_s(\beta)} - a_s(\beta)^2 = 1 - \frac{s-1}{\beta} a_s(\beta) - a_s(\beta)^2 \leq \frac{a_s(\beta)}{\beta}, \\
a_s(\beta)^2 + \frac{s}{\beta} a_s(\beta) - 1 \geq 0, \\
a_s(\beta) \geq \frac{\sqrt{s^2 + 4\beta^2} - s}{2\beta} = \frac{2\beta}{s + \sqrt{s^2 + 4\beta^2}}.
\end{aligned} \tag{C.64}$$

The last step uses $a_s(\beta) > 0$ to select the positive root. Applying this estimate in dimension $d+4$ to the recurrence (C.62) gives

$$\rho(\beta) = a_{d+2}(\beta) = \frac{\beta}{d+2+\beta a_{d+4}(\beta)} \le \frac{\beta}{d+2+\frac{1}{2}\big(\sqrt{(d+4)^2+4\beta^2}-(d+4)\big)} = \frac{2\beta}{d+\sqrt{(d+4)^2+4\beta^2}}. \tag{C.65}$$

At $\beta = \bar\beta_d = d + 8/5$, the last expression is below $7/11$. This comparison is equivalent to $22\bar\beta_d - 7d < 7\sqrt{(d+4)^2+4\bar\beta_d^2}$. Both sides are positive, so it suffices to verify

$$72\bar\beta_d^2 - 77d\bar\beta_d - 98d - 196 = -5d^2 + \frac{46}{5}d - \frac{292}{25} \le -\frac{4}{5}d - \frac{292}{25} < 0, \qquad d \ge 2. \tag{C.66}$$

Monotonicity of $\rho$ from (C.25), together with (C.63) and (C.65), now gives

$$\rho(\beta(t)) \le \rho(\bar\beta_d) \le \frac{2\bar\beta_d}{d+\sqrt{(d+4)^2+4\bar\beta_d^2}} < \frac{7}{11}. \tag{C.67}$$

Along pure-linear training, $ca < 1$. Hence the quadratic-response error is bounded away from zero, and (C.8) yields

$$\begin{aligned} 1 - c(t)b(\beta(t)) &= 1 - c(t)a(\beta(t))\rho(\beta(t)) > 1 - \rho(\beta(t)) > \frac{4}{11}, \\ R_0^{(\infty)}(\boldsymbol{\theta}(t)) &= (1-p_0)\{c(t)a(\beta(t))-1\}^2 + p_0\{c(t)b(\beta(t))-1\}^2 \\ &\ge p_0\left(\frac{4}{11}\right)^2 = \frac{16p_0}{121} > \frac{p_0}{8}. \end{aligned} \tag{C.68}$$

The bound holds uniformly over all $d \ge 2$ and finite training times $t \ge 0$. □

### C.3 Finite-horizon approximation by SGD

We next compare the population recursion with finite-context SGD over a prescribed finite number of generations. To measure the error between laws of bounded continuous task functions, define

$$\mathcal{D}(P,Q) = \inf_{\pi\in\Pi(P,Q)} \Big(\mathbb{E}_{(f,g)\sim\pi}[\|f-g\|_\infty^2]\Big)^{1/2}.$$

**Proposition C.1** (Finite-horizon approximation by SGD)**.** *Suppose Assumption 4.2 holds. Fix a finite generation horizon $R$, positive training times $T_1, \dots, T_R$, a tolerance $\epsilon > 0$, and $\delta \in (0,1)$. Let $(\bar{\boldsymbol{\theta}}_r, \bar P_r)$ denote the population-context recursion with full-parameter gradient flow from $(\boldsymbol{I}_d, \boldsymbol{I}_d, \mathbf{0}, 0)$ and global unit-energy normalization. There exist finite context lengths $n_r$, seed lengths $m_r$, batch sizes $B_r$, update counts $K_r$, and an initialization radius $\zeta > 0$ such that the following holds. At generation $r$, initialize independently within distance $\zeta$ of the reference initialization and run unprojected full-parameter SGD with fresh independent batches and step size $\eta_r = T_r/K_r$. Generate tasks from fresh independent seeds and normalize by their exact population energy. With probability at least $1-\delta$, all normalizers are positive and, simultaneously for $1 \le r \le R$,*

$$\begin{aligned} \big\|\boldsymbol{\theta}_r - \bar{\boldsymbol{\theta}}_r\big\| &\le \epsilon, & \mathcal{D}(P_r, \bar P_r) &\le \epsilon, \\ \Big|R_0^{(n_r)}(\boldsymbol{\theta}_r) - R_0^{(\infty)}(\bar{\boldsymbol{\theta}}_r)\Big| &\le \epsilon. \end{aligned} \tag{C.69}$$

The resources in this proposition may depend on the prescribed horizon and all training times. It does not assert approximation as $R \to \infty$ with fixed resources.

*Proof.* We separate four sources of error: finite training contexts, stochastic gradients, time discretization, and finite generation seeds. First we establish bounds on one compact parameter region. We then compare one training trajectory, propagate the task-law error through generation and normalization, and choose resources over the fixed horizon.

**The generated task law.** Finite-context generation may leave the linear–quadratic family, so we work with laws of bounded continuous functions.

To compare finite and population contexts at general parameters, let $f$ be bounded and continuous. The population-context output and loss are

$$F_{\boldsymbol{\theta},\infty}(f,\boldsymbol{x}) = c\frac{\int f(\boldsymbol{z})e^{(\boldsymbol{Q}\boldsymbol{x})^\top(\boldsymbol{K}\boldsymbol{z}+\boldsymbol{u}f(\boldsymbol{z}))}\,d\nu(\boldsymbol{z})}{\int e^{(\boldsymbol{Q}\boldsymbol{x})^\top(\boldsymbol{K}\boldsymbol{z}+\boldsymbol{u}f(\boldsymbol{z}))}\,d\nu(\boldsymbol{z})},$$
$$L_P^{(\infty)}(\boldsymbol{\theta}) = \frac{1}{2}\mathbb{E}_{f\sim P,\boldsymbol{x}_q\sim\nu}[(F_{\boldsymbol{\theta},\infty}(f,\boldsymbol{x}_q) - f(\boldsymbol{x}_q))^2]. \tag{C.70}$$

For a finite query tuple $X = (\boldsymbol{x}_1,\dots,\boldsymbol{x}_s)$, seed tuple $Z = (\boldsymbol{z}_1,\dots,\boldsymbol{z}_m)$, and label vector $\boldsymbol{y}$, define the vector of predictions by

$$[G_{\boldsymbol{\theta}}(X,Z,\boldsymbol{y})]_j = c\frac{\sum_{i=1}^m y_i e^{(\boldsymbol{Q}\boldsymbol{x}_j)^\top(\boldsymbol{K}\boldsymbol{z}_i+\boldsymbol{u}y_i)}}{\sum_{i=1}^m e^{(\boldsymbol{Q}\boldsymbol{x}_j)^\top(\boldsymbol{K}\boldsymbol{z}_i+\boldsymbol{u}y_i)}}. \tag{C.71}$$

We can use this prediction map to describe the full sequence of functions descended from an original task. Condition on the trained models and all generation normalizers, collectively $\mathcal{H}_r$. Draw a fresh original task and independent seed tuples $Z_1,\dots,Z_r$, independently of this history. To represent the original task, fix orthonormal bases of the linear and trace-free quadratic function spaces, let $\boldsymbol{a}_0$ denote its coefficients, and let $\boldsymbol{\Phi}_Z$ stack the basis evaluations at $Z$. These bases provide coordinates for the task law. The generated functions are then specified recursively by

$$\begin{aligned}\boldsymbol{y}_1 &= \boldsymbol{\Phi}_{Z_1}\boldsymbol{a}_0,\\ \boldsymbol{y}_j &= f_{j-1}(Z_j) = G_{\boldsymbol{\theta}_{j-1}}(Z_j,Z_{j-1},\boldsymbol{y}_{j-1}), \qquad 2\le j\le r,\\ f_r(X) &= G_{\boldsymbol{\theta}_r}(X,Z_r,\boldsymbol{y}_r).\end{aligned} \tag{C.72}$$

This construction uses one parent function to label every input in a seed tuple. We absorb each generation normalizer into the corresponding gain when generating the next task law; clean evaluation continues to use the trained model's original gain.

**Uniform bounds for prediction and training.** We work on a bounded parameter region containing the reference trajectories. The constants below depend only on this region and a uniform bound on task values.

Fix $R_\theta \ge 1$ and tasks with $\|f\|_\infty \le M$. Consider the convex parameter region

$$\|\boldsymbol{Q}\|_{\mathrm{F}} \le R_\theta, \qquad \|[\boldsymbol{K}\ \boldsymbol{u}]\|_{\mathrm{F}} \le R_\theta, \qquad |c| \le R_\theta. \tag{C.73}$$

Write $H_0 = \sqrt{1+M^2}$, $\Lambda = R_\theta^2 H_0$, and $V = \sqrt{2}R_\theta H_0$ for bounds on the augmented input norm, absolute score, and score-gradient norm, respectively. With $\boldsymbol{B} = [\boldsymbol{K}\ \boldsymbol{u}]$, $\boldsymbol{h}_i = [\boldsymbol{x}_i; y_i]$, and $s_i = (\boldsymbol{Q}\boldsymbol{x}_q)^\top\boldsymbol{B}\boldsymbol{h}_i$, we have

$$\begin{aligned}|s_i| &\le \|\boldsymbol{Q}\|_{\mathrm{op}}\|\boldsymbol{x}\|_q\|\boldsymbol{B}\|_{\mathrm{op}}\|\boldsymbol{h}\|_i \le R_\theta^2 H_0 = \Lambda,\\ \|\nabla_{\boldsymbol{Q},\boldsymbol{B}}s_i\|^2 &= \left\|\boldsymbol{B}\boldsymbol{h}_i\boldsymbol{x}_q^\top\right\|_{\mathrm{F}}^2 + \left\|\boldsymbol{Q}\boldsymbol{x}_q\boldsymbol{h}_i^\top\right\|_{\mathrm{F}}^2 \le R_\theta^2H_0^2 + R_\theta^2H_0^2 = V^2.\end{aligned} \tag{C.74}$$

For unit parameter directions $\boldsymbol{v} = (\Delta\boldsymbol{Q}_v, \Delta\boldsymbol{B}_v)$ and $\boldsymbol{w} = (\Delta\boldsymbol{Q}_w, \Delta\boldsymbol{B}_w)$, differentiating the bilinear score twice gives

$$\begin{aligned}D^2 s_i[\boldsymbol{v},\boldsymbol{w}] &= (\Delta\boldsymbol{Q}_v\boldsymbol{x}_q)^\top\Delta\boldsymbol{B}_w\boldsymbol{h}_i + (\Delta\boldsymbol{Q}_w\boldsymbol{x}_q)^\top\Delta\boldsymbol{B}_v\boldsymbol{h}_i,\\ |D^2 s_i[\boldsymbol{v},\boldsymbol{w}]| &\le H_0(\|\Delta\boldsymbol{Q}_v\|_{\mathrm{F}}\|\Delta\boldsymbol{B}_w\|_{\mathrm{F}} + \|\Delta\boldsymbol{Q}_w\|_{\mathrm{F}}\|\Delta\boldsymbol{B}_v\|_{\mathrm{F}})\\ &\le H_0\sqrt{\|\Delta\boldsymbol{Q}_v\|_{\mathrm{F}}^2 + \|\Delta\boldsymbol{B}_v\|_{\mathrm{F}}^2}\sqrt{\|\Delta\boldsymbol{Q}_w\|_{\mathrm{F}}^2 + \|\Delta\boldsymbol{B}_w\|_{\mathrm{F}}^2} = H_0.\end{aligned} \tag{C.75}$$

The final inequality follows from Cauchy–Schwarz applied to the two pairs of block norms in reversed order.

For the predictor, write $\pi_i = e^{s_i}/\sum_j e^{s_j}$, $t = \sum_i \pi_i y_i$, and $F = ct$. Differentiating the softmax weights and their weighted label with the labels fixed gives

$$d\pi_i = \pi_i\,ds_i - \pi_i\sum_j \pi_j\,ds_j = \pi_i\left(ds_i - \sum_j \pi_j\,ds_j\right), \tag{C.76}$$

$$dt = \sum_i y_i\,d\pi_i = \sum_i \pi_i(y_i - t)\,ds_i. \tag{C.77}$$

In particular, for a logit direction $\boldsymbol{h}$,

$$\|D\,\mathrm{softmax}[\boldsymbol{h}]\|_1 = \sum_i \pi_i \left| h_i - \sum_j \pi_j h_j \right| \le \sum_i \pi_i |h_i| + \left| \sum_j \pi_j h_j \right| \le 2\,\|\boldsymbol{h}\|_\infty . \tag{C.78}$$

For second derivatives, subscripts on the score denote the indicated parameter directions. For $t = \mathbb{E}_\pi[y]$, use $D(\mathbb{E}_\pi[g])[\boldsymbol{w}] = \mathbb{E}_\pi[Dg[\boldsymbol{w}]] + \mathbb{E}_\pi[g(s_w - \mathbb{E}_\pi[s_w])]$. Differentiating (C.77) and combining the centered terms yields

$$\begin{aligned} D^2 t[\boldsymbol{v},\boldsymbol{w}] = &-Dt[\boldsymbol{w}]\mathbb{E}_\pi[s_v] + \mathbb{E}_\pi[(y-t)s_{vw}] \\ &+ \mathbb{E}_\pi[(y-t)s_v(s_w - \mathbb{E}_\pi[s_w])] = \mathbb{E}_\pi[(y-t)s_{vw}] \\ &+ \mathbb{E}_\pi[(y-t)(s_v - \mathbb{E}_\pi[s_v])(s_w - \mathbb{E}_\pi[s_w])]. \end{aligned} \tag{C.79}$$

Since $|y-t| \le 2M$ and $|s_v - \mathbb{E}_\pi[s_v]| \le 2V$, this proves

$$\begin{aligned} |t| \le M, \qquad \|\nabla t\| &\le 2MV, \\ \left\|\nabla^2 t\right\|_{\mathrm{op}} &\le 2MH_0 + 8MV^2, \\ \|\nabla F\| &\le M + 2R_\theta MV \le M(1 + 4R_\theta^2 H_0) =: J, \\ \left\|\nabla^2 F\right\|_{\mathrm{op}} &\le R_\theta(2MH_0 + 8MV^2) + 4MV \le M(8R_\theta H_0 + 16R_\theta^3 H_0^2) =: H. \end{aligned} \tag{C.80}$$

The term $4MV$ in the Hessian bound accounts for the two $c$–score mixed blocks. These predictor estimates give gradient and Hessian bounds for the sample loss $\ell = \frac{1}{2}(F - y_q)^2$. By the chain rule,

$$\nabla_{\boldsymbol{\theta}} \ell = (F - y_q)\nabla_{\boldsymbol{\theta}} F. \tag{C.81}$$

Differentiating once more gives

$$\begin{aligned} \nabla^2 \ell &= (\nabla F)(\nabla F)^\top + (F - y_q)\nabla^2 F, \\ \|\nabla \ell\| &\le (R_\theta + 1)MJ =: G, \\ \left\|\nabla^2 \ell\right\|_{\mathrm{op}} &\le J^2 + (R_\theta + 1)MH =: L. \end{aligned} \tag{C.82}$$

These deterministic bounds hold for both finite and integral contexts. They also justify differentiation under the population expectations, so the same Hessian bound controls the Lipschitz constant of the population gradient.

**Bias from finite training contexts.** We control the numerator and denominator of each weighted average. Condition on the task, query, and current parameter. Let $U = e^s$, $\boldsymbol{v} = \nabla_{\boldsymbol{Q},\boldsymbol{B}} s$, and use a bar for the average over $n$ independent context inputs. Both $\overline{U}$ and $\mathbb{E}[U]$ are at least $e^{-\Lambda}$. For any scalar or vector $h$ with $\|h\| \le C$, the centered-average identity (A.25), applied to $Uh - \mathbb{E}[Uh]$, gives

$$\mathbb{E}[\left\|\overline{Uh} - \mathbb{E}[Uh]\right\|^2] = \frac{1}{n}\mathbb{E}[\|Uh - \mathbb{E}[Uh]\|^2] \le \frac{1}{n}\mathbb{E}[\|Uh\|^2] \le \frac{C^2 e^{2\Lambda}}{n}. \tag{C.83}$$

The case $h = 1$ gives $\mathbb{E}[|\overline{U} - \mathbb{E}[U]|^2] \le e^{2\Lambda}/n$. Let $\mu_h = \mathbb{E}[Uh]/\mathbb{E}[U]$, so $\|\mu_h\| \le C$. The quotient error is

$$\frac{\overline{Uh}}{\overline{U}} - \mu_h = \frac{\overline{Uh} - \mathbb{E}[Uh]}{\overline{U}} + \mu_h \frac{\mathbb{E}[U] - \overline{U}}{\overline{U}}. \tag{C.84}$$

Using $\overline{U} \ge e^{-\Lambda}$, squaring, and applying (C.83) gives

$$\mathbb{E}[\left\| \frac{\overline{Uh}}{\overline{U}} - \mu_h \right\|^2] \le 2e^{2\Lambda}\left\{\mathbb{E}[\left\|\overline{Uh} - \mathbb{E}[Uh]\right\|^2] + C^2\mathbb{E}[|\overline{U} - \mathbb{E}[U]|^2]\right\} \le \frac{4C^2 e^{4\Lambda}}{n}. \tag{C.85}$$

Cauchy–Schwarz therefore gives the first-moment estimate

$$\mathbb{E}[\left\| \frac{\overline{Uh}}{\overline{U}} - \mu_h \right\|] \le \left(\mathbb{E}[\left\| \frac{\overline{Uh}}{\overline{U}} - \mu_h \right\|^2]\right)^{1/2} \le \frac{2Ce^{2\Lambda}}{\sqrt{n}}. \tag{C.86}$$

Apply this bound to $h = y$, $h = \boldsymbol{v}$, and $h = y\boldsymbol{v}$. Writing $\mathbb{E}_n[\cdot]$ and $\mathbb{E}_\infty[\cdot]$ for the empirical and population tilted means, respectively, their gradients satisfy

$$\begin{aligned}\nabla t_n - \nabla t_\infty &= \mathbb{E}_n[y\boldsymbol{v}] - t_n\mathbb{E}_n[\boldsymbol{v}] - \mathbb{E}_\infty[y\boldsymbol{v}] + t_\infty\mathbb{E}_\infty[\boldsymbol{v}] \\ &= \{\mathbb{E}_n[y\boldsymbol{v}] - \mathbb{E}_\infty[y\boldsymbol{v}]\} - (t_n - t_\infty)\mathbb{E}_n[\boldsymbol{v}] \\ &\quad - t_\infty\{\mathbb{E}_n[\boldsymbol{v}] - \mathbb{E}_\infty[\boldsymbol{v}]\}. \end{aligned} \tag{C.87}$$

Applying the quotient estimate to $y$, $\boldsymbol{v}$, and $y\boldsymbol{v}$ gives

$$\begin{aligned}\mathbb{E}[\|\nabla t_n - \nabla t_\infty\|] &\le \mathbb{E}[\|\mathbb{E}_n[y\boldsymbol{v}] - \mathbb{E}_\infty[y\boldsymbol{v}]\|] + V\mathbb{E}[|t_n - t_\infty|] \\ &\quad + M\mathbb{E}[\|\mathbb{E}_n[\boldsymbol{v}] - \mathbb{E}_\infty[\boldsymbol{v}]\|] \\ &\le \frac{2MVe^{2\Lambda}}{\sqrt{n}} + V\frac{2Me^{2\Lambda}}{\sqrt{n}} + M\frac{2Ve^{2\Lambda}}{\sqrt{n}} = \frac{6MVe^{2\Lambda}}{\sqrt{n}}. \end{aligned} \tag{C.88}$$

The prediction difference obeys

$$\mathbb{E}[|F_n - F_\infty|] = |c|\mathbb{E}[|t_n - t_\infty|] \le \frac{2R_\theta Me^{2\Lambda}}{\sqrt{n}}. \tag{C.89}$$

Including the $c$-derivative in the full prediction gradient gives

$$\begin{aligned}\mathbb{E}[\|\nabla F_n - \nabla F_\infty\|] &\le \mathbb{E}[|t_n - t_\infty|] + |c|\mathbb{E}[\|\nabla t_n - \nabla t_\infty\|] \le \frac{2Me^{2\Lambda}}{\sqrt{n}} + \frac{6R_\theta MVe^{2\Lambda}}{\sqrt{n}} \\ &= \frac{(2M + 6R_\theta MV)e^{2\Lambda}}{\sqrt{n}}. \end{aligned} \tag{C.90}$$

Adding and subtracting $(F_\infty - y_q)\nabla F_n$ gives

$$\begin{aligned}\nabla\ell_n - \nabla\ell_\infty &= (F_n - F_\infty)\nabla F_n + (F_\infty - y_q)(\nabla F_n - \nabla F_\infty), \\ \left\|\nabla L_P^{(n)} - \nabla L_P^{(\infty)}\right\| &\le J\mathbb{E}[|F_n - F_\infty|] + (R_\theta + 1)M\mathbb{E}[\|\nabla F_n - \nabla F_\infty\|] \\ &\le \frac{M^2e^{2\Lambda}}{\sqrt{n}}\{2R_\theta(1 + 4R_\theta^2H_0) + (R_\theta + 1)(2 + 6\sqrt{2}R_\theta^2H_0)\} \\ &\le \frac{D}{\sqrt{n}}, \qquad D = 32M^2(1 + R_\theta)^3(1 + H_0)e^{2\Lambda}. \end{aligned} \tag{C.91}$$

Finally, factor the difference of squared losses:

$$\begin{aligned}|L_P^{(n)} - L_P^{(\infty)}| &\le \frac{1}{2}\mathbb{E}[|F_n - F_\infty|\,|F_n + F_\infty - 2y_q|] \le (R_\theta + 1)M\mathbb{E}[|F_n - F_\infty|] \\ &\le \frac{C_L}{\sqrt{n}}, \qquad C_L = 2R_\theta(R_\theta + 1)M^2e^{2\Lambda}. \end{aligned} \tag{C.92}$$

The constants $J, H, G, L, D, C_L$ depend only on the bounds on parameters and task values. Because each SGD step uses fresh contexts, these estimates also hold conditional on the current random iterate.

**SGD over one generation.** Let $\bar{\boldsymbol{\theta}}(t)$ solve population-context gradient flow for $0 \le t \le T$. Choose a radius $h > 0$ such that the closed neighborhood of that radius around the trajectory lies inside the parameter region. Run unprojected SGD with constant step size $\eta$, $K\eta = T$, batch size $B$, and initial distance at most $\epsilon_0$.

We first separate the stochastic gradient into its conditional mean and centered error. Let $\mathcal{F}_k$ contain the initialization and all batches preceding update $k$, so $\boldsymbol{\theta}_k$ is fixed conditional on $\mathcal{F}_k$. For the current update, write $\boldsymbol{g}^{(b)} := \nabla\ell(\boldsymbol{\theta}_k; \xi_k^{(b)})$ for the gradient of episode $b$ in the fresh batch, suppressing $k$ only on individual batch entries. Define

$$\begin{aligned}\widehat{\boldsymbol{g}}_k &:= \frac{1}{B}\sum_{b=1}^{B}\boldsymbol{g}^{(b)}, \\ \boldsymbol{\xi}_k &:= \widehat{\boldsymbol{g}}_k - \nabla L_P^{(n)}(\boldsymbol{\theta}_k), \\ \boldsymbol{b}_k &:= \nabla L_P^{(n)}(\boldsymbol{\theta}_k) - \nabla L_P^{(\infty)}(\boldsymbol{\theta}_k). \end{aligned} \tag{C.93}$$

For $\boldsymbol{\xi}^{(b)} := \boldsymbol{g}^{(b)} - \nabla L_P^{(n)}(\boldsymbol{\theta}_k)$, differentiation under expectation, justified by (C.82), gives $\mathbb{E}[\boldsymbol{\xi}^{(b)} \mid \mathcal{F}_k] = \mathbf{0}$. The centered episode gradients are conditionally independent, so the same centered-average identity used in (C.83) applies conditional on $\mathcal{F}_k$:

$$\mathbb{E}[\boldsymbol{\xi}_k \mid \mathcal{F}_k] = \mathbf{0},$$
$$\mathbb{E}[\|\boldsymbol{\xi}_k\|^2 \mid \mathcal{F}_k] = \frac{1}{B}\mathbb{E}[\left\|\boldsymbol{\xi}^{(1)}\right\|^2 \mid \mathcal{F}_k] = \frac{1}{B}\left\{\mathbb{E}[\left\|\boldsymbol{g}^{(1)}\right\|^2 \mid \mathcal{F}_k] - \left\|\nabla L_P^{(n)}(\boldsymbol{\theta}_k)\right\|^2\right\} \le \frac{G^2}{B}. \quad \text{(C.94)}$$

While the iterate lies in the parameter region, these estimates bound the conditional variance of the stochastic gradient. The finite-context bias estimate (C.91), evaluated at the same iterate, gives $\|\boldsymbol{b}_k\| \le D/\sqrt{n}$.

To use the local bounds without imposing a constraint on SGD, let $\tau$ be the first iterate outside the chosen neighborhood. Retain increments only at $k < \tau$, a decision measurable before drawing batch $k$, and set $\boldsymbol{m}_j = -\eta \sum_{k<j} \boldsymbol{\xi}_k \mathbf{1}_{\{k<\tau\}}$. For $i < j$, both the earlier increment and $\mathbf{1}_{\{j<\tau\}}$ are $\mathcal{F}_j$-measurable. Hence the cross term vanishes:

$$\mathbb{E}[\langle \boldsymbol{\xi}_i \mathbf{1}_{\{i<\tau\}}, \boldsymbol{\xi}_j \mathbf{1}_{\{j<\tau\}} \rangle] = \mathbb{E}\left[\mathbf{1}_{\{i<\tau\}}\mathbf{1}_{\{j<\tau\}} \langle \boldsymbol{\xi}_i, \mathbb{E}[\boldsymbol{\xi}_j \mid \mathcal{F}_j]\rangle\right] = 0. \quad \text{(C.95)}$$

Expanding the square and then conditioning each diagonal term gives

$$\mathbb{E}[\|\boldsymbol{m}_K\|^2] = \eta^2 \sum_{k=0}^{K-1} \mathbb{E}[\|\boldsymbol{\xi}_k\|^2 \mathbf{1}_{\{k<\tau\}}] = \eta^2 \sum_{k=0}^{K-1} \mathbb{E}\left[\mathbf{1}_{\{k<\tau\}}\mathbb{E}[\|\boldsymbol{\xi}_k\|^2 \mid \mathcal{F}_k]\right] \le \eta^2 \sum_{k=0}^{K-1} \frac{G^2}{B}$$
$$= \frac{\eta^2 K G^2}{B} = \frac{\eta T G^2}{B}. \quad \text{(C.96)}$$

Doob's $L^2$ maximal inequality gives

$$\mathbb{E}\left[\max_{0\le j\le K} \|\boldsymbol{m}_j\|^2\right] \le 4\mathbb{E}[\|\boldsymbol{m}_K\|^2] \le \frac{4\eta T G^2}{B}. \quad \text{(C.97)}$$

Markov's inequality therefore yields

$$\Pr\left(\max_{0\le j\le K} \|\boldsymbol{m}_j\| > 2G\sqrt{\frac{T\eta}{B\delta}}\right) \le \frac{\mathbb{E}[\max_{0\le j\le K} \|\boldsymbol{m}_j\|^2]}{4G^2 T\eta/(B\delta)} \le \delta. \quad \text{(C.98)}$$

Next bound the discretization error. The gradient is $L$-Lipschitz and $\left\|\dot{\bar{\boldsymbol{\theta}}}\right\| \le G$ in the parameter region. Integrating the ODE over one step gives

$$\bar{\boldsymbol{\theta}}((k+1)\eta) = \bar{\boldsymbol{\theta}}(k\eta) - \eta\nabla L_P^{(\infty)}(\bar{\boldsymbol{\theta}}(k\eta)) + \boldsymbol{r}_k,$$
$$\boldsymbol{r}_k = -\int_{k\eta}^{(k+1)\eta} [\nabla L_P^{(\infty)}(\bar{\boldsymbol{\theta}}(s)) - \nabla L_P^{(\infty)}(\bar{\boldsymbol{\theta}}(k\eta))]ds,$$
$$\|\boldsymbol{r}_k\| \le L\int_{k\eta}^{(k+1)\eta} \left\|\bar{\boldsymbol{\theta}}(s) - \bar{\boldsymbol{\theta}}(k\eta)\right\| ds \le LG\int_{k\eta}^{(k+1)\eta} (s - k\eta)ds = \tfrac{1}{2}LG\eta^2. \quad \text{(C.99)}$$

Write $\boldsymbol{e}_k := \boldsymbol{\theta}_k - \bar{\boldsymbol{\theta}}(k\eta)$. Subtracting the ODE step from the SGD update gives

$$\boldsymbol{e}_{k+1} = \boldsymbol{\theta}_k - \eta\widehat{\boldsymbol{g}}_k - \bar{\boldsymbol{\theta}}(k\eta) + \eta\nabla L_P^{(\infty)}(\bar{\boldsymbol{\theta}}(k\eta)) - \boldsymbol{r}_k$$
$$= \boldsymbol{e}_k - \eta\left\{\nabla L_P^{(\infty)}(\boldsymbol{\theta}_k) - \nabla L_P^{(\infty)}(\bar{\boldsymbol{\theta}}(k\eta))\right\}$$
$$- \eta\boldsymbol{b}_k - \eta\boldsymbol{\xi}_k - \boldsymbol{r}_k. \quad \text{(C.100)}$$

For $j \le \min\{\tau, K\}$, summing these identities gives

$$\boldsymbol{e}_j = \boldsymbol{e}_0 - \eta\sum_{k=0}^{j-1}\left\{\nabla L_P^{(\infty)}(\boldsymbol{\theta}_k) - \nabla L_P^{(\infty)}(\bar{\boldsymbol{\theta}}(k\eta))\right\}$$
$$- \eta\sum_{k=0}^{j-1}\boldsymbol{b}_k + \boldsymbol{m}_j - \sum_{k=0}^{j-1}\boldsymbol{r}_k. \quad \text{(C.101)}$$

On the event $\max_{0\le j\le K}\|\boldsymbol{m}_j\| \le 2G\sqrt{T\eta/(B\delta)}$, which has probability at least $1-\delta$ by (C.98), the triangle inequality and the three error bounds yield

$$\begin{aligned}\|\boldsymbol{e}_j\| &\le \|\boldsymbol{e}_0\| + L\eta\sum_{k=0}^{j-1}\|\boldsymbol{e}_k\| + \eta\sum_{k=0}^{j-1}\|\boldsymbol{b}_k\| + \|\boldsymbol{m}_j\| + \sum_{k=0}^{j-1}\|\boldsymbol{r}_k\| \\ &\le \epsilon_0 + L\eta\sum_{k=0}^{j-1}\|\boldsymbol{e}_k\| + \frac{j\eta D}{\sqrt{n}} + 2G\sqrt{\frac{T\eta}{B\delta}} + \frac{1}{2}jLG\eta^2 \\ &\le \epsilon_0 + \frac{DT}{\sqrt{n}} + \frac{1}{2}LGT\eta + 2G\sqrt{\frac{T\eta}{B\delta}} + L\eta\sum_{k=0}^{j-1}\|\boldsymbol{e}_k\|. \end{aligned} \tag{C.102}$$

Discrete Gronwall's inequality, applied through the first possible exit, gives

$$\max_{0\le j\le \min\{\tau,K\}}\left\|\boldsymbol{\theta}_j - \bar{\boldsymbol{\theta}}(j\eta)\right\| \le e^{LT}\left[\epsilon_0 + \frac{DT}{\sqrt{n}} + \frac{1}{2}LGT\eta + 2G\sqrt{\frac{T\eta}{B\delta}}\right] =: \epsilon_\theta. \tag{C.103}$$

If $\epsilon_\theta < h$, the bound at a first exit would contradict the definition of that exit. Thus $\tau > K$, and (C.103) holds for all $0 \le j \le K$ along the actual SGD trajectory. For a random initialization, the argument applies conditional on its value within the specified neighborhood.

For clean risk, evaluation uses $P_0$, which may differ from the training law $P$, so choose the task-value bound $M$ to cover both laws. The mean-value theorem in parameters and (C.92) applied to the clean law give

$$\begin{aligned}|R_0^{(n)}(\boldsymbol{\theta}_K) - R_0^{(\infty)}(\bar{\boldsymbol{\theta}}(T))| &\le \frac{2}{\mathcal{E}(P_0)}|L_{P_0}^{(n)}(\boldsymbol{\theta}_K) - L_{P_0}^{(n)}(\bar{\boldsymbol{\theta}}(T))| \\ &\quad + \frac{2}{\mathcal{E}(P_0)}|L_{P_0}^{(n)}(\bar{\boldsymbol{\theta}}(T)) - L_{P_0}^{(\infty)}(\bar{\boldsymbol{\theta}}(T))| \le \frac{2G\epsilon_\theta}{\mathcal{E}(P_0)} + \frac{2C_L}{\mathcal{E}(P_0)\sqrt{n}}. \end{aligned} \tag{C.104}$$

If evaluation uses an independent test context length, it replaces $n$ only in the last term.

**Stability with respect to the task law.** The task law changes at each generation. We therefore bound the effect of a task perturbation on both prediction and training.

Fix coupled tasks $f, g$ with sup norms at most $M$, use the same context and query inputs, and put $\Delta = \|f-g\|_\infty$. Only the label-dependent part of the key changes the scores, so

$$\begin{aligned}|s_i(f) - s_i(g)| &= |(\boldsymbol{Q}\boldsymbol{x}_q)^\top \boldsymbol{u}\{f(\boldsymbol{x}_i) - g(\boldsymbol{x}_i)\}| \le R_\theta^2\Delta, \\ \|\boldsymbol{\pi}(f) - \boldsymbol{\pi}(g)\|_1 &\le 2R_\theta^2\Delta, \\ t(f) - t(g) &= \sum_i \pi_i(f)\{f(\boldsymbol{x}_i) - g(\boldsymbol{x}_i)\} + \sum_i\{\pi_i(f) - \pi_i(g)\}g(\boldsymbol{x}_i), \\ |t(f) - t(g)| &\le (1 + 2MR_\theta^2)\Delta =: C_t\Delta, \\ |F(f) - F(g)| &\le R_\theta C_t\Delta =: C_f\Delta. \end{aligned} \tag{C.105}$$

The softmax estimate follows by integrating (C.78) along the segment between the two score vectors. The argument also applies to integral contexts.

For the training gradient, let $\boldsymbol{v}_i = \nabla_{\boldsymbol{Q},\boldsymbol{B}} s_i$. Each parameter block changes by at most $R_\theta\Delta$, and hence $\|\boldsymbol{v}_i(f) - \boldsymbol{v}_i(g)\| \le \sqrt{2}R_\theta\Delta =: V_1\Delta$. Expanding the difference of the three factors in $\nabla t = \sum_i \pi_i(y_i - t)\boldsymbol{v}_i$ gives

$$\begin{aligned}\nabla t(f) - \nabla t(g) &= \sum_i(\pi_i(f) - \pi_i(g))(y_i(f) - t(f))\boldsymbol{v}_i(f) \\ &\quad + \sum_i \pi_i(g)[y_i(f) - t(f) - y_i(g) + t(g)]\boldsymbol{v}_i(f) \\ &\quad + \sum_i \pi_i(g)(y_i(g) - t(g))[\boldsymbol{v}_i(f) - \boldsymbol{v}_i(g)], \\ \|\nabla t(f) - \nabla t(g)\| &\le [4MR_\theta^2 V + (1 + C_t)V + 2MV_1]\Delta =: D_t\Delta, \\ \|\nabla F(f) - \nabla F(g)\| &\le (C_t + R_\theta D_t)\Delta =: J_f\Delta. \end{aligned} \tag{C.106}$$

To pass from the predictor gradient to the loss gradient, include the change in the query target in the residual bound: $|(F(f) - f(\boldsymbol{x}_q)) - (F(g) - g(\boldsymbol{x}_q))| \le (R_\theta C_t + 1)\Delta$. Writing $r_f = F(f) - f(\boldsymbol{x}_q)$ and $r_g = F(g) - g(\boldsymbol{x}_q)$, add and subtract $r_g \nabla F(f)$:

$$\nabla \ell(f) - \nabla \ell(g) = r_f \nabla F(f) - r_g \nabla F(g) = (r_f - r_g)\nabla F(f) + r_g\{\nabla F(f) - \nabla F(g)\}. \quad \text{(C.107)}$$

The resulting norm bound is

$$\begin{aligned} \|\nabla \ell(f) - \nabla \ell(g)\| &\le (R_\theta C_t + 1)\Delta J + (R_\theta + 1) M J_f \Delta \\ &= L_f \Delta, \qquad L_f := (R_\theta C_t + 1)J + (R_\theta + 1)MJ_f. \end{aligned} \quad \text{(C.108)}$$

To compare task laws, take any coupling $\pi$ of the two laws and use common context inputs and a common query in the two gradients. The expectations below include these shared inputs. Jensen's inequality gives

$$\begin{aligned} \left\| \nabla L_P^{(\infty)} - \nabla L_Q^{(\infty)} \right\| &= \|\mathbb{E}_\pi[\nabla \ell(f) - \nabla \ell(g)]\| \le \mathbb{E}_\pi[\|\nabla \ell(f) - \nabla \ell(g)\|] \\ &\le L_f \mathbb{E}_\pi[\|f - g\|_\infty] \le L_f (\mathbb{E}_\pi[\|f - g\|_\infty^2])^{1/2}, \\ \left\| \nabla L_P^{(\infty)} - \nabla L_Q^{(\infty)} \right\| &\le L_f \mathcal{D}(P, Q). \end{aligned} \quad \text{(C.109)}$$

Taking couplings arbitrarily close to the infimum gives the last line. The estimate applies to the gradient in all trained coordinates, including the label-key column.

**Error from finite generation seeds.** We next compare finite-seed and population generation in sup norm. At a fixed query, apply (C.85) with $h = y$, $C = M$, and $n = m$, then multiply by the squared gain:

$$\mathbb{E}_Z[|F_{\boldsymbol{\theta},Z} f(\boldsymbol{x}) - F_{\boldsymbol{\theta},\infty} f(\boldsymbol{x})|^2] \le \frac{4c^2 M^2 e^{4\Lambda}}{m} \le \frac{4R_\theta^2 M^2 e^{4\Lambda}}{m}. \quad \text{(C.110)}$$

To obtain a bound under $\mathcal{D}$, we strengthen the fixed-query estimate to a sup-norm estimate. Holding the seed points and their task values fixed and differentiating with respect to the query gives

$$\begin{aligned} \nabla_{\boldsymbol{x}} s_i &= \boldsymbol{Q}^\top \boldsymbol{B} \boldsymbol{h}_i, \qquad \|\nabla_{\boldsymbol{x}} s_i\| \le R_\theta^2 H_0, \\ \|\nabla_{\boldsymbol{x}} F\| &= \left\| c \sum_i \pi_i (y_i - t) \nabla_{\boldsymbol{x}} s_i \right\| \le 2R_\theta^3 H_0 M. \end{aligned} \quad \text{(C.111)}$$

The same derivative bound holds for the integral output and throughout the unit ball. Thus the output difference is $4R_\theta^3 H_0 M$-Lipschitz in Euclidean query distance. Because differentiation acts only on the query, boundedness of the task values suffices for this estimate.

We use a finite covering of the sphere to combine the query-Lipschitz bound with the pointwise second-moment estimate. A maximal $\epsilon$-separated subset is an $\epsilon$-net. Balls of radius $\epsilon/2$ around its points are disjoint and lie in a ball of radius $1 + \epsilon/2$, so their number is at most $(1 + 2/\epsilon)^d \le (3/\epsilon)^d$ for $0 < \epsilon \le 1$. For the output difference $h_Z$ and this net $\mathcal{N}$,

$$\begin{aligned} \|h_Z\|_\infty &\le \max_{\boldsymbol{z} \in \mathcal{N}} |h_Z(\boldsymbol{z})| + 4R_\theta^3 H_0 M \epsilon, \\ \mathbb{E}[\|h_Z\|_\infty^2] &\le 2 \sum_{\boldsymbol{z} \in \mathcal{N}} \mathbb{E}[|h_Z(\boldsymbol{z})|^2] + 32 R_\theta^6 H_0^2 M^2 \epsilon^2 \le \frac{8R_\theta^2 M^2 e^{4\Lambda}}{m} (3/\epsilon)^d + 32 R_\theta^6 H_0^2 M^2 \epsilon^2. \end{aligned} \quad \text{(C.112)}$$

Taking $\epsilon = m^{-1/(d+2)}$ balances $m^{-1}\epsilon^{-d}$ and $\epsilon^2$ and gives

$$\begin{aligned} \mathbb{E}[\|h_Z\|_\infty^2] &\le S^2 m^{-2/(d+2)}, \\ S &= M \sqrt{8R_\theta^2 3^d e^{4\Lambda} + 32 R_\theta^6 H_0^2}. \end{aligned} \quad \text{(C.113)}$$

**Propagation across generations.** We now combine the estimates across generations. Fix a finite generation horizon $R$ and reference times $T_r$. Let $\bar{P}_r, \bar{\boldsymbol{\theta}}_r$ denote the population recursion and $P_r, \boldsymbol{\theta}_r$ the finite-context SGD recursion. Choose a common finite task-value bound and compact neighborhoods of the parameter trajectories over this horizon, and allocate failure probabilities $\delta_r$ with $\sum_{r\le R}\delta_r \le \delta$. Let $\epsilon_0^{(r)}$ bound the initialization error at generation $r$. Starting from $d_0 = \mathcal{D}(P_0, \bar{P}_0)$, define deterministic bounds on the parameter and task-law errors by

$$e_r = e^{LT_r}\left[\epsilon_0^{(r)} + \frac{DT_r}{\sqrt{n_r}} + L_f T_r d_{r-1} + \frac{1}{2}LGT_r\eta_r + 2G\sqrt{\frac{T_r\eta_r}{B_r\delta_r}}\right], \tag{C.114}$$

$$d_r = C_f d_{r-1} + Je_r + Sm_r^{-1/(d+2)}. \tag{C.115}$$

For the induction step, suppose the preceding generations satisfy these error bounds. Conditional on that history, (C.109) bounds the difference between the actual and reference training drifts by $L_f d_{r-1}$. Integrating this difference over training time gives the third term in $e_r$; the other terms come from the one-generation estimate (C.103). If $e_r$ is smaller than the radius of the chosen parameter neighborhood, the parameter error is at most $e_r$ with conditional probability at least $1-\delta_r$.

To bound the generated-law error, couple the parent tasks and use the same fresh seed for both models. Decompose the function difference into the effects of parent-task error, parameter error, and finite-seed error:

$$\begin{aligned} F_{\boldsymbol{\theta}_r,Z}f - F_{\bar{\boldsymbol{\theta}}_r,\infty}\bar{f} = {} & \{F_{\boldsymbol{\theta}_r,Z}f - F_{\boldsymbol{\theta}_r,Z}\bar{f}\} \\ & + \{F_{\boldsymbol{\theta}_r,Z}\bar{f} - F_{\bar{\boldsymbol{\theta}}_r,Z}\bar{f}\} \\ & + \{F_{\bar{\boldsymbol{\theta}}_r,Z}\bar{f} - F_{\bar{\boldsymbol{\theta}}_r,\infty}\bar{f}\}. \end{aligned} \tag{C.116}$$

On the event where the parameter bound holds, apply the root-mean-square triangle inequality to this decomposition:

$$\begin{aligned} &\left(\mathbb{E}\left[\left\|F_{\boldsymbol{\theta}_r,Z}f - F_{\bar{\boldsymbol{\theta}}_r,\infty}\bar{f}\right\|_\infty^2\right]\right)^{1/2} \\ &\quad \le C_f\left(\mathbb{E}[\left\|f-\bar{f}\right\|_\infty^2]\right)^{1/2} + J\left\|\boldsymbol{\theta}_r - \bar{\boldsymbol{\theta}}_r\right\| + Sm_r^{-1/(d+2)}. \end{aligned} \tag{C.117}$$

Here the three terms use (C.105), the parameter derivative bound $J$, and (C.113), respectively. Taking parent couplings arbitrarily close to the infimum yields

$$\begin{aligned} \mathcal{D}(P_r, \bar{P}_r) &\le C_f\mathcal{D}(P_{r-1}, \bar{P}_{r-1}) + J\left\|\boldsymbol{\theta}_r - \bar{\boldsymbol{\theta}}_r\right\| + Sm_r^{-1/(d+2)} \\ &\le C_f d_{r-1} + Je_r + Sm_r^{-1/(d+2)} = d_r. \end{aligned} \tag{C.118}$$

For simultaneous control, let $\mathcal{A}_r$ denote success through generation $r$. The conditional failure bounds imply

$$\Pr(\mathcal{A}_R^c) \le \sum_{r=1}^R \Pr(\mathcal{A}_{r-1}\cap\mathcal{A}_r^c) \le \sum_{r=1}^R \delta_r \le \delta. \tag{C.119}$$

Starting with identical real laws gives $d_0 = 0$.

**Stability of global normalization.** It remains to account for global unit-energy normalization in the law recursion. For coupled laws $P, Q$ before normalization, write $s_P = \sqrt{\mathcal{E}(P)}$, $s_Q = \sqrt{\mathcal{E}(Q)}$ and assume $s_P, s_Q \ge a > 0$. The reverse triangle inequality in the joint $L^2$ space first bounds the difference between the normalization denominators:

$$|s_P - s_Q| = \left|\|f\|_{L^2(\pi\otimes\nu)} - \|g\|_{L^2(\pi\otimes\nu)}\right| \le \|f-g\|_{L^2(\pi\otimes\nu)} \le (\mathbb{E}_\pi[\|f-g\|_\infty^2])^{1/2}. \tag{C.120}$$

We then separate the task difference from the denominator difference in the normalized functions:

$$\begin{aligned} \frac{f}{s_P} - \frac{g}{s_Q} &= \frac{f-g}{s_P} + g\frac{s_Q - s_P}{s_P s_Q}, \\ \left(\mathbb{E}_\pi[\left\|\frac{f}{s_P} - \frac{g}{s_Q}\right\|_\infty^2]\right)^{1/2} &\le \frac{(\mathbb{E}_\pi[\|f-g\|_\infty^2])^{1/2}}{a} + \frac{M|s_P - s_Q|}{a^2}, \\ \mathcal{D}((f\mapsto f/s_P)_\#P, (g\mapsto g/s_Q)_\#Q) &\le \left(\frac{1}{a} + \frac{M}{a^2}\right)\mathcal{D}(P,Q). \end{aligned} \tag{C.121}$$

Thus normalization multiplies the bound for the unnormalized outputs in (C.115) by a finite stability factor at each generation. For fixed positive reference training times, all reference outputs have positive energy before normalization. Since the horizon is finite, their square-root energies have a positive minimum. Choosing the unnormalized-law error sufficiently small keeps each actual square-root energy above half this minimum. Uniform task-value bounds can then be constructed successively from $|c| \le R_\theta$ and these denominator lower bounds. This argument proves positivity of the normalizers as part of the approximation. It uses exact population energies; applying it to estimated normalizers, as in the experiments, would require an additional calibration-error bound.

**Choice of finite resources.** Assumption 4.2 bounds the initial task values:

$$\|f\|_\infty \le \|\boldsymbol{w}\| + \|\boldsymbol{A}\|_{\mathrm{op}} \le \sqrt{d(1-p_0)} + \sqrt{\tfrac{1}{2}d(d+2)p_0}. \tag{C.122}$$

Every reference trajectory is compact over its fixed training interval, and every reference output has positive energy before normalization. There are finitely many generations, so we can successively choose bounded parameter regions, neighborhoods of positive radius, finite task-value bounds, and positive lower bounds for the normalization denominators over the prescribed horizon.

Starting from $d_0 = 0$, (C.114)–(C.115), together with the normalization factor in (C.121), form finitely many continuous error recursions. Choose a sufficiently small initialization radius, sufficiently large finite $n_r, m_r$, and sufficiently small $\eta_r = T_r/K_r$ with finite integers $K_r$; choose finite $B_r$ large enough to control the stochastic terms. The resulting error bounds can simultaneously be made smaller than all neighborhood radii and any prescribed positive tolerance. The clean-risk estimate following (C.103) then gives the risk approximation in the proposition. The union bound establishes all conclusions with probability at least $1 - \delta$. □

# D PROOF OF THEOREM 4

The proof uses the energy-odds recursion from Appendix C. The only additional ingredient is the relation between training time and the attention parameter $\beta$. We state the needed bounds as Lemmas D.1–D.3, prove the theorem from them, and then establish the lemmas.

For a single generation, let $p \in [0,1]$ be the quadratic fraction of a unit-energy task law with the symmetries in Assumption 4.2. Write $(\beta(t;p), c(t;p))$ for its full-parameter gradient-flow trajectory from $(\boldsymbol{I}_d, \boldsymbol{I}_d, \mathbf{0}, 0)$, and abbreviate its loss by $L_p(t) := L_p(\beta(t;p), c(t;p))$. We also write $h(\beta) := -2\ln\rho(\beta)$ for the decrease in quadratic log odds in one generation. Positive constants denoted by $k, K$ may change from one inequality to the next; their dependencies are specified below.

**Lemma D.1** (Bounds on the component responses)**.** *For $\beta \geq 1$, each of $a(\beta) - b(\beta)$ and $h(\beta)$ is bounded above and below by positive dimension-only multiples of $(1+\beta)^{-1}$. The quantity $a(\beta)b'(\beta) - b(\beta)a'(\beta)$ is similarly comparable to $(1+\beta)^{-2}$. Moreover, the upper bounds in* (D.18) *hold with a dimension-only constant $D_d < \infty$.*

**Lemma D.2** (Uniform upper bound on learned selectivity)**.** *There is a constant $U_d < \infty$, depending only on $d$, such that $(1+\beta(T;p))^3 \leq U_d(1+T)$ for every $p \in [0,1]$ and $T \geq 0$.*

**Lemma D.3** (Training rates for interior mixtures)**.** *Fix $[p_-, p_+] \subset (0,1)$. There are constants $0 < k \leq K < \infty$, depending only on $d, p_-, p_+$, such that, for every $p \in [p_-, p_+]$ and $T \geq 0$,*

$$k(1+T)^{1/3} \leq 1 + \beta(T;p) \leq K(1+T)^{1/3},$$
$$L_p(T) \leq K(1+T)^{-2/3}.$$

## D.1 PROOF OF THEOREM 4

*Proof. Case $0 \leq \gamma \leq 3$.* Apply Lemma D.2 to each generation. Let $(\beta_r, c_r)$ denote the generation-$r$ parameters after time $T_r$, trained on the law $P_{r-1}$ with quadratic fraction $p_{r-1}$. The positivity result (C.48) makes normalization well defined for $T_r > 0$. Taking logarithms in (C.5) gives

$$\begin{aligned}\ln\frac{p_r}{1-p_r} &= \ln\frac{p_0}{1-p_0} + 2\sum_{j=1}^{r}\ln\rho(\beta_j)\\ &= \ln\frac{p_0}{1-p_0} - \sum_{j=1}^{r} h(\beta_j). \end{aligned} \tag{D.1}$$

The bound in Lemma D.2 is uniform over the training mixture. Combining it with Lemma D.1, even when $p_{j-1}$ approaches zero, gives

$$h(\beta_j) \geq \frac{k_d}{1+\beta_j} \geq \frac{k_d}{U_d^{1/3}(1+T_j)^{1/3}}. \tag{D.2}$$

For $T_j = Aj^\gamma$ with $\gamma \geq 0$ and $j \geq 1$,

$$\begin{aligned}1 + T_j &= 1 + Aj^\gamma \leq (1+A)j^\gamma,\\ h(\beta_j) &\geq \frac{k_d}{U_d^{1/3}(1+A)^{1/3}} j^{-\gamma/3},\\ \sum_{j=1}^{r} h(\beta_j) &\geq k\sum_{j=1}^{r} j^{-\gamma/3}. \end{aligned} \tag{D.3}$$

Here $k > 0$ depends only on $d, A$. The integral comparison gives

$$\begin{aligned}\sum_{j=1}^{r} j^{-\gamma/3} &\geq \int_1^{r+1} x^{-\gamma/3}\,dx\\ &= \begin{cases}\dfrac{(r+1)^{1-\gamma/3}-1}{1-\gamma/3}, & 0 \leq \gamma < 3,\\ \ln(r+1), & \gamma = 3,\end{cases}\\ &\longrightarrow \infty. \end{aligned} \tag{D.4}$$

Exponentiating (D.1) therefore yields

$$0 \le p_r \le \frac{p_r}{1-p_r} \le \frac{p_0}{1-p_0} e^{-k\sum_{j=1}^{r} j^{-\gamma/3}} \longrightarrow 0. \tag{D.5}$$

This proves the vanishing quadratic-energy fraction for $0 \le \gamma \le 3$. The argument concerns task composition and does not yield a clean-risk lower bound for these growing schedules.

*Case* $\gamma > 3$. For $\gamma > 3$, we keep the accumulated loss of quadratic log odds below the amount required to halve the initial fraction. Apply Lemma D.3 on $[p_0/2, p_0]$. For $T \ge 1$ and $p$ in this interval,

$$h(\beta(T;p)) \le \frac{K_d}{1+\beta(T;p)} \le K(1+T)^{-1/3} \le KT^{-1/3}, \tag{D.6}$$

where $K$ depends only on $d, p_0$. The decrease in log odds from $p_0$ to $p_0/2$ is

$$D_0 := \ln\frac{p_0}{1-p_0} - \ln\frac{p_0/2}{1-p_0/2} = \ln\frac{2-p_0}{1-p_0} > 0. \tag{D.7}$$

The training-time exponent gives a finite sum

$$S_\gamma := \sum_{j=1}^{\infty} j^{-\gamma/3} \le 1 + \int_1^\infty x^{-\gamma/3}\,dx = 1 + \frac{1}{\gamma/3-1} < \infty. \tag{D.8}$$

The following choice is sufficient:

$$A \ge A_0 := \max\left\{1, \left(\frac{2KS_\gamma}{D_0}\right)^3\right\}. \tag{D.9}$$

We prove $p_r \in [p_0/2, p_0]$ by induction. The initial fraction $p_0$ lies in this interval. Suppose $p_0, \dots, p_{r-1}$ lie there. Then the interior estimate applies to each of the first $r$ training generations, so

$$\begin{aligned}\sum_{j=1}^{r} h(\beta_j) &\le K\sum_{j=1}^{r} T_j^{-1/3} = KA^{-1/3}\sum_{j=1}^{r} j^{-\gamma/3} \\ &\le KA^{-1/3}S_\gamma \le \frac{D_0}{2} < D_0.\end{aligned} \tag{D.10}$$

Substituting into (D.1) gives

$$\begin{aligned}\ln\frac{p_r}{1-p_r} &= \ln\frac{p_0}{1-p_0} - \sum_{j=1}^{r} h(\beta_j) \\ &> \ln\frac{p_0}{1-p_0} - D_0 = \ln\frac{p_0}{2-p_0}.\end{aligned} \tag{D.11}$$

Since the log-odds function is strictly increasing, this implies $p_r > p_0/2$. Also $\rho(\beta_r) < 1$ makes the odds, and hence the fractions, decrease: $p_r \le p_{r-1} \le p_0$. This closes the induction and proves $p_r \ge p_0/2$ for every generation.

*Risk under the original task law.* The same interval also lets us compare training loss with the risk under $P_0$. Let $p = p_{r-1} \in [p_0/2, p_0]$, and write $e_1 = c_r a(\beta_r) - 1$ and $e_2 = c_r b(\beta_r) - 1$. The clean-risk formula (C.8) and the loss formula (C.38) differ only in their component energy weights. Since the original task energy equals one,

$$\begin{aligned}R_0^{(\infty)}(\boldsymbol{\theta}_r(T_r)) &= (1-p_0)e_1^2 + p_0 e_2^2 \\ &\le (1-p)e_1^2 + 2pe_2^2 \le 2\{(1-p)e_1^2 + pe_2^2\} \\ &= 4L_p(\beta_r, c_r) \le K(1+T_r)^{-2/3} \longrightarrow 0.\end{aligned} \tag{D.12}$$

The first inequality uses $p \le p_0$ for the linear term and $p \ge p_0/2$ for the quadratic term. The last bound follows from Lemma D.3, uniformly on the interval preserved above. Its constant depends only on $d, p_0$, and in particular can be chosen uniformly over $A \ge A_0$. Together with the induction, this proves (17).

□

### D.2 PROOFS OF THE SUPPORTING LEMMAS

We first collect the identities shared by the three lemmas. The population responses and full-parameter dynamics were derived in Appendix C. Equations (C.22) and (C.25) give

$$a(\beta) = \frac{Z_d'(\beta)}{Z_d(\beta)}, \qquad b(\beta) = 1 - \frac{da(\beta)}{\beta}, \qquad \rho(\beta) = \frac{b(\beta)}{a(\beta)} = a_{d+2}(\beta). \tag{D.13}$$

Taking $E_1 = 1 - p$ and $E_2 = p$ in (C.38) gives $L_p(\beta, c) = \frac{1}{2}(1-p)(ca-1)^2 + \frac{1}{2}p(cb-1)^2$. By (C.45)–(C.46), the actual full-parameter trajectory remains in $\boldsymbol{u} = \mathbf{0}$, $\boldsymbol{Q} = \boldsymbol{K} = \sqrt{\beta}\boldsymbol{I}_d$ and satisfies

$$\dot{c} = -\partial_c L_p, \qquad \dot{\beta} = -\frac{2\beta}{d}\partial_\beta L_p. \tag{D.14}$$

Together with the factor gradients in (C.43), these equations express loss dissipation in the original parameter coordinates:

$$\begin{aligned}\frac{d}{dt}L_p &= \partial_c L_p \dot{c} + \partial_\beta L_p \dot{\beta} \\ &= -(\partial_c L_p)^2 - \frac{2\beta}{d}(\partial_\beta L_p)^2 = -\left\|\nabla_{\boldsymbol{\theta}} L_p\right\|^2.\end{aligned} \tag{D.15}$$

*Proof of Lemma D.1.* The training-time bounds depend on the gap between the two responses and on the derivative of their ratio. We use their positivity and the logarithmic derivative bounds from (C.25) and (C.27). The expansions in (C.35) give

$$\begin{aligned}a - b &= \left(1 - \frac{d-1}{2\beta} + O(\beta^{-2})\right) - \left(1 - \frac{d}{\beta} + O(\beta^{-2})\right) = \frac{d+1}{2\beta} + O(\beta^{-2}), \\ ab' - ba' &= \left(1 + O(\beta^{-1})\right)\left(\frac{d}{\beta^2} + O(\beta^{-3})\right) \\ &\quad - \left(1 + O(\beta^{-1})\right)\left(\frac{d-1}{2\beta^2} + O(\beta^{-3})\right) \\ &= \frac{d+1}{2\beta^2} + O(\beta^{-3}).\end{aligned} \tag{D.16}$$

Both differences are strictly positive: $a - b > 0$ and $ab' - ba' = a^2\rho' > 0$. The continuous functions $(1+\beta)(a-b)$ and $(1+\beta)^2(ab'-ba')$ therefore have positive minima and finite maxima on compact subintervals of $[1, \infty)$, and positive finite limits at infinity. Combining compact and tail bounds gives dimension-only constants with

$$\begin{aligned}\frac{k_d}{1+\beta} \le a - b \le \frac{K_d}{1+\beta}, \\ \frac{k_d}{(1+\beta)^2} \le ab' - ba' \le \frac{K_d}{(1+\beta)^2}, \qquad \beta \ge 1.\end{aligned} \tag{D.17}$$

The same argument gives a finite dimension-only $D_d$ such that

$$\begin{aligned}1 - a,\ 1 - b \le \frac{D_d}{1+\beta}, \\ a',\ b' \le \frac{D_d}{(1+\beta)^2}, \qquad \beta \ge 1.\end{aligned} \tag{D.18}$$

For $h(\beta) = -2\ln\rho(\beta)$, the expansion of $\rho$ in (C.35) gives

$$h(\beta) = -2\ln\left(1 - \frac{d+1}{2\beta} + O(\beta^{-2})\right) = \frac{d+1}{\beta} + O(\beta^{-2}). \tag{D.19}$$

Positivity and the same compact-and-tail argument now imply

$$\frac{k_d}{1+\beta} \le h(\beta) \le \frac{K_d}{1+\beta}, \qquad \beta \ge 1. \tag{D.20}$$

□

*Proof of Lemma D.2.* Fix $p \in [0,1]$. We bound $\beta$ first while $c \le 1$, and then after $c$ reaches one. The constants will be independent of $p$, including at the two endpoints. Define

$$\begin{aligned} \alpha &:= (1-p)a + pb, & B &:= (1-p)a^2 + pb^2, \\ \alpha_1 &:= (1-p)a' + pb', & B_1 &:= (1-p)aa' + pbb', \\ c_* &:= \frac{\alpha}{B}, & c_\dagger &:= \frac{\alpha_1}{B_1}. \end{aligned} \tag{D.21}$$

All denominators are positive for $\beta \ge 1$. The invariant-region and global-existence result (C.53), with $E_1 = 1-p$, $E_2 = p$, and the prescribed initialization, gives, uniformly over $p \in [0,1]$,

$$\begin{aligned} \beta(t) \ge 1, \qquad \dot\beta(t) \ge 0, \\ 0 \le c(t) \le \min\{c_\dagger(\beta(t)), C_d\}, \qquad C_d := \frac{1}{b(1)}. \end{aligned} \tag{D.22}$$

We divide the trajectory at the first time the gain reaches one.

Before this time, $c \le 1$ and both residuals $1 - ca$ and $1 - cb$ are nonnegative. The logarithmic derivative bounds therefore give

$$\begin{aligned} \dot\beta &= \frac{2\beta c}{d}\{(1-p)a'(1-ca) + pb'(1-cb)\} \\ &\le \frac{2c}{d}\{(1-p)a(1-ca) + 2pb(1-cb)\} \\ &\le \frac{4c}{d}\{(1-p)a(1-ca) + pb(1-cb)\} = \frac{4c}{d}\dot c. \end{aligned} \tag{D.23}$$

Since $\beta(0) = 1$ and $c(0) = 0$, integrating up to any such time gives

$$\beta(t) - 1 = \int_0^t \dot\beta(s)\,ds \le \frac{4}{d}\int_0^t c(s)\dot c(s)\,ds = \frac{2}{d}(c(t)^2 - c(0)^2) \le \frac{2}{d}. \tag{D.24}$$

Let $t_0 := \inf\{t : c(t) = 1\}$. If $t_0 = \infty$, (D.24) bounds $\beta$ for all time. Otherwise, at $c = 1$,

$$\dot c = (1-p)a(1-a) + pb(1-b) > 0, \tag{D.25}$$

so $c(t) \ge 1$ for every $t \ge t_0$.

After $t_0$, $1 - c\lambda \le 1 - \lambda$ for $\lambda = a, b$. The nonnegative selectivity velocity and (D.18) give

$$\begin{aligned} 0 \le \dot\beta &\le \frac{2\beta C_d}{d}\{(1-p)a'(1-a) + pb'(1-b)\} \\ &\le \frac{2\beta C_d}{d}\left\{(1-p)\frac{D_d^2}{(1+\beta)^3} + p\frac{D_d^2}{(1+\beta)^3}\right\} \\ &= \frac{2C_d D_d^2}{d}\frac{\beta}{(1+\beta)^3}. \end{aligned} \tag{D.26}$$

Thus the cube of $1 + \beta$ grows at most linearly:

$$\frac{d}{dt}(1+\beta)^3 = 3(1+\beta)^2\dot\beta \le \frac{6C_d D_d^2}{d}\frac{\beta}{1+\beta} \le \frac{6C_d D_d^2}{d}. \tag{D.27}$$

Before $t_0$, (D.24) and $d \ge 2$ give $\beta \le 2$. Integrating the last inequality when $t_0 \le T < \infty$, and using this initial bound when $T < t_0$, proves

$$(1 + \beta(T;p))^3 \le 27 + \frac{6C_d D_d^2}{d}T \le U_d(1+T), \qquad U_d := 27 + \frac{6C_d D_d^2}{d}. \tag{D.28}$$

The resulting upper bound holds for the selectivity attained by full-parameter training, for every mixture and every $T \ge 0$.

□

*Proof of Lemma D.3. Loss at a fixed selectivity.* Fix $p \in [p_-, p_+]$. All constants in this proof depend only on $d, p_-, p_+$. Use the quantities $\alpha, B, \alpha_1, B_1, c_*$ defined in the proof of Lemma D.2, and write $s = 1 + \beta$ and $e = c - c_*(\beta)$. Completing the square separates the gain error from the loss at the optimal gain:

$$\begin{aligned} L_p(\beta, c) &= \frac{1}{2}\{(1-p)(c^2a^2 - 2ca + 1) + p(c^2b^2 - 2cb + 1)\} \\ &= \frac{1}{2}(Bc^2 - 2\alpha c + 1) = \frac{1}{2}B\left(c - \frac{\alpha}{B}\right)^2 + \frac{1}{2}\left(1 - \frac{\alpha^2}{B}\right) \\ &= \frac{1}{2}Be^2 + \frac{B - \alpha^2}{2B}. \end{aligned} \tag{D.29}$$

The remaining numerator is the variance of the two component responses:

$$\begin{aligned} B - \alpha^2 &= (1-p)a^2 + pb^2 - \left((1-p)^2a^2 + 2p(1-p)ab + p^2b^2\right) \\ &= p(1-p)a^2 - 2p(1-p)ab + p(1-p)b^2 = p(1-p)(a-b)^2. \end{aligned} \tag{D.30}$$

Thus the minimum over the gain and the full loss satisfy

$$\begin{aligned} L_*(\beta) &:= L_p(\beta, c_*(\beta)) = \frac{p(1-p)(a-b)^2}{2B}, \\ L_p(\beta, c) &= \frac{1}{2}Be^2 + L_*(\beta). \end{aligned} \tag{D.31}$$

Here $b(1)^2 \le B \le 1$ and $1 \le c_* \le 1/b(1)$: the latter upper bound follows from $B \ge b(1)\alpha$, and the lower bound from $B \le \alpha$. The actual gain is also bounded by (D.22). Using the two-sided bound for $a - b$ and $p(1-p) \ge p_-(1-p_+) > 0$ gives

$$\begin{aligned} L_*(\beta) &\ge \frac{p_-(1-p_+)k_d^2}{2s^2}, \\ L_*(\beta) &\le \frac{K_d^2}{8b(1)^2 s^2}, \\ ks^{-2} \le L_*(\beta) &\le Ks^{-2}, \\ L_p(\beta, c) &\le \frac{1}{2}e^2 + Ks^{-2} \le K(s^{-2} + e^2). \end{aligned} \tag{D.32}$$

*A lower bound on the full gradient.* We next show that either the gain error or the selectivity derivative forces loss to decrease. Let $q = \partial_\beta L_p = c(B_1 c - \alpha_1)$. The numerator identity (C.49) gives its value at $c_*$:

$$\begin{aligned} q_* := \partial_\beta L_p(\beta, c_*) &= c_*\left(B_1\frac{\alpha}{B} - \alpha_1\right) \\ &= -\frac{c_*}{B}(\alpha_1 B - \alpha B_1) = -\frac{c_*}{B}p(1-p)(a-b)(ab' - ba'). \end{aligned} \tag{D.33}$$

The bounds for the two response differences, together with the positive upper and lower bounds on the remaining factors, imply

$$ks^{-3} \le |q_*| \le Ks^{-3}. \tag{D.34}$$

To compare the derivatives at the trained gain $c$ and optimal gain $c_*$, expand

$$\begin{aligned} q - q_* &= B_1(c^2 - c_*^2) - \alpha_1(c - c_*) = [B_1(c + c_*) - \alpha_1]e, \\ |q - q_*| &\le \left(B_1(c + c_*) + \alpha_1\right)|e| \le Ks^{-2}|e|, \end{aligned} \tag{D.35}$$

where $B_1, \alpha_1 \le Ks^{-2}$ by (D.18). The gain derivative is exactly

$$\partial_c L_p = Bc - \alpha = B(c - c_*) = Be. \tag{D.36}$$

Write $\mathcal{J} = (\partial_c L_p)^2 + (2\beta/d)q^2$. It first controls the gain error:

$$\mathcal{J} = B^2e^2 + \frac{2\beta}{d}q^2 \ge B^2e^2 \ge b(1)^4e^2. \tag{D.37}$$

The following estimate also controls the derivative at the gain minimizer when $q$ and $q - q_*$ partially cancel:

$$\begin{aligned}\beta q_*^2 &= \beta\big(q - (q - q_*)\big)^2 \le 2\beta q^2 + 2\beta(q - q_*)^2 \\ &\le 2\beta q^2 + 2K^2\beta s^{-4}e^2 \\ &\le d\mathcal{J} + \frac{2K^2\beta s^{-4}}{b(1)^4}\mathcal{J} \le \left(d + \frac{2K^2}{b(1)^4}\right)\mathcal{J}. \end{aligned} \tag{D.38}$$

Since $\beta = s - 1 \ge s/2$ for $\beta \ge 1$, (D.34) implies $\beta q_*^2 \ge ks^{-5}$. Combining both estimates therefore gives

$$\begin{aligned} 2\mathcal{J} &\ge b(1)^4 e^2 + ks^{-5}, \\ \mathcal{J} &\ge k(e^2 + s^{-5}). \end{aligned} \tag{D.39}$$

*Integration along the training trajectory.* We now integrate the gradient bound to obtain the training-time rates. The bounds on $c, c_*$ give $|e| \le C_e$ for a dimension-only constant. From (D.32) and $(x+y)^{5/2} \le 2^{3/2}(x^{5/2} + y^{5/2})$ for $x, y \ge 0$,

$$\begin{aligned} L_p^{5/2} &\le K(s^{-2} + e^2)^{5/2} \le K(s^{-5} + |e|^5) \\ &\le K(s^{-5} + C_e^3 e^2) \le K(s^{-5} + e^2). \end{aligned} \tag{D.40}$$

The response and gain bounds hold from initialization and throughout $\beta \ge 1$. Consequently, the dissipation identity (D.15) and the gradient bound (D.39) imply, for all $t \ge 0$,

$$-\frac{d}{dt}L_p = \mathcal{J} \ge k(e^2 + s^{-5}) \ge kL_p^{5/2}. \tag{D.41}$$

For an interior mixture, $L_p \ge L_* > 0$ at finite $\beta$, so negative powers of $L_p$ are well defined. Differentiating and integrating them gives

$$\begin{aligned} \frac{d}{dt}L_p^{-3/2} &= -\frac{3}{2}L_p^{-5/2}\dot{L}_p \ge \frac{3}{2}k, \\ L_p(T)^{-3/2} - L_p(0)^{-3/2} &= \int_0^T \frac{d}{dt}L_p(t)^{-3/2}\,dt \ge \frac{3}{2}kT. \end{aligned} \tag{D.42}$$

The initial predictor is zero, so $L_p(0) = 1/2$. Inverting the last inequality yields

$$\begin{aligned} L_p(T)^{-3/2} &\ge 2^{3/2} + \frac{3}{2}kT, \\ L_p(T) &\le \left(2^{3/2} + \frac{3}{2}kT\right)^{-2/3} \le K(1+T)^{-2/3}. \end{aligned} \tag{D.43}$$

The lower loss bound from (D.32) then forces selectivity to increase:

$$\begin{aligned} \frac{k}{(1+\beta(T;p))^2} &\le L_p(T) \le K(1+T)^{-2/3}, \\ (1+\beta(T;p))^2 &\ge \frac{k}{K}(1+T)^{2/3}, \\ 1+\beta(T;p) &\ge \sqrt{\frac{k}{K}}(1+T)^{1/3}. \end{aligned} \tag{D.44}$$

Together with (D.28), this proves

$$k(1+T)^{1/3} \le 1 + \beta(T;p) \le K(1+T)^{1/3}. \tag{D.45}$$

These estimates apply along the full training trajectory; the optimal gain $c_*$ enters only through the loss decomposition.

Equations (D.43) and (D.45) prove the estimates in Lemma D.3. □

### D.3 Extension to a finite generation context

The following bound concerns one fixed model and one generation. It is separate from the population recursion in Theorem 4.

**Proposition D.1** (Single-view context sampling)**.** *Let $\nu$ be uniform on the unit sphere in $\mathbb{R}^d$, $d \geq 2$, and let $\|f\|_\infty \leq M$. For $\beta \geq 0$ and independent seed inputs $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m \sim \nu$, define*

$$(K_\beta f)(\boldsymbol{x}) = \frac{\int e^{\beta \boldsymbol{x}^\top \boldsymbol{z}} f(\boldsymbol{z})\, d\nu(\boldsymbol{z})}{Z_d(\beta)},$$
$$(\widehat{K}_{\beta,m} f)(\boldsymbol{x}) = \frac{\sum_{i=1}^m e^{\beta \boldsymbol{x}^\top \boldsymbol{z}_i} f(\boldsymbol{z}_i)}{\sum_{i=1}^m e^{\beta \boldsymbol{x}^\top \boldsymbol{z}_i}}. \tag{D.46}$$

*For every unit query $\boldsymbol{x}$ and $m \geq 1$,*

$$\mathbb{E}[|\widehat{K}_{\beta,m} f(\boldsymbol{x}) - K_\beta f(\boldsymbol{x})|^2] \leq \frac{32 M^2 N_d(\beta)}{m}, \qquad N_d(\beta) = \frac{Z_d(2\beta)}{Z_d(\beta)^2}. \tag{D.47}$$

*The same bound holds for the expected squared $L^2(\nu)$ error. Moreover, if $\beta \geq 1$ and $|c| \leq 1/b(1)$, there is a dimension-only constant $C_d < \infty$ such that*

$$\mathbb{E}[\left\| c\widehat{K}_{\beta,m} f - cK_\beta f \right\|_{L^2(\nu)}^2] \leq C_d \frac{M^2 \beta^{(d-1)/2}}{m}. \tag{D.48}$$

*In particular, this gives the sufficient context length in* (18) *for a fixed trained population model.*

*Proof.* Fix a unit query $\boldsymbol{x}$ and write $\mu = (K_\beta f)(\boldsymbol{x})$. Normalize the attention weights by their population mean, $U_i := e^{\beta \boldsymbol{x}^\top \boldsymbol{z}_i} / Z_d(\beta)$. Then

$$\mathbb{E}[U_i] = \frac{\mathbb{E}[e^{\beta \boldsymbol{x}^\top \boldsymbol{z}_i}]}{Z_d(\beta)} = 1,$$
$$\mathbb{E}[U_i^2] = \frac{\mathbb{E}[e^{2\beta \boldsymbol{x}^\top \boldsymbol{z}_i}]}{Z_d(\beta)^2} = \frac{Z_d(2\beta)}{Z_d(\beta)^2} =: N_d(\beta). \tag{D.49}$$

Define the empirical denominator and centered numerator by

$$D_m := \frac{1}{m} \sum_{i=1}^m U_i, \qquad A_m := \frac{1}{m} \sum_{i=1}^m U_i (f(\boldsymbol{z}_i) - \mu). \tag{D.50}$$

For the finite-context attention output $\widehat{\mu}$, subtraction of the population output gives

$$\widehat{\mu} - \mu = \frac{\sum_{i=1}^m U_i f(\boldsymbol{z}_i)}{\sum_{i=1}^m U_i} - \mu = \frac{\sum_{i=1}^m U_i (f(\boldsymbol{z}_i) - \mu)}{\sum_{i=1}^m U_i} = \frac{A_m}{D_m}. \tag{D.51}$$

The numerator summands are centered:

$$\mathbb{E}[U_i (f(\boldsymbol{z}_i) - \mu)] = \mathbb{E}[U_i f(\boldsymbol{z}_i)] - \mu \mathbb{E}[U_i] = (K_\beta f)(\boldsymbol{x}) - \mu = 0. \tag{D.52}$$

Independence and the second-moment identity (A.25) give

$$\mathbb{E}[A_m^2] = \frac{1}{m} \mathbb{E}[U_1^2 (f(\boldsymbol{z}_1) - \mu)^2]$$
$$\leq \frac{4M^2}{m} \mathbb{E}[U_1^2] = \frac{4M^2 N_d(\beta)}{m}. \tag{D.53}$$

The same identity applied to $U_i - 1$, whose mean is zero, gives

$$\mathbb{E}[(D_m - 1)^2] = \frac{1}{m} \mathbb{E}[(U_1 - 1)^2] = \frac{1}{m} \{\mathbb{E}[U_1^2] - 2\mathbb{E}[U_1] + 1\} = \frac{N_d(\beta) - 1}{m}. \tag{D.54}$$

On $\mathbb{A} = \{D_m \geq 1/2\}$, the squared error is at most $4A_m^2$. On $\mathbb{A}^c$, both weighted averages lie in $[-M, M]$, giving squared error at most $4M^2$. Chebyshev's inequality gives

$$\Pr(\mathbb{A}^c) = \Pr(D_m < 1/2) \leq \Pr(|D_m - 1| > 1/2) \leq 4\mathbb{E}[(D_m - 1)^2]. \tag{D.55}$$

Splitting the expectation over these two events now yields

$$\begin{aligned}\mathbb{E}[(\widehat{\mu}-\mu)^2] &= \mathbb{E}[(\widehat{\mu}-\mu)^2\mathbf{1}_{\mathbb{A}}] + \mathbb{E}[(\widehat{\mu}-\mu)^2\mathbf{1}_{\mathbb{A}^c}] \\ &\le 4\mathbb{E}[A_m^2\mathbf{1}_{\mathbb{A}}] + 4M^2\Pr(\mathbb{A}^c) \\ &\le 4\mathbb{E}[A_m^2] + 16M^2\mathbb{E}[(D_m-1)^2] \\ &\le \frac{16M^2N_d(\beta)}{m} + \frac{16M^2(N_d(\beta)-1)}{m} \le \frac{32M^2N_d(\beta)}{m}. \end{aligned} \tag{D.56}$$

The bound is independent of $\boldsymbol{x}$, so integration gives the same expected squared $L^2(\nu)$ error. It also holds at $\beta = 0$, where $N_d(0) = 1$.

To bound $N_d$, let $\kappa = (d-1)/2$. The partition-integral expansion in Appendix C gives $Z_d(\beta) = A_d e^\beta \beta^{-\kappa}(1 + O(\beta^{-1}))$, where $A_d = 2^{(d-3)/2}\Gamma(d/2)/\sqrt{\pi}$. Substitution yields

$$\begin{aligned} N_d(\beta) &= \frac{Z_d(2\beta)}{Z_d(\beta)^2} \\ &= \frac{A_d e^{2\beta}(2\beta)^{-\kappa}(1+O(\beta^{-1}))}{A_d^2 e^{2\beta}\beta^{-2\kappa}(1+O(\beta^{-1}))} \\ &= \frac{2^{-\kappa}}{A_d}\beta^\kappa(1+O(\beta^{-1})) = \frac{\sqrt{\pi}}{2^{d-2}\Gamma(d/2)}\beta^{(d-1)/2}(1+O(\beta^{-1})). \end{aligned} \tag{D.57}$$

Continuity gives $N_d(\beta) \le C_d'\beta^{(d-1)/2}$ for all $\beta \ge 1$. For a trained population model, $|c| \le 1/b(1)$ by (D.22). Thus

$$\begin{aligned} \mathbb{E}[\left\| c\widehat{K}_{\beta,m}f - cK_\beta f\right\|_{L^2(\nu)}^2] &= c^2\int \mathbb{E}[(\widehat{\mu}(\boldsymbol{x})-\mu(\boldsymbol{x}))^2]\,d\nu(\boldsymbol{x}) \\ &\le c^2\frac{32\,\|f\|_\infty^2\,N_d(\beta)}{m} \\ &\le \frac{32C_d'}{b(1)^2}\frac{\|f\|_\infty^2\,\beta^{(d-1)/2}}{m}. \end{aligned} \tag{D.58}$$

This proves (18) with $C_d = 32C_d'/b(1)^2$. □

# E PROOF OF THEOREM 5

The main argument is an operator identity: after calibration, the two attention views together multiply every task by the same positive scalar. Global normalization removes this scalar. We prove the resulting law and risk statements first, using Lemma E.1 below. The remaining subsections prove the claims in Section 4.4 concerning nonnegative weights and finite contexts.

## E.1 PROOF OF THEOREM 5

For the calculations below, denote the population attention average at inverse temperature $\gamma \geq 0$ by

$$(K_\gamma f)(\boldsymbol{x}) := \frac{\int f(\boldsymbol{z}) e^{\gamma \boldsymbol{x}^\top \boldsymbol{z}}\, d\nu(\boldsymbol{z})}{Z_d(\gamma)}, \tag{E.1}$$

where $Z_d$ is defined in (C.10). We write $F_{\boldsymbol{\theta}}(f, \boldsymbol{x}; \tau)$ for the population-context counterpart of (19) and abbreviate $a_j = a(\gamma_j)$, $b_j = b(\gamma_j)$, and $\Delta = a_1 b_2 - a_2 b_1$.

**Lemma E.1** (Calibrated population generation)**.** *Fix $0 < \gamma_1 < \gamma_2 < \infty$ and parameters $\boldsymbol{\theta}_r = (\sqrt{\beta_r}\boldsymbol{I}_d, \sqrt{\beta_r}\boldsymbol{I}_d, \mathbf{0}, c_r)$ with $\beta_r \geq 1$ and $c_r > 0$. The calibration equations* (21) *have a unique solution with $\lambda_1 < 0 < \lambda_2$. For every task $f = f_1 + f_2$ in the centered linear–quadratic family, the population version of Algorithm 1 produces the unnormalized function $\widetilde{f}_r = c_r f$.*

*Proof of Theorem 5.* Fix a unit-energy law $P$ satisfying Assumption 4.2. By (C.53) and (C.48), training on $P$ for $T > 0$ gives

$$\boldsymbol{Q}_r = \boldsymbol{K}_r = \sqrt{\beta_r}\boldsymbol{I}_d, \qquad \boldsymbol{u}_r = \mathbf{0}, \qquad \beta_r \geq 1, \qquad c_r > 0. \tag{E.2}$$

Write $G_r f = \widetilde{f}_r$ for the unnormalized generated task. By Lemma E.1, $G_r f = c_r f$. Since $P$ has unit energy and $c_r > 0$, the global normalizer is

$$\begin{aligned} s_r &:= \sqrt{\mathcal{E}(\widetilde{P}_r)} = \left(\mathbb{E}_{f\sim P}[\|G_r f\|^2_{L^2(\nu)}]\right)^{1/2} \\ &= |c_r| \left(\mathbb{E}_{f\sim P}[\|f\|^2_{L^2(\nu)}]\right)^{1/2} = c_r > 0. \end{aligned} \tag{E.3}$$

Consequently, the normalized generated task equals the sampled function $f$:

$$f_r(\boldsymbol{x}) = \frac{(G_r f)(\boldsymbol{x})}{s_r} = \frac{c_r f(\boldsymbol{x})}{c_r} = f(\boldsymbol{x}). \tag{E.4}$$

Starting with $P_0$, this identity gives $P_1 = P_0$. If $P_{r-1} = P_0$, the same argument applies to generation $r$ and gives $P_r = P_{r-1} = P_0$. Induction proves $P_r = P_0$ and $p_r = p_0$ for every $r \geq 1$.

It remains to compare the training trajectories and risks. The unchanged law and common initialization imply

$$\begin{aligned} L^{(\infty)}_{P_{r-1}}(\boldsymbol{\theta}) &= L^{(\infty)}_{P_0}(\boldsymbol{\theta}), \\ \dot{\boldsymbol{\theta}}_r(t) &= -\nabla L^{(\infty)}_{P_0}(\boldsymbol{\theta}_r(t)), \\ \boldsymbol{\theta}_r(0) &= (\boldsymbol{I}_d, \boldsymbol{I}_d, \mathbf{0}, 0). \end{aligned} \tag{E.5}$$

Uniqueness of the full-parameter gradient flow established in Appendix C therefore gives $\boldsymbol{\theta}_r(t) = \boldsymbol{\theta}_1(t)$ for $0 \leq t \leq T$. Finally, $\mathcal{E}(P_0) = 1$ and Lemma D.3 give

$$\begin{aligned} \sup_{r\geq 1} R^{(\infty)}_0(\boldsymbol{\theta}_r(T)) &= R^{(\infty)}_0(\boldsymbol{\theta}_1(T)) \\ &= 2L_{p_0}(\beta_1(T), c_1(T)) \leq C_{d,p_0}(1+T)^{-2/3}. \end{aligned} \tag{E.6}$$

This proves all conclusions of Theorem 5. □

*Proof of Lemma E.1.* The temperatures $\tau_{r,j} = \beta_r/\gamma_j$ in (21) are positive. Write $G_{r,j}f$ for the $j$-th prediction in Algorithm 1. Its scaled logits and output satisfy

$$\frac{(\boldsymbol{Q}_r\boldsymbol{x})^\top(\boldsymbol{K}_r\boldsymbol{z}+\boldsymbol{u}_r f(\boldsymbol{z}))}{\tau_{r,j}} = \frac{\gamma_j}{\beta_r}\beta_r\boldsymbol{x}^\top\boldsymbol{z} = \gamma_j\boldsymbol{x}^\top\boldsymbol{z},$$
$$(G_{r,j}f)(\boldsymbol{x}) := F_{\boldsymbol{\theta}_r}(f,\boldsymbol{x};\tau_{r,j}) = c_r\big(K_{\gamma_j}(f_1+f_2)\big)(\boldsymbol{x})$$
$$= c_r\big(a_j f_1(\boldsymbol{x}) + b_j f_2(\boldsymbol{x})\big), \tag{E.7}$$

where the last equality is (C.22). The strict increase of $\rho = b/a$ in (C.25) gives

$$\Delta = a_1b_2 - a_2b_1 = a_1a_2\{\rho(\gamma_2) - \rho(\gamma_1)\} > 0. \tag{E.8}$$

The response-matching equations (21) therefore have the unique solution

$$\lambda_1 = \frac{b_2-a_2}{\Delta}, \qquad \lambda_2 = \frac{a_1-b_1}{\Delta}. \tag{E.9}$$

The inequalities $0 < b_j < a_j$ imply $\lambda_1 < 0 < \lambda_2$. Substitution verifies that

$$\lambda_1a_1 + \lambda_2a_2 = \frac{a_1(b_2-a_2)+a_2(a_1-b_1)}{\Delta} = 1,$$
$$\lambda_1b_1 + \lambda_2b_2 = \frac{b_1(b_2-a_2)+b_2(a_1-b_1)}{\Delta} = 1. \tag{E.10}$$

Write $G_rf = \lambda_1G_{r,1}f + \lambda_2G_{r,2}f$ for the unnormalized generated task $\widetilde{f}_r$ in the algorithm. Combining the outputs yields the pointwise identity

$$G_rf = \lambda_1c_r(a_1f_1+b_1f_2) + \lambda_2c_r(a_2f_1+b_2f_2)$$
$$= c_r\{(\lambda_1a_1+\lambda_2a_2)f_1 + (\lambda_1b_1+\lambda_2b_2)f_2\} = c_rf. \tag{E.11}$$

□

### E.2 Why response matching requires signed weights

**Proposition E.1** (Nonnegative attention combinations)**.** *Let $J < \infty$, $\gamma_j > 0$, and $q_j \geq 0$, not all zero. On the centered linear–quadratic task family, $\sum_{j=1}^J q_jK_{\gamma_j}$ cannot multiply both components by the same positive scalar.*

*Proof.* The linear and quadratic response coefficients are $A_q = \sum_j q_ja(\gamma_j)$ and $B_q = \sum_j q_jb(\gamma_j)$, respectively. By Lemma 4.1,

$$A_q - B_q = \sum_{j=1}^J q_j\{a(\gamma_j) - b(\gamma_j)\} > 0, \qquad A_q > 0. \tag{E.12}$$

Thus $A_q = B_q > 0$ is impossible, so response matching requires a negative weight. □

### E.3 Finite-context sampling

**Proposition E.2** (Sampling error with shared or correlated views)**.** *Fix $0 < \gamma_1 < \gamma_2 < \infty$, $m \geq 1$, and a bounded task $f$ with $\|f\|_\infty \leq M$. For each view define*

$$\widehat{K}_{\gamma_j,m}f(\boldsymbol{x}) = \frac{\sum_{i=1}^m e^{\gamma_j\boldsymbol{x}^\top\boldsymbol{z}_i}f(\boldsymbol{z}_i)}{\sum_{i=1}^m e^{\gamma_j\boldsymbol{x}^\top\boldsymbol{z}_i}}, \qquad \boldsymbol{z}_i \overset{\text{i.i.d.}}{\sim} \nu. \tag{E.13}$$

*Use the same seed tuple across queries. For the two views $j = 1, 2$ and the weights in* (22)*,*

$$\mathbb{E}[\left\|\sum_{j=1}^2\lambda_j\widehat{K}_{\gamma_j,m}f - \sum_{j=1}^2\lambda_jK_{\gamma_j}f\right\|_{L^2(\nu)}^2] \leq \frac{32M^2}{m}\left(\sum_{j=1}^2|\lambda_j|\sqrt{N_d(\gamma_j)}\right)^2. \tag{E.14}$$

*Here $N_d(\gamma) = Z_d(2\gamma)/Z_d(\gamma)^2$. For the centered linear–quadratic family, the second sum equals $f$ by Lemma E.1. The teacher gain multiplies the bound by $c^2$. No independence between views is required; in particular, both views may share the same seed.*

*Proof.* We combine the single-view errors using the triangle inequality, which allows arbitrary dependence between views. The estimate (D.56) applies to each target scale $\gamma_j$. For $e_j = (\widehat{K}_{\gamma_j,m} - K_{\gamma_j})f$, integrating over the query gives

$$\mathbb{E}[\|e_j\|^2_{L^2(\nu)}] \le \frac{32M^2 N_d(\gamma_j)}{m}. \tag{E.15}$$

In the joint Hilbert space of seed randomness and query inputs,

$$\begin{aligned}\left(\mathbb{E}[\left\|\sum_{j=1}^{2}\lambda_j e_j\right\|^2_{L^2(\nu)}]\right)^{1/2} &\le \sum_{j=1}^{2}|\lambda_j|\left(\mathbb{E}[\|e_j\|^2_{L^2(\nu)}]\right)^{1/2}\\ &\le \frac{\sqrt{32}M}{\sqrt{m}}\sum_{j=1}^{2}|\lambda_j|\sqrt{N_d(\gamma_j)}.\end{aligned} \tag{E.16}$$

Squaring proves the stated bound. □

### E.4 A FIXED-HORIZON EXTENSION TO FULL-MATRIX SGD

**Proposition E.3** (Finite contexts approximate corrected recursion on fixed horizons)**.** *Fix the real law of Theorem 5, $T > 0$, $0 < \gamma_1 < \gamma_2 < \infty$, a finite generation horizon $R$, and tolerances $\epsilon, \delta > 0$. Let $\bar{\boldsymbol{\theta}}_1(t)$ denote the population-context gradient-flow trajectory on $L^{(\infty)}_{P_0}$ from the initialization in Section 3. Use two-temperature generation with the weights in* (22)*. In finite-context generation define*

$$\widehat{\beta}(\boldsymbol{\theta}) = \max\left\{\frac{1}{2}, \frac{1}{d}\operatorname{tr}(\boldsymbol{K}^\top \boldsymbol{Q})\right\}, \qquad s_j(\boldsymbol{\theta}) = \frac{\gamma_j}{\widehat{\beta}(\boldsymbol{\theta})}, \quad j = 1, 2. \tag{E.17}$$

*Here $s_j(\boldsymbol{\theta}) = 1/\tau_j(\boldsymbol{\theta})$ is the inverse temperature of view $j$. This scalar rule scales generation logits without projecting the trained matrices. Use exact population normalization of each generated function law. There exist finite common training and generation context lengths $n, m$, a finite batch size $B$, an integer $K$ with step size $\eta = T/K$, and an initialization radius, such that independently reset, unprojected full-matrix SGD satisfies, with probability at least $1 - \delta$,*

$$\begin{aligned}\max_{1\le r\le R}\left\|\boldsymbol{\theta}_r(T) - \bar{\boldsymbol{\theta}}_1(T)\right\| &\le \epsilon,\\ \max_{1\le r\le R}\mathcal{D}(P_r, P_0) &\le \epsilon,\\ \max_{1\le r\le R}\left|R_0^{(n)}(\boldsymbol{\theta}_r(T)) - R_0^{(\infty)}(\bar{\boldsymbol{\theta}}_1(T))\right| &\le \epsilon.\end{aligned} \tag{E.18}$$

*Here $\mathcal{D}(P, Q) = \inf_{\pi\in\Pi(P,Q)}(\mathbb{E}_\pi[\|f - g\|^2_\infty])^{1/2}$ is the function-law distance from Appendix C, the laws are conditional on the trained history, and each initialization lies within the selected radius of $(\boldsymbol{I}_d, \boldsymbol{I}_d, \mathbf{0}, 0)$. The choices may depend on $R, T, \epsilon, \delta$ and the target selectivities. No fixed-$(n, m, B)$ infinite-generation guarantee is asserted.*

*Proof.* The proof extends the finite-horizon comparison from Appendix C to the signed generator. The training loss and SGD updates are unchanged, so it suffices to establish continuity in the parent task and trained parameters, control context sampling, and verify stability of normalization.

*A common parameter neighborhood.* We first choose a common region around the reference trajectories. By Theorem 5, every reference trajectory is the same compact path $\bar{\boldsymbol{\theta}}(t) := \bar{\boldsymbol{\theta}}_1(t)$, $0 \le t \le T$. Its normalized trace is $\beta(t) \ge 1$, and its endpoint gain is $c_T > 0$. Choose a parameter box of radius $R_\theta \ge 1$ and a sufficiently small neighborhood of the path so that

$$\|\boldsymbol{Q}\|_{\mathrm{F}}, \ \|[\boldsymbol{K}\ \boldsymbol{u}]\|_{\mathrm{F}}, \ |c| \le R_\theta, \qquad d^{-1}\operatorname{tr}(\boldsymbol{K}^\top \boldsymbol{Q}) \ge 3/4 \tag{E.19}$$

throughout this neighborhood, and a convex ball with the same properties around the endpoint. On these sets the lower cutoff in (E.17) is inactive, so the temperature multipliers are smooth. With

$g_{\max} = \gamma_2$,

$$0 \le s_j(\boldsymbol{\theta}) \le s_{\max} := 4g_{\max}/3,$$
$$\nabla_{\boldsymbol{Q},\boldsymbol{K}} \widehat{\beta} = d^{-1}(\boldsymbol{K}, \boldsymbol{Q}), \qquad \left\|\nabla\widehat{\beta}\right\|^2 = d^{-2}(\|\boldsymbol{K}\|_{\mathrm{F}}^2 + \|\boldsymbol{Q}\|_{\mathrm{F}}^2) \le 2R_\theta^2/d^2,$$
$$\|\nabla s_j\| = \frac{\gamma_j}{\widehat{\beta}^2}\left\|\nabla\widehat{\beta}\right\| \le \frac{16\sqrt{2}g_{\max}R_\theta}{9d} =: C_s. \tag{E.20}$$

These bounds hold throughout the neighborhood, including at anisotropic SGD iterates.

*Continuity of the generation operator.* Use a common bound $M$ on task values over the finite horizon; its construction is given in the normalization argument below. Put $H_0 = \sqrt{1+M^2}$ and $\Lambda_\lambda = \sum_{j=1}^2 |\lambda_j|$. For an unscaled logit $\ell_i = (\boldsymbol{Q}\boldsymbol{x})^\top(\boldsymbol{K}\boldsymbol{z}_i + \boldsymbol{u}f(\boldsymbol{z}_i))$, the original derivative bounds and the product rule give

$$|\ell_i| \le R_\theta^2 H_0, \qquad \|\nabla\ell_i\| \le \sqrt{2}R_\theta H_0,$$
$$\|\nabla(s_j\ell_i)\| \le s_{\max}\sqrt{2}R_\theta H_0 + R_\theta^2 H_0 C_s =: V_s. \tag{E.21}$$

For a weighted label average $t_j$, applying (C.77) to these scaled logits yields

$$\nabla t_j = \sum_i \pi_{j,i}(f(\boldsymbol{z}_i) - t_j)\nabla(s_j\ell_i),$$
$$\|\nabla t_j\| \le 2MV_s, \qquad \|\nabla(ct_j)\| \le M + 2R_\theta M V_s,$$
$$\left\|\nabla\sum_j \lambda_j c t_j\right\| \le \Lambda_\lambda(M + 2R_\theta M V_s) =: J_{\mathrm{g}}. \tag{E.22}$$

The derivative bound includes the dependence of the logit multiplier on the trained matrix product. To control perturbations of the parent law, consider coupled parents $f, g$ and a common seed, and put $\Delta_f = \|f - g\|_\infty$. Temperature scaling is independent of the parent labels, so

$$|s_j\ell_i(f) - s_j\ell_i(g)| \le s_{\max}R_\theta^2\Delta_f,$$
$$\|\boldsymbol{\pi}_j(f) - \boldsymbol{\pi}_j(g)\|_1 \le 2s_{\max}R_\theta^2\Delta_f,$$
$$|t_j(f) - t_j(g)| \le \Delta_f + 2Ms_{\max}R_\theta^2\Delta_f,$$
$$\|\mathcal{G}_{\boldsymbol{\theta},Z}f - \mathcal{G}_{\boldsymbol{\theta},Z}g\|_\infty \le \Lambda_\lambda R_\theta(1 + 2Ms_{\max}R_\theta^2)\Delta_f =: C_{\mathrm{g}}\Delta_f. \tag{E.23}$$

Here $\mathcal{G}$ is the signed generator before normalization. The same estimates hold for integral contexts. Absolute values of the signed coefficients enter through the triangle inequality.

*Uniform control over query inputs.* We strengthen the pointwise sampling bound to the supremum norm required by the function-law distance. At the exact reference parameters, the scaled logits equal the fixed target logits. Set $V_\lambda = (\sum_{j=1}^2 |\lambda_j|\sqrt{N_d(\gamma_j)})^2$. Proposition E.2 gives pointwise second-moment error at most $32c_T^2M^2V_\lambda/m$. By (C.77), each empirical or population gain-free view has query derivative bounded by

$$\left\|\nabla_{\boldsymbol{x}}\widehat{K}_{\gamma_j,m}f(\boldsymbol{x})\right\| \le 2\gamma_j M, \qquad \left\|\nabla_{\boldsymbol{x}}K_{\gamma_j}f(\boldsymbol{x})\right\| \le 2\gamma_j M. \tag{E.24}$$

Thus the unnormalized signed difference $h_Z$ is $4c_T M\sum_j |\lambda_j|\gamma_j$-Lipschitz. Apply the sphere-net bound (C.112) with this Lipschitz constant, the pointwise moment bound above, and net radius $u = m^{-1/(d+2)}$. The two contributions give

$$\mathbb{E}[\|h_Z\|_\infty^2] \le \left[64c_T^2M^2V_\lambda 3^d + 32c_T^2M^2\left(\sum_j |\lambda_j|\gamma_j\right)^2\right] m^{-2/(d+2)} =: S_{\mathrm{g}}^2 m^{-2/(d+2)},$$
$$(\mathbb{E}[\|h_Z\|_\infty^2])^{1/2} \le S_{\mathrm{g}} m^{-1/(d+2)}. \tag{E.25}$$

This supremum-norm rate allows us to apply the training-law comparison in the metric used by the proposition.

*Normalization and bounds on task values.* All reference unnormalized output laws equal the law of $c_T f$ for $f \sim P_0$, so their energy square root is $c_T$. If a coupled error between unnormalized laws is $\widetilde{d} < c_T/2$, the reverse triangle inequality implies that the square root of the actual unnormalized energy is at least $c_T/2$. Write $\widetilde{P}, \widetilde{Q}$ for these unnormalized laws and $P, Q$ for their globally normalized versions. Let $a = c_T/2$ and let $\widetilde{M}$ bound the absolute values of both unnormalized outputs. The normalization estimate (C.121) then gives

$$\mathcal{D}(P, Q) \le \left(\frac{1}{a} + \frac{\widetilde{M}}{a^2}\right) \mathcal{D}(\widetilde{P}, \widetilde{Q}) =: H_{\mathrm{N}} \widetilde{d}. \tag{E.26}$$

Common deterministic bounds on task values can be selected in advance on this stopped event. If $M_0$ bounds the real tasks, a normalized child satisfies

$$\|f_r\|_\infty \le \frac{\sum_j |\lambda_j||c_r| \, \|f_{r-1}\|_\infty}{c_T/2} \le \frac{2\Lambda_\lambda R_\theta}{c_T} \|f_{r-1}\|_\infty . \tag{E.27}$$

Thus it suffices to choose $M = M_0 \max\{1, 2\Lambda_\lambda R_\theta / c_T\}^R$ and $\widetilde{M} \ge \max\{\Lambda_\lambda R_\theta M, c_T M_0\}$. These deterministic bounds are finite for each fixed $R$, with explicit dependence on the generation horizon.

*Propagation over the fixed generation horizon.* We now propagate the parameter and law errors across generations. On the common parameter box and under the common bound on task values, let $G, L, D, C_L, L_f$ be the finite constants for the training loss from Appendix C, (C.82), (C.91), (C.92), and (C.109). Let $\epsilon_0$ bound the initialization error, and allocate failure probability $\delta/R$ to each generation. Start with $d_0 = 0$ and define deterministic error bounds

$$e_r = e^{LT} \left[ \epsilon_0 + \frac{DT}{\sqrt{n}} + L_f T d_{r-1} + \frac{1}{2} LGT\eta + 2G\sqrt{\frac{T\eta R}{B\delta}} \right], \tag{E.28}$$

$$\widetilde{d}_r = C_{\mathrm{g}} d_{r-1} + J_{\mathrm{g}} e_r + S_{\mathrm{g}} m^{-1/(d+2)}, \tag{E.29}$$

$$d_r = H_{\mathrm{N}} \widetilde{d}_r. \tag{E.30}$$

Conditional on earlier successful generations, the stopped SGD comparison (C.103) and the law-gradient perturbation (C.109) bound the next parameter error by $e_r$. This controls every parameter entry, including the label-key column. For generation, use the three-term comparison in (C.115): change the parent, change the trained parameters, and replace the finite seed by an integral context. Under the common-parent coupling and common seed, (E.23), (E.22), and (E.25) give the respective contributions to $\widetilde{d}_r$. When $\widetilde{d}_r < c_T/2$, normalization stability in (E.26) yields the law-error bound $d_r$. The original predictor satisfies

$$|R_0^{(n)}(\boldsymbol{\theta}_r(T)) - R_0^{(\infty)}(\bar{\boldsymbol{\theta}}(T))| \le 2Ge_r + \frac{2C_L}{\sqrt{n}}, \tag{E.31}$$

since $\mathcal{E}(P_0) = 1$.

Finally, choose the finite sampling and optimization budgets. All constants above are fixed before this choice. Over finitely many generations, the recursion is a finite composition of continuous affine maps, and every error bound tends to zero with the additive errors. Choose a small initialization radius, large finite $n, m$, and $\eta = T/K$ with a sufficiently large integer $K$ and finite $B$, so that every $e_r$ lies within its chosen neighborhood and is at most $\epsilon$, every $\widetilde{d}_r < c_T/2$, every $d_r \le \epsilon$, and the right side of (E.31) is at most $\epsilon$. The first-exit argument removes the stopping conditions. Conditional failure bounds and a union bound give total failure probability at most $R(\delta/R) = \delta$, completing the proof. □

# F Experimental Details

## F.1 Models and evaluation

**Softmax attention.** We use $d = 3$, one layer and head, and dense $4 \times 4$ matrices $\boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V, \boldsymbol{W}_O$. For the tokens in (1), scores are $(\boldsymbol{W}_Q \boldsymbol{h}_q)^\top \boldsymbol{W}_K \boldsymbol{h}_i / 2$ and the prediction is the last coordinate of $\boldsymbol{W}_O \sum_i \pi_i \boldsymbol{W}_V \boldsymbol{h}_i$. There is no MLP, bias, positional encoding, or additional residual path. Joint SGD minimizes half the MSE with batch size 128, learning rate 0.03, and no momentum or weight decay. Each generation starts afresh. Training and evaluation use $n = 128$; generation uses $m = 256$, except for the two-temperature comparison below. We use independent Xavier-uniform initialization unless stated otherwise.

**Tasks and normalization.** We sample Assumption 4.2 with $p_0 = 1/2$ by normalizing Gaussian coefficients after projecting the matrix onto the symmetric trace-free subspace. Each generated function retains its full ancestral prediction, with its context fixed across queries as in (C.72). Global normalization uses 1,024 functions, 24 Gauss–Legendre nodes, and 48 equally spaced azimuthal angles; a disjoint 1,024-function bank measures component energies by orthogonal projection on the same grid. These projections are used only for diagnostics. Ratios are formed within each run before averaging.

**Evaluation.** Original-task MSE uses 4,096 independent $P_0$ episodes, with the same test stream across generations within a seed. Bands are pointwise 95% Student-$t$ intervals across complete runs, retaining every planned seed. Budget and composition comparisons use seeds $1, \ldots, 20$. The linear experiment instead uses $d = 8$, $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_8)$, $n = m \in \{16, 64, 128\}$, zero initialization, 500 updates, batch size 512, learning rate one, double precision, and seeds $1, \ldots, 100$. Its moments and MSE are computed exactly; $n = 16$ stops at generation 22 before numerical energy underflow, and the other trajectories end at 23.

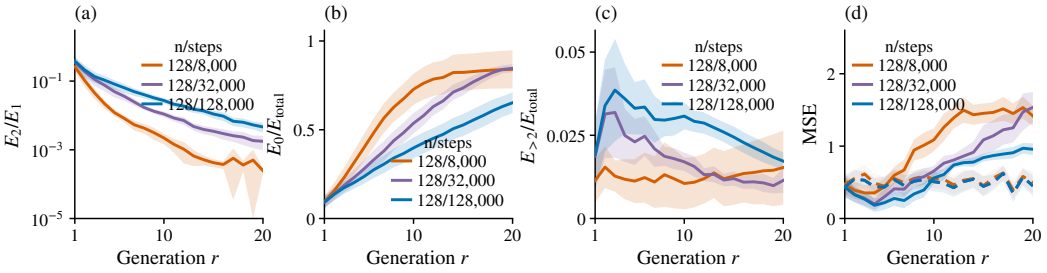


Figure 4: **Composition under normalized generation.** Component energies and original-task MSE across training budgets. In (d), solid lines show recursive training and dashed lines show matched fresh-task controls.

The runs in Figure 4 retain all components in subsequent training. The decreasing ratio $E_2/E_1$ isolates quadratic attenuation from growth of constant energy. The log-axis floor clips the negative lower confidence endpoint at generation 19 for 8,000 updates; the underlying statistics are untruncated.

### F.2 Generation and training comparisons

**Two-temperature implementation.** All three conditions in Figure 3 use 8,000 updates, $m = 512$, and seeds $301, \ldots, 320$ through generation 20. At each generation, draw four fresh Xavier-uniform matrices and initialize

$$\boldsymbol{W}_{Q,0} = \boldsymbol{W}_Q^{\mathrm{draw}}, \qquad \boldsymbol{W}_{K,0} = \frac{\left\|\boldsymbol{W}_K^{\mathrm{draw}}\right\|_{\mathrm{F}}}{\left\|\boldsymbol{W}_Q^{\mathrm{draw}}\right\|_{\mathrm{F}}} \boldsymbol{W}_Q^{\mathrm{draw}},$$

$$\boldsymbol{W}_{V,0} = \boldsymbol{W}_V^{\mathrm{draw}}, \qquad \boldsymbol{W}_{O,0} = \frac{\left\|\boldsymbol{W}_O^{\mathrm{draw}}\right\|_{\mathrm{F}}}{\left\|\boldsymbol{W}_V^{\mathrm{draw}}\right\|_{\mathrm{F}}} (\boldsymbol{W}_V^{\mathrm{draw}})^\top.$$

This pairing was selected in a separate three-seed pilot and is shared by all controls; the matrices subsequently train without tying or projection. Let $\boldsymbol{W}_{Q,x}, \boldsymbol{W}_{K,x}$ be their first three input columns. We set

$$\widehat{\beta} = \max\left\{\frac{1}{2}, \frac{1}{6}\operatorname{tr}(\boldsymbol{W}_{Q,x}^\top \boldsymbol{W}_{K,x})\right\}, \qquad \tau_j = \widehat{\beta}/\gamma_j, \qquad (\gamma_1, \gamma_2) = (4, 8).$$

The analytic weights from (22) are approximately $(-1.66517, 2.57142)$. Both views share one seed context, followed by global normalization. Calibration uses trained matrices and fixed response formulas, without fitting additional task labels.

**Comparisons and scope.** At generation $r$, the refresh baseline trains and generates using tasks from $D_{r-1}$, where $D_0 = P_0$ and $D_r = 0.1P_0 + 0.9P_r$. Each episode uses one task; composition is measured on $P_r$ before refresh. Training budgets match, while two-temperature generation requires an extra attention evaluation. All 20 runs remain included, notably seed 303's generation-18 response failure. A separate 32,000-update replay did not repair that case. These dense-model experiments support approximate retention, with finite contexts and estimated normalization.

**Training budgets.** Figure 5 uses seeds $1, \ldots, 20$. The single-generation sweep varies $p_0 \in \{0.1, 0.5, 0.9\}$ through 128,000 updates. Normalized recursion compares $K_r = 8{,}000 r^\gamma$, $\gamma \in \{2, 3, 4\}$, with a fixed 8,000-update budget. Larger schedules retain more quadratic energy over this five-generation horizon, which does not establish the asymptotic threshold in Theorem 4.

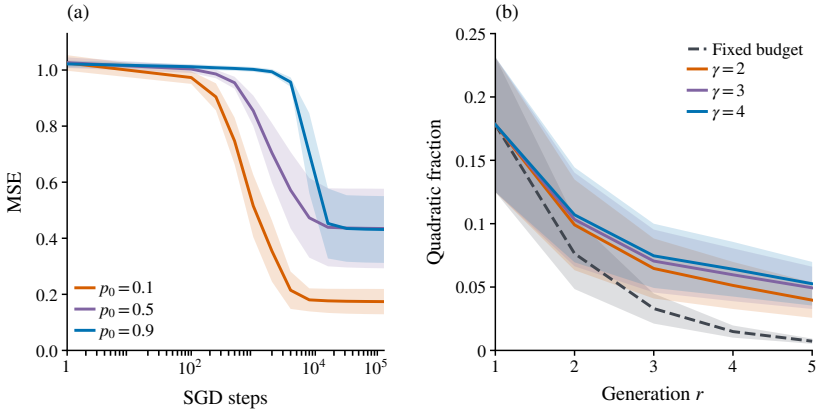


Figure 5: **Training budgets.** Single-generation MSE (left) and quadratic-energy fraction under fixed and growing budgets (right).

An earlier fitted-weight method used extra parent-label calibration, independent initialization, and $m = 256$. In its separate 20-seed, 50-generation follow-up, MSE was 1.5407 versus 1.1865 for normalized generation. This comparison concerns a different calibration method and is not pooled with the analytic two-temperature runs.

## G ATTENTION PARAMETERIZATION

This appendix derives the predictor (4) from the block weights (3). The query and key blocks determine the attention weights; the value block reads the labels; and the last coordinate of the output gives the prediction.

**Query and key projections.** For the tokens in (1), block multiplication gives

$$\begin{aligned}
\boldsymbol{W}_Q \boldsymbol{h}_q &= \begin{bmatrix} \boldsymbol{Q} & \boldsymbol{0} \\ \boldsymbol{0}^\top & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_q \\ 0 \end{bmatrix} = \begin{bmatrix} \boldsymbol{Q}\boldsymbol{x}_q \\ 0 \end{bmatrix}, \\
\boldsymbol{W}_K \boldsymbol{h}_i &= \begin{bmatrix} \boldsymbol{K} & \boldsymbol{u} \\ \boldsymbol{0}^\top & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_i \\ y_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{K}\boldsymbol{x}_i + \boldsymbol{u}y_i \\ 0 \end{bmatrix}, \\
(\boldsymbol{W}_K \boldsymbol{h}_i)^\top (\boldsymbol{W}_Q \boldsymbol{h}_q) &= (\boldsymbol{K}\boldsymbol{x}_i + \boldsymbol{u}y_i)^\top \boldsymbol{Q}\boldsymbol{x}_q = s_i. \qquad \text{(G.1)}
\end{aligned}$$

Thus the query contains only the input projection $\boldsymbol{Q}\boldsymbol{x}_q$, while key $i$ contains the input and label contributions $\boldsymbol{K}\boldsymbol{x}_i + \boldsymbol{u}y_i$. Taking their inner product gives exactly the score $s_i$ in Section 3, and hence the weight $\pi_i = e^{s_i} / \sum_{j=1}^n e^{s_j}$.

**Value projection and prediction.** The value block has only one nonzero entry, so its output contains the scaled label in the last coordinate. Substituting into the attention update (2) gives

$$\begin{aligned}
\boldsymbol{W}_V \boldsymbol{h}_i &= \begin{bmatrix} \boldsymbol{0}_{d\times d} & \boldsymbol{0} \\ \boldsymbol{0}^\top & c \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_i \\ y_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{0} \\ cy_i \end{bmatrix}, \\
\widetilde{\boldsymbol{h}}_q &= \boldsymbol{h}_q + \sum_{i=1}^n \pi_i \boldsymbol{W}_V \boldsymbol{h}_i = \begin{bmatrix} \boldsymbol{x}_q \\ 0 \end{bmatrix} + \sum_{i=1}^n \pi_i \begin{bmatrix} \boldsymbol{0} \\ cy_i \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_q \\ c\sum_{i=1}^n \pi_i y_i \end{bmatrix}, \\
F_{\boldsymbol{\theta}}(\mathcal{C}_n, \boldsymbol{x}_q) &= (\widetilde{\boldsymbol{h}}_q)_{d+1} = c\sum_{i=1}^n \pi_i y_i. \qquad \text{(G.2)}
\end{aligned}$$

The residual preserves the query input and adds zero to its label coordinate. Hence the last equality recovers (4) without any additional output projection. All trainable coordinates remain those specified in the preliminaries: $\boldsymbol{\theta} = (\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{u}, c)$.

**Relation to existing restricted attention models.** With $\boldsymbol{u} = \boldsymbol{0}$ and $c = 1$, (4) reduces to the label-weighted predictor of Collins et al. (2024), whose attention scores depend only on inputs. Cheng et al. (2023) also impose sparsity on query, key, and value matrices for nonlinear ICL; their query and key projections exclude labels, and their constrained value matrix reads only the label coordinate. Both constructions use structured attention weights to aggregate labels. Our parameterization additionally trains the label-dependent key vector $\boldsymbol{u}$ and the gain $c$. Consequently, the agreement at $\boldsymbol{u} = \boldsymbol{0}$ and $c = 1$ is an agreement of predictors; the training dynamics also depend on which coordinates are optimized.