How Does Chain-of-Thought Supervision Teach Models to Reason? A Theoretical Analysis

Anonymous authors
Paper under double-blind review

Abstract

Training with chain-of-thought (CoT) supervision is an effective approach for teaching models multi-step reasoning procedures. At inference time, however, models may need to continue reasoning beyond the finite teacher CoT sequences observed during training, and longer reasoning can instead degrade performance, a phenomenon known as overthinking. How the amount and structure of training-time CoT give rise to such failures remains poorly understood theoretically. To study this problem, we introduce an in-context weight prediction model trained on finite gradient descent trajectories for linear regression. Using high-dimensional asymptotic analysis, we show that the success and failure modes of long-horizon reasoning undergo phase transitions depending on the context length and the total amount of training CoT. We further distinguish two mechanisms of overthinking, error accumulation and error amplification, and show that when the training context is short, a model can learn to continue reasoning before it learns to correct errors. Experiments with fully trained linear and multi-head softmax attention further show that these qualitative behaviors persist. Our results provide a theoretical account linking properties of training-time CoT to inference-time failures and offer guidance for designing teacher CoT that supports stable reasoning.

1 Introduction

Chain-of-thought (CoT) (Wei et al., 2022) has become more than a test-time prompting technique: it is now widely used as training supervision to improve a model's reasoning ability itself. Whereas final-answer supervision provides only outcome-level feedback, supervising intermediate steps can help models learn general reasoning procedures. Such CoT-based supervision has produced substantial gains on complex tasks such as mathematics and logical reasoning (Ho et al., 2023; Zelikman et al., 2022; Lightman et al., 2024).

At inference, however, a model must often continue reasoning beyond the finite trajectories observed during training while maintaining its performance. In practice, extending reasoning can instead degrade performance, a phenomenon widely known as overthinking, so additional reasoning steps do not necessarily improve accuracy (Bansal et al., 2022; Su et al., 2025; Hassid et al., 2025). Recent empirical studies suggest that overthinking is a problem involving the correctness and stability of the reasoning procedure itself. For example, partial acquisition of reasoning capabilities may temporarily degrade long-horizon reasoning performance (Wei et al., 2023; Wu & Lo, 2025; Ren et al., 2026). Moreover, systems with similar short-horizon performance may exhibit sharply different long-horizon failure modes: errors may simply accumulate across reasoning steps, or they may be amplified as the system becomes increasingly confident in and builds on its earlier mistakes (Block et al., 2024; Ainslie et al., 2026). These findings show that locally plausible steps do not guarantee global convergence or error correction under repeated inference.

Despite the complexity of these failure mechanisms, most existing studies focus on observing models after training. In large language models, pretraining, post-training, model size, data structure, and optimization jointly affect observed performance and are difficult to separate experimentally (Biderman et al., 2023; Li et al., 2024). This makes it difficult to isolate experimentally how the amount and structure of training CoT determine the learned dynamics and their long-time stability.

To make this relationship analytically tractable, we introduce an in-context weight prediction model trained on gradient descent (GD) trajectories as teacher CoT. Specifically, we use GD trajectories for

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.

linear-regression tasks as teacher CoT and train a linear attention model on each transition. Previous work has shown that sufficient CoT supervision allows a model to approximately implement multi-step GD (Huang et al., 2025), but a theoretical understanding of long-horizon success conditions and failure regimes remains limited. More recently, phase transitions between successful inference and overthinking have been identified in a setting without training-time CoT (Takanami & Pehlevan, 2026). We extend this framework to finite teacher CoT trajectories and ask whether recursive application of the learned model can reproduce the correct update rule beyond the training horizon. Through this analysis, we characterize how the number of training tasks, context length, and teacher CoT length affect the inference dynamics.

Our contributions are as follows:

- We derive a phase diagram in the asymptotic limit showing that context length and the total amount of training CoT divide long-horizon reasoning into four regimes: exact GD recovery, bounded error, exponential divergence, and quadratic divergence. (Section 4)
- We show that reasoning can extend beyond the finite-length teacher CoT. The number of pretraining tasks required for this behavior decreases as the amount of teacher CoT increases, while shorter contexts require more teacher CoT. (Section 5.1)
- We identify two distinct mechanisms of overthinking: error accumulation, where errors introduced at each reasoning step remain uncorrected, and error amplification, where unstable reasoning dynamics repeatedly magnify existing errors, leading to quadratic and exponential MSE growth, respectively. (Section 5.2)
- We show that with short contexts, adding teacher CoT need not improve reasoning monotonically: a model can learn to continue reasoning before it learns to correct errors. (Section 5.3)
- We empirically show that the same qualitative mechanisms persist in more complex architectures, including fully trained linear attention and multi-head softmax attention. (Section 6)

Our results provide a unified theoretical understanding of the role of finite teacher CoT, appropriate CoT design as training data, and the mechanisms that generate failure.

2 Related Work

CoT supervision and overthinking. CoT data are increasingly used as training signals to improve reasoning and distill it into smaller models (Hsieh et al., 2023; Lightman et al., 2024; Agarwal et al., 2024; Guo et al., 2025; Zhu et al., 2025a; Pengmei et al., 2025). At the same time, longer reasoning or recurrent computation can reduce performance, a phenomenon often described as overthinking (Bansal et al., 2022; Block et al., 2024; Su et al., 2025; Hassid et al., 2025; Ainslie et al., 2026). However, most existing work focuses on improving training methods or observing trained models, leaving unclear how the properties of training CoT determine test-time behavior. We study this relation in a solvable model and characterize when training CoT supports correct inference and when it makes the test-time reasoning unstable.

Theoretical studies of CoT. Theoretical studies of CoT have mainly focused on its inference-time benefits, showing that CoT can enable efficient reasoning on otherwise difficult tasks (Abbe et al., 2024; Kim & Suzuki, 2025; Huang et al., 2025). However, the conditions under which such learning fails remain poorly understood. To study this question theoretically, recent work extends analytically tractable models of in-context linear regression (Garg et al., 2022; Akyürek et al., 2023; Von Oswald et al., 2023; Zhang et al., 2025; Lu et al., 2025) to iterative CoT, investigating how limited pretraining data or context and task mismatch can lead to overthinking (Javanmard et al., 2026; Takanami & Pehlevan, 2026). However, these studies mainly analyze the test-time behavior of trained models and do not characterize the effect of providing finite teacher CoT as training data. We therefore focus on training-time CoT and ask how far a model can continue to reason correctly beyond the reasoning steps observed during training and how this generalization fails.

3 MODEL

We study an in-context weight-prediction model for linear regression. In this model, training CoT consists of GD trajectories, and the pre-trained model is recursively applied as test-time CoT to reproduce GD in context on new tasks.

3.1 DATA GENERATION

We consider M linear-regression tasks indexed by $\mu = 1, \dots, M$. For each task μ, we prepare the training data and teacher CoT trajectories as follows.

Training tasks. For each of the M tasks, we prepare L labeled examples $(\{\boldsymbol{x}_{\mu,l}, y_{\mu,l}\}_{l=1,\cdots,L})$, given by

$$\mathbf{X}_\mu := [\boldsymbol{x}_{\mu,1}, \dots, \boldsymbol{x}_{\mu,L}] \in \mathbb{R}^{D\times L}, \quad \boldsymbol{y}_\mu := \mathbf{X}_\mu^\top \boldsymbol{w}_\mu + \boldsymbol{\varepsilon}_\mu \in \mathbb{R}^L, \tag{3.1}$$

where $\boldsymbol{x}_{\mu,l} \sim \mathcal{N}(\mathbf{0}_D, \mathbf{I}_D)$, $\boldsymbol{w}_\mu \sim \mathcal{N}(\mathbf{0}_D, \mathbf{I}_D/D)$, and $\boldsymbol{\varepsilon}_\mu \sim \mathcal{N}(\mathbf{0}_L, \sigma^2\mathbf{I}_L)$, independently across all tasks and examples. We denote the task context by $Q_\mu := (\mathbf{X}_\mu, \boldsymbol{y}_\mu)$.

CoT as training data. We prepare a K-step trajectory of gradient descent (GD) $\{\boldsymbol{s}_{\mu,k}\}_{k=0,\cdots,K}$ with step size $\eta > 0$ on the empirical squared loss. Starting from zero, its states satisfy

$$\boldsymbol{s}_{\mu,k+1} := \boldsymbol{s}_{\mu,k} - \frac{\eta}{L}\mathbf{X}_\mu\Big(\mathbf{X}_\mu^\top \boldsymbol{s}_{\mu,k} - \boldsymbol{y}_\mu\Big) \qquad (k = 0, \dots, K-1), \tag{3.2}$$

and $\boldsymbol{s}_{\mu,0} := \mathbf{0}_D$.

3.2 PRE-TRAINING SETTING

During pre-training, the model is trained to predict the next step of the teacher CoT, $\boldsymbol{s}_{\mu,k+1}$, from the context and the teacher CoT at step k, i.e., $(Q_\mu, \boldsymbol{s}_{\mu,k})$. Specifically, the input data (prompt), architecture, and loss function are defined as follows.

Prompt construction. At training step k, the prompt contains the original L labeled examples and a slot for the current teacher state:

$$\mathbf{E}_{\mu,k} := \begin{pmatrix} \mathbf{X}_\mu & \mathbf{0}_D \\ \boldsymbol{y}_\mu^\top & 0 \\ \mathbf{0}_{D\times L} & \boldsymbol{s}_{\mu,k} \\ \mathbf{0}_L^\top & 1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_{\mu,1} & \cdots & \boldsymbol{x}_{\mu,L} & \mathbf{0}_D \\ y_{\mu,1} & \cdots & y_{\mu,L} & 0 \\ \mathbf{0}_D & \cdots & \mathbf{0}_D & \boldsymbol{s}_{\mu,k} \\ 0 & \cdots & 0 & 1 \end{pmatrix} \in \mathbb{R}^{(2D+2)\times(L+1)}. \tag{3.3}$$

The first L columns encode labeled examples. The final column is a state slot, and the marker 1 in the last row distinguishes it from an example token. The prompt for predicting $\mathbf{s}_{\mu,k+1}$ contains the task context Q_μ and the current state $\mathbf{s}_{\mu,k}$, but no earlier states. Thus, conditional on the task context, the one-step predictor is Markovian. Masking or downweighting earlier reasoning tokens in this way can improve length generalization (Abbe et al., 2024; Liu et al., 2023).

Linear attention. For a prompt with L context examples, we use a one-layer linear self-attention model with a residual connection, defined by

$$\mathsf{Atten}(\mathbf{E};\theta) := \mathbf{E} + \frac{1}{L}(\mathbf{W}_V\mathbf{E})(\mathbf{W}_K\mathbf{E})^\top\mathbf{W}_Q\mathbf{E}. \tag{3.4}$$

Here the trainable parameters are $\theta = (\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V)$, where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V \in \mathbb{R}^{(2D+2)\times(2D+2)}$ are the query, key, and value matrices, respectively. We read the state slot of the final column using

$$\mathsf{Readout}(\mathbf{E}) := \mathbf{E}_{D+2:2D+1,\,\mathrm{end}}. \tag{3.5}$$

The next-step predictor of the weight vector is therefore

$$\hat{\boldsymbol{s}}_{\mu,k+1} := \mathsf{Readout}(\mathsf{Atten}(\mathbf{E}_{\mu,k};\theta)). \tag{3.6}$$

Learning. We consider the squared one-step prediction error over all teacher transitions in pre-training data:

$$\mathcal{L}(\theta) := \frac{1}{MK} \sum_{\mu=1}^{M} \sum_{k=0}^{K-1} \|\hat{\boldsymbol{s}}_{\mu,k+1} - \boldsymbol{s}_{\mu,k+1}\|_2^2. \tag{3.7}$$

3.3 Effective model

For analytical simplicity, we impose the following constraints on the trainable parameters:

$$\mathbf{W}_V := \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ \mathbf{I}_D & \mathbf{0}_D & \mathbf{I}_D & \mathbf{0}_D \\ * & * & * & * \end{pmatrix}, \quad \mathbf{W}_K^\top \mathbf{W}_Q := \begin{pmatrix} * & * & \mathbf{A} & \mathbf{0}_D \\ * & * & \mathbf{0}_D^\top & \eta \\ * & * & \mathbf{0}_{D\times D} & \mathbf{0}_D \\ * & * & \mathbf{0}_D^\top & 0 \end{pmatrix}, \tag{3.8}$$

where the rows from top to bottom and the columns from left to right are both partitioned into blocks of dimensions $(D, 1, D, 1)$. Here $*$ denotes a block that does not affect the readout, while $\mathbf{A} \in \mathbb{R}^{D\times D}$ is the only learnable parameter. This restriction is less severe than it may appear: each constrained block is justified by one of three considerations: (i) the rotational symmetry of the data, which motivates choosing the fixed blocks so that they do not select preferred directions; (ii) the fact that fixing it entails no loss of generality in the setting studied here; or (iii) preventing the model from advancing its inference without consulting the context (see Appendix A for details). These reductions concentrate the nontrivial parameter governing inference in the single matrix $\mathbf{A}$. Moreover, gradient-flow analyses of related models show that training from suitable initializations leads to analogous block structures (Huang et al., 2025; Javanmard et al., 2026). Under this parameterization, Equation (3.6) is given by

$$\hat{\boldsymbol{s}}_{\mu,k+1} = \boldsymbol{s}_{\mu,k} + \frac{1}{L} \mathbf{X}_\mu \Big(\mathbf{X}_\mu^\top \mathbf{A} \boldsymbol{s}_{\mu,k} + \eta \boldsymbol{y}_\mu \Big), \tag{3.9}$$

and the effective learning problem derived from Equation (3.7) is

$$\mathcal{L}_{\text{eff}}(\mathbf{A}) := \frac{1}{MK} \sum_{\mu=1}^{M} \sum_{k=0}^{K-1} \|\hat{\boldsymbol{s}}_{\mu,k+1} - \boldsymbol{s}_{\mu,k+1}\|_2^2 + \frac{\lambda}{D} \|\mathbf{A}\|_{\text{F}}^2, \tag{3.10}$$

where we introduced the regularization parameter $\lambda > 0$. For $\lambda > 0$, the objective is strongly convex, so $\hat{\mathbf{A}} := \operatorname{argmin}_{\mathbf{A}} \mathcal{L}_{\text{eff}}(\mathbf{A})$ is unique.

3.4 Inference setting

At inference time, we model test-time CoT by fixing the parameters of the trained model and applying the model recursively. Despite its simplicity, this recurrent form of CoT has recently regained attention because it enables flexible reasoning without relying on language tokens (Hao et al., 2025; Geiping et al., 2025; Shen et al., 2025; Zhu et al., 2025b).

Test-time CoT. Consider an independent task $Q^{\text{in}} = (\mathbf{X}^{\text{in}}, \boldsymbol{y}^{\text{in}})$, drawn from the same distribution as the training tasks. Explicitly, its examples and labels are generated by

$$\mathbf{X}^{\text{in}} := [\boldsymbol{x}_1^{\text{in}}, \dots, \boldsymbol{x}_L^{\text{in}}] \in \mathbb{R}^{D\times L}, \qquad \boldsymbol{y}^{\text{in}} := (\mathbf{X}^{\text{in}})^\top \boldsymbol{w}^{\text{in}} + \boldsymbol{\varepsilon}^{\text{in}} \in \mathbb{R}^L, \tag{3.11}$$

where $\boldsymbol{x}_l^{\text{in}} \sim \mathcal{N}(\mathbf{0}_D, \mathbf{I}_D)$, $\boldsymbol{w}^{\text{in}} \sim \mathcal{N}(\mathbf{0}_D, \mathbf{I}_D/D)$, and $\boldsymbol{\varepsilon}^{\text{in}} \sim \mathcal{N}(\mathbf{0}_L, \sigma^2 \mathbf{I}_L)$. At inference step t, the prompt is constructed in the same way as in Equation (3.3):

$$P_t := \begin{pmatrix} \mathbf{X}^{\text{in}} & \mathbf{0}_D \\ (\boldsymbol{y}^{\text{in}})^\top & 0 \\ \mathbf{0}_{D\times L} & \hat{\boldsymbol{w}}_t \\ \mathbf{0}_L^\top & 1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{x}_1^{\text{in}} & \cdots & \boldsymbol{x}_L^{\text{in}} & \mathbf{0}_D \\ y_1^{\text{in}} & \cdots & y_L^{\text{in}} & 0 \\ \mathbf{0}_D & \cdots & \mathbf{0}_D & \hat{\boldsymbol{w}}_t \\ 0 & \cdots & 0 & 1 \end{pmatrix} \in \mathbb{R}^{(2D+2)\times(L+1)}, \tag{3.12}$$

where $\hat{\boldsymbol{w}}_t$ is the current estimate of $\boldsymbol{w}^{\text{in}}$, and $\hat{\boldsymbol{w}}_0 = \mathbf{0}_D$. Then the test-time CoT recursion is

$$\hat{\boldsymbol{w}}_{t+1} = \mathsf{Readout}\Big(\mathsf{Atten}(P_t; \hat{\mathbf{A}})\Big) = \hat{\boldsymbol{w}}_t + \frac{1}{L} \mathbf{X}^{\text{in}} \Big[(\mathbf{X}^{\text{in}})^\top \hat{\mathbf{A}} \hat{\boldsymbol{w}}_t + \eta \boldsymbol{y}^{\text{in}} \Big]. \tag{3.13}$$

Evaluation. We evaluate the estimate at inference time t by its mean squared error:

$$\mathcal{E}(t) := \mathbb{E}\big\|\hat{\boldsymbol{w}}_t - \boldsymbol{w}^{\mathrm{in}}\big\|_2^2, \tag{3.14}$$

where the expectation is over the training and test data. For comparison, we define the generalization error when GD is successfully reproduced as

$$\mathcal{E}_{\mathrm{GD}}(t) := \mathbb{E}\big\|\boldsymbol{s}_t^{\mathrm{GD}} - \boldsymbol{w}^{\mathrm{in}}\big\|_2^2, \tag{3.15}$$

where $\boldsymbol{s}_t^{\mathrm{GD}}$ denotes the GD trajectory at step t on the test context, with initial condition $\boldsymbol{s}_0^{\mathrm{GD}} = \mathbf{0}_D$. When the model perfectly learns to generate the teacher-provided CoT, we have $\mathcal{E}(t) = \mathcal{E}_{\mathrm{GD}}(t)$.

3.5 High-dimensional limit

To obtain a nontrivial regime in which contextual information and training-task diversity remain comparable to the problem dimension, we take the proportional limit $D, L, M \longrightarrow \infty$ with

$$\frac{L}{D} \longrightarrow \alpha, \quad \frac{M}{D} \longrightarrow \tau, \qquad K, t = O(1). \tag{3.16}$$

Here, α and τ measure the numbers of context examples and training tasks per dimension, respectively, while the training and test-time CoT depths remain finite. We are particularly interested in $t > K$, where the model must continue inference beyond the GD steps observed in the teacher CoT.

4 Precise Characterization of Test-Time CoT Dynamics

The central question is how the generalization error $\mathcal{E}(t)$ evolves with test-time depth: when does the CoT inference reproduce GD, and what form does its failure take otherwise? To this end, we first seek a precise characterization of $\mathcal{E}(t)$. We introduce *training CoT budget* as

$$\varphi := \tau(K-1). \tag{4.1}$$

Since it measures the total number of CoT transitions in the training data, φ can be interpreted as the computational budget used to prepare the training CoT trajectories.

In the limit $\lambda \to 0$, the CoT inference introduced in Equation (3.13) gives

$$\hat{\boldsymbol{w}}_t = \mathbf{Q}_t \mathbf{X}^{\mathrm{in}} \boldsymbol{y}^{\mathrm{in}} / L, \tag{4.2}$$

where we introduced

$$\mathbf{H} := \frac{1}{L}\mathbf{X}^{\mathrm{in}}(\mathbf{X}^{\mathrm{in}})^\top, \quad \mathbf{Q}_t := \eta \sum_{j=0}^{t-1} (\mathbf{I}_D + \mathbf{H}\hat{\mathbf{A}})^j. \tag{4.3}$$

Consequently, the generalization error in Equation (3.14) is given by

$$\mathcal{E}(t) = 1 - \frac{2}{D}\mathbb{E}\,\mathrm{tr}(\mathbf{Q}_t\mathbf{H}) + \frac{1}{D}\mathbb{E}\|\mathbf{Q}_t\mathbf{H}\|_{\mathrm{F}}^2 + \frac{\sigma^2}{L}\mathbb{E}\,\mathrm{tr}(\mathbf{Q}_t\mathbf{H}\mathbf{Q}_t^\top). \tag{4.4}$$

The last two terms in Equation (4.4) mix $\mathbf{Q}_t$ and $\mathbf{Q}_t^\top$. Because $\mathbf{Q}_t$ is a polynomial in $\mathbf{H}\hat{\mathbf{A}}$, evaluating these terms at arbitrary depth t requires not only its eigenvalue distribution but the full family of mixed $*$-moments involving $\mathbf{H}\hat{\mathbf{A}}$ and $\hat{\mathbf{A}}^\top\mathbf{H}$. To overcome this difficulty, Takanami & Pehlevan (2026) introduced two-point generating functions and computed them by iteratively applying the cavity method in random matrix theory. In the present setting, the training CoT trajectories couple successive transitions, so the resulting cavity equations do not close under the same recursion. At finite depth, however, the two-point problem reduces to one-point Marchenko–Pastur averages when the learned update recovers full GD or, for $\alpha > 1$, performs GD only on the subspace learned during pretraining. We refer to the latter as projected GD. This reduction yields the following result.

Result 4.1 (Finite-depth error in GD or projected-GD regimes)**.** *Consider $\lambda \to 0$. For $\rho > 0$, let ν_ρ denote the Marchenko–Pastur law with sample-to-dimension ratio ρ, and define*

$$\langle f \rangle_\rho := \int f(x)\,\mathrm{d}\nu_\rho(x), \quad q_t(x) := \eta \sum_{j=0}^{t-1}(1-\eta x)^j, \quad \kappa_{j,t} := (\alpha-\varphi)(\eta t)^j + \varphi\langle q_t^j\rangle_{\alpha/\varphi}. \tag{4.5}$$

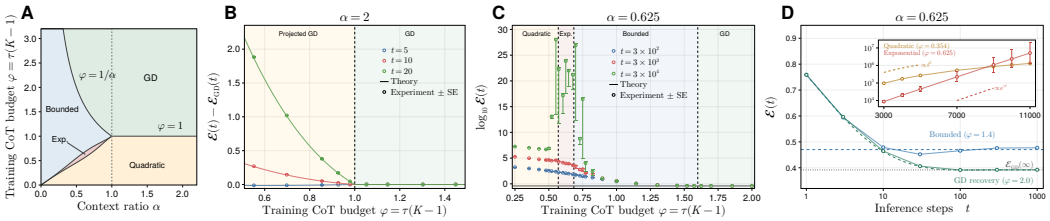


Figure 1: **Training CoT budget and context length control GD recovery and test-time error growth.** (A) Theoretical phase diagram in the context ratio α and training CoT budget $\varphi = \tau(K-1)$. The four regions correspond to GD recovery, bounded error, exponential divergence, and quadratic divergence. (B) Numerical experiments (circles) and theoretical prediction (solid line) on excess generalization error $\mathcal{E}(t) - \mathcal{E}_{\mathrm{GD}}(t)$ at $\alpha = 2$. (C) Numerical experiments of the generalization error at $\alpha = 0.625$. (D) Typical test-time error dynamics in the four phases; the dotted lines show the theoretical long-time predictions, and the inset highlights the quadratic and exponential regimes at large t. (B,C) The vertical dotted lines show the theoretical phase boundaries. Parameters: (B,C) $D = 96$, $K = 2$, $\eta = 0.15$ and $\sigma^2 = 0.01$. (B-D) Error bars represent one standard error across eight independent runs.

If $\varphi > \max\{1, 1/\alpha\}$*, the learned update is GD and*

$$\mathcal{E}(t) = \mathcal{E}_{\mathrm{GD}}(t) = \langle (1 - x q_t)^2 \rangle_\alpha + \frac{\sigma^2}{\alpha} \langle x q_t^2 \rangle_\alpha. \tag{4.6}$$

If $\alpha > 1$ *and* $0 < \varphi < 1$*, the learned update is projected GD and*

$$\begin{aligned} \mathcal{E}(t) =& 1 - \varphi + \varphi \langle (1 - x q_t)^2 \rangle_{\alpha/\varphi} + \frac{\varphi[2(1-\varphi) + \sigma^2]}{\alpha} \langle x q_t^2 \rangle_{\alpha/\varphi} \\ & - \frac{2(1-\varphi)}{\alpha} \kappa_{1,t} + \frac{(1-\varphi)(1-\varphi+\sigma^2)}{\alpha^2} \kappa_{2,t} + \frac{1-\varphi}{\alpha^2} \kappa_{1,t}^2. \end{aligned} \tag{4.7}$$

This result characterizes the regimes in which the learned update recovers full GD or projected GD, enabling a complete analytical characterization of its test-time dynamics. In the remaining regime, $0 < \alpha < 1$ and $0 < \phi < 1/\alpha$, evaluating Equation (4.4) at finite depth still requires the two-point functions. To isolate the failure mechanisms of interest, we instead study $\mathcal{E}(t)$ in the long-time limit $t \to \infty$. In this limit, whether $\mathcal{E}(t)$ saturates, grows polynomially, or diverges exponentially is determined only by the eigenvalue distribution. These one-point spectral quantities can be obtained from the cavity method, yielding the following four phases in the limit $t \to \infty$.

Result 4.2 (Long-time test-time CoT dynamics)**.** *For each* $0 < \alpha < 1$*, there exist critical values* $\varphi_-(\alpha)$ *and* $\varphi_+(\alpha)$ *such that, for sufficiently small* $\eta > 0$ *and as* $t \to \infty$*,*

$$\mathcal{E}(t) = \begin{cases} \mathcal{E}_{\mathrm{GD}}(t), & \varphi > \max\{1, 1/\alpha\} \\ O(1), & 0 < \alpha < 1, \quad \varphi_+ < \varphi < 1/\alpha \\ \exp(\Theta(\eta t)), & 0 < \alpha < 1, \quad \varphi_- < \varphi < \varphi_+ \\ \Theta(t^2), & (0 < \alpha < 1,\ 0 < \varphi < \varphi_-) \text{ or } (\alpha > 1,\ 0 < \varphi < 1). \end{cases} \tag{4.8}$$

Appendix B lists the corresponding derivations. Figure 1 summarizes the phase diagram and its agreement with numerical experiments.

Taken together, Results 4.1 and 4.2 show that the learned update and its long-time behavior change qualitatively with the training CoT budget and the context length. With sufficient budget, the learned update recovers GD, whereas below this threshold the error can remain bounded or diverge polynomially or exponentially. The next section discusses the implications of this phase diagram.

5 Insights from the Test-Time CoT Phase Diagram

In the previous section, we theoretically derived the test-time dynamics of the generalization error in the regimes where the learned update recovers GD or projected GD, and identified the phase

transitions governing the long-time inference limit. In this section, we discuss what these results reveal about the effects of training CoT, their practical implications, and related empirical observations.

5.1 Recovering GD from Finite-Length CoT

In this subsection, we discuss the GD recovery phase and the conditions that determine its boundary.

Generalization beyond the training horizon. The first insight is that finite-length teacher CoT can teach the full GD update rule. Although training trajectories contain only K steps, Result 4.1 shows that, in the recovery phase, the test-time recursion matches GD beyond the training horizon. From a practical perspective, increasing the training CoT length K reduces the number of tasks needed to recover GD. This provides a practical trade-off in training-data design: when collecting diverse tasks and in-context examples is costly, longer teacher CoT can be used instead.

CoT budget matters more than task diversity or CoT length alone. In our model, Results 4.1 and 4.2 show that the recovery condition depends on the number of training tasks and the training CoT length K only through the CoT budget $\varphi = \tau(K-1)$. To understand this mechanism intuitively, consider $\alpha > 1$: there are D^2 parameters to learn and $M(K-1)$ teacher steps, each providing D scalar constraints. This gives the recovery condition

$$M(K-1)D > D^2 \quad \Longleftrightarrow \quad \varphi = M(K-1)/D > 1. \tag{5.1}$$

Because each teacher step adds a linearly independent state direction until the accumulated state matrix has rank D, the recovery condition depends on τ and K only through the CoT budget φ.

More generally, our result suggests that the value of teacher CoT is determined not by the number of steps, but by the diversity of states used to constrain the update rule. If successive teacher states are redundant, shorter teacher CoT from more diverse tasks may be more informative. Conversely, a longer teacher CoT may compensate for fewer tasks when it contains sufficiently diverse states.

Context length determines the CoT budget required. For $0 < \alpha < 1$, however, the CoT budget required for recovery depends on the context length ($\varphi > 1/\alpha$). This dependence can be understood through a simple analogy. Suppose the task is sorting, but the context contains only examples of lists without duplicates, such as $[2, 1, 3]$. In this case, sorting and removing duplicates before sorting produce the same output $[1, 2, 3]$, so matching the examples does not uniquely identify the intended rule. However, an example containing duplicates, such as $[1, 2, 2, 3]$, distinguishes the two rules: sorting returns $[1, 2, 2, 3]$, whereas removing duplicates before sorting returns $[1, 2, 3]$. Similarly, in our setting, a short context constrains the update rule only on a subspace, so a larger teacher CoT budget across contexts is required to identify it. In our model, this trade-off is reflected in the recovery condition $\alpha\varphi > 1$ in Result 4.2. Practically, this suggests that the teacher CoT budget should not be chosen independently of the context length.

5.2 Two Mechanisms of Overthinking

Figure 1A shows that, in the divergent regimes, test-time error grows either polynomially or exponentially, depending on the training CoT budget and context length. To explain these behaviors, we formulate inference as a state transition and characterize polynomial overthinking as *error accumulation* and exponential overthinking as *error amplification* (Ainslie et al., 2026). The test-time dynamics in Equation (3.13) can be written as

$$\hat{\boldsymbol{w}}_{t+1} = \mathbf{J}\hat{\boldsymbol{w}}_t + (\eta/L)\mathbf{X}^{\text{in}}\boldsymbol{y}^{\text{in}}, \tag{5.2}$$

where $\mathbf{J} = \mathbf{I} + \mathbf{H}\hat{\mathbf{A}}$. Here, the first term propagates the current state, while the second term injects input from the context at every step. Error evolution is determined by the eigenvalue structure of $\mathbf{J}$.

For error accumulation, consider the polynomial regime in Result 4.2. In this regime, we show that $\mathbf{J}$ has an eigenvalue at 1 that is excited by the context input (see Appendices D and E). Let z_t and $c \neq 0$ denote the state and input components along one such excited mode, respectively. Projecting Equation (5.2) onto this mode gives

$$z_{t+1} = z_t + \eta c. \tag{5.3}$$

Iterating this recursion from $z_0 = 0$ yields $z_t = \eta c t$. Here, the unit coefficient on z_t means that the previously accumulated state persists unchanged from one step to the next, so the input added at each step accumulates rather than being canceled. Consequently, the estimate grows in proportion to t, and the MSE diverges as t^2. In the projected-GD regime ($\alpha > 1$ and $\varphi < 1$), these unit-eigenvalue modes correspond to directions outside the subspace learned during pretraining.

For error amplification, consider the exponential regime in Result 4.2. In this regime, we show that $\mathbf{J}$ has an unstable eigenmode with eigenvalue r satisfying $|r| > 1$, and the context input excites this mode (see Proposition E.8). Let z_t and $c \neq 0$ denote, respectively, the state and input components associated with this mode. Projecting Equation (5.2) onto the mode gives

$$z_{t+1} = r z_t + \eta c. \tag{5.4}$$

Iterating this recursion from $z_0 = 0$ yields $z_t = \eta c\,(r^t - 1)/(r - 1)$, which grows exponentially with t because $|r| > 1$. In our model, this exponential growth arises not merely from a failure to learn the update in some directions, but from learning an incorrect update that actively amplifies errors.

However, unit or unstable modes do not cause divergence unless they are excited by the context input. When $\alpha < 1$, the context input is confined to a lower-dimensional subspace and may have zero component along these modes ($c = 0$). In that case, neither error accumulation nor amplification occurs, giving rise to the bounded regime.

These two mechanisms are also consistent with the empirical observations of Ainslie et al. (2026). In autoregressive dynamical-system prediction, they found that models with unstable eigenmodes ($|\lambda| > 1$) rapidly amplified errors and produced unphysical trajectories, whereas models with neutral modes ($|\lambda| \approx 1$) gradually accumulated prediction errors over time. Starting from the reasoning dynamics of a linear attention model, our study explains how the system undergoes sharp transitions among the three phases of error accumulation, error amplification, and bounded error depending on the balance between the training CoT budget and context length.

5.3 Non-monotonic Effects of the Training CoT Budget

Figure 1C shows that, for $\alpha < 1$, increasing the training CoT budget leads to successive transitions from quadratic divergence to exponential divergence, bounded error, and finally GD recovery. Notably, at intermediate budgets, adding teacher CoT can worsen long-time error growth. This non-monotonicity has a simple interpretation: the model can learn to adapt its next step to its current prediction before it can do so reliably. With very little teacher CoT, the model barely learns to adapt its use of the context to its current prediction. Repeated attention updates therefore add essentially the same context information again and again, causing errors to accumulate. At intermediate budgets, the model begins to adapt each update to its current prediction, but short training contexts do not provide enough information to make this adaptation reliable. An error at one step can then distort the next update, producing an even larger error at the following step. Repeating this process causes the error to grow exponentially. As the teacher CoT budget increases further, the model learns increasingly reliable updates and eventually recovers GD.

A similar non-monotonic pattern has been observed in practical CoT training. Ren et al. (2026) report that model performance initially declines and then recovers during supervised fine-tuning with long teacher CoT. Our theory offers a possible explanation for this behavior by showing that adding teacher CoT can temporarily worsen reasoning performance when short contexts do not sufficiently identify the update rule.

6 Experiments Beyond the Effective Model

This section evaluates whether the qualitative behaviors predicted by the effective model persist in more complex architectures. We retain the setup of Section 3 and vary only the architecture. Specifically, we evaluate two end-to-end trained setups: a single-head linear-attention model (optimizing $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$) and a two-head softmax-attention model (optimizing $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V, \mathbf{W}_O$). Both architectures incorporate a context-only mask: the current state serves as the query, whereas the labeled context examples supply the keys and values. Masking out the state token from the key–value sequence prevents state-only updates and ensures that every inference step depends on the context. Complete formulations, parameterizations, and optimization details are provided in Appendix F.1.

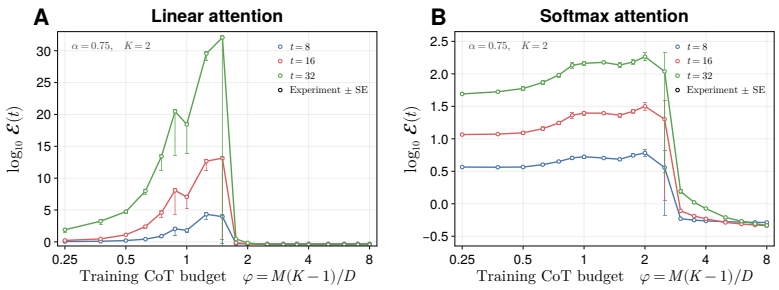


Figure 2: **Training CoT budget produces non-monotonic long-horizon performance in fully trained attention.** Numerical experiments of the test-time generalization error for (A) Single-head linear attention with learned Q/K/V and (B) Two-head softmax attention with learned Q/K/V/O. Parameters: (A,B) $D = 32$, $\alpha = 0.75$, $K = 2$, $\eta = 0.1$, and $\sigma = 0.1$. Error bars represent one standard error of the 8 independent experiments per point.

Figure 2 reports the test-time generalization error dynamics for linear and softmax attention. The first observation is that, given sufficient training CoT, both linear and softmax attention maintain stable inference beyond the CoT length used in training. This suggests that end-to-end trained attention models can learn from finite-length CoT to continue reasoning effectively beyond the training horizon, as discussed in Section 5.1.

More importantly, both models exhibit non-monotonic long-horizon performance as the training CoT budget increases: the error remains relatively low at small budgets, grows rapidly with inference depth at intermediate budgets, and then abruptly transitions to stable inference at larger budgets. This pattern suggests that the non-monotonic pattern and phase-transition-like behavior predicted in Section 5.3 persist in more complex attention models.

Beyond the overall non-monotonicity described above, Figure 2A shows a relatively sharp change within the failure region of linear attention. This is consistent with our finding in Section 5.2 that overthinking takes two distinct forms: error accumulation and error amplification. For softmax attention, the corresponding change in Figure 2B is more gradual, so the failure region is not clearly divided into these two regimes. To understand this, we write the state update for a fixed context Q as

$$\hat{\boldsymbol{w}}_{t+1} = F_Q(\hat{\boldsymbol{w}}_t) := \hat{\boldsymbol{w}}_t + \Delta_\theta(Q, \hat{\boldsymbol{w}}_t). \tag{6.1}$$

If the current state contains a small perturbation $\boldsymbol{\delta}_t$, its one-step evolution can be linearized as

$$\boldsymbol{\delta}_{t+1} := F_Q(\hat{\boldsymbol{w}}_t + \boldsymbol{\delta}_t) - F_Q(\hat{\boldsymbol{w}}_t) \simeq \mathbf{J}_t \boldsymbol{\delta}_t, \qquad \mathbf{J}_t := \left.\frac{\partial F_Q(\boldsymbol{w})}{\partial \boldsymbol{w}}\right|_{\boldsymbol{w}=\hat{\boldsymbol{w}}_t}. \tag{6.2}$$

Here, $\mathbf{J}_t$ describes the one-step evolution of a small state perturbation. For linear attention, the Jacobian is constant throughout inference for a fixed context. For softmax attention, it depends on the current state and changes along the inference trajectory. More generally, in nonlinear architectures, error propagation can vary along a single inference trajectory.

7 Conclusion

We studied how finite teacher CoT affects test-time reasoning and showed that finite trajectories can teach a GD update beyond the training horizon. When GD is not fully recovered, the model exhibits bounded error or two distinct forms of overthinking: polynomial error accumulation and exponential error amplification. In particular, with short contexts, increasing the training CoT budget can temporarily worsen long-horizon performance before ultimately recovering GD. Experiments with fully trained linear and softmax attention further reproduce these qualitative behaviors beyond the effective model. These results offer practical guidance for designing training CoT, while extending the theory to more complex models and tasks remains an important direction for future work.

Reproducibility statement

Detailed derivations are provided in the appendices, and code for reproducing the numerical experiments is provided in the supplementary materials.

AI use statement

Generative AI tools assisted with manuscript drafting, editing, translation, paper organization, numerical code, and the exploration and checking of mathematical derivations. All material was reviewed and verified by the authors, who take responsibility for the final manuscript.

References

Emmanuel Abbe, Samy Bengio, Aryo Lotfi, Colin Sandon, and Omid Saremi. How far can transformers reason? the globality barrier and inductive scratchpad. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 27850–27895. Curran Associates, Inc., 2024. doi: 10.52202/079017-0874. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/3107e4bdb658c79053d7ef59cbc804dd-Paper-Conference.pdf.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=3zKtaqxLhW.

Conrad Ainslie, Pedram Hassanzadeh, Michael W. Mahoney, and Ashesh Chattopadhyay. Eigenanalysis framework for autoregressive neural emulators of multi-scale chaotic dynamics, 2026. URL https://arxiv.org/abs/2608.16084.

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*, 2023. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Arpit Bansal, Avi Schwarzschild, Eitan Borgnia, Zeyad Emam, Furong Huang, Micah Goldblum, and Tom Goldstein. End-to-end algorithm synthesis with recurrent networks: Extrapolation without overthinking. In *Advances in Neural Information Processing Systems*, volume 35, pp. 20232–20242, 2022. doi: 10.52202/068431-1471. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/7f70331dbe58ad59d83941dfa7d975aa-Abstract-Conference.html.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Adam Block, Dylan Foster, Akshay Krishnamurthy, Max Simchowitz, and Cyril Zhang. Butterfly effects of SGD noise: Error amplification in behavior cloning and autoregression. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=CgPs04l9TO.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30583–30598, 2022. doi: 10.52202/068431-2217. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c529dba08a146ea8d6cf715ae8930cbe-Abstract-Conference.html.

Jonas Geiping, Sean Michael McLeish, Neel Jain, John Kirchenbauer, Siddharth Singh, Brian R. Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, and Tom Goldstein. Scaling

up test-time compute with latent reasoning: A recurrent depth approach. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1380. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3b01972cf31e6fa0fe29e4b8b5c2a0a1-Abstract-Conference.html.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space. In *The Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=Itxz7S4Ip3.

Michael Hassid, Gabriel Synnaeve, Yossi Adi, and Roy Schwartz. Don't overthink it. preferring shorter thinking chains for improved LLM reasoning, 2025. URL https://arxiv.org/abs/2505.17813.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14852–14882. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.830. URL https://aclanthology.org/2023.acl-long.830/.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-acl.507. URL https://aclanthology.org/2023.findings-acl.507/.

Jianhao Huang, Zixuan Wang, and Jason D. Lee. Transformers learn to implement multi-step gradient descent with chain of thought. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=r3DF5sOo5B.

Adel Javanmard, Baharan Mirzasoleiman, and Vahab Mirrokni. Understanding the role of training data in test-time scaling. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Y9FfDNa2nJ.

Juno Kim and Taiji Suzuki. Transformers provably solve parity efficiently with chain of thought. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=n2NidsYDop.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, et al. DataComp-LM: In search of the next generation of training sets for language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-0455. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/19e4ea30dded58259665db375885e412-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 39578–39601, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/aca97732e30bcf1303bc22ac3924fd16-Paper-Conference.pdf.

Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=De4FYqjFueZ.

Yue M. Lu, Mary Letey, Jacob A. Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *Proceedings of the National Academy of Sciences*, 122(28):e2502599122, 2025. doi: 10.1073/pnas.2502599122. URL https://doi.org/10.1073/pnas.2502599122.

Zihan Pengmei, Costas Mavromatis, Zhengyuan Shen, Yunyi Zhang, Vassilis N. Ioannidis, and Huzefa Rangwala. The kinetics of reasoning: How chain-of-thought shapes learning in transformers?, 2025. URL https://arxiv.org/abs/2510.25791.

Qihan Ren, Peng Wang, Ruikun Cai, Shuai Shao, Dadi Guo, Yuejin Xie, Yafu Li, Quanshi Zhang, Xia Hu, Jing Shao, and Dongrui Liu. Rethinking generalization in reasoning SFT: A conditional analysis on optimization, data, and model capability, 2026. URL https://arxiv.org/abs/2604.06628.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. CODI: Compressing chain-of-thought into continuous space via self-distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 677–693. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.36. URL https://aclanthology.org/2025.emnlp-main.36/.

Jinyan Su, Jennifer Healey, Preslav Nakov, and Claire Cardie. Between underthinking and overthinking: An empirical study of reasoning length and correctness in LLMs, 2025. URL https://arxiv.org/abs/2505.00127.

Kaito Takanami and Cengiz Pehlevan. An asymptotic theory of chain-of-thought in in-context learning, 2026. URL https://arxiv.org/abs/2606.03217.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174. PMLR, 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022. doi: 10.52202/068431-1800. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html.

Jason Wei, Najoung Kim, Yi Tay, and Quoc V. Le. Inverse scaling can become U-shaped. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15580–15591. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.963. URL https://aclanthology.org/2023.emnlp-main.963/.

Tung-Yu Wu and Melody Lo. U-shaped and inverted-U scaling behind emergent abilities of large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/f696200311b445af686c3ebd87ade1d7-Abstract-Conference.html.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. STaR: Bootstrapping reasoning with reasoning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 15476–15488, 2022. doi: 10.52202/068431-1126. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/639a9a172c044fbb64175b5fad42e9a5-Abstract-Conference.html.

Yedi Zhang, Aaditya K. Singh, Peter E. Latham, and Andrew M. Saxe. Training dynamics of in-context learning in linear attention. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 76047–76087. PMLR, 2025. URL https://proceedings.mlr.press/v267/zhang25br.html.

Dawei Zhu, Xiyu Wei, Guangxiang Zhao, Wenhao Wu, Haosheng Zou, Junfeng Ran, Xun Wang, Lin Sun, Xiangzheng Zhang, and Sujian Li. Chain-of-thought matters: Improving long-context language models with reasoning path supervision. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 3197–3211. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.findings-emnlp.170. URL https://aclanthology.org/2025.findings-emnlp.170/.

Hanlin Zhu, Shibo Hao, Zhiting Hu, Jiantao Jiao, Stuart Russell, and Yuandong Tian. Reasoning by superposition: A theoretical perspective on chain of continuous thought. In *Advances in Neural Information Processing Systems*, volume 38, 2025b. doi: 10.52202/085713-2670. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/72c363c2a573ca2128bd176d3317696b-Abstract-Conference.html.

A Derivation of the Effective Parameterization

In this section, we explain the rationale behind the parameter simplification of the effective model in Section 3.3.

Readout-relevant blocks. Write

$$\mathbf{V} := \mathbf{W}_V, \qquad \mathbf{W} := \mathbf{W}_K^\top \mathbf{W}_Q, \tag{A.1}$$

and partition both matrices into blocks of dimensions $(D, 1, D, 1)$:

$$\mathbf{V} = \begin{pmatrix} \mathbf{V}_{11} & \mathbf{V}_{12} & \mathbf{V}_{13} & \mathbf{V}_{14} \\ \mathbf{V}_{21} & V_{22} & \mathbf{V}_{23} & V_{24} \\ \mathbf{V}_{31} & \mathbf{V}_{32} & \mathbf{V}_{33} & \mathbf{V}_{34} \\ \mathbf{V}_{41} & V_{42} & \mathbf{V}_{43} & V_{44} \end{pmatrix}, \qquad \mathbf{W} = \begin{pmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} & \mathbf{W}_{13} & \mathbf{W}_{14} \\ \mathbf{W}_{21} & W_{22} & \mathbf{W}_{23} & W_{24} \\ \mathbf{W}_{31} & \mathbf{W}_{32} & \mathbf{W}_{33} & \mathbf{W}_{34} \\ \mathbf{W}_{41} & W_{42} & \mathbf{W}_{43} & W_{44} \end{pmatrix}. \tag{A.2}$$

Here every block has the conformable dimension determined by its row and column; in particular, $\mathbf{V}_{32}, \mathbf{V}_{34}, \mathbf{W}_{14}, \mathbf{W}_{34}$ are column vectors, and $\mathbf{W}_{23}, \mathbf{W}_{43}$ are row vectors.

For notational simplicity, suppress the task and step indices and write a context token and the state token as

$$\boldsymbol{c}_l := \begin{pmatrix} \boldsymbol{x}_l \\ y_l \\ \mathbf{0}_D \\ 0 \end{pmatrix}, \qquad \boldsymbol{z}(\boldsymbol{s}) := \begin{pmatrix} \mathbf{0}_D \\ 0 \\ \boldsymbol{s} \\ 1 \end{pmatrix}, \qquad \mathbf{E} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_L, \boldsymbol{z}(\boldsymbol{s})). \tag{A.3}$$

The state readout selects only the third block row of $\mathbf{V}$. Moreover, the query $\boldsymbol{z}(\boldsymbol{s})$ has support only in its third and fourth blocks, so only the third and fourth block columns of $\mathbf{W}$ enter the attention scores. Direct multiplication gives

$$\begin{aligned} \mathsf{Readout}(\mathsf{Atten}(\mathbf{E}; \theta)) = \boldsymbol{s} &+ \frac{1}{L}\big(\mathbf{V}_{31}\mathbf{X} + \mathbf{V}_{32}\boldsymbol{y}^\top\big)\Big[\mathbf{X}^\top(\mathbf{W}_{13}\boldsymbol{s} + \mathbf{W}_{14}) + \boldsymbol{y}(\mathbf{W}_{23}\boldsymbol{s} + W_{24})\Big] \\ &+ \frac{1}{L}(\mathbf{V}_{33}\boldsymbol{s} + \mathbf{V}_{34})\Big[\boldsymbol{s}^\top \mathbf{W}_{33}\boldsymbol{s} + \boldsymbol{s}^\top \mathbf{W}_{34} + \mathbf{W}_{43}\boldsymbol{s} + W_{44}\Big]. \end{aligned} \tag{A.4}$$

Consequently, all blocks outside the third block row of $\mathbf{V}$, and the first two block columns of $\mathbf{W}$, are exactly irrelevant to the state readout. They may be arbitrary, which gives the $*$ blocks in Equation (3.8).

Consequences of rotational symmetry. The data distribution is rotationally symmetric in the feature space. Motivated by this symmetry, we assume that the fixed blocks do not introduce preferred directions. The blocks $\mathbf{V}_{32}, \mathbf{V}_{34}, \mathbf{W}_{14}, \mathbf{W}_{34}$ would otherwise introduce fixed vectors in $\mathbb{R}^D$, while $\mathbf{W}_{23}, \mathbf{W}_{43}$ would select fixed directions from the state. We therefore set

$$\mathbf{V}_{32} = \mathbf{V}_{34} = \mathbf{W}_{14} = \mathbf{W}_{34} = \mathbf{0}_D, \qquad \mathbf{W}_{23} = \mathbf{W}_{43} = \mathbf{0}_D^\top. \tag{A.5}$$

For the same reason, we take the fixed square blocks to be proportional to $\mathbf{I}_D$, while leaving $\mathbf{W}_{13} = \mathbf{A}$ unrestricted. Thus,

$$\mathbf{V}_{31} = a\mathbf{I}_D, \qquad \mathbf{V}_{33} = b\mathbf{I}_D, \qquad \mathbf{W}_{33} = d\mathbf{I}_D, \qquad W_{24} = c, \qquad W_{44} = e. \tag{A.6}$$

Substitution into Equation (A.4) gives

$$\mathsf{Readout}(\mathsf{Atten}(\mathbf{E}; \theta)) = \boldsymbol{s} + \frac{a}{L}\mathbf{X}\Big(\mathbf{X}^\top \mathbf{A}\boldsymbol{s} + c\boldsymbol{y}\Big) + \frac{b}{L}\boldsymbol{s}\big(d\|\boldsymbol{s}\|_2^2 + e\big). \tag{A.7}$$

Scale convention and the label coefficient. The parameters in Equation (A.7) appear only through the products $a\mathbf{A}$, ac, bd, and be. We fix this scalar gauge by absorbing a into $\mathbf{A}$ and c, and set $a = 1$. The value of b becomes irrelevant after the state-only terms are removed below; we choose $b = 1$, giving $\mathbf{V}_{31} = \mathbf{V}_{33} = \mathbf{I}_D$.

The scalar c need not be fixed as an additional assumption. If it is included in the convex learning problem, the transition from $\boldsymbol{s}_{\mu,0} = \mathbf{0}_D$ satisfies

$$\hat{\boldsymbol{s}}_{\mu,1} = \frac{c}{L}\mathbf{X}_\mu \boldsymbol{y}_\mu, \qquad \boldsymbol{s}_{\mu,1} = \frac{\eta}{L}\mathbf{X}_\mu \boldsymbol{y}_\mu. \tag{A.8}$$

The teacher parameters interpolate all training transitions, so the $\lambda \to 0$ solution has zero training error. Equation (A.8) then requires $(c - \eta)\mathbf{X}_\mu \boldsymbol{y}_\mu = \mathbf{0}_D$. Since $\mathbf{X}_\mu \boldsymbol{y}_\mu \neq \mathbf{0}_D$ almost surely, every zero-loss solution has

$$c = \eta. \tag{A.9}$$

Thus fixing this scalar from the outset entails no loss of generality for the $\lambda \to 0$ results.

Excluding context-free inference. The last term of Equation (A.7) is equivariant but updates the estimate using only its current state without its context. We set

$$d = e = 0 \tag{A.10}$$

to prevent the model from advancing its inference without consulting the context. Combining Equations (A.5), (A.9), and (A.10) yields

$$\mathbf{V} = \begin{pmatrix} * & * & * & * \\ * & * & * & * \\ \mathbf{I}_D & \mathbf{0}_D & \mathbf{I}_D & \mathbf{0}_D \\ * & * & * & * \end{pmatrix}, \qquad \mathbf{W} = \begin{pmatrix} * & * & \mathbf{A} & \mathbf{0}_D \\ * & * & \mathbf{0}_D^\top & \eta \\ * & * & \mathbf{0}_{D\times D} & \mathbf{0}_D \\ * & * & \mathbf{0}_D^\top & 0 \end{pmatrix}. \tag{A.11}$$

The unrestricted attention output therefore reduces to

$$\mathsf{Readout}(\mathsf{Atten}(\mathbf{E};\theta)) = \boldsymbol{s} + \frac{1}{L}\mathbf{X}\Big(\mathbf{X}^\top \mathbf{A}\boldsymbol{s} + \eta \boldsymbol{y}\Big), \tag{A.12}$$

which is Equation (3.9). In particular, $\mathbf{A} = -\eta \mathbf{I}_D$ recovers the teacher GD update exactly.

B Summary of the Parameter Regimes

Table 1 summarizes the finite-time and long-time behavior in every open region of the (α, φ) plane.

Table 1: Summary of the parameter regimes.

Regime	Finite-time behavior (Result 4.1)	Long-time behavior (Result 4.2)	Derivation
$\alpha > 1,\ \varphi > 1$	GD	GD	Appendix C
$\alpha > 1,\ 0 < \varphi < 1$	Projected GD	$\Theta(t^2)$	Appendix D
$0 < \alpha < 1,\ 0 < \varphi < \varphi_-$	–	$\Theta(t^2)$	Proposition E.8
$0 < \alpha < 1,\ \varphi_- < \varphi < \varphi_+$	–	$\exp(\Theta(\eta t))$	Proposition E.8
$0 < \alpha < 1,\ \varphi_+ < \varphi < 1/\alpha$	–	$O(1)$	Proposition E.8
$0 < \alpha < 1,\ \alpha\varphi > 1$	GD	GD	Proposition E.4

The two boundaries for $0 < \alpha < 1$ are derived in Propositions E.6 and E.7.

C The $\alpha > 1,\ \varphi > 1$ Case

In this section, we derive the finite-depth generalization error in the GD-recovery regime of Result 4.1, especially in the $\alpha > 1$ case. The projected-GD error is evaluated separately in Appendix D. We write the task and inference covariances as

$$\mathbf{H}_\mu := \frac{\mathbf{X}_\mu \mathbf{X}_\mu^\top}{L}, \qquad \mathbf{H}^{\text{in}} := \frac{\mathbf{X}^{\text{in}}(\mathbf{X}^{\text{in}})^\top}{L}. \tag{C.1}$$

First, we verify that the GD parameters satisfy $\mathcal{L}_{\text{eff}}(\mathbf{A}_\star) = 0$. This is a direct substitution into the model update, after which only uniqueness remains to be shown.

Proposition C.1 (The GD parameters satisfy $\mathcal{L}_{\text{eff}} = 0$)**.** *Suppose $\lambda \to 0$. For every $\alpha > 0$, the parameter $\mathbf{A}_\star = -\eta \mathbf{I}_D$ satisfies*

$$\mathcal{L}_{\text{eff}}(\mathbf{A}_\star) = \frac{1}{MK} \sum_{\mu=1}^{M} \sum_{k=0}^{K-1} \|\hat{\boldsymbol{s}}_{\mu,k+1} - \boldsymbol{s}_{\mu,k+1}\|_2^2 \Big|_{\mathbf{A}=\mathbf{A}_\star} = 0. \tag{C.2}$$

Proof. Substituting $\mathbf{A}_\star = -\eta \mathbf{I}_D$ into the model update in Equation (3.9) gives

$$\hat{\boldsymbol{s}}_{\mu,k+1} = \boldsymbol{s}_{\mu,k} + \frac{1}{L} \mathbf{X}_\mu \left[-\eta \mathbf{X}_\mu^\top \boldsymbol{s}_{\mu,k} + \eta \boldsymbol{y}_\mu \right] \tag{C.3}$$

$$= \boldsymbol{s}_{\mu,k} - \frac{\eta}{L} \mathbf{X}_\mu \left(\mathbf{X}_\mu^\top \boldsymbol{s}_{\mu,k} - \boldsymbol{y}_\mu \right) \tag{C.4}$$

$$= \boldsymbol{s}_{\mu,k+1}, \tag{C.5}$$

where the last equality is the teacher update in Equation (3.2). Substituting Equation (C.5) into Equation (3.10) proves Equation (C.2) in the limit $\lambda \to 0$. □

Proposition C.1 establishes that $\mathcal{L}_{\text{eff}}(\mathbf{A}_\star) = 0$. We now assume $\alpha > 1$ and show that $\mathbf{A}_\star$ is the unique matrix satisfying $\mathcal{L}_{\text{eff}}(\mathbf{A}) = 0$ when $\varphi > 1$. Because every task covariance is then invertible, the proof reduces to the rank of one matrix containing all training states.

Proposition C.2 (Uniqueness for full-rank contexts)**.** *Suppose $\alpha > 1$, $\varphi > 1$, and $\lambda \to 0$. Then $\mathbf{A}_\star$ is the unique solution that satisfies $\mathcal{L}_{\text{eff}}(\mathbf{A}) = 0$, and thus*

$$\hat{\mathbf{A}} = -\eta \mathbf{I}_D. \tag{C.6}$$

Proof. Subtracting the teacher update in Equation (3.2) from the model update in Equation (3.9) gives

$$\hat{\boldsymbol{s}}_{\mu,k+1} - \boldsymbol{s}_{\mu,k+1} = \mathbf{H}_\mu (\mathbf{A} + \eta \mathbf{I}_D) \boldsymbol{s}_{\mu,k}. \tag{C.7}$$

Since $\boldsymbol{s}_{\mu,0} = \mathbf{0}$, the $k = 0$ term in $\mathcal{L}_{\text{eff}}(\mathbf{A})$ vanishes for every $\mathbf{A}$. Equation (C.7) therefore shows that $\mathcal{L}_{\text{eff}}(\mathbf{A}) = 0$ requires

$$\mathbf{H}_\mu (\mathbf{A} + \eta \mathbf{I}_D) \boldsymbol{s}_{\mu,k} = \mathbf{0}_D \qquad (\mu = 1, \dots, M,\ k = 1, \dots, K-1). \tag{C.8}$$

Because $\alpha > 1$, every $\mathbf{H}_\mu$ is invertible almost surely. Collect all noninitial teacher states in

$$\mathbf{S} := (\boldsymbol{s}_{1,1}, \dots, \boldsymbol{s}_{1,K-1}, \dots, \boldsymbol{s}_{M,1}, \dots, \boldsymbol{s}_{M,K-1}). \tag{C.9}$$

Then Equation (C.8) is equivalent to

$$\mathcal{L}_{\text{eff}}(\mathbf{A}) = 0 \quad \Longleftrightarrow \quad (\mathbf{A} + \eta \mathbf{I}_D) \mathbf{S} = \mathbf{0}. \tag{C.10}$$

The teacher recursion in Equation (3.2) yields

$$\boldsymbol{s}_{\mu,k} = \frac{1}{L} q_k(\mathbf{H}_\mu) \mathbf{X}_\mu \boldsymbol{y}_\mu, \tag{C.11}$$

where $q_k(x)$ is defined in Equation (4.5). Write

$$\mathbf{H}_\mu = \mathbf{U}_\mu \operatorname{diag}(x_{\mu,1}, \dots, x_{\mu,D}) \mathbf{U}_\mu^\top. \tag{C.12}$$

Equation (C.11) then gives

$$\mathbf{U}_\mu^\top (\boldsymbol{s}_{\mu,1}, \dots, \boldsymbol{s}_{\mu,K-1}) = \operatorname{diag}\left(\frac{1}{L} \mathbf{U}_\mu^\top \mathbf{X}_\mu \boldsymbol{y}_\mu \right) \begin{pmatrix} q_1(x_{\mu,1}) & \cdots & q_{K-1}(x_{\mu,1}) \\ \vdots & & \vdots \\ q_1(x_{\mu,D}) & \cdots & q_{K-1}(x_{\mu,D}) \end{pmatrix}. \tag{C.13}$$

The polynomial q_k has degree $k-1$ and leading coefficient $(-1)^{k-1}\eta^k$. Therefore, the determinant of the first $K-1$ rows of the polynomial matrix in Equation (C.13) is

$$\det[q_k(x_{\mu,i})]_{i,k=1}^{K-1} = \left[\prod_{k=1}^{K-1} (-1)^{k-1} \eta^k \right] \prod_{1 \le i < j \le K-1} (x_{\mu,j} - x_{\mu,i}). \tag{C.14}$$

The positive Gaussian-Wishart eigenvalues are distinct almost surely, and every entry of $\mathbf{U}_\mu^\top \mathbf{X}_\mu \boldsymbol{y}_\mu / L$ is nonzero almost surely. Hence Equation (C.14) is nonzero, and each task contributes exactly $K-1$ linearly independent states. Because the tasks are independent and Gaussian, their state spans are independent random subspaces and are in general position almost surely. Consequently,

$$\operatorname{rank} \mathbf{S} = \min\{D, M(K-1)\}, \qquad \frac{\operatorname{rank} \mathbf{S}}{D} \longrightarrow \min\{1, \varphi\}. \tag{C.15}$$

When $\varphi > 1$, Equation (C.15) gives $\operatorname{rank} \mathbf{S} = D$. Equation (C.10) then gives $\mathbf{A} = -\eta \mathbf{I}_D = \mathbf{A}_\star$. Hence $\mathbf{A}_\star$ is the unique matrix satisfying $\mathcal{L}_{\text{eff}}(\mathbf{A}) = 0$, and the $\lambda \to 0$ estimator obeys $\hat{\mathbf{A}} = -\eta \mathbf{I}_D$. □

Propositions C.1 and C.2 establish GD recovery for $\alpha > 1$ and $\varphi > 1$. Finally, we compute the generalization error after GD recovery. This calculation applies whenever $\hat{\mathbf{A}} = -\eta \mathbf{I}_D$.

Proposition C.3 (Generalization error after GD recovery)**.** *Suppose $\lambda \to 0$ and $\hat{\mathbf{A}} = -\eta \mathbf{I}_D$. For every fixed test-time depth t,*

$$\mathcal{E}(t) = \mathcal{E}_{\text{GD}}(t) = \left\langle (1 - x q_t)^2 \right\rangle_\alpha + \frac{\sigma^2}{\alpha} \left\langle x q_t^2 \right\rangle_\alpha . \tag{C.16}$$

Proof. Define

$$\mathbf{Q}_t := \eta \sum_{j=0}^{t-1} (\mathbf{I}_D + \mathbf{H}^{\text{in}} \hat{\mathbf{A}})^j. \tag{C.17}$$

Iterating the learned update in Equation (3.13) and substituting the test labels gives

$$\hat{\boldsymbol{w}}_t - \boldsymbol{w}^{\text{in}} = (\mathbf{Q}_t \mathbf{H}^{\text{in}} - \mathbf{I}_D) \boldsymbol{w}^{\text{in}} + \frac{1}{L} \mathbf{Q}_t \mathbf{X}^{\text{in}} \boldsymbol{\varepsilon}^{\text{in}}. \tag{C.18}$$

The two terms in Equation (C.18) are conditionally independent and centered. Their covariances give

$$\mathbb{E}\left[\|\hat{\boldsymbol{w}}_t - \boldsymbol{w}^{\text{in}}\|_2^2 \,\middle|\, \hat{\mathbf{A}}, \mathbf{X}^{\text{in}} \right] = 1 - \frac{2}{D} \operatorname{tr}(\mathbf{Q}_t \mathbf{H}^{\text{in}}) + \frac{1}{D} \|\mathbf{Q}_t \mathbf{H}^{\text{in}}\|_{\text{F}}^2 + \frac{\sigma^2}{L} \operatorname{tr}(\mathbf{Q}_t \mathbf{H}^{\text{in}} \mathbf{Q}_t^\top). \tag{C.19}$$

Substituting $\hat{\mathbf{A}} = -\eta \mathbf{I}_D$ into Equation (C.17) gives

$$\mathbf{Q}_t = \eta \sum_{j=0}^{t-1} (\mathbf{I}_D - \eta \mathbf{H}^{\text{in}})^j = q_t(\mathbf{H}^{\text{in}}). \tag{C.20}$$

Let $x_{1,D}, \ldots, x_{D,D}$ be the eigenvalues of $\mathbf{H}^{\text{in}}$. For fixed t, q_t is a polynomial, so the Marchenko–Pastur law applies term by term in Equation (C.19). Using Equation (C.20),

$$\frac{1}{D} \operatorname{tr}(\mathbf{Q}_t \mathbf{H}^{\text{in}}) = \frac{1}{D} \sum_{i=1}^{D} x_{i,D} q_t(x_{i,D}) \longrightarrow \langle x q_t \rangle_\alpha , \tag{C.21}$$

$$\frac{1}{D} \|\mathbf{Q}_t \mathbf{H}^{\text{in}}\|_{\text{F}}^2 = \frac{1}{D} \operatorname{tr}\left[(\mathbf{H}^{\text{in}})^2 q_t(\mathbf{H}^{\text{in}})^2 \right] \tag{C.22}$$

$$= \frac{1}{D} \sum_{i=1}^{D} x_{i,D}^2 q_t(x_{i,D})^2 \longrightarrow \left\langle x^2 q_t^2 \right\rangle_\alpha , \tag{C.23}$$

$$\frac{\sigma^2}{L} \operatorname{tr}(\mathbf{Q}_t \mathbf{H}^{\text{in}} \mathbf{Q}_t^\top) = \sigma^2 \frac{D}{L} \frac{1}{D} \operatorname{tr}\left[\mathbf{H}^{\text{in}} q_t(\mathbf{H}^{\text{in}})^2 \right] \tag{C.24}$$

$$= \sigma^2 \frac{D}{L} \frac{1}{D} \sum_{i=1}^{D} x_{i,D} q_t(x_{i,D})^2 \longrightarrow \frac{\sigma^2}{\alpha} \left\langle x q_t^2 \right\rangle_\alpha . \tag{C.25}$$

Substitution into Equation (C.19) gives

$$\mathcal{E}_{\text{GD}}(t) = 1 - 2 \langle x q_t \rangle_\alpha + \left\langle x^2 q_t^2 \right\rangle_\alpha + \frac{\sigma^2}{\alpha} \left\langle x q_t^2 \right\rangle_\alpha \tag{C.26}$$

$$= \left\langle (1 - x q_t)^2 \right\rangle_\alpha + \frac{\sigma^2}{\alpha} \left\langle x q_t^2 \right\rangle_\alpha , \tag{C.27}$$

which proves Equation (C.16). □

This completes the $\alpha > 1$ and $\varphi > 1$ case.

D THE $\alpha > 1,\ 0 < \varphi < 1$ CASE

In this section, we compute the generalization error for $\alpha > 1$ and $0 < \varphi < 1$. First, we state the finite-time error and derive both the learned matrix and its risk.

Proposition D.1 (Finite-time error of projected GD)**.** *Suppose $\alpha > 1$, $0 < \varphi < 1$, and $\lambda \to 0$. For every fixed test-time depth t, the quantities defined in Equation (4.5) give*

$$\mathcal{E}(t) = 1 - \varphi + \varphi\langle(1 - xq_t)^2\rangle_{\alpha/\varphi} + \frac{\varphi[2(1-\varphi) + \sigma^2]}{\alpha}\langle xq_t^2\rangle_{\alpha/\varphi} - \frac{2(1-\varphi)}{\alpha}\kappa_{1,t} + \frac{(1-\varphi)(1-\varphi+\sigma^2)}{\alpha^2}\kappa_{2,t} + \frac{1-\varphi}{\alpha^2}\kappa_{1,t}^2. \tag{D.1}$$

Proof. Define the projector onto the training-state span:

$$\mathbf{P} := \mathbf{S}\mathbf{S}^\dagger. \tag{D.2}$$

Every noninitial training state is a column of $\mathbf{S}$, so $\mathbf{P}\boldsymbol{s}_{\mu,k} = \boldsymbol{s}_{\mu,k}$ for $k = 1, \dots, K-1$. Using Equation (C.7),

$$(\hat{\boldsymbol{s}}_{\mu,k+1} - \boldsymbol{s}_{\mu,k+1})\big|_{\mathbf{A}=-\eta\mathbf{P}} = \eta\mathbf{H}_\mu(\mathbf{I}_D - \mathbf{P})\boldsymbol{s}_{\mu,k} = \mathbf{0}_D. \tag{D.3}$$

The $k = 0$ residual also vanishes because $\boldsymbol{s}_{\mu,0} = \mathbf{0}_D$. Consequently,

$$\mathcal{L}_{\text{eff}}(-\eta\mathbf{P}) = \frac{\lambda\eta^2}{D}\|\mathbf{P}\|_{\text{F}}^2 = \frac{\lambda\eta^2}{D}\operatorname{tr}\mathbf{P} \longrightarrow 0 \qquad (\lambda \to 0). \tag{D.4}$$

Multiplying Equation (C.10) by $\mathbf{S}^\dagger$ gives

$$\mathbf{A}\mathbf{P} = -\eta\mathbf{P}. \tag{D.5}$$

Every matrix satisfying Equation (D.5) obeys

$$\mathbf{A} = -\eta\mathbf{P} + \mathbf{A}(\mathbf{I}_D - \mathbf{P}), \tag{D.6}$$

$$\|\mathbf{A}\|_{\text{F}}^2 = \eta^2 \operatorname{tr}\mathbf{P} + \|\mathbf{A}(\mathbf{I}_D - \mathbf{P})\|_{\text{F}}^2. \tag{D.7}$$

Because $\lambda \to 0$ selects the minimum-Frobenius-norm matrix satisfying $\mathcal{L}_{\text{eff}}(\mathbf{A}) = 0$, Equation (D.7) gives

$$\hat{\mathbf{A}} = -\eta\mathbf{P}. \tag{D.8}$$

Let $r := M(K-1)$. Because $0 < \varphi < 1$, Equation (C.15) gives $\operatorname{rank}\mathbf{P} = r$ and $r/D \to \varphi$. The inference data are independent of $\mathbf{P}$, so rotational invariance allows us to choose a basis in which

$$\mathbf{P} = \begin{pmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \qquad \mathbf{X}^{\text{in}} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}, \qquad \mathbf{B} := \frac{\mathbf{X}_1\mathbf{X}_1^\top}{L}, \tag{D.9}$$

where $\mathbf{X}_1 \in \mathbb{R}^{r\times L}$ and $\mathbf{X}_2 \in \mathbb{R}^{(D-r)\times L}$ are independent standard Gaussian matrices. Substituting $\hat{\mathbf{A}} = -\eta\mathbf{P}$ into Equation (C.17) gives

$$\mathbf{Q}_t = q_t(\mathbf{H}^{\text{in}}\mathbf{P}). \tag{D.10}$$

The identity

$$(\mathbf{H}^{\text{in}}\mathbf{P})^j\mathbf{X}^{\text{in}} = \mathbf{X}^{\text{in}}\left[\frac{(\mathbf{X}^{\text{in}})^\top\mathbf{P}\mathbf{X}^{\text{in}}}{L}\right]^j \tag{D.11}$$

and the block decomposition in Equation (D.9) give

$$\mathbf{Q}_t\mathbf{X}^{\text{in}} = \mathbf{X}^{\text{in}}q_t\left(\frac{\mathbf{X}_1^\top\mathbf{X}_1}{L}\right) = \begin{pmatrix} q_t(\mathbf{B})\mathbf{X}_1 \\ \mathbf{X}_2\mathbf{K}_t \end{pmatrix}, \tag{D.12}$$

where we introduced

$$\mathbf{K}_t := q_t\left(\frac{\mathbf{X}_1^\top\mathbf{X}_1}{L}\right). \tag{D.13}$$

Multiplying Equation (D.12) by $(\mathbf{X}^{\text{in}})^\top/L$ gives

$$\mathbf{Q}_t\mathbf{H}^{\text{in}} = \begin{pmatrix} \mathbf{B}q_t(\mathbf{B}) & q_t(\mathbf{B})\mathbf{X}_1\mathbf{X}_2^\top/L \\ \mathbf{X}_2\mathbf{X}_1^\top q_t(\mathbf{B})/L & \mathbf{X}_2\mathbf{K}_t\mathbf{X}_2^\top/L \end{pmatrix}. \tag{D.14}$$

We next average the $\mathbf{X}_2$ blocks in Equations (D.12) and (D.14) conditional on $\mathbf{X}_1$. The matrix $\mathbf{K}_t$ is symmetric and independent of $\mathbf{X}_2$. For independent standard Gaussian rows $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ of $\mathbf{X}_2$, Wick's identity gives

$$\mathbb{E}(\boldsymbol{z}_i^\top\mathbf{K}_t\boldsymbol{z}_i) = \operatorname{tr}\mathbf{K}_t, \tag{D.15}$$

$$\mathbb{E}(\boldsymbol{z}_i^\top\mathbf{K}_t\boldsymbol{z}_i)^2 = 2\operatorname{tr}(\mathbf{K}_t^2) + (\operatorname{tr}\mathbf{K}_t)^2, \tag{D.16}$$

$$\mathbb{E}(\boldsymbol{z}_i^\top\mathbf{K}_t\boldsymbol{z}_j)^2 = \operatorname{tr}(\mathbf{K}_t^2) \qquad (i \neq j). \tag{D.17}$$

Summing the diagonal and off-diagonal row pairs in Equation (D.17) gives

$$\mathbb{E}_{\mathbf{X}_2}\operatorname{tr}(\mathbf{X}_2\mathbf{K}_t\mathbf{X}_2^\top) = (D-r)\operatorname{tr}\mathbf{K}_t, \tag{D.18}$$

$$\mathbb{E}_{\mathbf{X}_2}\|\mathbf{X}_2\mathbf{K}_t\mathbf{X}_2^\top\|_{\text{F}}^2 = (D-r)(D-r+1)\operatorname{tr}(\mathbf{K}_t^2) + (D-r)(\operatorname{tr}\mathbf{K}_t)^2. \tag{D.19}$$

The off-diagonal block in Equation (D.14) satisfies

$$\mathbb{E}_{\mathbf{X}_2}\left\|q_t(\mathbf{B})\frac{\mathbf{X}_1\mathbf{X}_2^\top}{L}\right\|_{\text{F}}^2 = \frac{D-r}{L^2}\operatorname{tr}\Big[q_t(\mathbf{B})\mathbf{X}_1\mathbf{X}_1^\top q_t(\mathbf{B})\Big] = \frac{D-r}{L}\operatorname{tr}\big[\mathbf{B}q_t(\mathbf{B})^2\big]. \tag{D.20}$$

The source response in Equation (D.12) satisfies

$$\mathbb{E}_{\mathbf{X}_2}\|\mathbf{Q}_t\mathbf{X}^{\text{in}}\|_{\text{F}}^2 = \|q_t(\mathbf{B})\mathbf{X}_1\|_{\text{F}}^2 + \mathbb{E}_{\mathbf{X}_2}\|\mathbf{X}_2\mathbf{K}_t\|_{\text{F}}^2 \tag{D.21}$$

$$= L\operatorname{tr}\big[\mathbf{B}q_t(\mathbf{B})^2\big] + (D-r)\operatorname{tr}(\mathbf{K}_t^2). \tag{D.22}$$

Substituting Equations (D.19), (D.20), and (D.22) into Equation (C.19) gives

$$\begin{aligned}\mathbb{E}\big[\|\hat{\boldsymbol{w}}_t - \boldsymbol{w}^{\text{in}}\|_2^2 \,\big|\, \mathbf{X}_1\big] = 1 &- \frac{2}{D}\left[\operatorname{tr}(\mathbf{B}q_t(\mathbf{B})) + \frac{D-r}{L}\operatorname{tr}\mathbf{K}_t\right] \\ &+ \frac{1}{D}\operatorname{tr}\big[\mathbf{B}^2q_t(\mathbf{B})^2\big] + \left[\frac{2(D-r)}{DL} + \frac{\sigma^2}{L}\right]\operatorname{tr}\big[\mathbf{B}q_t(\mathbf{B})^2\big] \\ &+ \frac{(D-r)(D-r+1+\sigma^2D)}{DL^2}\operatorname{tr}(\mathbf{K}_t^2) + \frac{D-r}{DL^2}(\operatorname{tr}\mathbf{K}_t)^2.\end{aligned} \tag{D.23}$$

Finally, we evaluate the traces in Equation (D.23) by the Marchenko–Pastur law. The block covariance $\mathbf{B}$ has aspect ratio

$$\frac{L}{r} = \frac{L/D}{r/D} \longrightarrow \frac{\alpha}{\varphi}. \tag{D.24}$$

Therefore, for every fixed polynomial f,

$$\frac{1}{D}\operatorname{tr} f(\mathbf{B}) = \frac{r}{D}\frac{1}{r}\operatorname{tr} f(\mathbf{B}) \longrightarrow \varphi\langle f\rangle_{\alpha/\varphi}. \tag{D.25}$$

The matrices $\mathbf{X}_1^\top\mathbf{X}_1/L$ and $\mathbf{B}$ have the same r nonzero eigenvalues. The matrix $\mathbf{X}_1^\top\mathbf{X}_1/L$ also has $L-r$ zero eigenvalues, and Equation (4.5) gives $q_t(0) = \eta t$. Hence, for $j = 1, 2$,

$$\frac{1}{D}\operatorname{tr}(\mathbf{K}_t^j) = \frac{L-r}{D}(\eta t)^j + \frac{r}{D}\frac{1}{r}\operatorname{tr}(q_t(\mathbf{B})^j) \tag{D.26}$$

$$\longrightarrow (\alpha-\varphi)(\eta t)^j + \varphi\langle q_t^j\rangle_{\alpha/\varphi} = \kappa_{j,t}. \tag{D.27}$$

Substituting Equations (D.25) and (D.27) into Equation (D.23) gives Equation (D.1), which is Equation (4.7) in Result 4.1. □

Proposition D.1 gives the exact fixed-time error without a cavity approximation. We now derive its long-time growth for $\alpha > 1$.

Proposition D.2 (Quadratic long-time growth of projected GD)**.** *Suppose* $\alpha > 1$, $0 < \varphi < 1$, *and*

$$0 < \eta < \frac{2}{(1+\sqrt{\varphi/\alpha})^2}. \tag{D.28}$$

Then

$$\lim_{t\to\infty} \frac{\mathcal{E}(t)}{t^2} = \frac{\eta^2(1-\varphi)(\alpha-\varphi)(\alpha+1-2\varphi+\sigma^2)}{\alpha^2} > 0. \tag{D.29}$$

Proof. Because $\alpha/\varphi > 1$, the Marchenko–Pastur support is separated from zero. Under Equation (D.28), the nonzero modes are stable, and therefore

$$\langle q_t \rangle_{\alpha/\varphi} = O(1), \qquad \langle q_t^2 \rangle_{\alpha/\varphi} = O(1), \qquad t \to \infty. \tag{D.30}$$

Equations (4.5) and (D.30) give

$$\kappa_{1,t} = \eta t(\alpha-\varphi) + O(1), \qquad \kappa_{2,t} = \eta^2 t^2(\alpha-\varphi) + O(1). \tag{D.31}$$

Substituting Equation (D.31) into Equation (D.1) gives

$$\lim_{t\to\infty} \frac{\mathcal{E}(t)}{t^2} = \frac{\eta^2(1-\varphi)(1-\varphi+\sigma^2)(\alpha-\varphi)}{\alpha^2} + \frac{\eta^2(1-\varphi)(\alpha-\varphi)^2}{\alpha^2} \tag{D.32}$$

$$= \frac{\eta^2(1-\varphi)(\alpha-\varphi)(\alpha+1-2\varphi+\sigma^2)}{\alpha^2}, \tag{D.33}$$

which proves Equation (D.29). □

The rank deficit produces a unit-eigenvalue sector in the learned iteration. Each inference step adds a source component in that sector, so the state grows linearly and the squared error grows quadratically. Together with full-GD recovery for $\varphi > 1$, Proposition D.2 completes the $\alpha > 1$ part of Result 4.2.

E THE $0 < \alpha < 1$ CASE

In this section, we derive the $\alpha < 1$ part of Result 4.2 by the cavity method and combine it with the exact $\alpha > 1$ calculation.

E.1 STATISTICAL PROPERTIES OF LEARNED MATRIX

First, we compute the moments of the learned matrix below the GD-recovery threshold using the cavity method in random matrix theory.

Proposition E.1 (Moments of the learned matrix)**.** *Suppose* $0 < \alpha < 1$, $0 < \varphi < 1/\alpha$, *and* $\lambda \to 0$. *With the large-D replacements stated explicitly in the proof,*

$$-\lim_{D\to\infty} \frac{\operatorname{tr}\hat{\mathbf{A}}}{\eta D} = \lim_{D\to\infty} \frac{\|\hat{\mathbf{A}}\|_{\mathrm{F}}^2}{\eta^2 D} = \frac{\varphi}{1+(1-\alpha)\varphi}, \tag{E.1}$$

$$\lim_{D\to\infty} \frac{\|\hat{\mathbf{A}} + \eta \mathbf{I}_D\|_{\mathrm{F}}^2}{\eta^2 D} = \frac{1-\alpha\varphi}{1+(1-\alpha)\varphi}, \tag{E.2}$$

and

$$\lim_{D\to\infty} \frac{\operatorname{tr}(\hat{\mathbf{A}}^2)}{\eta^2 D} = \frac{\varphi[1+(1-2\alpha)\varphi]}{[1+(1-\alpha)\varphi]^2(1-\alpha^2\varphi)}. \tag{E.3}$$

Proof. Proposition C.1 shows that $\mathcal{L}_{\mathrm{eff}}(\mathbf{A}_\star) = 0$ for every α. Because the training objective is quadratic, its $\lambda \to 0$ limit selects the minimum-norm matrix among those satisfying $\mathcal{L}_{\mathrm{eff}}(\mathbf{A}) = 0$. The training contexts are then rank deficient, so the constraints need not identify the update. Define

$$\mathbf{P}_\mu := \mathbf{X}_\mu(\mathbf{X}_\mu^\top \mathbf{X}_\mu)^{-1}\mathbf{X}_\mu^\top, \tag{E.4}$$

$$\mathbf{Q}_\mu := \mathbf{S}_\mu(\mathbf{S}_\mu^\top \mathbf{S}_\mu)^{-1}\mathbf{S}_\mu^\top, \tag{E.5}$$

$$\mathbf{S}_\mu := (\boldsymbol{s}_{\mu,1}, \ldots, \boldsymbol{s}_{\mu,K-1}). \tag{E.6}$$

By Equation (E.6), the teacher states lie in the context span and $\mathbf{P}_\mu \mathbf{Q}_\mu = \mathbf{Q}_\mu$. The condition $\mathcal{L}_{\text{eff}}(\mathbf{A}) = 0$ is equivalent to

$$\mathbf{H}_\mu(\mathbf{A} + \eta \mathbf{I}_D)\mathbf{S}_\mu = \mathbf{0} \iff \mathbf{P}_\mu(\mathbf{A} + \eta \mathbf{I}_D)\mathbf{Q}_\mu = \mathbf{0} \tag{E.7}$$

$$\iff (\mathbf{Q}_\mu \otimes \mathbf{P}_\mu)\,\mathrm{vec}(\mathbf{A} + \eta \mathbf{I}_D) = \mathbf{0}. \tag{E.8}$$

The minimum-norm solution is unchanged if the original loss is replaced by the squared norm of these equivalent constraints. With an auxiliary regularizer $\epsilon > 0$, define

$$\mathcal{R}_\epsilon := \left[\sum_{\mu=1}^{M} \mathbf{Q}_\mu \otimes \mathbf{P}_\mu + \epsilon \mathbf{I}_{D^2} \right]^{-1}. \tag{E.9}$$

Vectorizing Equation (E.8) and using Equation (E.9) gives

$$\mathrm{vec}(\hat{\mathbf{A}}) = -\eta \lim_{\epsilon \to 0} \mathcal{R}_\epsilon \left[\sum_{\mu} \mathbf{Q}_\mu \otimes \mathbf{P}_\mu \right] \mathrm{vec}\, \mathbf{I}_D \tag{E.10}$$

$$= -\eta\, \mathrm{vec}\, \mathbf{I}_D + \eta \lim_{\epsilon \to 0} \epsilon \mathcal{R}_\epsilon\, \mathrm{vec}\, \mathbf{I}_D. \tag{E.11}$$

Equation (E.11) gives

$$\frac{\mathrm{vec}(\hat{\mathbf{A}})}{\eta} = -\,\mathrm{vec}\, \mathbf{I}_D + \left[\lim_{\epsilon \to 0} \epsilon \mathcal{R}_\epsilon \right] \mathrm{vec}\, \mathbf{I}_D. \tag{E.12}$$

The matrix $\lim_{\epsilon \to 0} \epsilon \mathcal{R}_\epsilon$ is an orthogonal projector, and hence its transpose and its square are equal to itself. Therefore,

$$-\frac{\mathrm{tr}\, \hat{\mathbf{A}}}{\eta D} = 1 - \frac{1}{D} (\mathrm{vec}\, \mathbf{I}_D)^\top \left[\lim_{\epsilon \to 0} \epsilon \mathcal{R}_\epsilon \right] \mathrm{vec}\, \mathbf{I}_D, \tag{E.13}$$

$$\frac{\|\hat{\mathbf{A}}\|_{\mathrm{F}}^2}{\eta^2 D} = \frac{1}{D} (\mathrm{vec}\, \mathbf{I}_D)^\top \left[\mathbf{I}_{D^2} - \lim_{\epsilon \to 0} \epsilon \mathcal{R}_\epsilon \right]^2 \mathrm{vec}\, \mathbf{I}_D \tag{E.14}$$

$$= 1 - \frac{1}{D} (\mathrm{vec}\, \mathbf{I}_D)^\top \left[\lim_{\epsilon \to 0} \epsilon \mathcal{R}_\epsilon \right] \mathrm{vec}\, \mathbf{I}_D, \tag{E.15}$$

$$\frac{\mathrm{tr}(\hat{\mathbf{A}}^2)}{\eta^2 D} = 1 - \frac{2}{D} (\mathrm{vec}\, \mathbf{I}_D)^\top \left[\lim_{\epsilon \to 0} \epsilon \mathcal{R}_\epsilon \right] \mathrm{vec}\, \mathbf{I}_D \tag{E.16}$$

$$+ \frac{1}{D} (\mathrm{vec}\, \mathbf{I}_D)^\top \left[\lim_{\epsilon \to 0} \epsilon \mathcal{R}_\epsilon \right] \mathrm{T}_D \left[\lim_{\epsilon \to 0} \epsilon \mathcal{R}_\epsilon \right] \mathrm{vec}\, \mathbf{I}_D, \tag{E.17}$$

where $\mathrm{T}_D : \mathbb{R}^{D^2} \to \mathbb{R}^{D^2}$ denotes the transpose operator defined by

$$\mathrm{T}_D\, \mathrm{vec}(\mathbf{B}) = \mathrm{vec}(\mathbf{B}^\top). \tag{E.18}$$

It therefore suffices to calculate the following two scalars separately:

$$\frac{1}{D} (\mathrm{vec}\, \mathbf{I}_D)^\top \left[\lim_{\epsilon \to 0} \epsilon \mathcal{R}_\epsilon \right] \mathrm{vec}\, \mathbf{I}_D, \tag{E.19}$$

and

$$\frac{1}{D} (\mathrm{vec}\, \mathbf{I}_D)^\top \left[\lim_{\epsilon \to 0} \epsilon \mathcal{R}_\epsilon \right] \mathrm{T}_D \left[\lim_{\epsilon \to 0} \epsilon \mathcal{R}_\epsilon \right] \mathrm{vec}\, \mathbf{I}_D. \tag{E.20}$$

Calculation of Equation (E.19). First, we calculate Equation (E.19). Define the order parameter:

$$\chi := -\lim_{D \to \infty} \frac{\mathrm{tr}\, \hat{\mathbf{A}}}{\eta D}. \tag{E.21}$$

Equations (E.17) and (E.19) give

$$\frac{1}{D} (\mathrm{vec}\, \mathbf{I}_D)^\top \left[\lim_{\epsilon \to 0} \epsilon \mathcal{R}_\epsilon \right] \mathrm{vec}\, \mathbf{I}_D = 1 - \chi. \tag{E.22}$$

Since $\mathcal{R}_\epsilon$ is the resolvent of a sum of independent task contributions, we use a leave-one-task cavity calculation, in which the removed task is independent of the resolvent formed by the remaining tasks. In this approach, deleting task μ defines

$$\mathcal{R}_\epsilon^{-\mu} := \left[\sum_{\nu\neq\mu} \mathbf{Q}_\nu \otimes \mathbf{P}_\nu + \epsilon \mathbf{I}_{D^2}\right]^{-1}. \tag{E.23}$$

Because $\mathbf{P}_\mu$ and $\mathbf{Q}_\mu$ are projectors,

$$(\mathbf{Q}_\mu \otimes \mathbf{P}_\mu)^2 = (\mathbf{Q}_\mu^2 \otimes \mathbf{P}_\mu^2) = \mathbf{Q}_\mu \otimes \mathbf{P}_\mu \tag{E.24}$$

$$(\operatorname{vec} \mathbf{I}_D)^\top (\mathbf{Q}_\mu \otimes \mathbf{P}_\mu) \operatorname{vec} \mathbf{I}_D = \operatorname{tr}(\mathbf{P}_\mu \mathbf{Q}_\mu) = K - 1. \tag{E.25}$$

Applying the Woodbury formula to Equations (E.9) and (E.23), and using Equation (E.25), gives

$$(\mathbf{Q}_\mu \otimes \mathbf{P}_\mu)\mathcal{R}_\epsilon = \left[\epsilon \mathbf{I}_{D^2} + (\mathbf{Q}_\mu \otimes \mathbf{P}_\mu)\epsilon \mathcal{R}_\epsilon^{-\mu} (\mathbf{Q}_\mu \otimes \mathbf{P}_\mu)\right]^{-1} (\mathbf{Q}_\mu \otimes \mathbf{P}_\mu)\epsilon \mathcal{R}_\epsilon^{-\mu}. \tag{E.26}$$

We first evaluate the matrix inside the inverse in Equation (E.26). The matrix $\mathcal{R}_\epsilon^{-\mu}$ depends only on the tasks $\nu \neq \mu$, and is therefore independent of $(\mathbf{P}_\mu, \mathbf{Q}_\mu)$. Its compression by task μ concentrates around its conditional average, so

$$(\mathbf{Q}_\mu \otimes \mathbf{P}_\mu) \left[\lim_{\epsilon\to 0} \epsilon \mathcal{R}_\epsilon^{-\mu}\right] (\mathbf{Q}_\mu \otimes \mathbf{P}_\mu) \simeq \frac{1}{D^2} \operatorname{tr}\left[\lim_{\epsilon\to 0} \epsilon \mathcal{R}_\epsilon^{-\mu}\right] (\mathbf{Q}_\mu \otimes \mathbf{P}_\mu). \tag{E.27}$$

To calculate the normalized trace in this expression, let $\lambda_1, \ldots, \lambda_{D^2}$ be the eigenvalues of

$$\sum_{\nu\neq\mu} \mathbf{Q}_\nu \otimes \mathbf{P}_\nu. \tag{E.28}$$

The eigenvalues of $\mathcal{R}_\epsilon^{-\mu}$ are $1/(\lambda_j + \epsilon)$, and hence the eigenvalues of $\epsilon \mathcal{R}_\epsilon^{-\mu}$ satisfy

$$\frac{\epsilon}{\lambda_j + \epsilon} \to \begin{cases} 0, & \lambda_j > 0, \\ 1, & \lambda_j = 0. \end{cases} \tag{E.29}$$

Therefore,

$$\operatorname{tr}\left[\lim_{\epsilon\to 0} \epsilon \mathcal{R}_\epsilon^{-\mu}\right] = \#\{j : \lambda_j = 0\} = D^2 - \operatorname{rank}\left(\sum_{\nu\neq\mu} \mathbf{Q}_\nu \otimes \mathbf{P}_\nu\right). \tag{E.30}$$

For each task,

$$\operatorname{rank}(\mathbf{Q}_\nu \otimes \mathbf{P}_\nu) = \operatorname{rank}(\mathbf{Q}_\nu) \operatorname{rank}(\mathbf{P}_\nu) = L(K-1). \tag{E.31}$$

Since the constraints from different tasks are linearly independent,

$$\operatorname{rank}\left(\sum_{\nu\neq\mu} \mathbf{Q}_\nu \otimes \mathbf{P}_\nu\right) = (M-1)L(K-1). \tag{E.32}$$

It follows that

$$\frac{1}{D^2} \operatorname{tr}\left[\lim_{\epsilon\to 0} \epsilon \mathcal{R}_\epsilon^{-\mu}\right] = 1 - \frac{(M-1)L(K-1)}{D^2} \tag{E.33}$$

$$= 1 - \frac{L}{D}\frac{(M-1)(K-1)}{D} \tag{E.34}$$

$$\to 1 - \alpha\varphi. \tag{E.35}$$

Consequently, the compression in Equation (E.26) satisfies

$$(\mathbf{Q}_\mu \otimes \mathbf{P}_\mu) \left[\lim_{\epsilon\to 0} \epsilon \mathcal{R}_\epsilon^{-\mu}\right] (\mathbf{Q}_\mu \otimes \mathbf{P}_\mu) \simeq (1 - \alpha\varphi)(\mathbf{Q}_\mu \otimes \mathbf{P}_\mu). \tag{E.36}$$

Multiplying Equation (E.26) by $\operatorname{vec}\mathbf{I}_D$ and then by $(\operatorname{vec}\mathbf{I}_D)^\top/(K-1)$ from the left gives

$$\lim_{\epsilon\to 0}\frac{1}{K-1}(\operatorname{vec}\mathbf{I}_D)^\top(\mathbf{Q}_\mu\otimes\mathbf{P}_\mu)\mathcal{R}_\epsilon\operatorname{vec}\mathbf{I}_D \tag{E.37}$$

$$=\lim_{\epsilon\to 0}\frac{1}{K-1}(\operatorname{vec}\mathbf{I}_D)^\top\left[\epsilon\mathbf{I}_{D^2}+(\mathbf{Q}_\mu\otimes\mathbf{P}_\mu)\epsilon\mathcal{R}_\epsilon^{-\mu}(\mathbf{Q}_\mu\otimes\mathbf{P}_\mu)\right]^{-1}(\mathbf{Q}_\mu\otimes\mathbf{P}_\mu)\epsilon\mathcal{R}_\epsilon^{-\mu}\operatorname{vec}\mathbf{I}_D \tag{E.38}$$

$$\simeq\lim_{\epsilon\to 0}\frac{1}{K-1}(\operatorname{vec}\mathbf{I}_D)^\top\left[\epsilon\mathbf{I}_{D^2}+(1-\alpha\varphi)(\mathbf{Q}_\mu\otimes\mathbf{P}_\mu)\right]^{-1}(\mathbf{Q}_\mu\otimes\mathbf{P}_\mu)\epsilon\mathcal{R}_\epsilon^{-\mu}\operatorname{vec}\mathbf{I}_D \tag{E.39}$$

$$=\lim_{\epsilon\to 0}\frac{1}{(K-1)(\epsilon+1-\alpha\varphi)}(\operatorname{vec}\mathbf{I}_D)^\top(\mathbf{Q}_\mu\otimes\mathbf{P}_\mu)\epsilon\mathcal{R}_\epsilon^{-\mu}\operatorname{vec}\mathbf{I}_D \tag{E.40}$$

$$=\frac{1}{(K-1)(1-\alpha\varphi)}(\operatorname{vec}\mathbf{I}_D)^\top(\mathbf{Q}_\mu\otimes\mathbf{P}_\mu)\left[\lim_{\epsilon\to 0}\epsilon\mathcal{R}_\epsilon^{-\mu}\right]\operatorname{vec}\mathbf{I}_D \tag{E.41}$$

$$\simeq\frac{1}{1-\alpha\varphi}\frac{1}{D}(\operatorname{vec}\mathbf{I}_D)^\top\left[\lim_{\epsilon\to 0}\epsilon\mathcal{R}_\epsilon^{-\mu}\right]\operatorname{vec}\mathbf{I}_D \tag{E.42}$$

$$\simeq\frac{1}{1-\alpha\varphi}\frac{1}{D}(\operatorname{vec}\mathbf{I}_D)^\top\left[\lim_{\epsilon\to 0}\epsilon\mathcal{R}_\epsilon\right]\operatorname{vec}\mathbf{I}_D \tag{E.43}$$

$$=\frac{1-\chi}{1-\alpha\varphi}. \tag{E.44}$$

Here the first $\simeq$ uses Equation (E.36). The following equality uses

$$\left[\epsilon\mathbf{I}_{D^2}+(1-\alpha\varphi)(\mathbf{Q}_\mu\otimes\mathbf{P}_\mu)\right]^{-1}(\mathbf{Q}_\mu\otimes\mathbf{P}_\mu)=\frac{1}{\epsilon+1-\alpha\varphi}(\mathbf{Q}_\mu\otimes\mathbf{P}_\mu), \tag{E.45}$$

which follows from $(\mathbf{Q}_\mu\otimes\mathbf{P}_\mu)^2=\mathbf{Q}_\mu\otimes\mathbf{P}_\mu$. The second $\simeq$ uses the following large-D concentration for the contraction with the independent task μ:

$$\frac{1}{K-1}(\operatorname{vec}\mathbf{I}_D)^\top(\mathbf{Q}_\mu\otimes\mathbf{P}_\mu)\left[\lim_{\epsilon\to 0}\epsilon\mathcal{R}_\epsilon^{-\mu}\right]\operatorname{vec}\mathbf{I}_D\simeq\frac{1}{D}(\operatorname{vec}\mathbf{I}_D)^\top\left[\lim_{\epsilon\to 0}\epsilon\mathcal{R}_\epsilon^{-\mu}\right]\operatorname{vec}\mathbf{I}_D. \tag{E.46}$$

On the other hand, averaging the left-hand side over μ gives

$$\frac{1}{M(K-1)}(\operatorname{vec}\mathbf{I}_D)^\top\left[\sum_{\mu=1}^{M}\mathbf{Q}_\mu\otimes\mathbf{P}_\mu\right]\mathcal{R}_\epsilon\operatorname{vec}\mathbf{I}_D=\frac{1}{M(K-1)}(\operatorname{vec}\mathbf{I}_D)^\top(\mathbf{I}_{D^2}-\epsilon\mathcal{R}_\epsilon)\operatorname{vec}\mathbf{I}_D\to\frac{\chi}{\varphi}. \tag{E.47}$$

Equating Equations (E.44) and (E.47) gives the following self-consistent equation:

$$\frac{\chi}{\varphi}=\frac{1-\chi}{1-\alpha\varphi}, \tag{E.48}$$

which gives

$$\chi=\frac{\varphi}{1+(1-\alpha)\varphi}. \tag{E.49}$$

The estimator is the orthogonal projection of $-\eta\mathbf{I}_D$ onto the row space of the constraints. Therefore

$$\lim_{D\to\infty}\frac{\|\hat{\mathbf{A}}\|_{\mathrm{F}}^2}{\eta^2 D}=-\lim_{D\to\infty}\frac{\operatorname{tr}\hat{\mathbf{A}}}{\eta D}=\chi=\frac{\varphi}{1+(1-\alpha)\varphi}. \tag{E.50}$$

Equation (E.11) also gives

$$\lim_{D\to\infty}\frac{\|\hat{\mathbf{A}}+\eta\mathbf{I}_D\|_{\mathrm{F}}^2}{\eta^2 D}=\lim_{D\to\infty}\frac{1}{D}(\operatorname{vec}\mathbf{I}_D)^\top\left[\lim_{\epsilon\to 0}\epsilon\mathcal{R}_\epsilon\right]^2\operatorname{vec}\mathbf{I}_D \tag{E.51}$$

$$=\lim_{D\to\infty}\frac{1}{D}(\operatorname{vec}\mathbf{I}_D)^\top\left[\lim_{\epsilon\to 0}\epsilon\mathcal{R}_\epsilon\right]\operatorname{vec}\mathbf{I}_D \tag{E.52}$$

$$=1-\chi=\frac{1-\alpha\varphi}{1+(1-\alpha)\varphi}. \tag{E.53}$$

The second equality uses the projector identity $[\lim_{\epsilon\to 0}\epsilon\mathcal{R}_\epsilon]^2=\lim_{\epsilon\to 0}\epsilon\mathcal{R}_\epsilon$. The third equality uses Equation (E.22).

Calculation of Equation (E.20). Second, we calculate Equation (E.20) by deriving a second self-consistent equation from the same Woodbury identity. Define the order parameters:

$$m_{\mathrm{F}} := \lim_{D\to\infty} \frac{1}{D}\|\hat{\mathbf{A}} + \eta\chi\mathbf{I}_D\|_{\mathrm{F}}^2, \tag{E.54}$$

$$m_{\mathrm{T}} := \lim_{D\to\infty} \frac{1}{D}\operatorname{tr}(\hat{\mathbf{A}} + \eta\chi\mathbf{I}_D)^2. \tag{E.55}$$

By Equations (E.11) and (E.22), the second target satisfies

$$\frac{m_{\mathrm{T}}}{\eta^2} = \lim_{D\to\infty}\left\{\frac{1}{D}(\operatorname{vec}\mathbf{I}_D)^\top\left[\lim_{\epsilon\to 0}\epsilon\mathcal{R}_\epsilon\right]\mathrm{T}_D\left[\lim_{\epsilon\to 0}\epsilon\mathcal{R}_\epsilon\right]\operatorname{vec}\mathbf{I}_D - (1-\chi)^2\right\}. \tag{E.56}$$

The trace and Frobenius identities obtained above give

$$m_{\mathrm{F}} = \lim_{D\to\infty}\left[\frac{\|\hat{\mathbf{A}}\|_{\mathrm{F}}^2}{D} + \frac{2\eta\chi}{D}\operatorname{tr}\hat{\mathbf{A}} + \eta^2\chi^2\right] \tag{E.57}$$

$$=\eta^2\chi - 2\eta^2\chi^2 + \eta^2\chi^2 \tag{E.58}$$

$$=\eta^2(\chi - \chi^2). \tag{E.59}$$

We now start again from the Woodbury identity (Equation (E.26)):

$$(\mathbf{Q}_\mu\otimes\mathbf{P}_\mu)\mathcal{R}_\epsilon\operatorname{vec}\mathbf{I}_D = \left[\epsilon\mathbf{I}_{D^2} + (\mathbf{Q}_\mu\otimes\mathbf{P}_\mu)\epsilon\mathcal{R}_\epsilon^{-\mu}(\mathbf{Q}_\mu\otimes\mathbf{P}_\mu)\right]^{-1}(\mathbf{Q}_\mu\otimes\mathbf{P}_\mu)\epsilon\mathcal{R}_\epsilon^{-\mu}\operatorname{vec}\mathbf{I}_D. \tag{E.60}$$

Equation (E.11) can also be written as

$$\frac{\operatorname{vec}(\hat{\mathbf{A}} + \eta\chi\mathbf{I}_D)}{\eta} =\chi\operatorname{vec}\mathbf{I}_D - \lim_{\epsilon\to 0}\sum_{\mu=1}^{M}(\mathbf{Q}_\mu\otimes\mathbf{P}_\mu)\mathcal{R}_\epsilon\operatorname{vec}\mathbf{I}_D. \tag{E.61}$$

Applying $\eta^2 D^{-1}(\cdot)^\top\mathrm{T}_D(\cdot)$ to both sides of Equation (E.61) gives

$$\lim_{D\to\infty}\frac{1}{D}\operatorname{vec}(\hat{\mathbf{A}} + \eta\chi\mathbf{I}_D)^\top\mathrm{T}_D\operatorname{vec}(\hat{\mathbf{A}} + \eta\chi\mathbf{I}_D) \tag{E.62}$$

$$= \lim_{D\to\infty}\frac{\eta^2}{D}\left[\chi\operatorname{vec}\mathbf{I}_D - \lim_{\epsilon\to 0}\sum_{\mu=1}^{M}(\mathbf{Q}_\mu\otimes\mathbf{P}_\mu)\mathcal{R}_\epsilon\operatorname{vec}\mathbf{I}_D\right]^\top\mathrm{T}_D$$
$$\times\left[\chi\operatorname{vec}\mathbf{I}_D - \lim_{\epsilon\to 0}\sum_{\nu=1}^{M}(\mathbf{Q}_\nu\otimes\mathbf{P}_\nu)\mathcal{R}_\epsilon\operatorname{vec}\mathbf{I}_D\right] \tag{E.63}$$

$$= \lim_{D\to\infty}\frac{\eta^2}{D}\left\{\chi^2(\operatorname{vec}\mathbf{I}_D)^\top\mathrm{T}_D\operatorname{vec}\mathbf{I}_D - 2\chi\lim_{\epsilon\to 0}\sum_{\mu=1}^{M}(\operatorname{vec}\mathbf{I}_D)^\top\mathrm{T}_D(\mathbf{Q}_\mu\otimes\mathbf{P}_\mu)\mathcal{R}_\epsilon\operatorname{vec}\mathbf{I}_D\right.$$
$$\left.+\lim_{\epsilon\to 0}\sum_{\mu,\nu=1}^{M}(\operatorname{vec}\mathbf{I}_D)^\top\mathcal{R}_\epsilon(\mathbf{Q}_\mu\otimes\mathbf{P}_\mu)\mathrm{T}_D(\mathbf{Q}_\nu\otimes\mathbf{P}_\nu)\mathcal{R}_\epsilon\operatorname{vec}\mathbf{I}_D\right\} \tag{E.64}$$

$$= \lim_{D\to\infty}\frac{\eta^2}{D}\left\{\chi^2(\operatorname{vec}\mathbf{I}_D)^\top\mathrm{T}_D\operatorname{vec}\mathbf{I}_D - 2\chi\lim_{\epsilon\to 0}\sum_{\mu=1}^{M}(\operatorname{vec}\mathbf{I}_D)^\top\mathrm{T}_D(\mathbf{Q}_\mu\otimes\mathbf{P}_\mu)\mathcal{R}_\epsilon\operatorname{vec}\mathbf{I}_D\right.$$
$$+\lim_{\epsilon\to 0}\sum_{\mu=\nu}(\operatorname{vec}\mathbf{I}_D)^\top\mathcal{R}_\epsilon(\mathbf{Q}_\mu\otimes\mathbf{P}_\mu)\mathrm{T}_D(\mathbf{Q}_\nu\otimes\mathbf{P}_\nu)\mathcal{R}_\epsilon\operatorname{vec}\mathbf{I}_D$$
$$\left.+\lim_{\epsilon\to 0}\sum_{\mu\neq\nu}(\operatorname{vec}\mathbf{I}_D)^\top\mathcal{R}_\epsilon(\mathbf{Q}_\mu\otimes\mathbf{P}_\mu)\mathrm{T}_D(\mathbf{Q}_\nu\otimes\mathbf{P}_\nu)\mathcal{R}_\epsilon\operatorname{vec}\mathbf{I}_D\right\}. \tag{E.65}$$

The left-hand side of Equation (E.65) is

$$\lim_{D\to\infty}\frac{1}{D}\operatorname{vec}(\hat{\mathbf{A}} + \eta\chi\mathbf{I}_D)^\top\mathrm{T}_D\operatorname{vec}(\hat{\mathbf{A}} + \eta\chi\mathbf{I}_D) \tag{E.66}$$

$$= \lim_{D\to\infty}\frac{1}{D}\operatorname{tr}(\hat{\mathbf{A}} + \eta\chi\mathbf{I}_D)^2 \tag{E.67}$$

$$=m_{\mathrm{T}}. \tag{E.68}$$

For each task contribution on the right-hand side of Equation (E.65), we substitute Equation (E.60). Next, to calculate the right-hand side, we introduce the following proposition.

Proposition E.2 (Moments of the task projectors)**.** *Let* $\mathbf{B} \in \mathbb{R}^{D\times D}$ *be any matrix independent of task* μ*.*

$$\mathbb{E}_\mu[\mathbf{P}_\mu \mathbf{B} \mathbf{P}_\mu] \simeq \alpha^2 \mathbf{B} + \frac{\alpha(1-\alpha)}{D}\left(\mathbf{B}^\top + \mathrm{tr}(\mathbf{B})\mathbf{I}_D\right), \tag{E.69}$$

$$\mathbb{E}_\mu[\mathbf{Q}_\mu \mathbf{B} \mathbf{Q}_\mu] \simeq \frac{(K-1)^2}{D^2}\mathbf{B} + \frac{K-1}{D^2}\left(\mathbf{B}^\top + \mathrm{tr}(\mathbf{B})\mathbf{I}_D\right). \tag{E.70}$$

Proof. Marginally, $\mathbf{P}_\mu$ and $\mathbf{Q}_\mu$ are rotationally invariant orthogonal projectors of ranks L and $K-1$, respectively. The entrywise second moment of either projector has the form

$$\mathbb{E}_\mu[(\mathbf{\Pi}_\mu)_{ij}(\mathbf{\Pi}_\mu)_{kl}] = a_{D,r}\delta_{ij}\delta_{kl} + b_{D,r}(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}), \tag{E.71}$$

where $\mathbf{\Pi}_\mu = \mathbf{P}_\mu$ for $r = L$ and $\mathbf{\Pi}_\mu = \mathbf{Q}_\mu$ for $r = K-1$. Using $(\mathrm{tr}\,\mathbf{\Pi}_\mu)^2 = r^2$ and $\mathrm{tr}(\mathbf{\Pi}_\mu^2) = r$ gives

$$D^2 a_{D,r} + 2D b_{D,r} = r^2, \tag{E.72}$$

$$D a_{D,r} + (D^2 + D) b_{D,r} = r. \tag{E.73}$$

Solving these equations gives

$$a_{D,r} = \frac{r[r(D+1)-2]}{D(D-1)(D+2)}, \tag{E.74}$$

$$b_{D,r} = \frac{r(D-r)}{D(D-1)(D+2)}. \tag{E.75}$$

Therefore,

$$\mathbb{E}_\mu[(\mathbf{\Pi}_\mu \mathbf{B} \mathbf{\Pi}_\mu)_{il}] = \sum_{j,k}(\mathbf{B})_{jk}\mathbb{E}_\mu[(\mathbf{\Pi}_\mu)_{ij}(\mathbf{\Pi}_\mu)_{kl}] \tag{E.76}$$

$$= a_{D,r}\mathbf{B}_{il} + b_{D,r}\mathbf{B}_{li} + b_{D,r}\,\mathrm{tr}(\mathbf{B})\delta_{il}. \tag{E.77}$$

Substituting $r = L = \alpha D$ and $r = K-1$, and expanding the coefficients at their retained large-D orders, gives Equation (E.70). □

After substituting Equation (E.60), we use concentration of the normalized contractions around their conditional task averages. For $\mu \neq \nu$, independence of the two tasks gives

$$\mathbb{E}\,\mathrm{tr}(\mathbf{P}_\mu \mathbf{P}_\nu) = \mathrm{tr}\left(\mathbb{E}\mathbf{P}_\mu \mathbb{E}\mathbf{P}_\nu\right) = \mathrm{tr}\left(\frac{L}{D}\mathbf{I}_D \frac{L}{D}\mathbf{I}_D\right) = \frac{L^2}{D}, \tag{E.78}$$

$$\mathbb{E}\,\mathrm{tr}(\mathbf{Q}_\mu \mathbf{Q}_\nu) = \mathrm{tr}\left(\mathbb{E}\mathbf{Q}_\mu \mathbb{E}\mathbf{Q}_\nu\right) = \mathrm{tr}\left(\frac{K-1}{D}\mathbf{I}_D \frac{K-1}{D}\mathbf{I}_D\right) = \frac{(K-1)^2}{D}. \tag{E.79}$$

We now calculate the two cases separately. For $\mu = \nu$, Equation (E.70) gives

$$\left\{\frac{1}{D^2}\,\mathrm{tr}\left[\lim_{\epsilon\to 0}\epsilon \mathcal{R}_\epsilon^{-\mu}\right]\right\} m_{\mathrm{F}} = \left[1 - \frac{(M-1)L(K-1)}{D^2}\right] m_{\mathrm{F}} \tag{E.80}$$

$$\to (1 - \alpha\varphi) m_{\mathrm{F}}. \tag{E.81}$$

For $\mu \neq \nu$, independence of the two tasks gives

$$\sum_{\nu\neq\mu} \frac{1}{K-1}\mathbb{E}\,\mathrm{tr}(\mathbf{Q}_\mu \mathbf{Q}_\nu)\frac{1}{D}\mathbb{E}\,\mathrm{tr}(\mathbf{P}_\mu \mathbf{P}_\nu) m_{\mathrm{T}} \tag{E.82}$$

$$= (M-1)\frac{1}{K-1}\frac{(K-1)^2}{D}\frac{1}{D}\frac{L^2}{D} m_{\mathrm{T}} \tag{E.83}$$

$$= (M-1)\frac{K-1}{D}\frac{L^2}{D^2} m_{\mathrm{T}} \tag{E.84}$$

$$\to \alpha^2 \varphi m_{\mathrm{T}}. \tag{E.85}$$

Equating the two sides of Equation (E.65) gives the self-consistent equation

$$m_{\mathrm{T}} = (1-\alpha\varphi)m_{\mathrm{F}} + \alpha^2\varphi m_{\mathrm{T}}, \tag{E.86}$$

which gives

$$\frac{m_{\mathrm{T}}}{\eta^2} = \frac{\varphi(1-\alpha\varphi)^2}{[1+(1-\alpha)\varphi]^2(1-\alpha^2\varphi)}. \tag{E.87}$$

Since the centered matrix has zero normalized trace,

$$\lim_{D\to\infty} \frac{\mathrm{tr}(\hat{\mathbf{A}}^2)}{\eta^2 D} = \chi^2 + \frac{m_{\mathrm{T}}}{\eta^2} = \frac{\varphi[1+(1-2\alpha)\varphi]}{[1+(1-\alpha)\varphi]^2(1-\alpha^2\varphi)}. \tag{E.88}$$

This proves Equation (E.3). □

E.2 Effective expression

Proposition E.1 determines the normalized trace, Frobenius norm, and transpose moment of $\hat{\mathbf{A}}$. To simplify the remaining calculation, we replace $\hat{\mathbf{A}}$ by an effective random matrix that reproduces these three quantities. It is motivated by the fact that, after conditioning on the state projectors, the centered part of $\hat{\mathbf{A}}$ is a sum of $M = O(D)$ independent task contributions. We therefore model this centered part by a Gaussian matrix. The coefficients below are chosen so that this construction reproduces the three moments in Proposition E.1. We do not prove here that it preserves the full spectrum or every normalized resolvent.

Let

$$r = \mathrm{rank}\left(\sum_{\mu=1}^{M} \mathbf{Q}_\mu\right) = \min\{D, M(K-1)\}. \tag{E.89}$$

Define $\mathbf{O}_\parallel \in \mathbb{R}^{D\times r}$ as the matrix whose columns are orthonormal eigenvectors of $\sum_{\mu=1}^{M}\mathbf{Q}_\mu$ corresponding to its nonzero eigenvalues. Define $\mathbf{O}_\perp \in \mathbb{R}^{D\times(D-r)}$ as an orthonormal basis of the null space of $\sum_{\mu=1}^{M}\mathbf{Q}_\mu$. Then we have

$$\sum_{\mu=1}^{M}\mathbf{Q}_\mu = \mathbf{O}_\parallel \,\mathrm{diag}(x_1,\dots,x_r)\mathbf{O}_\parallel^\top, \qquad x_1,\dots,x_r>0, \qquad (\mathbf{O}_\parallel,\mathbf{O}_\perp)^\top(\mathbf{O}_\parallel,\mathbf{O}_\perp) = \mathbf{I}_D. \tag{E.90}$$

Proposition E.3 (Moment-matched effective expression)**.** *Suppose* $0<\alpha<1$, $0<\varphi<1/\alpha$, *and* $\lambda\to 0$. *Define* $\hat{\mathbf{A}}_{\mathrm{eff}}$ *by*

$$\frac{\hat{\mathbf{A}}_{\mathrm{eff}}\mathbf{O}_\parallel}{\eta} = -\mathbf{O}_\parallel \,\mathrm{diag}_j \frac{x_j}{u_\varphi + v_\varphi x_j} + \frac{\mathbf{W}}{\sqrt{D}}\,\mathrm{diag}_j \frac{\sqrt{c_\varphi x_j}}{u_\varphi + v_\varphi x_j}, \tag{E.91}$$

$$\hat{\mathbf{A}}_{\mathrm{eff}}\mathbf{O}_\perp = \mathbf{0}, \tag{E.92}$$

where $\mathbf{W}\in\mathbb{R}^{D\times r}$ *has independent* $N(0,1)$ *entries, and*

$$u_\varphi = (1-\alpha)(1-\alpha\varphi)[1+(1-\alpha)\varphi], \tag{E.93}$$

$$v_\varphi = \alpha[1+(1-\alpha)\varphi], \tag{E.94}$$

$$c_\varphi = \alpha(1-\alpha)\varphi(1-\alpha\varphi). \tag{E.95}$$

Then

$$-\lim_{D\to\infty}\frac{\mathrm{tr}\,\hat{\mathbf{A}}_{\mathrm{eff}}}{\eta D} = \lim_{D\to\infty}\frac{\|\hat{\mathbf{A}}_{\mathrm{eff}}\|_{\mathrm{F}}^2}{\eta^2 D} = \frac{\varphi}{1+(1-\alpha)\varphi}, \tag{E.96}$$

$$\lim_{D\to\infty}\frac{\mathrm{tr}(\hat{\mathbf{A}}_{\mathrm{eff}}^2)}{\eta^2 D} = \frac{\varphi[1+(1-2\alpha)\varphi]}{[1+(1-\alpha)\varphi]^2(1-\alpha^2\varphi)}. \tag{E.97}$$

Proof. For fixed K, the nonzero eigenvalues x_j of $\sum_\mu \mathbf{Q}_\mu$ obey the Marchenko–Pastur law $\nu_{1/\varphi}$ defined in Equation (4.5). Thus, for $f(0) = 0$,

$$\frac{1}{D}\sum_{j=1}^{r} f(x_j) \longrightarrow \varphi\langle f\rangle_{1/\varphi}. \tag{E.98}$$

For $\varphi > 1$, $\nu_{1/\varphi}$ has a zero atom of mass $1 - 1/\varphi$.

Since $\hat{\mathbf{A}}_{\text{eff}}\mathbf{O}_\perp = \mathbf{0}$, Equation (E.92) gives

$$\frac{\hat{\mathbf{A}}_{\text{eff}}}{\eta} = -\,\mathbf{O}_\parallel \operatorname{diag}_j \frac{x_j}{u_\varphi + v_\varphi x_j}\mathbf{O}_\parallel^\top + \frac{\mathbf{W}}{\sqrt{D}} \operatorname{diag}_j \frac{\sqrt{c_\varphi x_j}}{u_\varphi + v_\varphi x_j}\mathbf{O}_\parallel^\top. \tag{E.99}$$

Therefore, the normalized trace satisfies

$$-\frac{\operatorname{tr}\hat{\mathbf{A}}_{\text{eff}}}{\eta D} = \frac{1}{D}\sum_{j=1}^{r}\frac{x_j}{u_\varphi + v_\varphi x_j} - \frac{1}{D\sqrt{D}}\operatorname{tr}\left[\mathbf{W}\operatorname{diag}_j \frac{\sqrt{c_\varphi x_j}}{u_\varphi + v_\varphi x_j}\mathbf{O}_\parallel^\top\right] \tag{E.100}$$

$$\to \varphi\left\langle \frac{x}{u_\varphi + v_\varphi x}\right\rangle_{1/\varphi}. \tag{E.101}$$

Expanding the trace of the square gives

$$\frac{\operatorname{tr}(\hat{\mathbf{A}}_{\text{eff}}^2)}{\eta^2 D} = \frac{1}{D}\operatorname{tr}\left[\left(-\mathbf{O}_\parallel \operatorname{diag}_j \frac{x_j}{u_\varphi + v_\varphi x_j}\mathbf{O}_\parallel^\top + \frac{\mathbf{W}}{\sqrt{D}}\operatorname{diag}_j \frac{\sqrt{c_\varphi x_j}}{u_\varphi + v_\varphi x_j}\mathbf{O}_\parallel^\top\right)^2\right]$$

$$= \frac{1}{D}\sum_{j=1}^{r}\frac{x_j^2}{(u_\varphi + v_\varphi x_j)^2} - \frac{2}{D\sqrt{D}}\operatorname{tr}\left[\mathbf{O}_\parallel \operatorname{diag}_j \frac{x_j}{u_\varphi + v_\varphi x_j}\mathbf{O}_\parallel^\top \mathbf{W}\operatorname{diag}_j \frac{\sqrt{c_\varphi x_j}}{u_\varphi + v_\varphi x_j}\mathbf{O}_\parallel^\top\right] \tag{E.102}$$

$$+ \frac{1}{D^2}\operatorname{tr}\left[\mathbf{W}\operatorname{diag}_j \frac{\sqrt{c_\varphi x_j}}{u_\varphi + v_\varphi x_j}\mathbf{O}_\parallel^\top \mathbf{W}\operatorname{diag}_j \frac{\sqrt{c_\varphi x_j}}{u_\varphi + v_\varphi x_j}\mathbf{O}_\parallel^\top\right] \tag{E.103}$$

$$= \frac{1}{D}\sum_{j=1}^{r}\frac{x_j^2}{(u_\varphi + v_\varphi x_j)^2} + o_{\mathbb{P}}(1)$$

$$\to \varphi\left\langle \frac{x^2}{(u_\varphi + v_\varphi x)^2}\right\rangle_{1/\varphi}. \tag{E.104}$$

Similarly, expanding the Frobenius norm gives

$$\frac{\|\hat{\mathbf{A}}_{\text{eff}}\|_{\text{F}}^2}{\eta^2 D} = \frac{1}{D}\operatorname{tr}\left[\begin{array}{l}\left(-\mathbf{O}_\parallel \operatorname{diag}_j \frac{x_j}{u_\varphi + v_\varphi x_j}\mathbf{O}_\parallel^\top + \frac{\mathbf{W}}{\sqrt{D}}\operatorname{diag}_j \frac{\sqrt{c_\varphi x_j}}{u_\varphi + v_\varphi x_j}\mathbf{O}_\parallel^\top\right)^\top \\ \quad \times \left(-\mathbf{O}_\parallel \operatorname{diag}_j \frac{x_j}{u_\varphi + v_\varphi x_j}\mathbf{O}_\parallel^\top + \frac{\mathbf{W}}{\sqrt{D}}\operatorname{diag}_j \frac{\sqrt{c_\varphi x_j}}{u_\varphi + v_\varphi x_j}\mathbf{O}_\parallel^\top\right)\end{array}\right]$$

$$= \frac{1}{D}\sum_{j=1}^{r}\frac{x_j^2}{(u_\varphi + v_\varphi x_j)^2} - \frac{2}{D\sqrt{D}}\operatorname{tr}\left[\operatorname{diag}_j \frac{x_j}{u_\varphi + v_\varphi x_j}\mathbf{O}_\parallel^\top \mathbf{W}\operatorname{diag}_j \frac{\sqrt{c_\varphi x_j}}{u_\varphi + v_\varphi x_j}\right] \tag{E.105}$$

$$+ \frac{1}{D^2}\operatorname{tr}\left[\operatorname{diag}_j \frac{\sqrt{c_\varphi x_j}}{u_\varphi + v_\varphi x_j}\mathbf{W}^\top \mathbf{W}\operatorname{diag}_j \frac{\sqrt{c_\varphi x_j}}{u_\varphi + v_\varphi x_j}\right] \tag{E.106}$$

$$= \frac{1}{D}\sum_{j=1}^{r}\frac{x_j^2}{(u_\varphi + v_\varphi x_j)^2} + \frac{1}{D}\sum_{j=1}^{r}\frac{c_\varphi x_j}{(u_\varphi + v_\varphi x_j)^2} + o_{\mathbb{P}}(1)$$

$$= \frac{1}{D}\sum_{j=1}^{r}\frac{x_j^2 + c_\varphi x_j}{(u_\varphi + v_\varphi x_j)^2} + o_{\mathbb{P}}(1)$$

$$\to \varphi\left\langle \frac{x^2 + c_\varphi x}{(u_\varphi + v_\varphi x)^2}\right\rangle_{1/\varphi}. \tag{E.107}$$

Here we used $\mathbf{O}_\parallel^\top \mathbf{O}_\parallel = \mathbf{I}_r$ and the cyclicity of the trace to combine the two cross terms. Conditional on $(x_j, \mathbf{O}_\parallel)$, the entries of $\mathbf{O}_\parallel^\top \mathbf{W}$ are independent standard Gaussians. Consequently, the cross terms in both displays and the final Gaussian term in the trace-of-the-square display are $o_{\mathbb{P}}(1)$. In the Frobenius display, $D^{-1}\mathbf{W}^\top \mathbf{W}$ instead contributes $\mathbf{I}_r$ inside the normalized trace. Equation (E.98) then gives the two displayed limits.

The Marchenko–Pastur equation gives

$$\left\langle \frac{1}{(1-\alpha)(1-\alpha\varphi)+\alpha x} \right\rangle_{1/\varphi} = \frac{1}{1-\alpha\varphi}. \tag{E.108}$$

Consequently,

$$\left\langle \frac{x}{u_\varphi + v_\varphi x} \right\rangle_{1/\varphi} = \frac{1}{v_\varphi}\left[1 - \frac{u_\varphi}{v_\varphi}\left\langle \frac{1}{u_\varphi/v_\varphi + x} \right\rangle_{1/\varphi}\right] \tag{E.109}$$

$$= \frac{1}{v_\varphi}[1-(1-\alpha)] \tag{E.110}$$

$$= \frac{\alpha}{v_\varphi} \tag{E.111}$$

$$= \frac{1}{1+(1-\alpha)\varphi}. \tag{E.112}$$

The MP equation and its derivative then give

$$\left\langle \frac{x^2}{(u_\varphi + v_\varphi x)^2} \right\rangle_{1/\varphi} = \frac{1+(1-2\alpha)\varphi}{[1+(1-\alpha)\varphi]^2(1-\alpha^2\varphi)}, \tag{E.113}$$

$$\left\langle \frac{x}{(u_\varphi + v_\varphi x)^2} \right\rangle_{1/\varphi} = \frac{1}{[1+(1-\alpha)\varphi]^2(1-\alpha^2\varphi)}. \tag{E.114}$$

Substitution into Equation (E.101) gives

$$-\lim_{D\to\infty} \frac{\operatorname{tr}\hat{\mathbf{A}}_{\text{eff}}}{\eta D} = \frac{\varphi}{1+(1-\alpha)\varphi}. \tag{E.115}$$

Substitution into Equation (E.104) gives

$$\lim_{D\to\infty} \frac{\operatorname{tr}(\hat{\mathbf{A}}_{\text{eff}}^2)}{\eta^2 D} = \frac{\varphi[1+(1-2\alpha)\varphi]}{[1+(1-\alpha)\varphi]^2(1-\alpha^2\varphi)}. \tag{E.116}$$

Finally, substitution into Equation (E.107) gives

$$\lim_{D\to\infty} \frac{\|\hat{\mathbf{A}}_{\text{eff}}\|_{\text{F}}^2}{\eta^2 D} = \frac{\varphi[1+(1-2\alpha)\varphi + c_\varphi]}{[1+(1-\alpha)\varphi]^2(1-\alpha^2\varphi)} \tag{E.117}$$

$$= \frac{\varphi[1+(1-\alpha)\varphi](1-\alpha^2\varphi)}{[1+(1-\alpha)\varphi]^2(1-\alpha^2\varphi)} \tag{E.118}$$

$$= \frac{\varphi}{1+(1-\alpha)\varphi}. \tag{E.119}$$

Hence the three moments agree with Proposition E.1. □

Proposition E.4 (GD recovery for rank-deficient contexts)**.** *Suppose* $0 < \alpha < 1$, $\alpha\varphi > 1$, *and* $\lambda \to 0$. *Under the same cavity approximation as in Proposition E.1, the effective learned matrix satisfies*

$$\hat{\mathbf{A}}_{\text{eff}} = -\eta \mathbf{I}_D. \tag{E.120}$$

Consequently,

$$\mathcal{E}(t) = \mathcal{E}_{\text{GD}}(t). \tag{E.121}$$

Proof. Let

$$\mathbf{Z}_0 := \lim_{\epsilon \to 0} \epsilon \mathcal{R}_\epsilon, \tag{E.122}$$

where $\mathcal{R}_\epsilon$ is defined in Equation (E.9). Equation (E.11) gives

$$\operatorname{vec}(\hat{\mathbf{A}}_{\text{eff}} + \eta \mathbf{I}_D) = \eta \mathbf{Z}_0 \operatorname{vec} \mathbf{I}_D. \tag{E.123}$$

Keeping $\epsilon > 0$ in the leave-one-task cavity calculation gives the scalar resolvent $\rho_\epsilon := D^{-2} \operatorname{tr} \mathcal{R}_\epsilon$ satisfies

$$\epsilon \rho_\epsilon^2 - (1 - \alpha\varphi)\rho_\epsilon - a_{K,\alpha} = o(1), \qquad a_{K,\alpha} > 0. \tag{E.124}$$

Its positive solution is

$$\rho_\epsilon = \frac{1 - \alpha\varphi + \sqrt{(1-\alpha\varphi)^2 + 4\epsilon a_{K,\alpha}}}{2\epsilon} + o(1). \tag{E.125}$$

When $\alpha\varphi > 1$, the two terms in the numerator cancel at leading order and $\rho_\epsilon \to a_{K,\alpha}/(\alpha\varphi - 1)$. Therefore

$$\lim_{D\to\infty} \frac{1}{D^2} \operatorname{tr} \mathbf{Z}_0 = \lim_{\epsilon \to 0} \epsilon \rho_\epsilon = 0. \tag{E.126}$$

Since $\mathbf{Z}_0$ is the orthogonal projector onto the kernel of $\sum_\mu \mathbf{Q}_\mu \otimes \mathbf{P}_\mu$, this gives $\mathbf{Z}_0 = \mathbf{0}$ in the effective large-D description. Equation (E.123) then gives Equation (E.120). Proposition C.3 completes the proof. □

For the remainder of the calculation, we make the replacement

$$\hat{\mathbf{A}} \longrightarrow \hat{\mathbf{A}}_{\text{eff}}. \tag{E.127}$$

To avoid additional notation, we write $\hat{\mathbf{A}}$ for $\hat{\mathbf{A}}_{\text{eff}}$ after this point.

E.3 Resolvent calculation

We now use Equation (E.92) to calculate the two resolvents needed to determine φ_- and φ_+. As a preparation, we calculate two scalar functions below:

$$r_\varphi(y) := \lim_{D\to\infty} \frac{1}{D} \operatorname{tr}(\mathbf{I}_D - y\hat{\mathbf{A}}/\eta)^{-1}, \tag{E.128}$$

$$G_\varphi(y) := \lim_{D\to\infty} \frac{1}{D} \|(\mathbf{I}_D - y\hat{\mathbf{A}}/\eta)^{-1}\|_{\text{F}}^2. \tag{E.129}$$

These resolvent quantities are given by the following proposition.

Proposition E.5 (Resolvents of the effective matrix)**.** *For sufficiently small y,*

$$r_\varphi(y) = 1 - \varphi + \varphi \left\langle \frac{u_\varphi + v_\varphi x}{u_\varphi + (v_\varphi + y)x} \right\rangle_{1/\varphi}, \tag{E.130}$$

$$G_\varphi(y) = \frac{1 - \varphi + \varphi \left\langle (u_\varphi + v_\varphi x)^2/[u_\varphi + (v_\varphi + y)x]^2 \right\rangle_{1/\varphi}}{1 - \varphi y^2 c_\varphi \left\langle x/[u_\varphi + (v_\varphi + y)x]^2 \right\rangle_{1/\varphi}}. \tag{E.131}$$

Proof.

Calculation of $r_\varphi(y)$. The effective matrix vanishes on the subspace spanned by $\mathbf{O}_\perp$, so the inverse is the identity on this subspace. It remains to calculate the inverse on the subspace spanned by $\mathbf{O}_\parallel$. For sufficiently small y, factoring the deterministic diagonal matrix in Equation (E.92) gives the convergent expansion

$$\begin{aligned} &\mathbf{O}_\parallel^\top (\mathbf{I}_D - y\hat{\mathbf{A}}/\eta)^{-1} \mathbf{O}_\parallel \\ &= \operatorname{diag}_j \left(\frac{u_\varphi + v_\varphi x_j}{u_\varphi + (v_\varphi + y)x_j} \right) \sum_{n=0}^{\infty} \left[\frac{y}{\sqrt{D}} \mathbf{O}_\parallel^\top \mathbf{W} \operatorname{diag}_j \left(\frac{\sqrt{c_\varphi x_j}}{u_\varphi + (v_\varphi + y)x_j} \right) \right]^n. \end{aligned} \tag{E.132}$$

For each fixed $n \geq 1$, every term with the maximal number of free index sums contains a Gaussian column only once after all other columns have been fixed, so its conditional mean is zero. Terms in which every Gaussian column is repeated have one fewer free index sum and vanish after division by D. Consequently,

$$\lim_{D\to\infty} \frac{1}{D}\mathbb{E}\,\mathrm{tr}\left\{\mathrm{diag}_j\left(\frac{u_\varphi + v_\varphi x_j}{u_\varphi + (v_\varphi + y)x_j}\right)\left[\frac{y}{\sqrt{D}}\mathbf{O}_\parallel^\top \mathbf{W}\,\mathrm{diag}_j\left(\frac{\sqrt{c_\varphi x_j}}{u_\varphi + (v_\varphi + y)x_j}\right)\right]^n\right\}$$
$$=0, \qquad n \geq 1. \tag{E.133}$$

Thus only the $n = 0$ term remains in the normalized trace. The inactive state directions contribute an identity block, so

$$r_\varphi(y) = (1-\varphi)_+ + \lim_{D\to\infty}\frac{1}{D}\sum_{j=1}^{r}\frac{u_\varphi + v_\varphi x_j}{u_\varphi + (v_\varphi + y)x_j} \tag{E.134}$$

$$= 1 - \varphi + \varphi\left\langle \frac{u_\varphi + v_\varphi x}{u_\varphi + (v_\varphi + y)x}\right\rangle_{1/\varphi}. \tag{E.135}$$

For $\varphi > 1$, the zero atom contributes $\varphi - 1$, cancelling the negative term $1 - \varphi$. Thus the same formula also holds for $\varphi > 1$.

Calculation of $G_\varphi(y)$. The squared Frobenius norm contains products of two Gaussian factors, so these terms do not vanish. We calculate them by the cavity method. Define

$$\mathbf{R}_y := (\mathbf{I}_D - y\hat{\mathbf{A}}/\eta)^{-1}. \tag{E.136}$$

Let $\boldsymbol{w}_j$ and $\boldsymbol{o}_j$ be the j-th columns of $\mathbf{W}$ and $\mathbf{O}_\parallel$, respectively. The cavity step removes $\boldsymbol{w}_j$ from $\mathbf{W}$ and defines $\mathbf{R}_y^{(j)}$ as the resulting inverse. Then $\mathbf{R}_y^{(j)}$ is independent of $\boldsymbol{w}_j$, and

$$\mathbf{R}_y^{-1} = (\mathbf{R}_y^{(j)})^{-1} - \frac{y}{\sqrt{D}}\frac{\sqrt{c_\varphi x_j}}{u_\varphi + v_\varphi x_j}\boldsymbol{w}_j\boldsymbol{o}_j^\top. \tag{E.137}$$

Restoring the removed column with the Woodbury formula gives

$$\mathbf{R}_y = \mathbf{R}_y^{(j)} + \frac{\frac{y}{\sqrt{D}}\frac{\sqrt{c_\varphi x_j}}{u_\varphi + v_\varphi x_j}\mathbf{R}_y^{(j)}\boldsymbol{w}_j\boldsymbol{o}_j^\top\mathbf{R}_y^{(j)}}{1 - \frac{y}{\sqrt{D}}\frac{\sqrt{c_\varphi x_j}}{u_\varphi + v_\varphi x_j}\boldsymbol{o}_j^\top\mathbf{R}_y^{(j)}\boldsymbol{w}_j}. \tag{E.138}$$

For sufficiently small y, the norm of $\mathbf{R}_y^{(j)}$ is bounded, so $\boldsymbol{o}_j^\top\mathbf{R}_y^{(j)}\boldsymbol{w}_j = O_\mathbb{P}(1)$. Therefore the denominator is $1 + O_\mathbb{P}(D^{-1/2})$ and converges to one. Because the cavity matrix is independent of $\boldsymbol{w}_j$, its conditional average satisfies

$$\frac{1}{D}\mathbb{E}\left[\boldsymbol{w}_j^\top\mathbf{R}_y^{(j)}\mathbf{R}_y^{(j)\top}\boldsymbol{w}_j \,\middle|\, \mathbf{R}_y^{(j)}\right] = \frac{1}{D}\,\mathrm{tr}(\mathbf{R}_y^{(j)}\mathbf{R}_y^{(j)\top}) \longrightarrow G_\varphi(y). \tag{E.139}$$

Substituting Equation (E.138) into $D^{-1}\,\mathrm{tr}(\mathbf{R}_y\mathbf{R}_y^\top)$, taking the conditional average, and summing over j gives the self-consistent equation

$$G_\varphi(y) = 1 - \varphi + \varphi\left\langle\frac{(u_\varphi + v_\varphi x)^2}{[u_\varphi + (v_\varphi + y)x]^2}\right\rangle_{1/\varphi}$$
$$+ \varphi y^2 c_\varphi\left\langle\frac{x}{[u_\varphi + (v_\varphi + y)x]^2}\right\rangle_{1/\varphi} G_\varphi(y) \tag{E.140}$$

$$= \frac{1 - \varphi + \varphi\left\langle(u_\varphi + v_\varphi x)^2/[u_\varphi + (v_\varphi + y)x]^2\right\rangle_{1/\varphi}}{1 - \varphi y^2 c_\varphi\left\langle x/[u_\varphi + (v_\varphi + y)x]^2\right\rangle_{1/\varphi}}. \tag{E.141}$$

The first equality is the self-consistent equation for $G_\varphi(y)$, and the second equality solves it. This proves the second formula in Proposition E.5. □

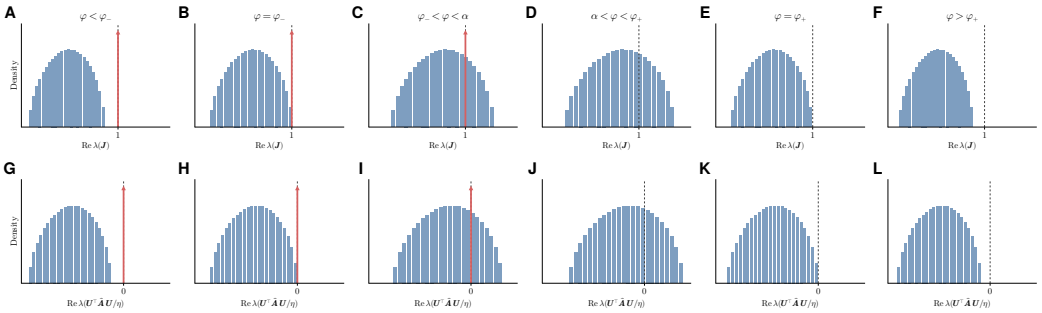


Figure 3: **Schematic spectral evolution across the long-time phases.** (A–F) Densities of the real parts of the eigenvalues of $\mathbf{I}_D + \mathbf{H}^{\text{in}}\hat{\mathbf{A}}$. (G–L) Densities of the real parts of the eigenvalues of $\mathbf{A}_c/\eta$. Red arrows in (A–C) and (G–I) mark the rank-deficient atoms at eigenvalues 1 and 0, respectively.

E.4 The spectrum conditions

In this subsection, we first express the long-time behavior in terms of the spectral radius of the inference matrix. Define $\boldsymbol{e}_t := \hat{\boldsymbol{w}}_t - \boldsymbol{w}^{\text{in}}$. For a fixed test context, Equation (3.13) gives the inference recursion

$$\boldsymbol{e}_{t+1} = \mathbf{J}\boldsymbol{e}_t + \mathbf{H}^{\text{in}}(\hat{\mathbf{A}} + \eta\mathbf{I}_D)\boldsymbol{w}^{\text{in}} + \frac{\eta}{L}\mathbf{X}^{\text{in}}\boldsymbol{\varepsilon}^{\text{in}}, \tag{E.142}$$

where

$$\mathbf{J} := \mathbf{I}_D + \mathbf{H}^{\text{in}}\hat{\mathbf{A}}. \tag{E.143}$$

Therefore,

$$\boldsymbol{e}_t = \mathbf{J}^t\boldsymbol{e}_0 + \sum_{s=0}^{t-1}\mathbf{J}^s\left[\mathbf{H}^{\text{in}}(\hat{\mathbf{A}} + \eta\mathbf{I}_D)\boldsymbol{w}^{\text{in}} + \frac{\eta}{L}\mathbf{X}^{\text{in}}\boldsymbol{\varepsilon}^{\text{in}}\right]. \tag{E.144}$$

The long-time behavior is consequently determined by

$$\rho(\mathbf{J}) := \max_{\lambda\in\text{spec}(\mathbf{J})}|\lambda|. \tag{E.145}$$

More precisely,

$$\begin{cases} \rho(\mathbf{J}) < 1 & \Longrightarrow \quad \mathbf{J}^t \to \mathbf{0} \text{ and the error remains bounded,} \\ \rho(\mathbf{J}) > 1 & \Longrightarrow \quad \mathbf{J}^t \text{ and the error grow exponentially,} \\ \rho(\mathbf{J}) = 1 & \Longrightarrow \quad \text{the limiting case.} \end{cases} \tag{E.146}$$

When $\rho(\mathbf{J}) = 1$, this gives

$$\sum_{s=0}^{t-1}\mathbf{J}^s\left[\mathbf{H}^{\text{in}}(\hat{\mathbf{A}} + \eta\mathbf{I}_D)\boldsymbol{w}^{\text{in}} + \frac{\eta}{L}\mathbf{X}^{\text{in}}\boldsymbol{\varepsilon}^{\text{in}}\right] = O(t), \tag{E.147}$$

so the state grows linearly and its squared error grows as $\Theta(t^2)$.

We next rewrite the conditions under which an eigenvalue of $\mathbf{J}$ reaches 1 as conditions on $\mathbf{A}_c$. Define $\mathbf{U}$ and $\mathbf{A}_c$ by

$$\mathbf{U} := \mathbf{X}^{\text{in}}[(\mathbf{X}^{\text{in}})^\top\mathbf{X}^{\text{in}}]^{-1/2}, \qquad \mathbf{A}_c := \mathbf{U}^\top\hat{\mathbf{A}}\mathbf{U}. \tag{E.148}$$

The top row of Figure 3 shows the corresponding changes of the spectrum of $\mathbf{J}$. In Figure 3A and B, $\rho(\mathbf{J}) = 1$. In Figure 3C and D, an eigenvalue lies beyond 1, and hence $\rho(\mathbf{J}) > 1$. In Figure 3E, the edge returns to 1, while Figure 3F has $\rho(\mathbf{J}) < 1$. The bottom row of Figure 3 shows the eigenvalues of $\mathbf{A}_c/\eta$. Since $\mathbf{U}^\top\mathbf{U} = \mathbf{I}_L$ and $\mathbf{H}^{\text{in}} = \mathbf{X}^{\text{in}}(\mathbf{X}^{\text{in}})^\top/L$, we have

$$\mathbf{H}^{\text{in}} = \mathbf{U}\frac{(\mathbf{X}^{\text{in}})^\top\mathbf{X}^{\text{in}}}{L}\mathbf{U}^\top, \tag{E.149}$$

$$\text{spec}(\mathbf{H}^{\text{in}}\hat{\mathbf{A}}) \setminus \{0\} = \text{spec}\left(\frac{(\mathbf{X}^{\text{in}})^\top\mathbf{X}^{\text{in}}}{L}\mathbf{A}_c\right) \setminus \{0\}. \tag{E.150}$$

The matrix $(\mathbf{X}^{\mathrm{in}})^\top \mathbf{X}^{\mathrm{in}}/L$ is positive definite for $0 < \alpha < 1$. Consequently,

$$\begin{aligned} &\text{an additional eigenvalue of } \mathbf{A}_c \text{ reaches } 0 \\ \iff &\text{an additional eigenvalue of } \mathbf{H}^{\mathrm{in}}\hat{\mathbf{A}} \text{ reaches } 0 \\ \iff &\text{an additional eigenvalue of } \mathbf{I}_D + \mathbf{H}^{\mathrm{in}}\hat{\mathbf{A}} \text{ reaches } 1. \end{aligned} \tag{E.151}$$

Let

$$r_D := \mathrm{rank}(\hat{\mathbf{A}}) = \min\{D, M(K-1)\}. \tag{E.152}$$

If the singular values are ordered as

$$\sigma_1(\mathbf{A}_c) \geq \cdots \geq \sigma_L(\mathbf{A}_c) \geq 0, \tag{E.153}$$

we write

$$\sigma_{\min}^{\mathrm{nonzero}}(\mathbf{A}_c) := \sigma_{\min\{L, r_D\}}(\mathbf{A}_c). \tag{E.154}$$

For $0 < \varphi < \alpha$, the red arrows in Figure 3A–C and G–I show the $(\alpha-\varphi)D + o(D)$ eigenvalues fixed at 1 and 0, respectively. These fixed eigenvalues do not define the lower boundary. After excluding them, the lower-boundary condition is

$$\lim_{D\to\infty} \sigma_{\min}^{\mathrm{nonzero}}(\mathbf{A}_c) = 0. \tag{E.155}$$

This is shown in Figure 3B and H, where the edge of the remaining eigenvalues reaches 1 in the top row and 0 in the bottom row. For $\alpha < \varphi < 1/\alpha$, there are no fixed zero eigenvalues. The upper-boundary condition is therefore

$$\lim_{D\to\infty} \sigma_L(\mathbf{A}_c) = 0. \tag{E.156}$$

This is shown in Figure 3E and K, where the edge reaches 1 in the top row and 0 in the bottom row.

E.5 Lower boundary

In this subsection, we determine the lower boundary $\varphi_-(\alpha)$. In this regime, $\mathbf{A}_c$ already has $(\alpha - \varphi)D + o(D)$ zero eigenvalues because of its rank deficit. Therefore, $\varphi_-(\alpha)$ is the value at which an additional eigenvalue reaches zero.

Equation (E.92) gives the effective matrix, and Equations (E.130) and (E.131) give the two resolvents. Next, we determine the lower boundary while retaining the Gaussian test covariance.

Proposition E.6 (Lower boundary)**.** *Suppose $0 < \alpha < 1$. There is a unique $\varphi_-(\alpha) \in (0, \alpha)$ such that*

$$\lim_{D\to\infty} \sigma_{\min}^{\mathrm{nonzero}}(\mathbf{A}_c) = 0 \quad \iff \quad \varphi = \varphi_-(\alpha), \qquad 0 < \varphi < \alpha. \tag{E.157}$$

It satisfies

$$(\alpha - \varphi_-)(1 - \varphi_-) = \alpha(1-\alpha)\varphi_-^2(1 - \alpha\varphi_-)(\alpha + 1 - 2\varphi_-), \tag{E.158}$$

Proof. We derive Equation (E.158) for $0 < \varphi < \alpha$ and then show that it has a unique solution. The matrix $\mathbf{H}^{\mathrm{in}}\hat{\mathbf{A}}$ has $(\alpha - \varphi)D + o(D)$ zero eigenvalues in this regime, so replacing $\mathbf{H}^{\mathrm{in}}$ by an orthogonal projector would change its remaining eigenvalues. In the orthogonal basis $(\mathbf{O}_\parallel, \mathbf{O}_\perp)$ defined above, define $\mathbf{X}_1$ and $\mathbf{X}_2$ by

$$\begin{pmatrix} \mathbf{O}_\parallel^\top \\ \mathbf{O}_\perp^\top \end{pmatrix} \mathbf{X}^{\mathrm{in}} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix}. \tag{E.159}$$

Equation (E.92) gives

$$\frac{1}{\eta} \begin{pmatrix} \mathbf{O}_\parallel^\top \\ \mathbf{O}_\perp^\top \end{pmatrix} \hat{\mathbf{A}} \begin{pmatrix} \mathbf{O}_\parallel & \mathbf{O}_\perp \end{pmatrix} = \begin{pmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{Z} & \mathbf{0} \end{pmatrix}. \tag{E.160}$$

Here $\mathbf{X}_1 \in \mathbb{R}^{r\times L}$, $\mathbf{X}_2 \in \mathbb{R}^{(D-r)\times L}$, $\mathbf{C} \in \mathbb{R}^{r\times r}$, and $\mathbf{Z} \in \mathbb{R}^{(D-r)\times r}$, where $r/D \to \varphi < \alpha$. The matrix $\mathbf{C}$ is invertible almost surely, so define

$$\mathbf{R} := \mathbf{Z}\mathbf{C}^{-1}. \tag{E.161}$$

Since $r < L$, the matrix

$$\mathbf{B} := \frac{\mathbf{X}_1\mathbf{X}_1^\top}{L} \tag{E.162}$$

is also invertible almost surely. Direct multiplication gives

$$\frac{1}{\eta}\begin{pmatrix}\mathbf{O}_\parallel^\top\\ \mathbf{O}_\perp^\top\end{pmatrix}\mathbf{H}^{\text{in}}\hat{\mathbf{A}}\begin{pmatrix}\mathbf{O}_\parallel & \mathbf{O}_\perp\end{pmatrix} = \frac{1}{L}\begin{pmatrix}\mathbf{X}_1\mathbf{X}_1^\top & \mathbf{X}_1\mathbf{X}_2^\top\\ \mathbf{X}_2\mathbf{X}_1^\top & \mathbf{X}_2\mathbf{X}_2^\top\end{pmatrix}\begin{pmatrix}\mathbf{C} & \mathbf{0}\\ \mathbf{Z} & \mathbf{0}\end{pmatrix} \tag{E.163}$$

$$= \frac{1}{L}\begin{pmatrix}\mathbf{X}_1\mathbf{X}_1^\top\mathbf{C} + \mathbf{X}_1\mathbf{X}_2^\top\mathbf{Z} & \mathbf{0}\\ \mathbf{X}_2\mathbf{X}_1^\top\mathbf{C} + \mathbf{X}_2\mathbf{X}_2^\top\mathbf{Z} & \mathbf{0}\end{pmatrix}. \tag{E.164}$$

Therefore,

$$\det\left(\lambda\mathbf{I}_D - \frac{\mathbf{H}^{\text{in}}\hat{\mathbf{A}}}{\eta}\right) = \lambda^{D-r}\det\left[\lambda\mathbf{I}_r - \frac{1}{L}\left(\mathbf{X}_1\mathbf{X}_1^\top\mathbf{C} + \mathbf{X}_1\mathbf{X}_2^\top\mathbf{Z}\right)\right]. \tag{E.165}$$

Hence the nonzero eigenvalues of $\mathbf{H}^{\text{in}}\hat{\mathbf{A}}/\eta$ are those of the matrix below, which can be rewritten as

$$\frac{1}{L}(\mathbf{X}_1\mathbf{X}_1^\top\mathbf{C} + \mathbf{X}_1\mathbf{X}_2^\top\mathbf{Z}) = \mathbf{B}\mathbf{C} + \frac{\mathbf{X}_1\mathbf{X}_2^\top}{L}\mathbf{Z} \tag{E.166}$$

$$= \mathbf{B}\mathbf{C} + \frac{\mathbf{X}_1\mathbf{X}_2^\top}{L}\mathbf{R}\mathbf{C} \tag{E.167}$$

$$= \left(\mathbf{B} + \frac{\mathbf{X}_1\mathbf{X}_2^\top}{L}\mathbf{R}\right)\mathbf{C} \tag{E.168}$$

$$= \mathbf{B}\left(\mathbf{I}_r + \mathbf{B}^{-1}\frac{\mathbf{X}_1\mathbf{X}_2^\top}{L}\mathbf{R}\right)\mathbf{C}. \tag{E.169}$$

Expanding the inverse of the upper-left block in Equation (E.92), every term containing an odd number of centered Gaussian factors has zero conditional average. Each additional pair multiplies the preceding contribution by $\varphi c_\varphi \langle x^{-1}\rangle_{1/\varphi}$. Therefore

$$\lim_{D\to\infty}\frac{\|\mathbf{R}\|_{\text{F}}^2}{D} = (1-\varphi)\varphi c_\varphi\langle x^{-1}\rangle_{1/\varphi}\sum_{j=0}^{\infty}[\varphi c_\varphi\langle x^{-1}\rangle_{1/\varphi}]^j \tag{E.170}$$

$$= \varphi c_\varphi\sum_{j=0}^{\infty}\left(\frac{\varphi c_\varphi}{1-\varphi}\right)^j \tag{E.171}$$

$$= \frac{\varphi c_\varphi(1-\varphi)}{1-\varphi-\varphi c_\varphi}. \tag{E.172}$$

The denominator is positive for $0 < \varphi \le \alpha$. Indeed, $\varphi + \varphi c_\varphi$ is increasing there and

$$1 - \alpha - \alpha c_\alpha = (1-\alpha)[1-\alpha^3(1-\alpha^2)] > 0. \tag{E.173}$$

Conditional on $\mathbf{X}_1$ and the learned matrix, $\mathbf{B}^{-1}\mathbf{X}_1\mathbf{X}_2^\top\mathbf{R}/L$ is centered Gaussian. For any matrix $\mathbf{F}$ independent of $\mathbf{X}_2$, direct averaging over $\mathbf{X}_2$ gives

$$\mathbb{E}_{\mathbf{X}_2}\left[\mathbf{B}^{-1}\frac{\mathbf{X}_1\mathbf{X}_2^\top}{L}\mathbf{R}\mathbf{F}\mathbf{R}^\top\frac{\mathbf{X}_2\mathbf{X}_1^\top}{L}\mathbf{B}^{-1}\right] = \frac{\mathbf{B}^{-1}}{L}\operatorname{tr}(\mathbf{R}^\top\mathbf{R}\mathbf{F}). \tag{E.174}$$

The right-hand side is proportional to $\mathbf{B}^{-1}$. Substituting $\mathbf{F} = \mathbf{B}^{-1}$ into the same equality gives

$$\frac{\mathbf{B}^{-1}}{L}\operatorname{tr}(\mathbf{R}^\top\mathbf{R}\mathbf{B}^{-1}) = \left\{\frac{1}{L}\operatorname{tr}(\mathbf{B}^{-1}\mathbf{R}^\top\mathbf{R})\right\}\mathbf{B}^{-1}. \tag{E.175}$$

Thus each additional pair of centered Gaussian factors multiplies the preceding term by $L^{-1}\operatorname{tr}(\mathbf{B}^{-1}\mathbf{R}^\top\mathbf{R})$. Rotational independence and the Wishart inverse moment give

$$\frac{1}{L}\operatorname{tr}(\mathbf{B}^{-1}\mathbf{R}^\top\mathbf{R}) \longrightarrow \frac{D}{L}\left(\frac{1}{r}\operatorname{tr}\mathbf{B}^{-1}\right)\left(\frac{1}{D}\operatorname{tr}\mathbf{R}^\top\mathbf{R}\right) \tag{E.176}$$

$$= \frac{1}{\alpha}\frac{\alpha}{\alpha-\varphi}\frac{\varphi c_\varphi(1-\varphi)}{1-\varphi-\varphi c_\varphi} \tag{E.177}$$

$$= \frac{\varphi c_\varphi(1-\varphi)}{(\alpha-\varphi)(1-\varphi-\varphi c_\varphi)}. \tag{E.178}$$

The resulting geometric series becomes singular when the last displayed quantity equals one. The lower boundary is therefore obtained by setting the last displayed quantity to one. Clearing the positive denominator gives

$$\begin{aligned}(\alpha-\varphi)(1-\varphi-\varphi c_\varphi) &= \varphi c_\varphi(1-\varphi),\\ (\alpha-\varphi)(1-\varphi) &= \varphi c_\varphi(\alpha+1-2\varphi)\\ &= \alpha(1-\alpha)\varphi^2(1-\alpha\varphi)(\alpha+1-2\varphi).\end{aligned}$$

Thus the boundary equation is Equation (E.158).

To check uniqueness directly, set $\varphi = \alpha z/(1+z)$. Multiplying the left side minus the right side of Equation (E.158) by $(1+z)^4$ gives

$$\begin{aligned}0 =& \alpha + \alpha(3-\alpha)z + \alpha[(1-\alpha)(3+\alpha)+\alpha^4]z^2\\ &+ \alpha(1-\alpha)[(1-\alpha^2)^2+\alpha^5]z^3 - \alpha^3(1-\alpha)^3(1+\alpha)z^4.\end{aligned} \tag{E.179}$$

There is exactly one sign change, so Descartes' rule gives one positive root. Since $z>0$ parametrizes $0<\varphi<\alpha$, the lower boundary is unique in that interval. □

E.6 Upper boundary

In this subsection, we determine the upper boundary $\varphi_+(\alpha)$. In this regime, $\mathbf{A}_c$ is generically full rank. Therefore, $\varphi_+(\alpha)$ is the value at which its smallest singular value reaches zero, or equivalently at which an eigenvalue reaches zero.

The preceding calculation gives the value at which another zero eigenvalue appears for $0<\varphi<\alpha$. We next calculate the corresponding value for $\alpha<\varphi<1/\alpha$ by reducing the two resolvent conditions to one scalar MP equation.

Proposition E.7 (Upper boundary)**.** *Suppose $0<\alpha<1$. Let z_+ be the unique positive solution of*

$$\begin{aligned}0 =& (1-\alpha)^2 z_+^4 + (1-\alpha)[3(1-\alpha^2)+\alpha^3]z_+^3\\ &+ (1-\alpha)[3+3\alpha(1-\alpha)+\alpha^4]z_+^2\\ &+ [(1-\alpha)(1+2\alpha)+\alpha^5]z_+ - \alpha^4(1-\alpha^2).\end{aligned} \tag{E.180}$$

Then

$$\varphi_+(\alpha) := \frac{\alpha^2 + z_+ + (1-\alpha)z_+^2}{\alpha(1+z_+)[1+(1-\alpha)z_+]}. \tag{E.181}$$

is the unique value in $(\alpha, 1/\alpha)$ such that

$$\lim_{D\to\infty}\sigma_L(\mathbf{A}_c) = 0 \quad\Longleftrightarrow\quad \varphi = \varphi_+(\alpha), \qquad \alpha<\varphi<1/\alpha. \tag{E.182}$$

Moreover, $\alpha<\varphi_+(\alpha)<1$.

Proof. For $\alpha<\varphi<1/\alpha$, we use that $\mathbf{A}_c$ has full rank except at the boundary calculated below. The positive-definite matrix $(\mathbf{X}^{\text{in}})^\top\mathbf{X}^{\text{in}}/L$ is boundedly invertible. Therefore $((\mathbf{X}^{\text{in}})^\top\mathbf{X}^{\text{in}}/L)\mathbf{A}_c$ has a zero eigenvalue exactly when $\mathbf{A}_c$ has a zero eigenvalue. We now obtain the two scalar boundary

equations by a leave-one-column cavity calculation. Let $\mathbf{U}^{-\ell}$ be $\mathbf{U}$ with column $\boldsymbol{u}_\ell$ removed, and define

$$\mathcal{G}_\zeta := \left[\zeta \mathbf{I}_L - \mathbf{U}^\top \frac{\hat{\mathbf{A}}}{\eta} \mathbf{U}\right]^{-1}, \tag{E.183}$$

$$\mathcal{G}_\zeta^{-\ell} := \left[\zeta \mathbf{I}_{L-1} - (\mathbf{U}^{-\ell})^\top \frac{\hat{\mathbf{A}}}{\eta} \mathbf{U}^{-\ell}\right]^{-1}. \tag{E.184}$$

Ordering $\boldsymbol{u}_\ell$ last gives

$$\mathcal{G}_\zeta^{-1} = \begin{pmatrix} (\mathcal{G}_\zeta^{-\ell})^{-1} & -(\mathbf{U}^{-\ell})^\top (\hat{\mathbf{A}}/\eta)\boldsymbol{u}_\ell \\ -\boldsymbol{u}_\ell^\top (\hat{\mathbf{A}}/\eta)\mathbf{U}^{-\ell} & \zeta - \boldsymbol{u}_\ell^\top (\hat{\mathbf{A}}/\eta)\boldsymbol{u}_\ell \end{pmatrix}. \tag{E.185}$$

Restoring the deleted column with the Woodbury formula gives

$$[\mathcal{G}_\zeta]_{\ell\ell} = \left[\zeta - \boldsymbol{u}_\ell^\top \frac{\hat{\mathbf{A}}}{\eta} \boldsymbol{u}_\ell - \boldsymbol{u}_\ell^\top \frac{\hat{\mathbf{A}}}{\eta} \mathbf{U}^{-\ell} \mathcal{G}_\zeta^{-\ell} (\mathbf{U}^{-\ell})^\top \frac{\hat{\mathbf{A}}}{\eta} \boldsymbol{u}_\ell\right]^{-1}. \tag{E.186}$$

We use the large-D replacement that $\boldsymbol{u}_\ell$ can be averaged independently of $\mathcal{G}_\zeta^{-\ell}$. Repeatedly applying Equation (E.186) to the columns of $\mathbf{U}^{-\ell}$ produces a scalar coefficient $y > 0$ and the matrix $(\mathbf{I}_D - y\hat{\mathbf{A}}/\eta)^{-1}$. The normalization $L/D \to \alpha$ gives the first self-consistent equation

$$\alpha = -\lim_{D\to\infty} \frac{1}{D} \operatorname{tr}\left[y \frac{\hat{\mathbf{A}}}{\eta} (\mathbf{I}_D - y\hat{\mathbf{A}}/\eta)^{-1}\right] \tag{E.187}$$

$$= -\lim_{D\to\infty} \frac{1}{D} \operatorname{tr}\left[(\mathbf{I}_D - y\hat{\mathbf{A}}/\eta)^{-1} - \mathbf{I}_D\right] \tag{E.188}$$

$$= 1 - r_\varphi(y). \tag{E.189}$$

Hence

$$r_\varphi(y) = 1 - \alpha. \tag{E.190}$$

To determine when the zero-eigenvalue solution first appears, we square the linear change produced by restoring $\boldsymbol{u}_\ell$ in Equation (E.186) and take its conditional average. At the boundary, its multiplier is one, which gives the second self-consistent equation

$$\alpha = \lim_{D\to\infty} \frac{1}{D} \left\| y \frac{\hat{\mathbf{A}}}{\eta} (\mathbf{I}_D - y\hat{\mathbf{A}}/\eta)^{-1} \right\|_{\mathrm{F}}^2 \tag{E.191}$$

$$= \lim_{D\to\infty} \frac{1}{D} \left\| (\mathbf{I}_D - y\hat{\mathbf{A}}/\eta)^{-1} - \mathbf{I}_D \right\|_{\mathrm{F}}^2 \tag{E.192}$$

$$= G_\varphi(y) - 2r_\varphi(y) + 1. \tag{E.193}$$

Substituting Equation (E.190) gives

$$G_\varphi(y) = 1 - \alpha. \tag{E.194}$$

Thus the upper boundary is determined by

$$r_\varphi(y) = 1 - \alpha, \qquad G_\varphi(y) = 1 - \alpha, \qquad y > 0. \tag{E.195}$$

For the scalar elimination, put

$$\beta := v_\varphi + y, \qquad m := \varphi \left\langle \frac{\beta x}{u_\varphi + \beta x} \right\rangle_{1/\varphi}. \tag{E.196}$$

The first resolvent equation reduces to

$$1 - \alpha = 1 - \varphi + \varphi \left\langle 1 - \frac{yx}{u_\varphi + \beta x} \right\rangle_{1/\varphi} \tag{E.197}$$

$$= 1 - \frac{y}{\beta} m, \tag{E.198}$$

hence $\frac{y}{\beta} = \frac{\alpha}{m}$. The MP identity is

$$\frac{u_\varphi}{\beta} m = (\varphi - m)(1 - m). \tag{E.199}$$

Using $y/\beta = \alpha/m$, it becomes

$$(\varphi - m)(1 - m) = \frac{u_\varphi}{v_\varphi}(m - \alpha) = \frac{(1-\alpha)(1-\alpha\varphi)}{\alpha}(m - \alpha) \tag{E.200}$$

for $\alpha < m < \min\{\varphi, 1\}$.

We next reduce the two-copy equation explicitly. Differentiating the MP identity at fixed u_φ gives

$$\beta\partial_\beta m = \frac{(u_\varphi/\beta)m}{u_\varphi/\beta + \varphi + 1 - 2m} = \frac{m(\varphi - m)(1 - m)}{\varphi - m^2}, \tag{E.201}$$

$$\varphi \left\langle \frac{(\beta x)^2}{(u_\varphi + \beta x)^2} \right\rangle_{1/\varphi} = m - \beta\partial_\beta m. \tag{E.202}$$

Subtracting $1 - \alpha$ from Equation (E.131) and multiplying by its denominator yields

$$\left[1 - \varphi y^2 c_\varphi \left\langle \frac{x}{(u_\varphi + \beta x)^2} \right\rangle_{1/\varphi}\right] [G_\varphi(y) - (1 - \alpha)]$$

$$= -\alpha + \frac{\alpha^2}{m} - \frac{\alpha^2}{m^2}\beta\partial_\beta m + \frac{(1-\alpha)\alpha^2 c_\varphi}{m^2(u_\varphi/\beta)}\beta\partial_\beta m \tag{E.203}$$

$$= \frac{\alpha}{\varphi - m^2}[m^2 - 2\alpha m + \alpha - (1 - \alpha)\varphi + \alpha(1 - \alpha)c_\varphi] \tag{E.204}$$

$$= -\frac{\alpha(1-\alpha)}{\varphi - m^2}\left[\varphi - \alpha - \alpha c_\varphi - \frac{(m - \alpha)^2}{1 - \alpha}\right]. \tag{E.205}$$

Thus the upper boundary satisfies the scalar equation

$$\varphi - \alpha = \alpha c_\varphi + \frac{(m - \alpha)^2}{1 - \alpha}. \tag{E.206}$$

Set $z = (m - \alpha)/(1 - m) > 0$, so that $m = (\alpha + z)/(1 + z)$. Solving Equation (E.200) for φ gives

$$\varphi = \frac{\alpha m(1 - m) + (1 - \alpha)(m - \alpha)}{\alpha[(1 - m) + (1 - \alpha)(m - \alpha)]} \tag{E.207}$$

$$= \frac{\alpha^2 + z + (1 - \alpha)z^2}{\alpha(1 + z)[1 + (1 - \alpha)z]}. \tag{E.208}$$

Substitution into Equation (E.206) gives

$$\begin{aligned} 0 =& (1 - \alpha)^2 z^4 + (1 - \alpha)[3(1 - \alpha^2) + \alpha^3]z^3 \\ &+ (1 - \alpha)[3 + 3\alpha(1 - \alpha) + \alpha^4]z^2 \\ &+ [(1 - \alpha)(1 + 2\alpha) + \alpha^5]z - \alpha^4(1 - \alpha^2). \end{aligned}$$

This is Equation (E.180). Every nonconstant coefficient is positive and the constant term is negative. The polynomial is therefore strictly increasing on $z > 0$ and has exactly one positive root. Substituting this root into Equation (E.208) gives Equation (E.181).

The parametrization also gives

$$\frac{1}{\alpha} - \varphi = \frac{(1 - \alpha)(1 + \alpha + z)}{\alpha(1 + z)[1 + (1 - \alpha)z]}. \tag{E.209}$$

The right side decreases from $1/\alpha - \alpha$ to zero, so $z > 0$ parametrizes the full interval $\alpha < \varphi < 1/\alpha$. The linear term of the polynomial bounds its root by

$$z_+ < \frac{\alpha^4(1 - \alpha^2)}{(1 - \alpha)(1 + 2\alpha) + \alpha^5} < \alpha. \tag{E.210}$$

On the other hand,

$$\varphi = 1 \iff z(1 + z) = \frac{\alpha}{1 - \alpha}, \tag{E.211}$$

whose positive solution is larger than α. Hence $\alpha < \varphi_+ < 1$, and these equations have no further solution before recovery. □

E.7 Long-time behavior

Propositions E.6 and E.7 determine the two boundaries. Finally, we use the eigenvalues of the iteration matrix to obtain the large-t behavior.

Proposition E.8 (Long-time behavior)**.** *Suppose $0 < \alpha < 1$, $\lambda \to 0$, and $\eta > 0$ is sufficiently small. Using the conditional averages in Proposition E.1, assume that the largest real part among the nonzero eigenvalues of $((\mathbf{X}^{\mathrm{in}})^\top \mathbf{X}^{\mathrm{in}}/(\eta L))\mathbf{A}_c$ changes from negative to positive at φ_-, changes from positive to negative at φ_+, and does not change sign elsewhere. Assume also that the normalized traces below have finite limits and that the inverse matrices in the bounded region have uniformly bounded normalized norms.*

$$\mathcal{E}(t) = \begin{cases} \Theta(t^2), & 0 < \varphi < \varphi_-(\alpha), \\ \exp\big(\Theta(\eta t)\big), & \varphi_-(\alpha) < \varphi < \varphi_+(\alpha), \\ O(1), & \varphi_+(\alpha) < \varphi < 1/\alpha, \\ \mathcal{E}_{\mathrm{GD}}(t), & \varphi > 1/\alpha. \end{cases} \tag{E.212}$$

Proof. As $\varphi \to (1/\alpha)^-$, $u_\varphi, c_\varphi \to 0$, $v_\varphi \to 1$, and $\hat{\mathbf{A}}/\eta \to -\mathbf{I}_D$, so all nonzero eigenvalues have negative real part near $1/\alpha$. Together with Propositions E.6 and E.7, the spectral assumption in the statement implies that the nonzero eigenvalues have negative real part below φ_-, positive real part between φ_- and φ_+, and negative real part above φ_+. The discussion in Subsection E.4 then gives $\mathcal{E}(t) = \Theta(t^2)$ for $0 < \varphi < \varphi_-$, $\mathcal{E}(t) = \exp(\Theta(\eta t))$ for $\varphi_- < \varphi < \varphi_+$, and $\mathcal{E}(t) = O(1)$ for $\varphi_+ < \varphi < 1/\alpha$. Equation (E.2) shows that the GD matrix has not yet been recovered in the last regime.

Finally, if $\varphi > 1/\alpha$, then $\alpha\varphi > 1$, and Proposition E.4 gives $\mathcal{E}(t) = \mathcal{E}_{\mathrm{GD}}(t)$. These arguments prove Equation (E.212). □

Finally, $0 < \varphi_- < \alpha < \varphi_+ < 1 < 1/\alpha$ partitions the entire $\alpha < 1$ half-plane into the four open intervals in Result 4.2. For $\alpha > 1$, Proposition C.2 gives the exact recovery condition $\varphi > 1$, while Proposition D.2 gives the quadratic phase for $0 < \varphi < 1$. Together, these propositions give every region of Result 4.2.

F Experimental Details for Fully Trained Attention

This appendix provides the full protocol for the experiments in Section 6 and Figure 2.

F.1 Data and architectures

We use the Gaussian regression task, teacher GD trajectories, prompt coordinates, and generalization error defined in Section 3. The dimension is $D = 32$, the context ratio is $\alpha = 0.75$, the training CoT length is $K = 2$, the teacher step size is $\eta = 0.1$, and the observation-noise standard deviation is $\sigma = 0.1$.

Both architectures have total token width $2D + 2$. The context-only mask in Section 6 means that the query is formed from the final state column,

$$\boldsymbol{z}_k := (\mathbf{E}_{\mu,k})_{:,L+1} = \begin{pmatrix} \mathbf{0}_D \\ 0 \\ \boldsymbol{s}_{\mu,k} \\ 1 \end{pmatrix}, \tag{F.1}$$

whereas the keys and values are formed only from the first L context columns,

$$(\boldsymbol{c}_1 \quad \cdots \quad \boldsymbol{c}_L) := (\mathbf{E}_{\mu,k})_{:,1:L} = \begin{pmatrix} \mathbf{X}_\mu \\ \boldsymbol{y}_\mu^\top \\ \mathbf{0}_{D\times L} \\ \mathbf{0}_L^\top \end{pmatrix} \in \mathbb{R}^{(2D+2)\times L}. \tag{F.2}$$

For each head, the three bias-free projections are therefore

$$\boldsymbol{q}_k^h = \mathbf{W}_Q^h \boldsymbol{z}_k, \qquad \boldsymbol{k}_\ell^h = \mathbf{W}_K^h \boldsymbol{c}_\ell, \qquad \boldsymbol{v}_\ell^h = \mathbf{W}_V^h \boldsymbol{c}_\ell. \tag{F.3}$$

Only $\boldsymbol{z}_k$ enters the query projection, and only $\boldsymbol{c}_1, \dots, \boldsymbol{c}_L$ enter the key and value projections. The resulting head output is

$$\begin{aligned} \boldsymbol{u}_k^h &= \frac{1}{L} \sum_{\ell=1}^{L} \big((\boldsymbol{q}_k^h)^\top \boldsymbol{k}_\ell^h\big) \boldsymbol{v}_\ell^h && \text{(linear attention)}, \\ \boldsymbol{u}_k^h &= \sum_{\ell=1}^{L} \frac{\exp\big((\boldsymbol{q}_k^h)^\top \boldsymbol{k}_\ell^h / \sqrt{d_h}\big)}{\sum_{j=1}^{L} \exp\big((\boldsymbol{q}_k^h)^\top \boldsymbol{k}_j^h / \sqrt{d_h}\big)} \boldsymbol{v}_\ell^h && \text{(softmax attention)}. \end{aligned} \tag{F.4}$$

Thus the state token is absent from the key/value sequence, rather than present with a zero key and value.

Linear attention has one head with $d_1 = 2D + 2$, learns the bias-free matrices $\mathbf{W}_Q^1, \mathbf{W}_K^1, \mathbf{W}_V^1 \in \mathbb{R}^{(2D+2)\times(2D+2)}$, and has no learned output projection. A fixed readout selects the D state coordinates of the head output, and the residual connection adds this increment to the current state:

$$\hat{\boldsymbol{s}}_{\mu,k+1} = \boldsymbol{s}_{\mu,k} + (\boldsymbol{u}_k^1)_{D+2:2D+1}. \tag{F.5}$$

Its Q/K/V weights use the default PyTorch linear-layer initialization, which samples each entry uniformly from $[-1/\sqrt{2D+2}, 1/\sqrt{2D+2}]$, followed by multiplication by 0.1.

Softmax attention has two heads with $d_1 = d_2 = D + 1$. It learns the bias-free $\mathbf{W}_Q^h, \mathbf{W}_K^h, \mathbf{W}_V^h \in \mathbb{R}^{(D+1)\times(2D+2)}$ and $\mathbf{W}_O \in \mathbb{R}^{(2D+2)\times(2D+2)}$. The two head outputs are concatenated, mapped by $\mathbf{W}_O$, restricted to the same D state coordinates, and added to the current state:

$$\hat{\boldsymbol{s}}_{\mu,k+1} = \boldsymbol{s}_{\mu,k} + \left[\mathbf{W}_O \begin{pmatrix} \boldsymbol{u}_k^1 \\ \boldsymbol{u}_k^2 \end{pmatrix}\right]_{D+2:2D+1}. \tag{F.6}$$

All four softmax projections use the same default PyTorch initialization without additional rescaling.

F.2 Optimization and evaluation

Each model is trained for 5,000 AdamW updates with initial learning rate 0.003, cosine decay to 0.00003, weight decay 10^{-4}, gradient-norm clipping at 10, and minibatches of at most 128 transitions. For each seed, evaluation uses 192 independently generated test contexts and a recursive horizon of 128 steps.