An Asymptotic Theory of Chain-of-Thought in In-Context Learning

Kaito Takanami
Department of Physics, Graduate School of Science, The University of Tokyo
Tokyo, Japan
John A. Paulson School of Engineering and Applied Sciences, Harvard University
Cambridge, MA, USA
takanami255@g.ecc.u-tokyo.ac.jp

Cengiz Pehlevan
John A. Paulson School of Engineering and Applied Sciences, Harvard University
Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University
Center for Brain Science, Harvard University
Cambridge, MA, USA

Abstract

Chain-of-thought (CoT) reasoning has become a widely used mechanism for eliciting multi-step reasoning in large language models by generating intermediate reasoning steps at inference time. Yet the scaling behavior of generalization with CoT depth remains poorly understood. To address this question, we study a theoretically solvable model of CoT for in-context weight prediction in linear regression, where test-time reasoning is represented as an iterative refinement of the weight-parameter estimate. Using tools from random matrix theory under high-dimensional asymptotics, we derive an exact formula for the generalization error as a function of reasoning depth, pretraining data amount, and context length. Our analysis reveals a sharp phase transition separating exponential and polynomial improvement, saturation, and overthinking, and characterizes how the optimal reasoning depth scales. We further show that deeper reasoning is most effective with sufficiently rich pretraining and in-context information, whereas limited pretraining or context makes longer reasoning prone to error amplification or saturation. We also validate these predictions through experiments on fully learned linear attention and softmax attention models. Our results provide a unified theoretical account of how test-time CoT depth affects generalization.

1 Introduction

Chain-of-thought (CoT) reasoning refers to a test-time procedure in which a model generates intermediate reasoning steps before producing its final answer [1]. In large language models (LLMs), CoT has become a widely used prompting strategy because it often improves performance on tasks that require multi-step reasoning beyond direct one-shot prediction [1, 2, 3]. From a practical perspective, an important advantage of CoT is that it enables performance gains through test-time computation, which is increasingly valuable as continued scaling through pretraining alone faces growing constraints in both data and training resources [4, 5].

Despite its empirical effectiveness, how test-time CoT depth affects generalization remains poorly understood. In particular, recent empirical studies have shown that increasing the number of reasoning steps does not lead to uniformly monotone gains: depending on the setting, deeper reasoning may

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

improve performance, saturate, or even amplify errors through overthinking [6, 7]. A central open question is what governs these qualitatively different regimes, and in particular how they depend on the properties of the data seen during pretraining and on the structure of the information available at test time. While empirical studies can reveal these phenomena, they cannot by themselves identify the principles that govern them, which motivates the study of a theoretically tractable model in which test-time reasoning dynamics can be analyzed explicitly and related to generalization.

To address these questions, we study a theoretically tractable model of CoT in an in-context learning (ICL) problem for linear regression, where the task is to predict the underlying regression weight from a sequence of contextual examples. This setting provides a minimal framework that still captures the essential structure needed to analyze the effect of test-time reasoning depth [8, 9, 10]. During pretraining, the model directly predicts the regression weight from the context. At inference time, CoT is modeled as a sequence of intermediate updates that progressively refine the weight estimate before the final prediction. This formulation enables us to derive a closed-form expression for the generalization error in the high-dimensional limit and to analyze test-time reasoning dynamics explicitly.

Our main results are summarized as follows:

- We introduce an in-context weight prediction model for pretraining and test-time CoT, and derive an exact characterization of its test-time error dynamics in a controlled high-dimensional limit, where the task-parameter dimension D, number of pretraining tasks M, and number of in-context examples L go to infinity simultaneously with fixed ratios (Section 4).
- We identify a phase-transition structure in test-time CoT across data regimes, separating four regimes: an exponential-improvement regime, a polynomial-improvement regime, a saturation regime, and an overthinking regime in which deeper reasoning amplifies error (Section 5).
- The phase transition further reveals how pretraining task diversity and in-context examples jointly govern test-time scaling. Pretraining task diversity determines the stability of iterative refinement and the optimal reasoning depth, while in-context examples control the error decay rate and the asymptotic information limit at test time (Sections 5.1–5.3).
- We validate these theoretical predictions in nonlinear softmax-attention models and find that the same qualitative behavior persists beyond the solvable linear setting (Section 6).

These results provide unified theoretical insights into how test-time CoT depth affects generalization, and yield principled understandings for when deeper reasoning helps, saturates, or becomes harmful[1].

Impact statement. This work is theoretical and studies simple attention models in a synthetic setting. Its potential positive impact is a better understanding of test-time reasoning, and we do not identify direct negative societal impacts.

2 Related Works

CoT and test-time reasoning in LLMs. Recent work has brought increasing attention to test-time scaling in LLMs. A central reason is that additional inference-time computation can improve reasoning capabilities that are not easily unlocked by scaling pretraining alone [1, 3, 11]. This practical importance is reflected in recent frontier reasoning-oriented systems, which have adopted test-time reasoning as a core design principle rather than a purely prompt-level heuristic [12, 13]. At the same time, empirical studies have shown that increasing reasoning depth is not uniformly beneficial: longer reasoning can yield diminishing returns and saturation [14], or even harm performance through overthinking [15, 6, 7, 16]. However, these works are largely empirical and do not provide a general account of what governs the transition between improvement, saturation, and harmful overthinking. Moreover, much of the existing literature emphasizes problem difficulty, while paying comparatively less attention to the role of data properties. More broadly, theoretical understanding of test-time scaling remains relatively limited [17, 9, 18, 19, 20, 21]. Motivated by this gap, we theoretically investigate how pretraining quality and in-context information govern these regimes and their scaling laws.

[1]The code for reproducing the results is available at https://github.com/taka255/cot_asymptotics.

Theoretical studies of CoT and ICL. Theoretical studies of in-context learning have long used linear regression as a tractable benchmark for understanding how transformers can implement learning algorithms [22, 23, 24, 25]. One line of development has pushed this framework further through asymptotic analyses that make exact characterization possible, revealing richer phenomena such as double descent and phase transitions associated with the emergence of genuine in-context generalization beyond memorization [26, 27, 28, 29, 30]. A separate line of work has connected this broader ICL setting to CoT through in-context weight prediction models for linear regression, where transformers iteratively refine task-parameter estimates without learning [8, 9, 10]. In particular, [9] theoretically showed that increasing test-time compute can reduce the amount of contextual information required during training, but can also hurt performance when the training data fails to represent the input directions that are important for the downstream task. Our work builds on both lines of research in the finite-data regime, where overthinking can occur even under identical training and inference distributions. This allows us to view the qualitatively different effects of test-time reasoning depth as phase transitions, and to identify when deeper reasoning improves generalization, when it saturates, and when it becomes harmful through overthinking. We further show how these transitions are shaped by pretraining quality and in-context information.

3 Model

In this section, we introduce the model used in our analysis. We study a variant of in-context linear regression model [22]. While standard in-context linear regression aims to predict the output for a new query input given a prompt of labeled examples, our objective is to infer the regression weight itself, as proposed in [8, 10, 9]. During pretraining, the model learns an update rule from prompts of contextual examples. At inference time, CoT is generated by iteratively applying this learned update rule to a new task.

Training data. We consider M linear regression tasks indexed by $\mu = 1, \ldots, M$. For each task μ, the training examples $\{(\boldsymbol{x}_{\mu,l}, y_{\mu,l})\}_{l=1}^{L}$ are generated as

$$\boldsymbol{x}_{\mu,l} \sim \mathcal{N}\left(\mathbf{0}, \frac{I_D}{D}\right), \quad y_{\mu,l} = \boldsymbol{w}_\mu^\top \boldsymbol{x}_{\mu,l} + \epsilon_{\mu,l}, \quad \epsilon_{\mu,l} \sim \mathcal{N}(0, \sigma_\mu^2), \tag{1}$$

where L is the length of the context, and $\boldsymbol{w}_\mu \in \mathbb{R}^D$ denotes the task-specific ground-truth parameter vector, sampled from a Gaussian normal distribution.

Prompt construction. For each task μ, we form a prompt matrix from the labeled examples together with a dedicated slot that stores the current estimate of the task parameter. Let

$$\mathbf{X}_\mu = [\boldsymbol{x}_{\mu,1}, \ldots, \boldsymbol{x}_{\mu,L}] \in \mathbb{R}^{D\times L}, \qquad \boldsymbol{y}_\mu = [y_{\mu,1}, \ldots, y_{\mu,L}]^\top \in \mathbb{R}^L. \tag{2}$$

We define the embedding matrix as

$$\mathbf{E}_\mu = \begin{bmatrix} \mathbf{X}_\mu & \mathbf{0}_D \\ \boldsymbol{y}_\mu^\top & 0 \\ 0_{D\times L} & \hat{\boldsymbol{w}}_\mu^{\text{init}} \\ 0_{1\times L} & 1 \end{bmatrix} \in \mathbb{R}^{(2D+2)\times(L+1)}. \tag{3}$$

The first L columns encode the labeled examples, while the last column functions as a slot for the predicted task parameter. It is initialized with $\hat{\boldsymbol{w}}_\mu^{\text{init}}$ and updated during reasoning, unless otherwise specified, we take $\hat{\boldsymbol{w}}_\mu^{\text{init}} = \mathbf{0}$. The final row serves as a marker that distinguishes the example columns, which take value 0, from the parameter-estimate column, which takes value 1.

Architecture. We consider a single-layer linear self-attention model with residual connection. For an input matrix $\mathbf{E} \in \mathbb{R}^{(2D+2)\times T}$, the layer is defined as

$$\textbf{Atten}(\mathbf{E}; \mathbf{W}_K, \mathbf{W}_Q, \mathbf{W}_V) = \mathbf{E} + \frac{1}{L}(\mathbf{W}_V\mathbf{E})(\mathbf{W}_K\mathbf{E})^\top(\mathbf{W}_Q\mathbf{E}), \tag{4}$$

where $\mathbf{W}_K, \mathbf{W}_Q, \mathbf{W}_V \in \mathbb{R}^{(2D+2)\times(2D+2)}$. Defining $\mathbf{W} := \mathbf{W}_K^\top\mathbf{W}_Q$, $\mathbf{V} := \mathbf{W}_V$, and $\theta := (\mathbf{W}, \mathbf{V})$, we equivalently write

$$\textbf{Atten}(\mathbf{E}; \theta) = \mathbf{E} + \frac{1}{L}\mathbf{V}\mathbf{E}\mathbf{E}^\top\mathbf{W}\mathbf{E}. \tag{5}$$

The readout is taken from the weight slot of the final column, which is given by

$$\textbf{Read}(\mathbf{E}) = \mathbf{E}_{(D+2):(2D+1),T} \in \mathbb{R}^D. \tag{6}$$

The resulting estimate of the task parameter is thus given by $\textbf{Read}(\textbf{Atten}(\mathbf{E};\theta))$.

Simplified parameterization. To obtain a tractable model, we retain only the blocks that directly control the update of the parameter-estimate token. Accordingly, we use the block parameterization

$$\mathbf{W} = \begin{pmatrix} 0_{D\times D} & 0_{D\times 1} & I_{D\times D} & 0_{D\times 1} \\ 0_{1\times D} & 0_{1\times 1} & 0_{1\times D} & -1 \\ 0_{D\times D} & 0_{D\times 1} & 0_{D\times D} & 0_{D\times 1} \\ 0_{1\times D} & 0_{1\times 1} & 0_{1\times D} & 0_{1\times 1} \end{pmatrix}, \qquad \mathbf{V} = \begin{pmatrix} 0_{D\times D} & 0_{D\times 1} & 0_{D\times D} & 0_{D\times 1} \\ 0_{1\times D} & 0_{1\times 1} & 0_{1\times D} & 0_{1\times 1} \\ -D\mathbf{A} & 0_{D\times 1} & 0_{D\times D} & 0_{D\times 1} \\ 0_{1\times D} & 0_{1\times 1} & 0_{1\times D} & 0_{1\times 1} \end{pmatrix}. \tag{7}$$

Here $\mathbf{A} \in \mathbb{R}^{D\times D}$ is the only learnable parameter; all blocks of $\mathbf{W}$ and all other blocks of $\mathbf{V}$ are fixed. In both $\mathbf{W}$ and $\mathbf{V}$, the row and column indices are partitioned into four consecutive blocks of sizes $D, 1, D, 1$. This restriction is motivated by prior gradient-flow analyses of related in-context weight-prediction settings [8, 9]. These works show that, under suitable initializations, gradient flow converges to solutions with the same sparse block structure. Consistent with this motivation, our full-parameter training experiments suggest that these fixed blocks are learned close to the prescribed values or remain negligible, while the retained block $\mathbf{A}$ governs the nontrivial update of the parameter estimate at inference time. See Appendix A for empirical support for this simplification.

Under this parameterization, the attention layer maps the embedding matrix to

$$\textbf{Read}(\textbf{Atten}(\mathbf{E}_\mu;\theta)) = \hat{\boldsymbol{w}}_\mu^{\text{init}} - \frac{D}{L}\mathbf{A}\left(\mathbf{X}_\mu\mathbf{X}_\mu^\top \hat{\boldsymbol{w}}_\mu^{\text{init}} - \mathbf{X}_\mu \boldsymbol{y}_\mu\right). \tag{8}$$

Thus, the readout is an updated estimate of the task parameter obtained by a linear transformation of the empirical gradient. Detailed calculations and discussions are given in Appendix B.

Training objective. We learn $\mathbf{A}$ by ridge-regularized empirical risk minimization so that the readout matches the ground-truth task parameter. Under the above simplification, the empirical objective is

$$\mathcal{L}(\mathbf{A}) := \frac{1}{M}\sum_{\mu=1}^{M} \|\boldsymbol{w}_\mu - \textbf{Read}(\textbf{Atten}(\mathbf{E}_\mu;\theta))\|^2 + \frac{\lambda}{2}\|\mathbf{A}\|_F^2, \tag{9}$$

where $\lambda \geq 0$ is a ridge regularization parameter. We denote the minimizer of Eq. (9) by $\mathbf{A}^*$.

CoT in inference phase. At inference time, we consider a new linear regression task generated from the same distribution as in pre-training. Specifically, the task parameter $\boldsymbol{w} \in \mathbb{R}^D$ denotes the ground-truth parameter vector, and the examples $\{(\boldsymbol{x}_l, y_l)\}_{l=1}^L$ are generated without noise as $\boldsymbol{x}_l \sim \mathcal{N}(\mathbf{0}, I_D/D)$ and $y_l = \boldsymbol{w}^\top \boldsymbol{x}_l$, for $l = 1, \ldots, L$. Let $\mathbf{X} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L] \in \mathbb{R}^{D\times L}$ and $\boldsymbol{y} = [y_1, \ldots, y_L]^\top \in \mathbb{R}^L$. We then generate a CoT trajectory by repeatedly constructing a prompt from the observed examples and the current sequence of parameter estimates, and applying the learned attention layer. At step t, we define $\hat{\mathbf{W}}_t = [\hat{\boldsymbol{w}}_0, \hat{\boldsymbol{w}}_1, \ldots, \hat{\boldsymbol{w}}_t] \in \mathbb{R}^{D\times(t+1)}$, and construct the prompt

$$\mathbf{P}_t = \begin{bmatrix} \mathbf{X} & 0_{D\times(t+1)} \\ \boldsymbol{y}^\top & 0_{1\times(t+1)} \\ 0_{D\times L} & \hat{\mathbf{W}}_t \\ 0_{1\times L} & \mathbf{1}_{1\times(t+1)} \end{bmatrix} \in \mathbb{R}^{(2D+2)\times(L+t+1)}. \tag{10}$$

Starting from $\hat{\boldsymbol{w}}_0 = \mathbf{0}$, we define the inference-time CoT trajectory by $\hat{\boldsymbol{w}}_{t+1} := \textbf{Read}(\textbf{Atten}(\mathbf{P}_t;\theta^*))$, where θ^* denotes the parameter pair $(\mathbf{W}, \mathbf{V})$ with $\mathbf{A} = \mathbf{A}^*$. Here, t can be interpreted as the amount of test-time computation, since the computation required at inference is proportional to t.

At each step, the new parameter-estimate token is explicitly generated and appended to the prompt, where it conditions the next step. This captures the sequential generation of and conditioning on intermediate tokens central to CoT, while representing each token as a parameter estimate rather than a natural-language reasoning step. Our formulation is therefore related to scratchpad reasoning [31] and recurrent-depth or looped Transformers [32, 33, 34], but retains the intermediate estimates as explicit tokens in the context rather than only updating hidden states.

Evaluation. To study how the CoT depth at inference time t affects the final prediction performance, we evaluate the parameter estimation error on a new task. We then define the mean squared error (MSE) by $\mathcal{E}_t := \mathbb{E}[\|\boldsymbol{w} - \hat{\boldsymbol{w}}_t\|^2]/D$, where the expectation is taken over the pre-training and inference data distributions.

High-dimensional limit. We consider the high-dimensional limit in which $D, L, M \to \infty$ with the ratios

$$\frac{L}{D} \to \alpha, \quad \frac{M}{D} \to \tau, \tag{11}$$

where $\alpha, \tau \in (0, \infty)$ are fixed constants. Here, α represents the number of in-context examples per task (context length), and τ represents the amount of task diversity. The advantage of this regime is that the generalization error can be characterized quantitatively in terms of only a few macroscopic system parameters, with sample-to-sample fluctuations becoming asymptotically negligible.

4 Precise characterization of test-time CoT dynamics

Our goal is to characterize the generalization error at a finite CoT depth t in the high-dimensional limit. The key observation is that the test-time CoT iteration induces a linear recursion for the estimation error (see Appendix C.2), which can be written as

$$\boldsymbol{w} - \hat{\boldsymbol{w}}_t = \left(\mathbf{I} - \frac{1}{\alpha}\mathbf{A}^*\mathbf{S}\right)^t \boldsymbol{w}, \tag{12}$$

where $\mathbf{S} = \mathbf{X}\mathbf{X}^\top$ is the inference covariance matrix and $\mathbf{A}^*$ is the learned update matrix.

Operationally, this means that CoT behaves like a learned iterative algorithm: pretraining determines the update matrix $\mathbf{A}$, and inference applies this update repeatedly to the prompt examples so as to progressively refine the estimate of the underlying task parameter [9]. This representation implies that the MSE depends only on the matrix product generated by the CoT dynamics. In particular, the MSE is given by

$$\mathcal{E}_t = \frac{1}{D}\mathbb{E}\,\mathrm{Tr}\left[\left(\mathbf{I} - \frac{1}{\alpha}\mathbf{A}^*\mathbf{S}\right)^t \left(\mathbf{I} - \frac{1}{\alpha}\mathbf{S}\mathbf{A}^{*\top}\right)^t\right]. \tag{13}$$

Thus, evaluating the error reduces to understanding a family of mixed moments of $\mathbf{A}^*\mathbf{S}$ and $\mathbf{S}\mathbf{A}^{*\top}$. Since these two matrices generally do not commute, their mixed moments cannot be recovered from a one-point resolvent, which motivates the use of a two-point generating function. To organize these moments, we introduce the two-point generating function

$$F(u, v) := \frac{1}{D}\mathbb{E}\,\mathrm{Tr}\Big[(\mathbf{I} - u\mathbf{A}^*\mathbf{S})^{-1}(\mathbf{I} - v\mathbf{S}\mathbf{A}^{*\top})^{-1}\Big]. \tag{14}$$

A coefficient comparison shows that $\mathcal{E}_t$ is obtained from the derivatives of $F(u, v)$ at $u = v = 0$. Therefore, the problem of evaluating the generalization error is reduced to computing the large-D limit of the scalar function $F(u, v)$.

In the proportional high-dimensional limit, the generating function $F(u, v)$ with randomness converges to a deterministic limit, which we denote again by $F(u, v)$. Combined with the reduction above, this implies that the finite-depth generalization error is asymptotically determined by the derivatives of this limiting function at $u = v = 0$. The following result makes this characterization explicit and, in the case $\lambda = 0$, gives a closed-form expression for $F(u, v)$.

Result 1. *The generalization error at CoT depth t is asymptotically given by*

$$\mathcal{E}_t = \sum_{p=0}^{t}\sum_{q=0}^{t}\binom{t}{p}\binom{t}{q}\left(-\frac{1}{\alpha}\right)^{p+q}\frac{1}{p!\,q!}\frac{\partial^p}{\partial u^p}\frac{\partial^q}{\partial v^q}F(u, v)\Big|_{u=v=0}. \tag{15}$$

Here, especially when $\lambda = 0$, $F(u, v)$ is explicitly given by

$$F(u, v) = \frac{\alpha\, g(u)\, g(v)}{\alpha - (g(u) - 1)(g(v) - 1)\left[1 + \frac{1+\sigma^2}{\alpha(\tau-1)}(\alpha + g(u) + g(v) - 1)\right]}, \tag{16}$$

where $g(u)$ is the resolvent function of Wishart matrix, which is given by

$$g(u) = \frac{c + u(1-\alpha) - \sqrt{\big(c + u(1-\alpha)\big)^2 - 4cu}}{2u}. \tag{17}$$

with $c = 1 + (1+\sigma^2)/\alpha$.

Technically, the idea to reduce the finite-time error to a two-point correlation function is recently discussed in [35] for the case of stochastic gradient dynamics. The present setting is more delicate because the propagation operator $\mathbf{A}^*$ is itself learned and generally non-symmetric, so the resulting two-point object does not reduce to the simpler forms analyzed in prior work. To evaluate this two-point generating function, we combine the generating-function reduction with additional linearization [36] and multi-source cavity techniques [37] in random matrix theory. The detailed derivation, including the case $\lambda > 0$, is given in Appendix C. Also, its validity is supported by the good agreement with numerical experiments reported in Appendix E.

5 Phase transition of test-time scaling law

In this section, we extract the asymptotic test-time scaling law from the finite-depth characterization of the generalization error using Result 1. As $t \to \infty$, the behavior of $\mathcal{E}_t$ is controlled by two quantities: the limiting error $\mathcal{E}_\infty$ and the exponential rate $\Lambda(\alpha, \tau, \sigma^2)$. Their interaction yields a sharp phase transition, separating regimes of exponential improvement, polynomial improvement, information-limited saturation, and overthinking. More precisely, we obtain the following asymptotic test-time scaling law.

Result 2. *Assume $\tau > 1$. Asymptotically as $t \to \infty$, the test-time scaling law of the generalization error obeys*

$$\mathcal{E}_t - \mathcal{E}_\infty \sim K\, t^{-1/2}\, \Lambda(\alpha, \tau, \sigma^2)^t, \tag{18}$$

where K is a positive constant and the limiting value $\mathcal{E}_\infty$ is given by

$$\mathcal{E}_\infty = \begin{cases} \dfrac{(1-\alpha)(\tau-1)}{\tau - 2 - \sigma^2}, & 0 < \alpha < 1 \text{ and } \tau \geq \tau_c(\alpha, \sigma^2) \\ 0, & \text{otherwise}, \end{cases} \tag{19}$$

and the exponential rate satisfies

$$\Lambda(\alpha, \tau, \sigma^2) \begin{cases} > 1, & \tau < \tau_c(\alpha, \sigma^2), \\ = 1, & \tau = \tau_c(\alpha, \sigma^2), \\ < 1, & \tau > \tau_c(\alpha, \sigma^2). \end{cases} \tag{20}$$

Here, the critical value $\tau_c(\alpha, \sigma^2)$ is given by

$$\tau_c(\alpha, \sigma^2) = 1 + \frac{(1+\sigma^2)(\alpha + 1 + 2\sigma^2 - \sqrt{\Delta})(2\alpha + 2 + 2\sigma^2 - \sqrt{\Delta})}{2\alpha\sqrt{\Delta}}, \tag{21}$$

with $\Delta := (\alpha + 1 + 2\sigma^2)^2 - 4\alpha$.

The detailed derivation is given in Appendix F. The scaling law is derived under a condition on the relevant singularity that remains to be justified rigorously through multivariate singularity analysis, and agrees closely with numerical results in Appendix E.

Result 2 shows that the large-t behavior of test-time CoT is controlled by two quantities: the exponential factor $\Lambda(\alpha, \tau, \sigma^2)$, which determines whether the finite-depth error grows or decays, and the limiting error floor $\mathcal{E}_\infty$, which determines whether perfect asymptotic recovery is possible. Their combination yields four qualitatively distinct asymptotic regimes. When $\Lambda > 1$, deeper reasoning amplifies the error, giving the overthinking regime. At the critical boundary $\Lambda = 1$, the exponential improvement disappears and the error decays only polynomially in depth. When $\Lambda < 1$, the error decays exponentially fast; this corresponds to the exponential-improvement regime if $\mathcal{E}_\infty = 0$, and to the saturation regime if $\mathcal{E}_\infty > 0$, where the dynamics remains stable but converges to a nonzero error floor.

Figure 1A summarizes this phase structure. It shows that the effect of increasing test-time CoT depth is not governed by a single universal scaling law, but instead changes qualitatively across data regimes. Figure 1B and Figure 1C show representative trajectories of $\mathcal{E}_t$ across these regimes, illustrating how the finite-depth dynamics transition between error amplification, polynomial decay, exponential decay to zero, and exponential decay toward a nonzero limit. We now discuss the structure and behavior of these regimes in detail.

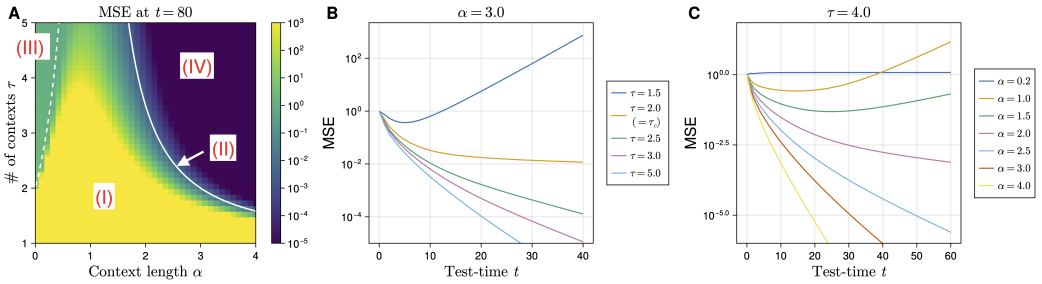

Figure 1: **Phase diagram of the test-time scaling law and representative error dynamics.** (A) Heatmap of theoretical prediction of the generalization error (MSE) at a fixed test-time depth $t = 80$. The diagram is divided into four regimes: **(I)** the *overthinking regime*, $\tau < \tau_c(\alpha, \sigma^2)$, where long test-time CoT amplifies the error; **(II)** the *polynomial-decay regime*, $\alpha > 1$ and $\tau = \tau_c(\alpha, \sigma^2)$, where the error decreases only polynomially; **(III)** the *saturation regime*, $0 < \alpha < 1$ and $\tau > \tau_c(\alpha, \sigma^2)$, where the dynamics is stable but the error converges to a nonzero limit; and **(IV)** the *exponential-decay regime*, $\alpha > 1$ and $\tau > \tau_c(\alpha, \sigma^2)$, where the error decays exponentially to zero. (B,C) Theoretical prediction of the generalization error $\mathcal{E}_t$ dynamics for (B) fixed α and (C) fixed τ. (A-C) Parameters: $\lambda = 10^{-5}, \sigma^2 = 0.01$.

5.1 Exponential/polynomial scaling regime

In the exponential/polynomial scaling regime, repeated test-time refinement remains beneficial and the generalization error vanishes asymptotically ($\mathcal{E}_\infty = 0$). When $\tau > \tau_c(\alpha, \sigma^2)$, the error decays exponentially fast ($\Lambda(\alpha, \tau, \sigma^2) < 1$), whereas at the critical boundary $\tau = \tau_c(\alpha, \sigma^2)$, the exponential factor disappears and only polynomial decay remains ($\Lambda(\alpha, \tau, \sigma^2) = 1$).

In this regime, repeated refinement remains stable, so increasing the CoT depth consistently reduces the generalization error rather than amplifying it. Moreover, the error converges to zero asymptotically, implying that perfect recovery is achievable in the large-depth limit. The following theorem gives a representative asymptotic scaling law in this favorable setting.

Theorem 3. *Assume the population-risk regime $\tau \to \infty$, ridgeless learning $\lambda = 0$, and $\alpha > 1$. Then, as $t \to \infty$,*

$$\mathcal{E}_t \asymp t^{-3/2} \left(\frac{2\sqrt{\alpha} + \sigma^2}{\alpha + 1 + \sigma^2} \right)^{2t}. \tag{22}$$

Especially, in the context-rich limit $\alpha \gg \sigma^2 + 1$, $\mathcal{E}_t \asymp t^{-3/2}(4/\alpha)^t$.

The proof is given in Appendix G.

The simplified form Theorem 3 clarifies that the decay rate is controlled by α, namely the effective amount of in-context examples. Larger α leads to faster exponential improvement, meaning that each refinement step becomes more informative when the context is richer.

The main implication is that, sufficiently accurate pretraining makes additional test-time compute genuinely useful for improving performance. Moreover, increasing the number of in-context examples in this regime further improves the efficiency of test-time scaling by accelerating the decay rate. In this sense, pretraining enables effective test-time scaling, while in-context information determines how rapidly its benefit is realized.

5.2 Overthinking regime

In the overthinking regime, the effect of increasing test-time CoT depth is non-monotone. At small depths, additional reasoning can still improve performance, but beyond a certain point it amplifies error instead of reducing it. This behavior is governed by task diversity τ, which determines the quality of pretraining. More specifically, when τ is small, the learned operator $\mathbf{A}^*$ retains systematic error, which is then amplified by repeated test-time refinement. As a result, the overthinking regime gives rise to a nontrivial optimal stopping depth. The following result characterizes its scaling near the phase boundary, together with the asymptotic error achieved at the optimum.

Result 4. *Assume that (α, τ, σ^2) is in the overthinking regime. Let $t^* := \operatorname{argmin}_{t \in \mathbb{Z}_{\geq 0}} \mathcal{E}_t$. Then, as $\tau \to \tau_c(\alpha, \sigma^2)$ from below,*

$$t^* = C_t(\alpha, \sigma^2)\big(\tau_c(\alpha, \sigma^2) - \tau\big)^{-1}\big(1 + o(1)\big), \tag{23}$$

for some positive function $C_t(\alpha, \sigma^2)$. Moreover, the generalization error at t^* satisfies

$$\mathcal{E}_{t^*} = C_E(\alpha, \sigma^2)\,(t^*)^{-1/2}\big(1 + o(1)\big), \tag{24}$$

for some positive function $C_E(\alpha, \sigma^2)$.

The detailed derivation is given in Appendix H.

Result 4 shows that, in the overthinking regime, the benefit of test-time reasoning is intrinsically finite-depth. As τ approaches $\tau_c(\alpha, \sigma^2)$ from below, the optimal stopping depth t^* diverges, meaning that deeper reasoning remains useful for increasingly longer horizons. At the same time, the best achievable error decreases only polynomially, as $(t^*)^{-1/2}$, rather than exponentially. Thus, improving pretraining extends the useful range of test-time CoT, but does not immediately restore the efficient exponential scaling.

This behavior reflects the origin of overthinking. When τ is small, pretraining task diversity is insufficient, and the learned update rule $\mathbf{A}^*$ is systematically imperfect. Repeated test-time refinement then amplifies this residual imperfection, so deeper reasoning eventually becomes harmful. The bottleneck is therefore not test-time compute itself, but the quality of the learned rule produced by pretraining.

The broader implication is that test-time scaling is conditional on pretraining quality. In the overthinking regime, increasing reasoning depth does not provide an unconditional path to better performance, because its benefit is confined to a finite useful horizon before error amplification takes over. Deeper reasoning therefore becomes reliably effective only when pretraining is strong enough to delay the onset of overthinking and sustain useful refinement over longer horizons.

5.3 Saturation regime

In the saturation regime, repeated test-time refinement remains stable and converges ($\Lambda(\alpha, \tau, \sigma^2) < 1$), but the generalization error does not vanish asymptotically ($\mathcal{E}_\infty > 0$). This nonzero limit reflects an information bottleneck at test time. Even when pretraining is sufficiently accurate, too few in-context examples leave the prompt itself incomplete, so some information required for prediction is simply unavailable at inference time. It cannot reconstruct information that is absent from the context.

This mechanism also explains a seemingly counterintuitive phenomenon: increasing the number of in-context examples does not always lead to better long-horizon performance. As shown in Fig. 1C, a setting with larger α may perform better at small depths, yet be overtaken at sufficiently large depths by the saturation regime because long-horizon behavior is governed by stability of the test-time refinement dynamics. This qualitative prediction is consistent with [38], which reports that adding more in-context CoT examples does not monotonically improve performance.

These results imply that, in this regime, performance is limited mainly by the amount of in-context information rather than by test-time compute. Therefore, increasing the CoT depth alone yields diminishing returns, while enriching prompt examples is the more effective way to improve performance. This identifies the saturation regime as an information-limited phase of test-time scaling.

6 Experiments

In the preceding sections, we theoretically analyzed how the effect of test-time CoT depends on the number of pretraining tasks and in-context examples using a simplified linear attention model. Here, we ask if the same qualitative predictions continue to hold in more expressive settings. To this end, we consider two extensions: a full linear attention model and a softmax attention model. As in the theoretical setting, L denotes the number of in-context examples per task and M denotes the number of training tasks. In both models, we use the same basic prompt format as in Section 3, initialize $\boldsymbol{w}_\mu^{\text{init}}$ as a random Gaussian vector, and read out the prediction from the parameter block of the final token. In both extensions, the attention mechanism is parameterized by query, key, and value matrices.

Because these expanded models permit more general token interactions than the simplified theory, we explicitly control which source tokens are visible to the final parameter-estimate token through an attention mask. Concretely, at each step, we treat the final parameter-estimate token as the query token and allow it to attend to the L in-context example tokens and all previously generated estimate tokens, while masking out the query token itself. This mask makes CoT an iterative update process anchored to the observed examples, while allowing previously generated estimates to be explicitly reused as intermediate states. These details are given in Appendix I.

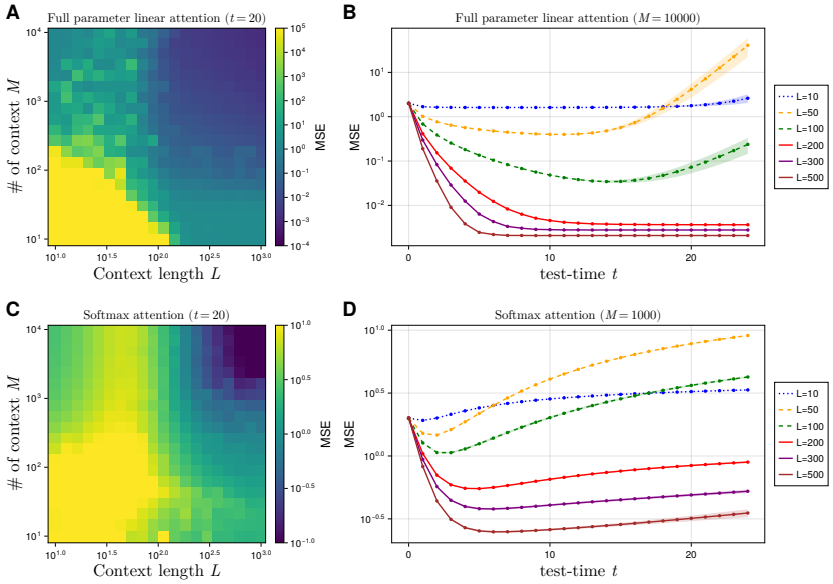

Figure 2: **CoT experiments in the fully learned linear attention and softmax attention models.** (A, C) Phase diagrams of the test-time generalization error at $t = 20$ as a function of the context length L and the number of training tasks M, for (A) the fully learned linear attention model and (C) the softmax attention model, respectively. (B, D) Test-time generalization error as a function of reasoning depth in the (B) fully learned linear attention model and (D) the softmax attention model, respectively. The solid curves correspond to an exponential decay regime, the dashed curves to an overthinking regime, and the dotted curves to a saturation regime. Parameters: (A-D) $\lambda = 10^{-5}$, $\sigma^2 = 0$, and $D = 50$. Error bars represent the standard error of the mean over 5 trials per point.

Figures 2A and C show the phase diagrams for long test-time CoT ($t = 20$) in the fully learned linear attention model and the softmax attention model, respectively. These phase diagrams suggest that the broad qualitative picture from the theory holds in more expressive models. In particular, when M is small, the overthinking regime appears more prominently, whereas as M increases, the behavior separates into saturation and decay regimes depending on the number of in-context examples L.

Figures 2B and D examine the corresponding test-time dynamics at representative values of M. The results also reproduce the three characteristic phases predicted by the theory at a qualitative level. The solid curves exhibit a decay regime; in the fully learned linear attention model, this decay is close to the exponential behavior predicted by the theory. In the softmax attention model, however, the

decrease appears more gradual over the displayed range, suggesting that while the same qualitative phase structure persists, the precise decay rate depends on the attention parameterization. A possible reason is that softmax normalization constrains the effective update size at each step, making the dynamics less sharply exponential than in linear attention. The dashed curves show an overthinking regime, where the error initially decreases but eventually increases as the number of reasoning steps increases. The dotted curves correspond to a saturation regime, in which the error improves little after an initial period and then remains nearly constant over a long range of reasoning steps.

Overall, the behavior of these regimes is consistent with the theoretical predictions on how L and M control the effectiveness of test-time CoT. First, the decay regime appears consistent with the prediction in Section 5.1: increasing the number of in-context examples accelerates the rate of error decay. Second, the overthinking regime is broadly consistent with the prediction in Section 5.2, in the sense that pretraining on more tasks makes longer test-time CoT beneficial over a wider range of reasoning steps. Finally, in the long-CoT limit, the overthinking regime can yield worse performance than the saturation regime with fewer in-context examples, which is also in line with the prediction in Section 5.3.

7 Conclusion

We studied the effect of test-time chain-of-thought depth on generalization in a solvable model of in-context learning for linear regression. Using a two-point generating function approach combined with random matrix theory, we derived an exact asymptotic formula for the test-time generalization error dynamics. This formula revealed a sharp phase transition separating four qualitatively distinct regimes of test-time scaling. When pretraining is sufficient, deeper reasoning reduces the error exponentially at a rate controlled by the in-context examples. When pretraining is insufficient, reasoning beyond a finite optimal depth amplifies error, and the useful range of test-time compute grows only as pretraining improves. When in-context information is limited, reasoning remains stable but saturates at a nonzero error floor, since it cannot recover information absent from the context. Our results provide a unified theoretical account of empirical test-time scaling phenomena, including improvement, saturation and overthinking, and clarify that the primary bottleneck shifts from pretraining quality to context information to test-time compute across these regimes. These findings not only lay a theoretical foundation for future research on test-time scaling, but also offer practical insights for the design of pretraining strategies and test-time reasoning in LLMs.

Limitations. While our analysis provides an insightful unified theoretical framework for understanding the test-time CoT dynamics, it is restricted to a linear in-context learner in which CoT is formalized as iterative weight refinements. Extending the theory beyond isotropic linear regression and to richer CoT structures and realistic architectures, together with broader empirical validation, is an important direction for future work.

Acknowledgments and Disclosure of Funding

We thank Mary Letey and Alex Atanasov for helpful discussions. K.T. acknowledges support from the 2025 Young Researchers Overseas Dispatch Program, DAIKIN Advanced Interdisciplinary Research (AIR)-Vision Mobility Grant. K.T. was also supported by JST BOOST NAIS Grant Number JPMJBS2418. C.P. is supported by an NSF CAREER Award (IIS-2239780), DARPA grants DIAL-FP-038 and AIQ-HR00112520041, the Simons Collaboration on the Physics of Learning and Neural Computation, and the William F. Milton Fund from Harvard University. This work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence. The authors declare no competing interests.

References

[1] 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 Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors,

Advances in Neural Information Processing Systems, 2022. URL https://openreview.net/forum?id=_VjQlMeSB_J.

[2] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 22199–22213. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

[3] Charlie Victor Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling parameters for reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=4FWAwZtd2n.

[4] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Position: Will we run out of data? limits of llm scaling based on human-generated data. In *Forty-first International Conference on Machine Learning*, 2024.

[5] Ben Cottier, Robi Rahman, Loredana Fattorini, Nestor Maslej, Tamay Besiroglu, and David Owen. The rising costs of training frontier ai models. *arXiv preprint arXiv:2405.21015*, 2024.

[6] Wenkai Yang, Shuming Ma, Yankai Lin, and Furu Wei. Towards thinking-optimal scaling of test-time compute for LLM reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=6ICFqmixlS.

[7] Ryan Liu, Jiayi Geng, Addison J. Wu, Ilia Sucholutsky, Tania Lombrozo, and Thomas L. Griffiths. Mind your step (by step): Chain-of-thought can reduce performance on tasks where thinking makes humans worse, 2025. URL https://openreview.net/forum?id=rpbzBXdo4x.

[8] 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.

[9] 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.

[10] Adel Javanmard, Baharan Mirzasoleiman, and Vahab Mirrokni. Theoretical perspectives on data quality and synergistic effects in pre- and post-training reasoning models, 2026. URL https://arxiv.org/abs/2603.01293.

[11] Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori B Hashimoto. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20286–20332, 2025.

[12] OpenAI. OpenAI o3 and o4-mini System Card, 2025. URL https://api.semanticscholar.org/CorpusID:277857808.

[13] 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.

[14] Jian Wang, Boyan Zhu, Chak Tou Leong, Yongqi Li, and Wenjie Li. Scaling over scaling: Exploring test-time scaling plateau in large reasoning models, 2025. URL https://arxiv.org/abs/2505.20522.

[15] 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.

[16] Michael Hassid, Gabriel Synnaeve, Yossi Adi, and Roy Schwartz. Don't overthink it. preferring shorter thinking chains for improved LLM reasoning, 2026. URL https://openreview.net/forum?id=nhUlA8iMkD.

[17] Indranil Halder and Cengiz Pehlevan. Demystifying llm-as-a-judge: Analytically tractable model for inference-time scaling. *arXiv preprint arXiv:2512.19905*, 2025.

[18] Rylan Schaeffer, Joshua Kazdan, John Hughes, Jordan Juravsky, Sara Price, Aengus Lynch, Erik Jones, Robert Kirk, Azalia Mirhoseini, and Sanmi Koyejo. How do large language monkeys get their power (laws)? *arXiv preprint arXiv:2502.17578*, 2025.

[19] Audrey Huang, Adam Block, Qinghua Liu, Nan Jiang, Akshay Krishnamurthy, and Dylan J Foster. Is best-of-n the best of them? coverage, scaling, and optimality in inference-time alignment. *arXiv preprint arXiv:2503.21878*, 2025.

[20] Noam Levi. A simple model of inference scaling laws. In *Proceedings of the 42nd International Conference on Machine Learning*, pages 33984–33998. ML Research Press, 2025.

[21] Noam Itzhak Levi. Learning shrinks the hard tail: Training-dependent inference scaling in a solvable linear model. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=KUNywR7nQx.

[22] Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=flNZJ2eOet.

[23] 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.

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

[25] Yedi Zhang, Aaditya K Singh, Peter E Latham, and Andrew Saxe. Training dynamics of in-context learning in linear attention. *arXiv preprint arXiv:2501.16265*, 2025.

[26] 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.

[27] Mary Letey, Jacob A Zavatone-Veth, Yue M. Lu, and Cengiz Pehlevan. Pretrain–test task alignment governs generalization in in-context learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=KZLeg0MQ2r.

[28] Alex Nguyen and Gautam Reddy. Differential learning kinetics govern the transition from memorization to generalization during in-context learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=INyi7qUdjZ.

[29] Blake Bordelon, Mary Letey, and Cengiz Pehlevan. Theory of scaling laws for in-context regression: Depth, width, context and time. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qA42mWsnbl.

[30] Kaito Takanami, Takashi Takahashi, and Yoshiyuki Kabashima. Learning linear regression with low-rank tasks in-context. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pages 4672–4680. PMLR, 2026. URL https://proceedings.mlr.press/v300/takanami26a.html.

[31] Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, Charles Sutton, and Augustus Odena. Show your work: Scratchpads for intermediate computation with language models, 2022. URL https://openreview.net/forum?id=iedYJm92o0a.

[32] 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 *ES-FoMo III: 3rd Workshop on Efficient Systems for Foundation Models*, 2025. URL https://openreview.net/forum?id=D6o6Bwtq7h.

[33] Hayden Prairie, Zachary Novack, Taylor Berg-Kirkpatrick, and Daniel Y. Fu. Parcae: Scaling laws for stable looped language models, 2026. URL https://arxiv.org/abs/2604.12946.

[34] Harsh Kohli, Srinivasan Parthasarathy, Huan Sun, and Yuekun Yao. Loop, think, & generalize: Implicit reasoning in recurrent-depth transformers, 2026. URL https://arxiv.org/abs/2604.07822.

[35] Alexander Atanasov, Blake Bordelon, Jacob A Zavatone-Veth, Courtney Paquette, and Cengiz Pehlevan. Two-point deterministic equivalence for stochastic gradient dynamics in linear models. *arXiv [cond-mat.dis-nn]*, 10 November 2025. doi: 10.48550/arXiv.2502.05074. URL http://dx.doi.org/10.48550/arXiv.2502.05074.

[36] J William Helton, Tobias Mai, and Roland Speicher. Applications of realizations (aka linearizations) to free probability. *Journal of Functional Analysis*, 274(1):1–79, 2018.

[37] David Clark and Haim Sompolinsky. Simplified derivations for high-dimensional convex learning problems. *SciPost Physics Lecture Notes*, page 105, 2025.

[38] Yuyao Ge, Shenghua Liu, Yiwei Wang, Lingrui Mei, Lizhe Chen, Baolong Bi, and Xueqi Cheng. Innate reasoning is not enough: In-context learning enhances reasoning large language models with less overthinking, 2025. URL https://arxiv.org/abs/2503.19602.

Appendix

In this appendix, we present a systematic asymptotic derivation of the high-dimensional theory underlying our results. Our analysis is based on standard cavity and resolvent methods for random matrices and high-dimensional random systems. The derivation is not fully rigorous in the mathematical sense: several steps rely on self-averaging, leave-one-out replacements, and deterministic-equivalent substitutions that we do not justify here in complete detail. However, these manipulations are standard in the random matrix literature, and they lead to a closed system of equations that is internally consistent. Moreover, the resulting predictions are in excellent quantitative agreement with our numerical experiments, which provides strong evidence that the theory derived below correctly captures the asymptotic behavior of the model.

A Justification of parameter simplification

This section explains the simplified parameterization in Section 3 and gives empirical support for this choice. The main text fixes all blocks of $\mathbf{W}$ and all non-$\mathbf{A}$ blocks of $\mathbf{V}$, leaving only $\mathbf{A}$ as a learnable parameter. This reduction is motivated by the structure learned by the full-parameter model.

Only the third and fourth block columns of $\mathbf{W}$ and the third block row of $\mathbf{V}$ can affect the readout. With block sizes $(D, 1, D, 1)$, write

$$\mathbf{W} = \begin{pmatrix} \star & \star & W_{13} & W_{14} \\ \star & \star & W_{23} & W_{24} \\ \star & \star & W_{33} & W_{34} \\ \star & \star & W_{43} & W_{44} \end{pmatrix}, \qquad \mathbf{V} = \begin{pmatrix} \star & \star & \star & \star \\ \star & \star & \star & \star \\ V_{31} & V_{32} & V_{33} & V_{34} \\ \star & \star & \star & \star \end{pmatrix}, \tag{25}$$

where $\star$ is any conformable block that does not affect the readout. Rotational equivariance of the isotropic task distribution motivates setting $W_{14} = W_{23} = V_{32} = V_{34} = 0$ and $W_{13} = cI_D$. Requiring the ground-truth estimate to be a fixed point when $\boldsymbol{y} = \mathbf{X}^\top \boldsymbol{w}$ gives $W_{24} = -c$. We absorb the common scale c into V_{31} and set $c = 1$. We also set $V_{33} = 0$ to remove a history-only update that could change the estimate after the context residual vanishes.

We train the full model, where all entries of $\mathbf{W}$ and $\mathbf{V}$ are learnable, and show the learned matrices in Figure 3. The results show that the learned $\mathbf{W}$ develops the prescribed routing structure: the block mapping the parameter-estimate coordinates to the feature coordinates becomes close to an identity map, while the marker coordinate learns the corresponding sign structure. Similarly, the learned $\mathbf{V}$ concentrates its mass on the block that maps feature coordinates into the parameter-estimate coordinates, which is precisely the block represented by $\mathbf{A}$ in the simplified model. The remaining blocks stay close to zero.

This behavior becomes clearer as the number of training tasks and examples increases. Thus, the simplified parameterization should not be viewed as imposing an arbitrary constraint. Rather, for analytical tractability, it fixes part of the model to the parameter structure that is naturally approached under full-parameter training, while keeping the essential component $\mathbf{A}$, which governs the test-time update of the parameter estimate, learnable. The detailed setup of the full-parameter model is given in Appendix I.

B Derivation of the readout formula

In this appendix, we derive the readout formula in Eq. (8). First, recall that, for task μ, the prompt embedding is

$$\mathbf{E}_\mu = \begin{bmatrix} \mathbf{X}_\mu & \mathbf{0}_D \\ \boldsymbol{y}_\mu^\top & 0 \\ 0_{D\times L} & \hat{\boldsymbol{w}}_\mu^{\text{init}} \\ 0_{1\times L} & 1 \end{bmatrix} \in \mathbb{R}^{(2D+2)\times(L+1)}, \tag{26}$$

where

$$\mathbf{X}_\mu \in \mathbb{R}^{D\times L}, \qquad \boldsymbol{y}_\mu \in \mathbb{R}^L, \qquad \hat{\boldsymbol{w}}_\mu^{\text{init}} \in \mathbb{R}^D. \tag{27}$$

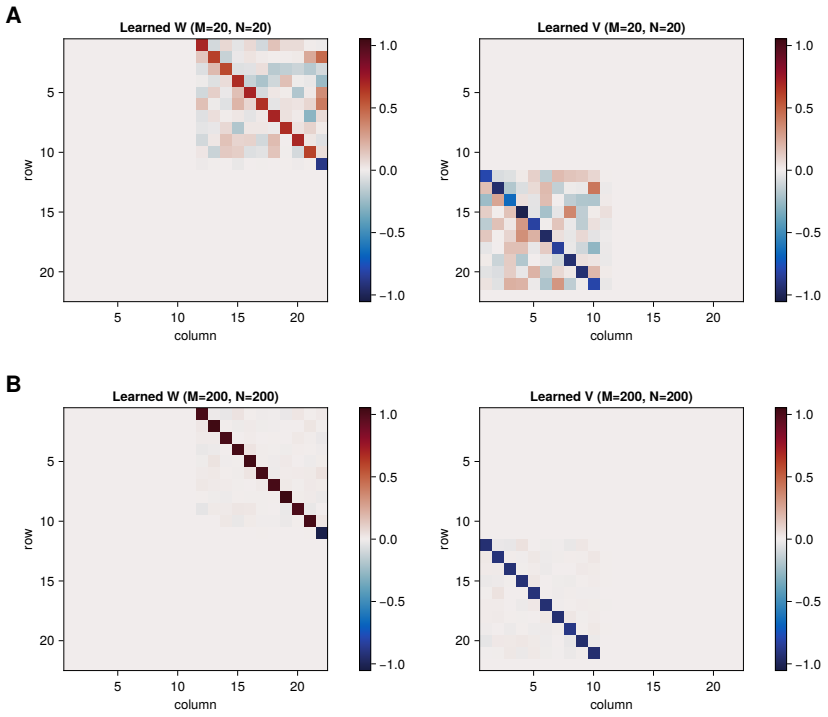

Figure 3: Heatmaps of the pretrained full-parameter matrices $\mathbf{W}$ and $\mathbf{V}$. Parameters: (A,B) $\lambda = 10^{-3}, \sigma = 0.1$ (A) $M = N = 20$, (B) $M = N = 200$.

We partition all matrices according to the block structure $(D, 1, D, 1)$, corresponding respectively to the feature block, the response block, the parameter-estimate block, and the marker block. The attention layer is

$$\textbf{Atten}(\mathbf{E}; \theta) = \mathbf{E} + \frac{1}{L}\mathbf{V}\mathbf{E}\mathbf{E}^\top\mathbf{W}\mathbf{E}, \tag{28}$$

and the readout extracts the parameter-estimate block from the final column:

$$\textbf{Read}(\mathbf{E}) := \mathbf{E}_{(D+2):(2D+1),(L+1)} \in \mathbb{R}^D. \tag{29}$$

To isolate the mechanism relevant to the parameter-estimate update, we retain only the blocks that are used in the parameterization of the main text and set the other potentially contributing blocks to zero. Specifically, we consider

$$\mathbf{W} = \begin{pmatrix} 0 & 0 & W_{13} & 0 \\ 0 & 0 & 0 & W_{24} \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \qquad \mathbf{V} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ V_{31} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}, \tag{30}$$

where $V_{31} \in \mathbb{R}^{D\times D}$, $W_{13} \in \mathbb{R}^{D\times D}$, and $W_{24} \in \mathbb{R}$.

Since the first L columns of $\mathbf{E}_\mu$ have nonzero entries only in the first two block rows, while the last column has nonzero entries only in the third and fourth block rows, the only part of $\mathbf{W}$ that affects the last column through the readout-relevant channel is

$$(\mathbf{W}\mathbf{E}_\mu)_{:,(L+1)} = \begin{bmatrix} W_{13}\hat{\boldsymbol{w}}_\mu^{\text{init}} \\ W_{24} \\ 0 \\ 0 \end{bmatrix}. \tag{31}$$

The lower two blocks are immaterial in what follows, because they do not contribute to the first block of $\mathbf{E}_\mu\mathbf{E}_\mu^\top\mathbf{W}\mathbf{E}_\mu$, which is the only part later selected by V_{31}. Multiplying by $\mathbf{E}_\mu^\top$ from the left, we

only need the contribution of the first two block rows of $(\mathbf{W}\mathbf{E}_\mu)_{:,(L+1)}$. This gives

$$\left(\mathbf{E}_\mu^\top \mathbf{W}\mathbf{E}_\mu\right)_{:,(L+1)} = \mathbf{E}_\mu^\top \begin{bmatrix} W_{13}\hat{\boldsymbol{w}}_\mu^{\text{init}} \\ W_{24} \\ 0 \\ 0 \end{bmatrix} \tag{32}$$

$$= \begin{bmatrix} \mathbf{X}_\mu^\top W_{13}\hat{\boldsymbol{w}}_\mu^{\text{init}} + \boldsymbol{y}_\mu W_{24} \\ 0 \end{bmatrix}. \tag{33}$$

Multiplying by $\mathbf{E}_\mu$ from the left, the first block of the last column becomes

$$\left(\mathbf{E}_\mu \mathbf{E}_\mu^\top \mathbf{W}\mathbf{E}_\mu\right)_{1,(L+1)} = \mathbf{X}_\mu \left(\mathbf{X}_\mu^\top W_{13}\hat{\boldsymbol{w}}_\mu^{\text{init}} + \boldsymbol{y}_\mu W_{24}\right). \tag{34}$$

Since only the first block is needed by V_{31}, this is the only term we keep. Because only the $(3,1)$ block of $\mathbf{V}$ contributes to the third row block, the update of the parameter-estimate slot is

$$\textbf{Read}(\textbf{Atten}(\mathbf{E}_\mu;\theta)) = \hat{\boldsymbol{w}}_\mu^{\text{init}} + \frac{1}{L} V_{31}\mathbf{X}_\mu \left(\mathbf{X}_\mu^\top W_{13}\hat{\boldsymbol{w}}_\mu^{\text{init}} + \boldsymbol{y}_\mu W_{24}\right). \tag{35}$$

This is the general readout formula induced by the three relevant blocks V_{31}, W_{13}, and W_{24}.

Here, we specialize to the parameterization:

$$V_{31} = -D\mathbf{A}, \qquad W_{13} = I_D, \qquad W_{24} = -1, \tag{36}$$

with $\mathbf{A} \in \mathbb{R}^{D\times D}$. Substituting these values into the previous expression gives

$$\textbf{Read}(\textbf{Atten}(\mathbf{E}_\mu;\theta)) = \hat{\boldsymbol{w}}_\mu^{\text{init}} - \frac{D}{L}\mathbf{A}\mathbf{X}_\mu \left(\mathbf{X}_\mu^\top \hat{\boldsymbol{w}}_\mu^{\text{init}} - \boldsymbol{y}_\mu\right) \tag{37}$$

$$= \hat{\boldsymbol{w}}_\mu^{\text{init}} - \frac{D}{L}\mathbf{A}\left(\mathbf{X}_\mu\mathbf{X}_\mu^\top \hat{\boldsymbol{w}}_\mu^{\text{init}} - \mathbf{X}_\mu\boldsymbol{y}_\mu\right), \tag{38}$$

which is Eq. (8). The choice $W_{13} = I_D$ means that the current estimate $\hat{\boldsymbol{w}}_\mu^{\text{init}}$ is passed to the interaction term without distortion. Likewise, $W_{24} = -1$ attaches the response vector with the opposite sign, so that the quantity inside the parentheses becomes

$$\mathbf{X}_\mu^\top \hat{\boldsymbol{w}}_\mu^{\text{init}} - \boldsymbol{y}_\mu, \tag{39}$$

which is exactly the vector of prediction residuals on the context examples. Therefore, with these two choices, the attention layer forms the standard regression error signal before applying the learned linear transform $\mathbf{A}$.

The remaining block $V_{31} = -D\mathbf{A}$ determines how this error signal is converted into an update of the parameter-estimate slot. The factor $\mathbf{X}_\mu$ maps residuals back to parameter space, while $\mathbf{A}$ acts as a learned preconditioner. In this sense, $W_{13} = I_D$ and $W_{24} = -1$ specify the canonical gradient-like structure, and the learnable content of the update is concentrated in $\mathbf{A}$.

Connection to gradient descent. Under this specialization, the term

$$\mathbf{X}_\mu\mathbf{X}_\mu^\top \hat{\boldsymbol{w}}_\mu^{\text{init}} - \mathbf{X}_\mu\boldsymbol{y}_\mu \tag{40}$$

is proportional to the empirical gradient of the squared loss

$$\mathcal{L}_\mu(\boldsymbol{w}) := \frac{1}{2L}\|\mathbf{X}_\mu^\top \boldsymbol{w} - \boldsymbol{y}_\mu\|_2^2, \tag{41}$$

since

$$\nabla\mathcal{L}_\mu(\boldsymbol{w}) = \frac{1}{L}\left(\mathbf{X}_\mu\mathbf{X}_\mu^\top \boldsymbol{w} - \mathbf{X}_\mu\boldsymbol{y}_\mu\right). \tag{42}$$

Hence,

$$\textbf{Read}(\textbf{Atten}(\mathbf{E}_\mu;\theta)) = \hat{\boldsymbol{w}}_\mu^{\text{init}} - D\mathbf{A}\nabla\mathcal{L}_\mu(\hat{\boldsymbol{w}}_\mu^{\text{init}}), \tag{43}$$

so the attention layer performs a preconditioned gradient step on the task parameter.

C Derivation of Result 1

In this appendix, we derive Result 1, which characterizes the asymptotic generalization error. The key object is the learned matrix $\mathbf{A}^* = \mathbf{B}\mathbf{C}^{-1}$, whose randomness is inherited from the pre-training tasks. Since the inference error depends on both the random matrix $\mathbf{A}^*$ and the inference-time sample covariance $\mathbf{S} = \mathbf{X}\mathbf{X}^\top$, the main challenge is to control their coupled effect in the high-dimensional limit.

Our derivation proceeds in four steps. First, in Appendix C.2, we rewrite the generalization error as a finite linear combination of derivatives of a two-variable generating function $F(u, v)$. This step reduces the original prediction problem to the evaluation of a resolvent-type quantity associated with the inference dynamics. Next, in Appendix C.3, we introduce an extended block matrix $\mathcal{M}(u, v)$ whose resolvent contains $F(u, v)$ as the normalized trace of a single block. This reformulation is essential because it removes the inverse $\mathbf{C}^{-1}$ appearing in $\mathbf{A}^* = \mathbf{B}\mathbf{C}^{-1}$, thereby converting the problem into a more tractable linear resolvent analysis.

The next task is to characterize the resolvent of $\mathcal{M}(u, v)$ through a finite collection of scalar order parameters. To this end, in Appendix C.4, we define the relevant normalized traces and derive the exact algebraic identities that follow from $\mathbf{G}\mathcal{M}(u, v) = \mathbf{I}$, where $\mathbf{G} = \mathcal{M}(u, v)^{-1}$. These identities alone do not close, because they involve mixed traces with insertions of $\mathbf{S}$, $\mathbf{B}$, $\mathbf{B}^\top$, and $\mathbf{C}$. We therefore derive cavity equations for these mixed quantities: Appendix C.5 treats the $\mathbf{S}$-dependent terms by a leave-one-sample-out argument, while Appendices C.6, C.7, and C.8 treat the $\mathbf{B}$-, $\mathbf{B}^\top$-, and $\mathbf{C}$-dependent terms by leave-one-task-out arguments. Combining these relations yields, in Appendix C.9, a closed deterministic self-consistent system for the primary order parameters.

Finally, in Appendix C.10, we solve this closed system explicitly in the limit $\lambda \to 0$. The solution is expressed in terms of the scalar resolvent $g(u)$ of a Wishart-type matrix, and in particular yields a closed-form expression for the block $m_{13}(u, v)$. Substituting this expression into the derivative formula obtained in Appendix C.2 gives the claimed representation of the asymptotic generalization error. In this way, the proof isolates the effect of pre-training through the finite-dimensional order-parameter system and makes the dependence of the final error on α, τ, and σ^2 fully explicit.

We begin by restating the result.

Result 5. *The generalization error at CoT depth t is asymptotically given by*

$$\mathcal{E}_t = \sum_{p=0}^{t}\sum_{q=0}^{t}\binom{t}{p}\binom{t}{q}\left(-\frac{1}{\alpha}\right)^{p+q}\frac{1}{p!\,q!}\frac{\partial^p}{\partial u^p}\frac{\partial^q}{\partial v^q}F(u,v)\Big|_{u=v=0}. \tag{15}$$

Here, especially when $\lambda = 0$, $F(u, v)$ is explicitly given by

$$F(u,v) = \frac{\alpha\, g(u)\, g(v)}{\alpha - (g(u)-1)(g(v)-1)\left[1 + \frac{1+\sigma^2}{\alpha(\tau-1)}(\alpha + g(u) + g(v) - 1)\right]}, \tag{16}$$

where $g(u)$ is the resolvent function of Wishart matrix, which is given by

$$g(u) = \frac{c + u(1-\alpha) - \sqrt{\big(c + u(1-\alpha)\big)^2 - 4cu}}{2u}. \tag{17}$$

with $c = 1 + (1+\sigma^2)/\alpha$.

We now derive this formula in a sequence of reductions.

C.1 Preliminaries

For clarity, we first introduce the notation used throughout the appendix. For each pre-training task $\mu = 1, \ldots, M$, define

$$\boldsymbol{v}_\mu := \frac{1}{\alpha}\mathbf{X}_\mu \boldsymbol{y}_\mu \in \mathbb{R}^D, \qquad \alpha = \frac{L}{D}, \qquad \tau = \frac{M}{D}, \tag{44}$$

and recall that, for $k = 0$, the empirical minimizer $\mathbf{A}^*$ is given by

$$\mathbf{A}^* = \left[\frac{1}{M}\sum_{\mu=1}^{M}\boldsymbol{w}_\mu \boldsymbol{v}_\mu^\top\right]\left[\frac{1}{M}\sum_{\mu=1}^{M}\boldsymbol{v}_\mu\boldsymbol{v}_\mu^\top + \frac{\lambda}{2}\mathbf{I}\right]^{-1}. \tag{45}$$

Accordingly, throughout this appendix we write

$$\mathbf{B} := \frac{1}{M}\sum_{\mu=1}^{M} \boldsymbol{w}_\mu \boldsymbol{v}_\mu^\top, \qquad \mathbf{C} := \frac{1}{M}\sum_{\mu=1}^{M} \boldsymbol{v}_\mu \boldsymbol{v}_\mu^\top + \frac{\lambda}{2}\mathbf{I}, \qquad \mathbf{A} := \mathbf{B}\mathbf{C}^{-1}, \tag{46}$$

so that $\mathbf{A} = \mathbf{A}^*$. For the inference-time task, we define

$$\mathbf{S} := \mathbf{X}\mathbf{X}^\top = \sum_{\ell=1}^{L} \boldsymbol{x}_\ell \boldsymbol{x}_\ell^\top \in \mathbb{R}^{D\times D}. \tag{47}$$

In the large-D analysis below, we repeatedly use the effective second-order statistics of the pre-training pair $(\boldsymbol{w}_\mu, \boldsymbol{v}_\mu)$. Since $\boldsymbol{v}_\mu = \alpha^{-1}\mathbf{X}_\mu \boldsymbol{y}_\mu$ and $\boldsymbol{y}_\mu = \mathbf{X}_\mu^\top \boldsymbol{w}_\mu + \boldsymbol{\epsilon}_\mu$, the isotropy of the Gaussian design implies

$$\frac{1}{D}\mathbb{E}\|\boldsymbol{w}_\mu\|^2 = 1, \qquad \frac{1}{D}\mathbb{E}\big[\boldsymbol{w}_\mu^\top \boldsymbol{v}_\mu\big] = 1, \qquad \frac{1}{D}\mathbb{E}\|\boldsymbol{v}_\mu\|^2 \to c, \tag{48}$$

where

$$c := 1 + \frac{1+\sigma^2}{\alpha}. \tag{49}$$

Equivalently, at the matrix level,

$$\mathbb{E}[\boldsymbol{w}_\mu \boldsymbol{w}_\mu^\top] = I_D, \qquad \mathbb{E}[\boldsymbol{w}_\mu \boldsymbol{v}_\mu^\top] = I_D, \qquad \frac{1}{D}\operatorname{Tr}\mathbb{E}[\boldsymbol{v}_\mu \boldsymbol{v}_\mu^\top] \to c. \tag{50}$$

Therefore, whenever $Q^{[\mu]} \in \mathbb{R}^{D\times D}$ is independent of $(\boldsymbol{w}_\mu, \boldsymbol{v}_\mu)$ and has bounded normalized trace norm, the corresponding quadratic forms self-average as

$$\begin{aligned} \frac{1}{D}\boldsymbol{w}_\mu^\top Q^{[\mu]} \boldsymbol{w}_\mu &\asymp \frac{1}{D}\operatorname{Tr} Q^{[\mu]}, \\ \frac{1}{D}\boldsymbol{v}_\mu^\top Q^{[\mu]} \boldsymbol{w}_\mu &\asymp \frac{1}{D}\operatorname{Tr} Q^{[\mu]}, \\ \frac{1}{D}\boldsymbol{v}_\mu^\top Q^{[\mu]} \boldsymbol{v}_\mu &\asymp c\,\frac{1}{D}\operatorname{Tr} Q^{[\mu]}. \end{aligned} \tag{51}$$

Later, $Q^{[\mu]}$ will typically be a block of a leave-one-task-out resolvent $\mathbf{G}^{[\mu]}$, so (51) is precisely the input used when replacing inner products such as $\boldsymbol{v}_\mu^\top \mathbf{G}_{ab}^{[\mu]} \boldsymbol{w}_\mu$ and $\boldsymbol{v}_\mu^\top \mathbf{G}_{ab}^{[\mu]} \boldsymbol{v}_\mu$ by normalized traces.

C.2 Reduction of the generalization error to a generating function

We first show that the generalization error can be expressed in terms of a two-variable generating function. This representation is the starting point of the random matrix analysis.

When $k = 0$, the pre-training prompt contains only the initial estimate $\hat{\boldsymbol{w}}_{\mu,0} = \mathbf{0}$, and the learned matrix $\mathbf{A} = \mathbf{A}^*$ is used at inference time to update the parameter estimate by

$$\hat{\boldsymbol{w}}_{t+1} = \hat{\boldsymbol{w}}_t - \frac{1}{\alpha}\mathbf{A}\left(\mathbf{S}\hat{\boldsymbol{w}}_t - \mathbf{X}\boldsymbol{y}\right), \qquad \hat{\boldsymbol{w}}_0 = \mathbf{0}, \tag{52}$$

where $\mathbf{S} = \mathbf{X}\mathbf{X}^\top$ and $\boldsymbol{y} = \mathbf{X}^\top \boldsymbol{w}$ in the noiseless inference task. Since $\mathbf{X}\boldsymbol{y} = \mathbf{X}\mathbf{X}^\top \boldsymbol{w} = \mathbf{S}\boldsymbol{w}$, this recursion becomes

$$\hat{\boldsymbol{w}}_{t+1} = \hat{\boldsymbol{w}}_t - \frac{1}{\alpha}\mathbf{A}\mathbf{S}(\hat{\boldsymbol{w}}_t - \boldsymbol{w}). \tag{53}$$

Therefore, the estimation error $\boldsymbol{e}_t := \boldsymbol{w} - \hat{\boldsymbol{w}}_t$ satisfies the linear recursion

$$\boldsymbol{e}_{t+1} = \left(\mathbf{I} - \frac{1}{\alpha}\mathbf{A}\mathbf{S}\right)\boldsymbol{e}_t, \qquad \boldsymbol{e}_0 = \boldsymbol{w}. \tag{54}$$

Iterating this relation yields

$$\boldsymbol{w} - \hat{\boldsymbol{w}}_t = \left(\mathbf{I} - \frac{1}{\alpha}\mathbf{A}\mathbf{S}\right)^t \boldsymbol{w}. \tag{55}$$

Using this identity, the generalization error can be written as

$$\mathcal{E}_t := \frac{1}{D}\mathbb{E}\big[\|\boldsymbol{w} - \hat{\boldsymbol{w}}_t\|^2\big] = \frac{1}{D}\mathbb{E}\left[\boldsymbol{w}^\top \left(\mathbf{I} - \frac{1}{\alpha}\mathbf{S}\mathbf{A}^\top\right)^t \left(\mathbf{I} - \frac{1}{\alpha}\mathbf{A}\mathbf{S}\right)^t \boldsymbol{w}\right]. \tag{56}$$

Since $\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_D)$ is independent of the inference design matrix $\mathbf{X}$, we may average over $\boldsymbol{w}$ first and obtain

$$\mathcal{E}_t = \frac{1}{D}\mathbb{E}\,\mathrm{Tr}\left[\left(\mathbf{I} - \frac{1}{\alpha}\mathbf{A}\mathbf{S}\right)^t \left(\mathbf{I} - \frac{1}{\alpha}\mathbf{S}\mathbf{A}^\top\right)^t\right]. \tag{57}$$

Here and below, $\mathbb{E}$ denotes the expectation over both the pre-training randomness defining $\mathbf{A}$ and the inference-time randomness defining $\mathbf{S}$.

To extract this quantity, we introduce the two-variable generating function

$$F(u, v) := \frac{1}{D}\mathbb{E}\,\mathrm{Tr}\Big[(\mathbf{I} - u\mathbf{A}\mathbf{S})^{-1}(\mathbf{I} - v\mathbf{S}\mathbf{A}^\top)^{-1}\Big]. \tag{58}$$

Expanding both resolvents around $u = v = 0$, we have the absolutely formal power series

$$F(u, v) = \sum_{p,q\geq 0} u^p v^q \,\frac{1}{D}\mathbb{E}\,\mathrm{Tr}\big[(\mathbf{A}\mathbf{S})^p(\mathbf{S}\mathbf{A}^\top)^q\big]. \tag{59}$$

On the other hand, expanding the powers in (57) by the binomial formula gives

$$\begin{aligned}\mathcal{E}_t &= \frac{1}{D}\mathbb{E}\,\mathrm{Tr}\left[\sum_{p=0}^{t}\binom{t}{p}\left(-\frac{1}{\alpha}\right)^p (\mathbf{A}\mathbf{S})^p \sum_{q=0}^{t}\binom{t}{q}\left(-\frac{1}{\alpha}\right)^q (\mathbf{S}\mathbf{A}^\top)^q\right] \\ &= \sum_{p=0}^{t}\sum_{q=0}^{t}\binom{t}{p}\binom{t}{q}\left(-\frac{1}{\alpha}\right)^{p+q} \frac{1}{D}\mathbb{E}\,\mathrm{Tr}\big[(\mathbf{A}\mathbf{S})^p(\mathbf{S}\mathbf{A}^\top)^q\big].\end{aligned} \tag{60}$$

Comparing this with (59), we conclude that

$$\mathcal{E}_t = \sum_{p=0}^{t}\sum_{q=0}^{t}\binom{t}{p}\binom{t}{q}\left(-\frac{1}{\alpha}\right)^{p+q} \frac{1}{p!\,q!}\partial_u^p\partial_v^q F(u, v)\Big|_{u=v=0}. \tag{61}$$

Equation (61) shows that the asymptotic evaluation of $\mathcal{E}_t$ reduces to the determination of the scalar function $F(u, v)$. The remaining task is therefore to compute the large-D limit of $F(u, v)$.

C.3 Extended matrix representation of the resolvent

We next rewrite $F(u, v)$ as a block trace of the resolvent of a larger matrix. This step is crucial because the matrix $\mathbf{A} = \mathbf{B}\mathbf{C}^{-1}$ contains the inverse $\mathbf{C}^{-1}$, which is inconvenient to manipulate directly. The extended matrix formulation removes this inverse by embedding $\mathbf{A}$ into a linear block system.

Recall that

$$\mathbf{A} = \mathbf{B}\mathbf{C}^{-1}, \qquad \mathbf{A}^\top = \mathbf{C}^{-1}\mathbf{B}^\top, \tag{62}$$

where

$$\mathbf{B} = \frac{1}{M}\sum_{\mu=1}^{M}\boldsymbol{w}_\mu\boldsymbol{v}_\mu^\top, \qquad \mathbf{C} = \frac{1}{M}\sum_{\mu=1}^{M}\boldsymbol{v}_\mu\boldsymbol{v}_\mu^\top + \frac{\lambda}{2}\mathbf{I}. \tag{63}$$

We introduce the extended block matrix

$$\mathcal{M}(u, v) := \begin{pmatrix} \mathbf{I} & -u\mathbf{B} & -\mathbf{I} & 0 \\ -\mathbf{S} & \mathbf{C} & 0 & 0 \\ 0 & 0 & \mathbf{I} & -v\mathbf{S} \\ 0 & 0 & -\mathbf{B}^\top & \mathbf{C} \end{pmatrix} \in \mathbb{R}^{4D\times 4D}, \tag{64}$$

and denote its inverse by

$$\mathbf{G}(u,v) := \mathcal{M}(u,v)^{-1}. \tag{65}$$

We write $\mathbf{G}$ in $D \times D$ block form as

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} & \mathbf{G}_{13} & \mathbf{G}_{14} \\ \mathbf{G}_{21} & \mathbf{G}_{22} & \mathbf{G}_{23} & \mathbf{G}_{24} \\ 0 & 0 & \mathbf{G}_{33} & \mathbf{G}_{34} \\ 0 & 0 & \mathbf{G}_{43} & \mathbf{G}_{44} \end{pmatrix}. \tag{66}$$

The lower-left zero block follows from the upper block-triangular structure of $\mathcal{M}(u,v)$.

To identify $\mathbf{G}_{13}$, it is convenient to decompose $\mathcal{M}(u,v)$ as

$$\mathcal{M}(u,v) = \begin{pmatrix} M_{11} & M_{12} \\ 0 & M_{22} \end{pmatrix}, \tag{67}$$

where

$$M_{11} := \begin{pmatrix} \mathbf{I} & -u\mathbf{B} \\ -\mathbf{S} & \mathbf{C} \end{pmatrix}, \qquad M_{12} := \begin{pmatrix} -\mathbf{I} & 0 \\ 0 & 0 \end{pmatrix}, \qquad M_{22} := \begin{pmatrix} \mathbf{I} & -v\mathbf{S} \\ -\mathbf{B}^\top & \mathbf{C} \end{pmatrix}. \tag{68}$$

Since $\mathcal{M}(u,v)$ is upper block triangular, its inverse is

$$\mathbf{G} = \mathcal{M}(u,v)^{-1} = \begin{pmatrix} M_{11}^{-1} & -M_{11}^{-1} M_{12} M_{22}^{-1} \\ 0 & M_{22}^{-1} \end{pmatrix}. \tag{69}$$

Hence the $(1,3)$-block of $\mathbf{G}$ is given by

$$\mathbf{G}_{13} = \left(-M_{11}^{-1} M_{12} M_{22}^{-1}\right)_{11} = \left(M_{11}^{-1}\right)_{11} \left(M_{22}^{-1}\right)_{11}. \tag{70}$$

We now evaluate these two factors by the Schur complement formula. For M_{11}, the $(1,1)$-block of the inverse is

$$\left(M_{11}^{-1}\right)_{11} = \Big(\mathbf{I} - (-u\mathbf{B})\mathbf{C}^{-1}(-\mathbf{S})\Big)^{-1} = (\mathbf{I} - u\mathbf{B}\mathbf{C}^{-1}\mathbf{S})^{-1} = (\mathbf{I} - u\mathbf{A}\mathbf{S})^{-1}. \tag{71}$$

Similarly, for M_{22}, we obtain

$$\left(M_{22}^{-1}\right)_{11} = \Big(\mathbf{I} - (-v\mathbf{S})\mathbf{C}^{-1}(-\mathbf{B}^\top)\Big)^{-1} = (\mathbf{I} - v\mathbf{S}\mathbf{C}^{-1}\mathbf{B}^\top)^{-1} = (\mathbf{I} - v\mathbf{S}\mathbf{A}^\top)^{-1}. \tag{72}$$

Substituting these identities into (70), we arrive at

$$\mathbf{G}_{13} = (\mathbf{I} - u\mathbf{A}\mathbf{S})^{-1}(\mathbf{I} - v\mathbf{S}\mathbf{A}^\top)^{-1}. \tag{73}$$

Taking the normalized trace and expectation, we obtain the desired representation of $F(u,v)$:

$$F(u,v) = \frac{1}{D}\mathbb{E}\,\mathrm{Tr}(\mathbf{G}_{13}). \tag{74}$$

Thus, the original problem has been reduced to the analysis of one block of the resolvent of the extended matrix $\mathcal{M}(u,v)$.

In the remainder of the proof, we study the large-D behavior of $\mathbf{G} = \mathcal{M}(u,v)^{-1}$ through a finite set of normalized traces, which we call order parameters. The key point is that the block structure of $\mathcal{M}(u,v)$ allows these order parameters to satisfy a closed deterministic system in the high-dimensional limit. Once that system is solved, (74) yields $F(u,v)$, and then (61) gives the asymptotic generalization error.

Similarly, considering the average error when μ is chosen uniformly at random from the set of trained tasks is given by

$$\mathcal{E}_t(\mathbf{A}) = \frac{1}{D}\mathbb{E}_X\ \mathrm{Tr}\left(\left(\mathbf{I} - \frac{D}{L}\mathbf{A}\mathbf{X}\mathbf{X}^\top\right)^t \boldsymbol{w}_\mu \boldsymbol{w}_\mu^\top \left(\mathbf{I} - \frac{D}{L}\mathbf{X}\mathbf{X}^\top\mathbf{A}^\top\right)^t\right) \tag{75}$$

$$= \frac{1}{D}\mathbb{E}_X\ \mathrm{Tr}\left(\left(\mathbf{I} - \frac{D}{L}\mathbf{A}\mathbf{X}\mathbf{X}^\top\right)^t \left(\frac{1}{M}\sum_\mu \boldsymbol{w}_\mu \boldsymbol{w}_\mu^\top\right)\left(\mathbf{I} - \frac{D}{L}\mathbf{X}\mathbf{X}^\top\mathbf{A}^\top\right)^t\right) \tag{76}$$

$$= \sum_{p=0}^{t}\sum_{q=0}^{t}\binom{t}{p}\binom{t}{q}\left(-\frac{1}{\alpha}\right)^{p+q}\frac{1}{p!\,q!}\,\partial_u^p \partial_v^q \tilde{F}(u,v)\Big|_{u=v=0} \tag{77}$$

where

$$\tilde{F}(u,v) = \frac{1}{D}\mathbb{E}_X \ \mathrm{Tr}\left[(\mathbf{I} - u\mathbf{A}\mathbf{S})^{-1}\left(\frac{1}{M}\sum_{\mu}\boldsymbol{w}_\mu\boldsymbol{w}_\mu^\top\right)(\mathbf{I} - v\mathbf{S}\mathbf{A}^\top)^{-1}\right] \tag{78}$$

$$= \frac{1}{D}\mathbb{E}_X \ \mathrm{tr}\left(\mathbf{G}_{13}\left(\frac{1}{M}\sum_{\mu}\boldsymbol{w}_\mu\boldsymbol{w}_\mu^\top\right)\right). \tag{79}$$

C.4 Order parameters and basic identities from $\mathbf{G}\mathcal{M}(u,v) = \mathbf{I}$

We now introduce the normalized traces that will be used to characterize the resolvent $\mathbf{G} = \mathcal{M}(u,v)^{-1}$. Our aim is to reduce the full matrix-valued problem to a finite set of scalar quantities that remain deterministic in the high-dimensional limit. These quantities are chosen so that the block matrix identity

$$\mathbf{G}\mathcal{M}(u,v) = \mathbf{I}_{4D} \tag{80}$$

closes into a finite system once combined with the cavity relations derived in the following subsections.

Order parameters. Recall that $\mathbf{G}$ is written in $D \times D$ blocks as

$$\mathbf{G} = \begin{pmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} & \mathbf{G}_{13} & \mathbf{G}_{14} \\ \mathbf{G}_{21} & \mathbf{G}_{22} & \mathbf{G}_{23} & \mathbf{G}_{24} \\ 0 & 0 & \mathbf{G}_{33} & \mathbf{G}_{34} \\ 0 & 0 & \mathbf{G}_{43} & \mathbf{G}_{44} \end{pmatrix}. \tag{81}$$

We first define the normalized traces of these blocks:

$$\begin{aligned} m_{11} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{11}), & m_{12} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{12}), & m_{13} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{13}), & m_{14} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{14}), \\ m_{21} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{21}), & m_{22} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{22}), & m_{23} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{23}), & m_{24} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{24}), \\ m_{33} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{33}), & m_{34} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{34}), & m_{43} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{43}), & m_{44} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{44}). \end{aligned} \tag{82}$$

Next, since the block equations generated by $\mathbf{G}\mathcal{M} = \mathbf{I}$ involve the matrices $\mathbf{S}$, $\mathbf{B}$, $\mathbf{B}^\top$, and $\mathbf{C}$, we also introduce mixed normalized traces in which these matrices are inserted. For the terms involving the inference covariance $\mathbf{S}$, we define

$$\begin{aligned} s_{12} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{12}\mathbf{S}), & s_{13} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{13}\mathbf{S}), & s_{22} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{22}\mathbf{S}), \\ s_{23} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{23}\mathbf{S}), & s_{33} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{33}\mathbf{S}), & s_{43} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{43}\mathbf{S}). \end{aligned} \tag{83}$$

For the terms involving the task-dependent matrix $\mathbf{B}$, we define

$$b_{11} := \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{11}\mathbf{B}), \qquad b_{21} := \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{21}\mathbf{B}), \tag{84}$$

and for the terms involving $\mathbf{B}^\top$, we define

$$\begin{aligned} \beta_{14} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{14}\mathbf{B}^\top), & \beta_{24} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{24}\mathbf{B}^\top), \\ \beta_{34} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{34}\mathbf{B}^\top), & \beta_{44} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{44}\mathbf{B}^\top). \end{aligned} \tag{85}$$

Finally, for the terms involving $\mathbf{C}$, we define

$$\begin{aligned} c_{12} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{12}\mathbf{C}), & c_{14} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{14}\mathbf{C}), \\ c_{22} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{22}\mathbf{C}), & c_{24} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{24}\mathbf{C}), \\ c_{34} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{34}\mathbf{C}), & c_{44} &:= \frac{1}{D}\mathrm{Tr}(\mathbf{G}_{44}\mathbf{C}). \end{aligned} \tag{86}$$

The quantity of ultimate interest is m_{13}, because by (74) we have

$$F(u,v) = \frac{1}{D}\mathbb{E}\,\mathrm{Tr}(\mathbf{G}_{13}) = \mathbb{E}[m_{13}]. \tag{87}$$

The remaining order parameters are auxiliary variables introduced only to obtain a closed system.

Block equations from $\mathbf{G}\mathcal{M} = \mathbf{I}$. We now derive the basic algebraic relations satisfied by these order parameters. Multiplying $\mathbf{G}$ and $\mathcal{M}(u,v)$, we obtain

$$\begin{aligned}
&\mathbf{G}\mathcal{M}(u,v) \\
&= \begin{pmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} & \mathbf{G}_{13} & \mathbf{G}_{14} \\ \mathbf{G}_{21} & \mathbf{G}_{22} & \mathbf{G}_{23} & \mathbf{G}_{24} \\ 0 & 0 & \mathbf{G}_{33} & \mathbf{G}_{34} \\ 0 & 0 & \mathbf{G}_{43} & \mathbf{G}_{44} \end{pmatrix} \begin{pmatrix} \mathbf{I} & -u\mathbf{B} & -\mathbf{I} & 0 \\ -\mathbf{S} & \mathbf{C} & 0 & 0 \\ 0 & 0 & \mathbf{I} & -v\mathbf{S} \\ 0 & 0 & -\mathbf{B}^\top & \mathbf{C} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{G}_{11} - \mathbf{G}_{12}\mathbf{S} & -u\mathbf{G}_{11}\mathbf{B} + \mathbf{G}_{12}\mathbf{C} & -\mathbf{G}_{11} + \mathbf{G}_{13} - \mathbf{G}_{14}\mathbf{B}^\top & -v\mathbf{G}_{13}\mathbf{S} + \mathbf{G}_{14}\mathbf{C} \\ \mathbf{G}_{21} - \mathbf{G}_{22}\mathbf{S} & -u\mathbf{G}_{21}\mathbf{B} + \mathbf{G}_{22}\mathbf{C} & -\mathbf{G}_{21} + \mathbf{G}_{23} - \mathbf{G}_{24}\mathbf{B}^\top & -v\mathbf{G}_{23}\mathbf{S} + \mathbf{G}_{24}\mathbf{C} \\ 0 & 0 & \mathbf{G}_{33} - \mathbf{G}_{34}\mathbf{B}^\top & -v\mathbf{G}_{33}\mathbf{S} + \mathbf{G}_{34}\mathbf{C} \\ 0 & 0 & \mathbf{G}_{43} - \mathbf{G}_{44}\mathbf{B}^\top & -v\mathbf{G}_{43}\mathbf{S} + \mathbf{G}_{44}\mathbf{C} \end{pmatrix}.
\end{aligned} \tag{88}$$

Since $\mathbf{G}\mathcal{M}(u,v) = \mathbf{I}_{4D}$, each block must match the corresponding block of the identity matrix. Taking the normalized trace of each nontrivial block yields the following scalar relations.

From the $(1,1)$, $(1,2)$, $(1,3)$, and $(1,4)$ blocks, we obtain

$$m_{11} - s_{12} = 1, \tag{89}$$
$$-u\, b_{11} + c_{12} = 0, \tag{90}$$
$$-m_{11} + m_{13} - \beta_{14} = 0, \tag{91}$$
$$-v\, s_{13} + c_{14} = 0. \tag{92}$$

From the $(2,1)$, $(2,2)$, $(2,3)$, and $(2,4)$ blocks, we obtain

$$m_{21} - s_{22} = 0, \tag{93}$$
$$-u\, b_{21} + c_{22} = 1, \tag{94}$$
$$-m_{21} + m_{23} - \beta_{24} = 0, \tag{95}$$
$$-v\, s_{23} + c_{24} = 0. \tag{96}$$

From the $(3,3)$, $(3,4)$, $(4,3)$, and $(4,4)$ blocks, we obtain

$$m_{33} - \beta_{34} = 1, \tag{97}$$
$$-v\, s_{33} + c_{34} = 0, \tag{98}$$
$$m_{43} - \beta_{44} = 0, \tag{99}$$
$$-v\, s_{43} + c_{44} = 1. \tag{100}$$

Equations (89)–(100) are exact identities that hold for every realization of the randomness. They express the primary traces m_{ab} in terms of the mixed traces involving $\mathbf{S}$, $\mathbf{B}$, $\mathbf{B}^\top$, and $\mathbf{C}$. At this stage the system is not yet closed, because the quantities s_{ab}, b_{ab}, β_{ab}, and c_{ab} still depend on the full matrix structure of $\mathbf{G}$.

The role of the cavity method is precisely to close this system. In the next subsection, we analyze the dependence on the inference samples $\{\boldsymbol{x}_\ell\}_{\ell=1}^L$ and express the s-variables in terms of the m-variables alone. Subsequently, by removing one pre-training task at a time, we derive analogous relations for b, β, and c. Once these relations are combined with (89)–(100), the order parameters satisfy a closed deterministic system in the high-dimensional limit.

C.5 Cavity equations for the S-dependent order parameters

We next derive closed equations for the quantities $s_{12}, s_{13}, s_{22}, s_{23}, s_{33}, s_{43}$, which contain the inference-time sample covariance $\mathbf{S} = \sum_{\ell=1}^L \boldsymbol{x}_\ell \boldsymbol{x}_\ell^\top$. The key idea is to remove one inference sample $\boldsymbol{x}_\ell$ from $\mathbf{S}$, compare the full resolvent with the leave-one-out resolvent, and then use concentration of quadratic forms. Since the randomness of $\mathbf{S}$ comes from the independent Gaussian vectors $\{\boldsymbol{x}_\ell\}_{\ell=1}^L$, this leave-one-out argument closes the s-variables in terms of the basic traces m_{ab}.

Leave-one-out decomposition. Fix $\ell \in \{1, \dots, L\}$, and define the leave-one-out sample covariance

$$\mathbf{S}^{[\ell]} := \sum_{j \neq \ell} \boldsymbol{x}_j \boldsymbol{x}_j^\top = \mathbf{S} - \boldsymbol{x}_\ell \boldsymbol{x}_\ell^\top. \tag{101}$$

Correspondingly, we introduce the leave-one-out extended matrix

$$\mathcal{M}^{[\ell]}(u, v) := \begin{pmatrix} \mathbf{I} & -u\mathbf{B} & -\mathbf{I} & 0 \\ -\mathbf{S}^{[\ell]} & \mathbf{C} & 0 & 0 \\ 0 & 0 & \mathbf{I} & -v\mathbf{S}^{[\ell]} \\ 0 & 0 & -\mathbf{B}^\top & \mathbf{C} \end{pmatrix}, \tag{102}$$

and its inverse

$$\mathbf{G}^{[\ell]} := \big(\mathcal{M}^{[\ell]}(u, v)\big)^{-1}. \tag{103}$$

Then the full matrix can be written as

$$\mathcal{M}(u, v) = \mathcal{M}^{[\ell]}(u, v) + \Delta_\ell, \tag{104}$$

where

$$\Delta_\ell = \begin{pmatrix} 0 & 0 & 0 & 0 \\ -\boldsymbol{x}_\ell \boldsymbol{x}_\ell^\top & 0 & 0 & 0 \\ 0 & 0 & 0 & -v\, \boldsymbol{x}_\ell \boldsymbol{x}_\ell^\top \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{105}$$

This perturbation has rank 2, and it is convenient to factorize it as

$$\Delta_\ell = U_\ell V_\ell^\top, \tag{106}$$

with

$$U_\ell := \begin{pmatrix} 0 & 0 \\ -\boldsymbol{x}_\ell & 0 \\ 0 & -\sqrt{v}\, \boldsymbol{x}_\ell \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{4D \times 2}, \qquad V_\ell := \begin{pmatrix} \boldsymbol{x}_\ell & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & \sqrt{v}\, \boldsymbol{x}_\ell \end{pmatrix} \in \mathbb{R}^{4D \times 2}. \tag{107}$$

Resolvent identity. Applying the Woodbury formula to $\mathcal{M} = \mathcal{M}^{[\ell]} + U_\ell V_\ell^\top$, we obtain

$$\mathbf{G} = \mathbf{G}^{[\ell]} - \mathbf{G}^{[\ell]} U_\ell \Big(\mathbf{I}_2 + V_\ell^\top \mathbf{G}^{[\ell]} U_\ell\Big)^{-1} V_\ell^\top \mathbf{G}^{[\ell]}. \tag{108}$$

We denote the 2×2 matrix in the middle by

$$\mathbf{K}_\ell := V_\ell^\top \mathbf{G}^{[\ell]} U_\ell. \tag{109}$$

Using the explicit forms (107), a direct block computation gives

$$\mathbf{K}_\ell = \begin{pmatrix} -\boldsymbol{x}_\ell^\top \mathbf{G}^{[\ell]}_{12} \boldsymbol{x}_\ell & -\sqrt{v}\, \boldsymbol{x}_\ell^\top \mathbf{G}^{[\ell]}_{13} \boldsymbol{x}_\ell \\ 0 & -v\, \boldsymbol{x}_\ell^\top \mathbf{G}^{[\ell]}_{43} \boldsymbol{x}_\ell \end{pmatrix}. \tag{110}$$

The advantage of the leave-one-out construction is that $\boldsymbol{x}_\ell$ is independent of $\mathbf{G}^{[\ell]}$, because $\mathbf{G}^{[\ell]}$ depends only on $\{\boldsymbol{x}_j\}_{j \neq \ell}$ and on the pre-training randomness. Since $\boldsymbol{x}_\ell \sim \mathcal{N}(\mathbf{0}, I_D/D)$, standard concentration of quadratic forms implies that, for any matrix $Q^{[\ell]}$ independent of $\boldsymbol{x}_\ell$ with bounded normalized trace norm,

$$\boldsymbol{x}_\ell^\top Q^{[\ell]} \boldsymbol{x}_\ell - \frac{1}{D} \operatorname{Tr}(Q^{[\ell]}) \;\longrightarrow\; 0 \tag{111}$$

in probability as $D \to \infty$. Applying this to each entry of (110), and using the fact that removing a single sample does not affect the normalized traces at leading order, we obtain the deterministic equivalent

$$\mathbf{K}_\ell \;\asymp\; \bar{K}(u, v) := \begin{pmatrix} -m_{12} & -\sqrt{v}\, m_{13} \\ 0 & -v\, m_{43} \end{pmatrix}, \tag{112}$$

where $\asymp$ means equality up to terms vanishing in the high-dimensional limit.

Extracting the S-inserted traces. To derive the equations for the s-variables, we must compute the quadratic forms that appear when $\mathbf{G}$ is sandwiched by $\boldsymbol{x}_\ell$. For this purpose, define

$$L_\ell^\top := \begin{pmatrix} \boldsymbol{x}_\ell^\top & 0 & 0 & 0 \\ 0 & \boldsymbol{x}_\ell^\top & 0 & 0 \\ 0 & 0 & \boldsymbol{x}_\ell^\top & 0 \\ 0 & 0 & 0 & \boldsymbol{x}_\ell^\top \end{pmatrix} \in \mathbb{R}^{4\times 4D}, \tag{113}$$

$$R_\ell := \begin{pmatrix} 0 & 0 \\ \boldsymbol{x}_\ell & 0 \\ 0 & \boldsymbol{x}_\ell \\ 0 & 0 \end{pmatrix} \in \mathbb{R}^{4D\times 2}. \tag{114}$$

Then

$$L_\ell^\top \mathbf{G} R_\ell = \begin{pmatrix} \boldsymbol{x}_\ell^\top \mathbf{G}_{12}\boldsymbol{x}_\ell & \boldsymbol{x}_\ell^\top \mathbf{G}_{13}\boldsymbol{x}_\ell \\ \boldsymbol{x}_\ell^\top \mathbf{G}_{22}\boldsymbol{x}_\ell & \boldsymbol{x}_\ell^\top \mathbf{G}_{23}\boldsymbol{x}_\ell \\ 0 & \boldsymbol{x}_\ell^\top \mathbf{G}_{33}\boldsymbol{x}_\ell \\ 0 & \boldsymbol{x}_\ell^\top \mathbf{G}_{43}\boldsymbol{x}_\ell \end{pmatrix}. \tag{115}$$

Averaging over ℓ, the definitions of the s-variables in (83) yield

$$\frac{1}{L}\sum_{\ell=1}^{L} L_\ell^\top G R_\ell = \begin{pmatrix} s_{12} & s_{13} \\ s_{22} & s_{23} \\ 0 & s_{33} \\ 0 & s_{43} \end{pmatrix}. \tag{116}$$

Indeed, for example,

$$\frac{1}{L}\sum_{\ell=1}^{L} \boldsymbol{x}_\ell^\top \mathbf{G}_{12}\boldsymbol{x}_\ell = \frac{1}{L}\operatorname{Tr}\left(\mathbf{G}_{12}\sum_{\ell=1}^{L}\boldsymbol{x}_\ell\boldsymbol{x}_\ell^\top\right) = \frac{1}{L}\operatorname{Tr}(\mathbf{G}_{12}\mathbf{S}) = \frac{1}{\alpha}\, s_{12}. \tag{117}$$

Thus one may equivalently work either with the average over ℓ or directly with the normalized traces. Below we present the final equations in terms of the normalized quantities s_{ab} defined in (83).

We now apply (108) between $L_\ell^\top$ and R_ℓ:

$$L_\ell^\top \mathbf{G} R_\ell = L_\ell^\top \mathbf{G}^{[\ell]} R_\ell - L_\ell^\top \mathbf{G}^{[\ell]} U_\ell (\mathbf{I}_2 + \mathbf{K}_\ell)^{-1} V_\ell^\top \mathbf{G}^{[\ell]} R_\ell. \tag{118}$$

Each term can now be evaluated by the same concentration argument as above.

First term. Since $\boldsymbol{x}_\ell$ is independent of $\mathbf{G}^{[\ell]}$, we have

$$L_\ell^\top \mathbf{G}^{[\ell]} R_\ell \asymp \frac{1}{D}\begin{pmatrix} \operatorname{Tr}(\mathbf{G}_{12}^{[\ell]}) & \operatorname{Tr}(\mathbf{G}_{13}^{[\ell]}) \\ \operatorname{Tr}(\mathbf{G}_{22}^{[\ell]}) & \operatorname{Tr}(\mathbf{G}_{23}^{[\ell]}) \\ 0 & \operatorname{Tr}(\mathbf{G}_{33}^{[\ell]}) \\ 0 & \operatorname{Tr}(\mathbf{G}_{43}^{[\ell]}) \end{pmatrix} \asymp \begin{pmatrix} m_{12} & m_{13} \\ m_{22} & m_{23} \\ 0 & m_{33} \\ 0 & m_{43} \end{pmatrix}. \tag{119}$$

Second term. Using (107), we compute

$$L_\ell^\top \mathbf{G}^{[\ell]} U_\ell \asymp \begin{pmatrix} -m_{12} & -\sqrt{v}\, m_{13} \\ -m_{22} & -\sqrt{v}\, m_{23} \\ 0 & -\sqrt{v}\, m_{33} \\ 0 & -\sqrt{v}\, m_{43} \end{pmatrix}, \tag{120}$$

and

$$V_\ell^\top \mathbf{G}^{[\ell]} R_\ell \asymp \begin{pmatrix} m_{12} & m_{13} \\ 0 & v\, m_{43} \end{pmatrix}. \tag{121}$$

Moreover, from (112),

$$I_2 + \bar{K} = \begin{pmatrix} 1 - m_{12} & -\sqrt{v}\, m_{13} \\ 0 & 1 - v m_{43} \end{pmatrix}, \tag{122}$$

so its inverse is

$$(I_2 + \bar{K})^{-1} = \begin{pmatrix} \dfrac{1}{1-m_{12}} & \dfrac{\sqrt{v}\, m_{13}}{(1-m_{12})(1-vm_{43})} \\ 0 & \dfrac{1}{1-vm_{43}} \end{pmatrix}. \tag{123}$$

Substituting these expressions into the second term of (118), we find after a direct matrix multiplication that

$$\begin{aligned} & L_\ell^\top G^{[\ell]} U_\ell (I_2 + \bar{K})^{-1} V_\ell^\top G^{[\ell]} R_\ell \\ & \asymp \begin{pmatrix} -\dfrac{(m_{12})^2}{1-m_{12}} & -\dfrac{m_{12}m_{13}}{(1-m_{12})(1-vm_{43})} - \dfrac{v\, m_{13}m_{43}}{1-vm_{43}} \\ -\dfrac{m_{12}m_{22}}{1-m_{12}} & -\dfrac{m_{22}m_{13}}{(1-m_{12})(1-vm_{43})} - \dfrac{v\, m_{23}m_{43}}{1-vm_{43}} \\ 0 & -\dfrac{v\, m_{33}m_{43}}{1-vm_{43}} \\ 0 & -\dfrac{v\,(m_{43})^2}{1-vm_{43}} \end{pmatrix}. \end{aligned} \tag{124}$$

Final equations. Combining (118), (119), and (124), and then summing over ℓ, we obtain the deterministic relations

$$\begin{pmatrix} s_{12} & s_{13} \\ s_{22} & s_{23} \\ 0 & s_{33} \\ 0 & s_{43} \end{pmatrix} = \alpha \begin{pmatrix} \dfrac{m_{12}}{1-m_{12}} & \dfrac{m_{13}}{(1-m_{12})(1-vm_{43})} \\ \dfrac{m_{22}}{1-m_{12}} & \dfrac{m_{23}}{1-vm_{43}} + \dfrac{m_{22}m_{13}}{(1-m_{12})(1-vm_{43})} \\ 0 & \dfrac{m_{33}}{1-vm_{43}} \\ 0 & \dfrac{m_{43}}{1-vm_{43}} \end{pmatrix}. \tag{125}$$

Equivalently, componentwise,

$$s_{12} = \alpha \frac{m_{12}}{1-m_{12}}, \tag{126}$$

$$s_{13} = \alpha \frac{m_{13}}{(1-m_{12})(1-vm_{43})}, \tag{127}$$

$$s_{22} = \alpha \frac{m_{22}}{1-m_{12}}, \tag{128}$$

$$s_{23} = \alpha \left(\frac{m_{23}}{1-vm_{43}} + \frac{m_{22}m_{13}}{(1-m_{12})(1-vm_{43})} \right), \tag{129}$$

$$s_{33} = \alpha \frac{m_{33}}{1-vm_{43}}, \tag{130}$$

$$s_{43} = \alpha \frac{m_{43}}{1-vm_{43}}. \tag{131}$$

These equations complete the treatment of the S-dependent order parameters. They show that the six quantities s_{ab} can be expressed entirely in terms of the primary traces m_{ab}.

C.6 Cavity equations for the B-dependent order parameters

We now turn to the order parameters b_{11} and b_{21}, which contain the pre-training matrix

$$\mathbf{B} = \frac{1}{M} \sum_{\mu=1}^{M} \boldsymbol{w}_\mu \boldsymbol{v}_\mu^\top.$$

Unlike the S-dependent quantities treated in the previous subsection, these variables depend on the randomness of the pre-training tasks. Accordingly, we now perform a leave-one-task-out analysis with respect to the index $\mu \in \{1, \dots, M\}$.

The main point is that both $\mathbf{B}$ and $\mathbf{C}$ depend on the same task variables $(\boldsymbol{w}_\mu, \boldsymbol{v}_\mu)$, so removing one task produces a perturbation that simultaneously affects two blocks of the extended matrix. This is why the corresponding cavity matrix has rank 4, rather than rank 2 as in the S-cavity analysis.

Leave-one-out decomposition. Fix $\mu \in \{1, \dots, M\}$, and define

$$\mathbf{B}^{[\mu]} := \frac{1}{M} \sum_{\nu \neq \mu} \boldsymbol{w}_\nu \boldsymbol{v}_\nu^\top, \tag{132}$$

$$\mathbf{C}^{[\mu]} := \frac{1}{M} \sum_{\nu \neq \mu} \boldsymbol{v}_\nu \boldsymbol{v}_\nu^\top + \frac{\lambda}{2} \mathbf{I}. \tag{133}$$

We then introduce the leave-one-out extended matrix

$$\mathcal{M}^{[\mu]}(u, v) := \begin{pmatrix} \mathbf{I} & -u\mathbf{B}^{[\mu]} & -\mathbf{I} & 0 \\ -\mathbf{S} & \mathbf{C}^{[\mu]} & 0 & 0 \\ 0 & 0 & \mathbf{I} & -v\mathbf{S} \\ 0 & 0 & -\big(\mathbf{B}^{[\mu]}\big)^\top & \mathbf{C}^{[\mu]} \end{pmatrix}, \tag{134}$$

and denote its inverse by

$$\mathbf{G}^{[\mu]} := \big(\mathcal{M}^{[\mu]}(u, v)\big)^{-1}. \tag{135}$$

Since

$$\mathbf{B} = \mathbf{B}^{[\mu]} + \frac{1}{M} \boldsymbol{w}_\mu \boldsymbol{v}_\mu^\top, \qquad \mathbf{C} = \mathbf{C}^{[\mu]} + \frac{1}{M} \boldsymbol{v}_\mu \boldsymbol{v}_\mu^\top, \tag{136}$$

the full matrix can be decomposed as

$$\mathcal{M}(u, v) = \mathcal{M}^{[\mu]}(u, v) + \Delta_\mu, \tag{137}$$

where

$$\Delta_\mu = \begin{pmatrix} 0 & -\frac{u}{M} \boldsymbol{w}_\mu \boldsymbol{v}_\mu^\top & 0 & 0 \\ 0 & \frac{1}{M} \boldsymbol{v}_\mu \boldsymbol{v}_\mu^\top & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{M} \boldsymbol{v}_\mu \boldsymbol{w}_\mu^\top & \frac{1}{M} \boldsymbol{v}_\mu \boldsymbol{v}_\mu^\top \end{pmatrix}. \tag{138}$$

It is convenient to write this perturbation in factorized form

$$\Delta_\mu = U_\mu V_\mu^\top, \tag{139}$$

with

$$U_\mu := \begin{pmatrix} -\sqrt{\frac{u}{M}}\, \boldsymbol{w}_\mu & 0 & 0 & 0 \\ 0 & \frac{1}{\sqrt{M}} \boldsymbol{v}_\mu & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & -\frac{1}{\sqrt{M}} \boldsymbol{v}_\mu & \frac{1}{\sqrt{M}} \boldsymbol{v}_\mu \end{pmatrix} \in \mathbb{R}^{4D \times 4}, \tag{140}$$

and

$$V_\mu := \begin{pmatrix} 0 & 0 & 0 & 0 \\ \sqrt{\frac{u}{M}}\, \boldsymbol{v}_\mu & \frac{1}{\sqrt{M}} \boldsymbol{v}_\mu & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{M}} \boldsymbol{w}_\mu & 0 \\ 0 & 0 & 0 & \frac{1}{\sqrt{M}} \boldsymbol{v}_\mu \end{pmatrix} \in \mathbb{R}^{4D \times 4}. \tag{141}$$

Woodbury formula and the cavity matrix. Applying the Woodbury formula to $\mathcal{M} = \mathcal{M}^{[\mu]} + U_\mu V_\mu^\top$, we obtain

$$\mathbf{G} = \mathbf{G}^{[\mu]} - \mathbf{G}^{[\mu]} U_\mu \Big(\mathbf{I}_4 + V_\mu^\top \mathbf{G}^{[\mu]} U_\mu \Big)^{-1} V_\mu^\top \mathbf{G}^{[\mu]}. \tag{142}$$

We denote the corresponding 4×4 cavity matrix by

$$\mathbf{J}_\mu := V_\mu^\top \mathbf{G}^{[\mu]} U_\mu. \tag{143}$$

Using (141), we compute

$$\mathbf{J}_\mu = \begin{pmatrix} -\frac{u}{M} \boldsymbol{v}_\mu^\top \mathbf{G}_{21}^{[\mu]} \boldsymbol{w}_\mu & \frac{\sqrt{u}}{M} \boldsymbol{v}_\mu^\top \mathbf{G}_{22}^{[\mu]} \boldsymbol{v}_\mu & -\frac{\sqrt{u}}{M} \boldsymbol{v}_\mu^\top \mathbf{G}_{24}^{[\mu]} \boldsymbol{v}_\mu & \frac{\sqrt{u}}{M} \boldsymbol{v}_\mu^\top \mathbf{G}_{24}^{[\mu]} \boldsymbol{v}_\mu \\ -\frac{\sqrt{u}}{M} \boldsymbol{v}_\mu^\top \mathbf{G}_{21}^{[\mu]} \boldsymbol{w}_\mu & \frac{1}{M} \boldsymbol{v}_\mu^\top \mathbf{G}_{22}^{[\mu]} \boldsymbol{v}_\mu & -\frac{1}{M} \boldsymbol{v}_\mu^\top \mathbf{G}_{24}^{[\mu]} \boldsymbol{v}_\mu & \frac{1}{M} \boldsymbol{v}_\mu^\top \mathbf{G}_{24}^{[\mu]} \boldsymbol{v}_\mu \\ 0 & 0 & -\frac{1}{M} \boldsymbol{w}_\mu^\top \mathbf{G}_{34}^{[\mu]} \boldsymbol{v}_\mu & \frac{1}{M} \boldsymbol{w}_\mu^\top \mathbf{G}_{34}^{[\mu]} \boldsymbol{v}_\mu \\ 0 & 0 & -\frac{1}{M} \boldsymbol{v}_\mu^\top \mathbf{G}_{44}^{[\mu]} \boldsymbol{v}_\mu & \frac{1}{M} \boldsymbol{v}_\mu^\top \mathbf{G}_{44}^{[\mu]} \boldsymbol{v}_\mu \end{pmatrix}. \tag{144}$$

To simplify this expression, we use the moment relations (48)–(51) from the preliminaries. As in the previous subsection, removing a single task does not affect normalized traces at leading order. Therefore $\mathbf{J}_\mu$ is asymptotically equivalent to the deterministic matrix

$$\bar{J} := \begin{pmatrix} -\frac{u}{\tau} m_{21} & \frac{\sqrt{u}}{\tau} c\, m_{22} & -\frac{\sqrt{u}}{\tau} c\, m_{24} & \frac{\sqrt{u}}{\tau} c\, m_{24} \\ -\frac{\sqrt{u}}{\tau} m_{21} & \frac{1}{\tau} c\, m_{22} & -\frac{1}{\tau} c\, m_{24} & \frac{1}{\tau} c\, m_{24} \\ 0 & 0 & -\frac{1}{\tau} m_{34} & \frac{1}{\tau} m_{34} \\ 0 & 0 & -\frac{c}{\tau} m_{44} & \frac{c}{\tau} m_{44} \end{pmatrix}, \tag{145}$$

where $\tau = M/D$.

Choice of test vectors. We now derive equations for b_{11} and b_{21}. By definition,

$$b_{11} = \frac{1}{D} \operatorname{Tr}(\mathbf{G}_{11} \mathbf{B}), \qquad b_{21} = \frac{1}{D} \operatorname{Tr}(\mathbf{G}_{21} \mathbf{B}). \tag{146}$$

Since $\mathbf{B}$ is a sum of rank-one terms $\boldsymbol{w}_\nu \boldsymbol{v}_\nu^\top / M$, these quantities can be extracted by testing $\mathbf{G}$ against the vectors $\boldsymbol{w}_\mu$ and $\boldsymbol{v}_\mu$. More precisely, define

$$L_\mu^{b\,\top} := \begin{pmatrix} \boldsymbol{v}_\mu^\top & 0 & 0 & 0 \\ 0 & \boldsymbol{v}_\mu^\top & 0 & 0 \end{pmatrix} \in \mathbb{R}^{2 \times 4D}, \tag{147}$$

$$R_\mu^b := \begin{pmatrix} \boldsymbol{w}_\mu \\ 0 \\ 0 \\ 0 \end{pmatrix} \in \mathbb{R}^{4D \times 1}. \tag{148}$$

Then

$$L_\mu^{b\,\top} \mathbf{G} R_\mu^b = \begin{pmatrix} \boldsymbol{v}_\mu^\top \mathbf{G}_{11} \boldsymbol{w}_\mu \\ \boldsymbol{v}_\mu^\top \mathbf{G}_{21} \boldsymbol{w}_\mu \end{pmatrix}, \tag{149}$$

so after normalization and averaging over μ, the limit is exactly $(b_{11}, b_{21})^\top$.

Applying (142) between $L_\mu^{b\,\top}$ and R_μ^b, we obtain

$$L_\mu^{b\,\top} \mathbf{G} R_\mu^b = L_\mu^{b\,\top} \mathbf{G}^{[\mu]} R_\mu^b - \big(L_\mu^{b\,\top} \mathbf{G}^{[\mu]} U_\mu \big) (\mathbf{I}_4 + \mathbf{J}_\mu)^{-1} \big(V_\mu^\top \mathbf{G}^{[\mu]} R_\mu^b \big). \tag{150}$$

Evaluation of each factor. We now evaluate the three factors on the right-hand side of (150).

First, by the quadratic-form concentration stated in (51) and the independence between $(\boldsymbol{w}_\mu, \boldsymbol{v}_\mu)$ and the leave-one-out resolvent $\mathbf{G}^{[\mu]}$,

$$\frac{1}{D} {L_\mu^b}^\top \mathbf{G}^{[\mu]} R_\mu^b \asymp \begin{pmatrix} m_{11} \\ m_{21} \end{pmatrix}. \tag{151}$$

Indeed, for the first component, $D^{-1}\boldsymbol{v}_\mu^\top \mathbf{G}_{11}^{[\mu]} \boldsymbol{w}_\mu \to D^{-1}\operatorname{Tr}(\mathbf{G}_{11}^{[\mu]})$, where we used $\mathbb{E}[\boldsymbol{w}_\mu \boldsymbol{v}_\mu^\top] = \mathbf{I}_D$ from (50). The second component is treated in the same way.

Next, a direct multiplication using (141) gives

$$\frac{1}{D} {L_\mu^b}^\top \mathbf{G}^{[\mu]} U_\mu \asymp \frac{1}{\sqrt{M}} \begin{pmatrix} -\sqrt{u}\, m_{11} & c\, m_{12} & -c\, m_{14} & c\, m_{14} \\ -\sqrt{u}\, m_{21} & c\, m_{22} & -c\, m_{24} & c\, m_{24} \end{pmatrix}, \tag{152}$$

and similarly

$$\frac{1}{D} V_\mu^\top \mathbf{G}^{[\mu]} R_\mu^b \asymp \frac{1}{\sqrt{M}} \begin{pmatrix} \sqrt{u}\, m_{21} \\ m_{21} \\ 0 \\ 0 \end{pmatrix}. \tag{153}$$

At this point, only the first two columns of $(\mathbf{I}_4 + \bar{J})^{-1}$ contribute, because the vector in (153) has vanishing third and fourth entries. Therefore only the upper-left 2×2 block matters. From (145), this block is

$$\begin{pmatrix} 1 - \dfrac{u}{\tau} m_{21} & \dfrac{\sqrt{u}}{\tau} c\, m_{22} \\ -\dfrac{\sqrt{u}}{\tau} m_{21} & 1 + \dfrac{c}{\tau} m_{22} \end{pmatrix}, \tag{154}$$

whose determinant is

$$\frac{1}{\tau}\Big(\tau - u\, m_{21} + c\, m_{22}\Big). \tag{155}$$

We therefore introduce the shorthand

$$D_1 := \tau - u\, m_{21} + c\, m_{22}. \tag{156}$$

A short calculation then shows that

$$\big({L_\mu^b}^\top \mathbf{G}^{[\mu]} U_\mu\big)(\mathbf{I}_4 + \bar{J})^{-1}\big(V_\mu^\top \mathbf{G}^{[\mu]} R_\mu^b\big) \asymp \frac{1}{\tau} \begin{pmatrix} \dfrac{m_{21}(c\, m_{12} - u\, m_{11})}{D_1} \\ \dfrac{m_{21}(c\, m_{22} - u\, m_{21})}{D_1} \end{pmatrix}. \tag{157}$$

Substituting this together with (151) into (150) yields the closed equations for b_{11} and b_{21}.

Final equations. We conclude that, in the high-dimensional limit,

$$b_{11} = m_{11} - \frac{m_{21}(c\, m_{12} - u\, m_{11})}{\tau - u\, m_{21} + c\, m_{22}}, \tag{158}$$

$$b_{21} = m_{21} - \frac{m_{21}(c\, m_{22} - u\, m_{21})}{\tau - u\, m_{21} + c\, m_{22}}. \tag{159}$$

These relations express the B-dependent order parameters entirely in terms of the primary traces m_{ab}. Together with the identities derived from $G\mathcal{M} = I$, they already determine the u-sector once the C-dependent quantities are also closed.

C.7 Cavity equations for the $B^\top$-dependent order parameters

We next derive closed equations for the quantities $\beta_{14}, \beta_{24}, \beta_{34}, \beta_{44}$, which contain the transpose matrix

$$\mathbf{B}^\top = \frac{1}{M}\sum_{\mu=1}^{M} \boldsymbol{v}_\mu \boldsymbol{w}_\mu^\top.$$

As in the previous subsection, the randomness enters through the pre-training tasks $\{(\boldsymbol{w}_\mu, \boldsymbol{v}_\mu)\}_{\mu=1}^M$, and we therefore use the same leave-one-task-out decomposition. The underlying mechanism is the same as for the b-variables, but now we must probe the third and fourth block columns of the resolvent, because $\mathbf{B}^\top$ appears in the $(4,3)$-block of the extended matrix $\mathcal{M}(u,v)$.

Setup. We keep the leave-one-task-out notation introduced in the previous subsection. For each fixed μ, we write

$$\mathcal{M}(u,v) = \mathcal{M}^{[\mu]}(u,v) + U_\mu V_\mu^\top, \tag{160}$$

with U_μ, V_μ, and $\mathbf{G}^{[\mu]} = \big(\mathcal{M}^{[\mu]}(u,v)\big)^{-1}$ defined in (141). The corresponding Woodbury formula is

$$\mathbf{G} = \mathbf{G}^{[\mu]} - \mathbf{G}^{[\mu]} U_\mu \big(\mathbf{I}_4 + \mathbf{J}_\mu\big)^{-1} V_\mu^\top \mathbf{G}^{[\mu]}, \qquad \mathbf{J}_\mu := V_\mu^\top \mathbf{G}^{[\mu]} U_\mu, \tag{161}$$

and, in the high-dimensional limit, $\mathbf{J}_\mu$ can be replaced by its deterministic equivalent $\bar{J}$ given in (145).

Our goal is to evaluate

$$\beta_{14} = \frac{1}{D}\operatorname{Tr}(\mathbf{G}_{14}\mathbf{B}^\top), \qquad \beta_{24} = \frac{1}{D}\operatorname{Tr}(\mathbf{G}_{24}\mathbf{B}^\top), \qquad \beta_{34} = \frac{1}{D}\operatorname{Tr}(\mathbf{G}_{34}\mathbf{B}^\top), \qquad \beta_{44} = \frac{1}{D}\operatorname{Tr}(\mathbf{G}_{44}\mathbf{B}^\top). \tag{162}$$

Since $\mathbf{B}^\top$ is a sum of rank-one terms $\boldsymbol{v}_\mu \boldsymbol{w}_\mu^\top / M$, these quantities can again be extracted by testing the resolvent against the vectors $\boldsymbol{w}_\mu$ and $\boldsymbol{v}_\mu$.

Choice of test vectors. Define

$$L_\mu^{\beta\,\top} := \begin{pmatrix} \boldsymbol{w}_\mu^\top & 0 & 0 & 0 \\ 0 & \boldsymbol{w}_\mu^\top & 0 & 0 \\ 0 & 0 & \boldsymbol{w}_\mu^\top & 0 \\ 0 & 0 & 0 & \boldsymbol{w}_\mu^\top \end{pmatrix} \in \mathbb{R}^{4\times 4D}, \tag{163}$$

$$R_\mu^\beta := \begin{pmatrix} 0 \\ 0 \\ 0 \\ \boldsymbol{v}_\mu \end{pmatrix} \in \mathbb{R}^{4D\times 1}. \tag{164}$$

Then

$$L_\mu^{\beta\,\top} \mathbf{G} R_\mu^\beta = \begin{pmatrix} \boldsymbol{w}_\mu^\top \mathbf{G}_{14} \boldsymbol{v}_\mu \\ \boldsymbol{w}_\mu^\top \mathbf{G}_{24} \boldsymbol{v}_\mu \\ \boldsymbol{w}_\mu^\top \mathbf{G}_{34} \boldsymbol{v}_\mu \\ \boldsymbol{w}_\mu^\top \mathbf{G}_{44} \boldsymbol{v}_\mu \end{pmatrix}, \tag{165}$$

so that, after normalization and averaging over μ, the limiting vector is precisely $(\beta_{14}, \beta_{24}, \beta_{34}, \beta_{44})^\top$.

Applying the Woodbury formula between $L_\mu^{\beta\,\top}$ and R_μ^β, we obtain

$$L_\mu^{\beta\,\top} \mathbf{G} R_\mu^\beta = L_\mu^{\beta\,\top} \mathbf{G}^{[\mu]} R_\mu^\beta - \big(L_\mu^{\beta\,\top} \mathbf{G}^{[\mu]} U_\mu\big)\big(\mathbf{I}_4 + \mathbf{J}_\mu\big)^{-1} \big(V_\mu^\top \mathbf{G}^{[\mu]} R_\mu^\beta\big). \tag{166}$$

Evaluation of the three factors. We now analyze each term on the right-hand side of (166).

First, by the same quadratic-form concentration (51) and the leave-one-out argument as before,

$$\frac{1}{D} L_\mu^{\beta\,\top} \mathbf{G}^{[\mu]} R_\mu^\beta \asymp \begin{pmatrix} m_{14} \\ m_{24} \\ m_{34} \\ m_{44} \end{pmatrix}. \tag{167}$$

Indeed, the μ-th task is independent of $\mathbf{G}^{[\mu]}$, and replacing quadratic forms by their normalized traces yields the corresponding block traces of the leave-one-out resolvent.

Next, using the explicit form of U_μ in (141), we obtain

$$\frac{1}{D} L_\mu^{\beta\top} \mathbf{G}^{[\mu]} U_\mu \asymp \frac{1}{\sqrt{M}} \begin{pmatrix} -\sqrt{u}\, m_{11} & m_{12} & -m_{14} & m_{14} \\ -\sqrt{u}\, m_{21} & m_{22} & -m_{24} & m_{24} \\ 0 & 0 & -m_{34} & m_{34} \\ 0 & 0 & -m_{44} & m_{44} \end{pmatrix}. \tag{168}$$

Here the first two columns come from the first two block columns of U_μ, which contain $\boldsymbol{w}_\mu$ and $\boldsymbol{v}_\mu$, while the last two columns reflect the two opposite contributions in the fourth block row of U_μ.

Similarly, using the explicit form of V_μ, we find

$$\frac{1}{D} V_\mu^\top \mathbf{G}^{[\mu]} R_\mu^\beta \asymp \frac{1}{\sqrt{M}} \begin{pmatrix} \sqrt{u}\, c\, m_{24} \\ c\, m_{24} \\ m_{34} \\ c\, m_{44} \end{pmatrix}. \tag{169}$$

The factor c appears whenever $\boldsymbol{v}_\mu$ is contracted with itself, reflecting the asymptotic norm $D^{-1}\|\boldsymbol{v}_\mu\|^2 \to c$.

Reduction of $(\mathbf{I}_4 + \bar{J})^{-1}$. To simplify the cavity correction in (166), we use the block upper-triangular structure of $\bar{J}$. Introduce the shorthand

$$D_1 := \tau - u\, m_{21} + c\, m_{22}, \tag{170}$$

$$D_2 := \tau - m_{34} + c\, m_{44}. \tag{171}$$

These quantities are the determinants of the relevant 2×2 blocks of $\mathbf{I}_4 + \bar{J}$, up to the common factor τ^{-1}. More precisely, the upper-left block of $\mathbf{I}_4 + \bar{J}$ has determinant D_1/τ, while the lower-right block has determinant D_2/τ.

Because the vector (169) has nonzero entries in both the first two and the last two components, both sectors contribute here. A direct but straightforward matrix multiplication using (168), (169), and the explicit inverse of $\mathbf{I}_4 + \bar{J}$ yields the cavity correction. After simplifying the resulting expressions, we obtain the following closed equations.

Final equations. In the high-dimensional limit,

$$\begin{aligned} \beta_{14} &= m_{14} - \frac{c\, m_{24}(m_{12} - u\, m_{11})}{D_1 D_2/\tau} - \frac{m_{14}(c\, m_{44} - m_{34})}{D_2} \\ &= m_{14} - \frac{\tau\, c\, m_{24}(m_{12} - u\, m_{11})}{D_1 D_2} - \frac{m_{14}(c\, m_{44} - m_{34})}{D_2}, \end{aligned} \tag{172}$$

$$\beta_{24} = m_{24} - \frac{\tau\, c\, m_{24}(m_{22} - u\, m_{21})}{D_1 D_2} - \frac{m_{24}(c\, m_{44} - m_{34})}{D_2}, \tag{173}$$

$$\beta_{34} = m_{34} - \frac{m_{34}(c\, m_{44} - m_{34})}{D_2}, \tag{174}$$

$$\beta_{44} = m_{44} - \frac{m_{44}(c\, m_{44} - m_{34})}{D_2}. \tag{175}$$

Equivalently, writing out the denominators explicitly,

$$\beta_{14} = m_{14} - \frac{c\, m_{24}(m_{12} - u\, m_{11})}{\tau\left(1 - \frac{u}{\tau} m_{21} + \frac{c}{\tau} m_{22}\right)\left(1 - \frac{1}{\tau} m_{34} + \frac{c}{\tau} m_{44}\right)} - \frac{m_{14}(c\, m_{44} - m_{34})}{\tau\left(1 - \frac{1}{\tau} m_{34} + \frac{c}{\tau} m_{44}\right)}, \tag{176}$$

$$\beta_{24} = m_{24} - \frac{c\, m_{24}(m_{22} - u\, m_{21})}{\tau\left(1 - \frac{u}{\tau} m_{21} + \frac{c}{\tau} m_{22}\right)\left(1 - \frac{1}{\tau} m_{34} + \frac{c}{\tau} m_{44}\right)} - \frac{m_{24}(c\, m_{44} - m_{34})}{\tau\left(1 - \frac{1}{\tau} m_{34} + \frac{c}{\tau} m_{44}\right)}, \tag{177}$$

$$\beta_{34} = m_{34} - \frac{m_{34}(c\, m_{44} - m_{34})}{\tau\left(1 - \frac{1}{\tau} m_{34} + \frac{c}{\tau} m_{44}\right)}, \tag{178}$$

$$\beta_{44} = m_{44} - \frac{m_{44}(c\, m_{44} - m_{34})}{\tau\left(1 - \frac{1}{\tau} m_{34} + \frac{c}{\tau} m_{44}\right)}. \tag{179}$$

These relations close the $\mathbf{B}^\top$-dependent order parameters in terms of the primary traces m_{ab}. Together with the equations for s and b, they determine the contributions coming from the blocks involving $\mathbf{S}$, $\mathbf{B}$, and $\mathbf{B}^\top$.

C.8 Cavity equations for the C-dependent order parameters

We finally derive closed equations for the quantities $c_{12}, c_{14}, c_{22}, c_{24}, c_{34}, c_{44}$, which contain the matrix

$$\mathbf{C} = \frac{1}{M}\sum_{\mu=1}^{M} \boldsymbol{v}_\mu \boldsymbol{v}_\mu^\top + \frac{\lambda}{2}\mathbf{I}.$$

This subsection completes the closure of the order-parameter system. As in the previous two subsections, we use a leave-one-task-out decomposition with respect to the pre-training index μ. The derivation closely parallels the one for the $\mathbf{B}^\top$-dependent quantities, but here both left and right test vectors involve $\boldsymbol{v}_\mu$, so the asymptotic norm factor c enters more systematically.

Setup. We keep the leave-one-task-out notation introduced above. For each fixed μ, recall that

$$\mathcal{M}(u,v) = \mathcal{M}^{[\mu]}(u,v) + U_\mu V_\mu^\top, \tag{180}$$

with U_μ, V_μ, and $\mathbf{G}^{[\mu]} = \big(\mathcal{M}^{[\mu]}(u,v)\big)^{-1}$ defined in (141). The associated Woodbury formula is

$$\mathbf{G} = \mathbf{G}^{[\mu]} - \mathbf{G}^{[\mu]} U_\mu \big(\mathbf{I}_4 + \mathbf{J}_\mu\big)^{-1} V_\mu^\top \mathbf{G}^{[\mu]}, \tag{181}$$

where $\mathbf{J}_\mu = V_\mu^\top \mathbf{G}^{[\mu]} U_\mu$, and in the high-dimensional limit we may replace $\mathbf{J}_\mu$ by the deterministic matrix $\bar{J}$ given in (145).

Our aim is to evaluate the mixed traces involving C. Since

$$\mathbf{C} = \frac{1}{M}\sum_{\mu=1}^{M} \boldsymbol{v}_\mu \boldsymbol{v}_\mu^\top + \frac{\lambda}{2}\mathbf{I},$$

each term $\mathrm{Tr}(\mathbf{G}_{ab}\mathbf{C})$ splits into a contribution from the empirical covariance of the $\boldsymbol{v}_\mu$'s and a deterministic ridge contribution $\frac{\lambda}{2}\mathrm{Tr}(\mathbf{G}_{ab})$. For this reason, the cavity equations naturally determine the combinations

$$c_{ab} - \frac{\lambda}{2} m_{ab}.$$

Choice of test vectors. To extract the C-dependent quantities, define

$$L_\mu^{c\,\top} := \begin{pmatrix} \boldsymbol{v}_\mu^\top & 0 & 0 & 0 \\ 0 & \boldsymbol{v}_\mu^\top & 0 & 0 \\ 0 & 0 & \boldsymbol{v}_\mu^\top & 0 \\ 0 & 0 & 0 & \boldsymbol{v}_\mu^\top \end{pmatrix} \in \mathbb{R}^{4\times 4D}, \tag{182}$$

$$R_\mu^c := \begin{pmatrix} 0 & 0 \\ \boldsymbol{v}_\mu & 0 \\ 0 & 0 \\ 0 & \boldsymbol{v}_\mu \end{pmatrix} \in \mathbb{R}^{4D\times 2}. \tag{183}$$

Then

$$L_\mu^{c\,\top} \mathbf{G} R_\mu^c = \begin{pmatrix} \boldsymbol{v}_\mu^\top \mathbf{G}_{12} \boldsymbol{v}_\mu & \boldsymbol{v}_\mu^\top \mathbf{G}_{14} \boldsymbol{v}_\mu \\ \boldsymbol{v}_\mu^\top \mathbf{G}_{22} \boldsymbol{v}_\mu & \boldsymbol{v}_\mu^\top \mathbf{G}_{24} \boldsymbol{v}_\mu \\ 0 & \boldsymbol{v}_\mu^\top \mathbf{G}_{34} \boldsymbol{v}_\mu \\ 0 & \boldsymbol{v}_\mu^\top \mathbf{G}_{44} \boldsymbol{v}_\mu \end{pmatrix}. \tag{184}$$

Averaging over μ, the empirical covariance part becomes

$$\frac{1}{M}\sum_{\mu=1}^{M} \boldsymbol{v}_\mu \boldsymbol{v}_\mu^\top = \mathbf{C} - \frac{\lambda}{2}\mathbf{I},$$

and therefore

$$\frac{1}{D}\frac{1}{M}\sum_{\mu=1}^{M} L_\mu^{c\top}\mathbf{G}R_\mu^c = \begin{pmatrix} c_{12}-\frac{\lambda}{2}m_{12} & c_{14}-\frac{\lambda}{2}m_{14} \\ c_{22}-\frac{\lambda}{2}m_{22} & c_{24}-\frac{\lambda}{2}m_{24} \\ 0 & c_{34}-\frac{\lambda}{2}m_{34} \\ 0 & c_{44}-\frac{\lambda}{2}m_{44} \end{pmatrix}. \tag{185}$$

Applying the Woodbury formula between $L_\mu^{c\top}$ and R_μ^c, we obtain

$$L_\mu^{c\top}\mathbf{G}R_\mu^c = L_\mu^{c\top}\mathbf{G}^{[\mu]}R_\mu^c - \big(L_\mu^{c\top}\mathbf{G}^{[\mu]}U_\mu\big)(\mathbf{I}_4+\mathbf{J}_\mu)^{-1}\big(V_\mu^\top\mathbf{G}^{[\mu]}R_\mu^c\big). \tag{186}$$

Evaluation of the three factors. We now evaluate each term on the right-hand side of (186).

First, since $\boldsymbol{v}_\mu$ is independent of the leave-one-out resolvent $\mathbf{G}^{[\mu]}$, concentration of quadratic forms gives

$$\frac{1}{D}L_\mu^{c\top}\mathbf{G}^{[\mu]}R_\mu^c \asymp \begin{pmatrix} c\,m_{12} & c\,m_{14} \\ c\,m_{22} & c\,m_{24} \\ 0 & c\,m_{34} \\ 0 & c\,m_{44} \end{pmatrix}. \tag{187}$$

Here each factor c comes from the asymptotic relation $D^{-1}\|\boldsymbol{v}_\mu\|^2 \to c$.

Next, using the explicit form of U_μ in (141), we find

$$\frac{1}{D}L_\mu^{c\top}\mathbf{G}^{[\mu]}U_\mu \asymp \frac{1}{\sqrt{M}}\begin{pmatrix} -\sqrt{u}\,m_{11} & c\,m_{12} & -c\,m_{14} & c\,m_{14} \\ -\sqrt{u}\,m_{21} & c\,m_{22} & -c\,m_{24} & c\,m_{24} \\ 0 & 0 & -c\,m_{34} & c\,m_{34} \\ 0 & 0 & -c\,m_{44} & c\,m_{44} \end{pmatrix}. \tag{188}$$

Similarly, using the explicit form of V_μ, we obtain

$$\frac{1}{D}V_\mu^\top\mathbf{G}^{[\mu]}R_\mu^c \asymp \frac{1}{\sqrt{M}}\begin{pmatrix} \sqrt{u}\,c\,m_{22} & \sqrt{u}\,c\,m_{24} \\ c\,m_{22} & c\,m_{24} \\ 0 & m_{34} \\ 0 & c\,m_{44} \end{pmatrix}. \tag{189}$$

As in the previous subsection, the inverse of $\mathbf{I}_4+\bar{J}$ is naturally expressed in terms of the two scalar combinations

$$D_1 := \tau - u\,m_{21} + c\,m_{22}, \qquad D_2 := \tau - m_{34} + c\,m_{44}. \tag{190}$$

The upper-left and lower-right 2×2 blocks of $\mathbf{I}_4+\bar{J}$ have determinants D_1/τ and D_2/τ, respectively. Using this structure and multiplying the matrices in (188) and (189), one obtains the cavity correction term explicitly.

Final equations. Substituting the resulting expression into (186), averaging over μ, and using (185), we arrive at the following closed system:

$$c_{12}-\frac{\lambda}{2}m_{12} = c\,m_{12} - \frac{1}{\tau}\frac{c\,m_{22}(c\,m_{12}-u\,m_{11})}{1-\frac{u}{\tau}m_{21}+\frac{c}{\tau}m_{22}}, \tag{191}$$

$$c_{22}-\frac{\lambda}{2}m_{22} = c\,m_{22} - \frac{1}{\tau}\frac{c\,m_{22}(c\,m_{22}-u\,m_{21})}{1-\frac{u}{\tau}m_{21}+\frac{c}{\tau}m_{22}}, \tag{192}$$

$$c_{14}-\frac{\lambda}{2}m_{14} = c\,m_{14} - \frac{1}{\tau}\left[\frac{c\,m_{24}(c\,m_{12}-u\,m_{11})}{\big(1-\frac{u}{\tau}m_{21}+\frac{c}{\tau}m_{22}\big)\big(1-\frac{1}{\tau}m_{34}+\frac{c}{\tau}m_{44}\big)} + \frac{c\,m_{14}(c\,m_{44}-m_{34})}{1-\frac{1}{\tau}m_{34}+\frac{c}{\tau}m_{44}}\right], \tag{193}$$

$$c_{24}-\frac{\lambda}{2}m_{24} = c\,m_{24} - \frac{1}{\tau}\left[\frac{c\,m_{24}(c\,m_{22}-u\,m_{21})}{\big(1-\frac{u}{\tau}m_{21}+\frac{c}{\tau}m_{22}\big)\big(1-\frac{1}{\tau}m_{34}+\frac{c}{\tau}m_{44}\big)} + \frac{c\,m_{24}(c\,m_{44}-m_{34})}{1-\frac{1}{\tau}m_{34}+\frac{c}{\tau}m_{44}}\right], \tag{194}$$

$$c_{34}-\frac{\lambda}{2}m_{34} = c\,m_{34} - \frac{1}{\tau}\frac{c\,m_{34}(c\,m_{44}-m_{34})}{1-\frac{1}{\tau}m_{34}+\frac{c}{\tau}m_{44}}, \tag{195}$$

$$c_{44}-\frac{\lambda}{2}m_{44} = c\,m_{44} - \frac{1}{\tau}\frac{c\,m_{44}(c\,m_{44}-m_{34})}{1-\frac{1}{\tau}m_{34}+\frac{c}{\tau}m_{44}}. \tag{196}$$

Equivalently, using $D_1 = \tau - u\,m_{21} + c\,m_{22}$ and $D_2 = \tau - m_{34} + c\,m_{44}$, the same equations can be written in the slightly shorter form

$$c_{12} - \frac{\lambda}{2} m_{12} = c\,m_{12} - \frac{c\,m_{22}(c\,m_{12} - u\,m_{11})}{D_1}, \tag{197}$$

$$c_{22} - \frac{\lambda}{2} m_{22} = c\,m_{22} - \frac{c\,m_{22}(c\,m_{22} - u\,m_{21})}{D_1}, \tag{198}$$

$$c_{14} - \frac{\lambda}{2} m_{14} = c\,m_{14} - \frac{\tau\,c\,m_{24}(c\,m_{12} - u\,m_{11})}{D_1 D_2} - \frac{c\,m_{14}(c\,m_{44} - m_{34})}{D_2}, \tag{199}$$

$$c_{24} - \frac{\lambda}{2} m_{24} = c\,m_{24} - \frac{\tau\,c\,m_{24}(c\,m_{22} - u\,m_{21})}{D_1 D_2} - \frac{c\,m_{24}(c\,m_{44} - m_{34})}{D_2}, \tag{200}$$

$$c_{34} - \frac{\lambda}{2} m_{34} = c\,m_{34} - \frac{c\,m_{34}(c\,m_{44} - m_{34})}{D_2}, \tag{201}$$

$$c_{44} - \frac{\lambda}{2} m_{44} = c\,m_{44} - \frac{c\,m_{44}(c\,m_{44} - m_{34})}{D_2}. \tag{202}$$

This completes the cavity analysis. At this point, the basic identities from $G\mathcal{M} = I$, together with the cavity equations for s, b, β, and c, form a closed deterministic system for the primary traces m_{ab}.

C.9 Closed self-consistent equations for the primary order parameters

We now combine the exact identities from $G\mathcal{M}(u, v) = I$ with the cavity equations derived in the previous subsections. The goal of this subsection is to eliminate the auxiliary quantities

$$s_{ab}, \qquad b_{ab}, \qquad \beta_{ab}, \qquad c_{ab},$$

and obtain a closed system for the primary traces

$$m_{11}, m_{12}, m_{21}, m_{22}, \qquad m_{33}, m_{34}, m_{43}, m_{44}, \qquad m_{13}, m_{14}, m_{23}, m_{24}.$$

A useful structural point is that this system is triangular. Indeed, the first group $(m_{11}, m_{12}, m_{21}, m_{22})$ depends only on u, the second group $(m_{33}, m_{34}, m_{43}, m_{44})$ depends only on v, and the mixed group $(m_{13}, m_{14}, m_{23}, m_{24})$ becomes linear once the first two groups are known. This structure is what later makes the explicit solution possible.

The u-sector. We begin with the equations involving $m_{11}, m_{12}, m_{21}, m_{22}$. From the basic identities (89), (90), (93), and (94), we have

$$m_{11} - s_{12} = 1, \tag{203}$$

$$-u\,b_{11} + c_{12} = 0, \tag{204}$$

$$m_{21} - s_{22} = 0, \tag{205}$$

$$-u\,b_{21} + c_{22} = 1. \tag{206}$$

We first substitute the S-cavity relations (126) and (128) into (203) and (205). This gives

$$m_{11} = 1 + \alpha\,\frac{m_{12}}{1 - m_{12}}, \tag{207}$$

$$m_{21} = \alpha\,\frac{m_{22}}{1 - m_{12}}. \tag{208}$$

Next we use the cavity equations for b and c. From (158) and (191), we obtain

$$b_{11} = m_{11} - \frac{m_{21}(c\,m_{12} - u\,m_{11})}{\tau - u\,m_{21} + c\,m_{22}}, \tag{209}$$

$$c_{12} = \frac{\lambda}{2} m_{12} + c\,m_{12} - \frac{c\,m_{22}(c\,m_{12} - u\,m_{11})}{\tau - u\,m_{21} + c\,m_{22}}. \tag{210}$$

Substituting these into (204) and collecting terms yields

$$-u\,m_{11} + \left(c + \frac{\lambda}{2}\right) m_{12} + \frac{(u\,m_{21} - c\,m_{22})(c\,m_{12} - u\,m_{11})}{\tau - u\,m_{21} + c\,m_{22}} = 0. \tag{211}$$

Similarly, substituting (159) and (192) into (206), we obtain

$$-u\,m_{21} + \left(c + \frac{\lambda}{2}\right) m_{22} + \frac{(u\,m_{21} - c\,m_{22})(c\,m_{22} - u\,m_{21})}{\tau - u\,m_{21} + c\,m_{22}} = 1. \tag{212}$$

Thus the u-sector is closed by (207)–(212).

The v-sector. We next consider $m_{33}, m_{34}, m_{43}, m_{44}$. From the basic identities (97), (98), (99), and (100), we have

$$m_{33} - \beta_{34} = 1, \tag{213}$$
$$-v\,s_{33} + c_{34} = 0, \tag{214}$$
$$m_{43} - \beta_{44} = 0, \tag{215}$$
$$-v\,s_{43} + c_{44} = 1. \tag{216}$$

Using the cavity relations (174) and (175), we can rewrite (213) and (215) as

$$m_{33} = 1 + m_{34} - \frac{m_{34}(c\,m_{44} - m_{34})}{\tau - m_{34} + c\,m_{44}}, \tag{217}$$
$$m_{43} = m_{44} - \frac{m_{44}(c\,m_{44} - m_{34})}{\tau - m_{34} + c\,m_{44}}. \tag{218}$$

Next, substituting the S-cavity relations (130) and (131), together with the C-cavity relations (195) and (196), into (214) and (216), we obtain

$$-v\,\alpha\,\frac{m_{33}}{1 - v\,m_{43}} + \left(c + \frac{\lambda}{2}\right) m_{34} - \frac{c\,m_{34}(c\,m_{44} - m_{34})}{\tau - m_{34} + c\,m_{44}} = 0, \tag{219}$$
$$-v\,\alpha\,\frac{m_{43}}{1 - v\,m_{43}} + \left(c + \frac{\lambda}{2}\right) m_{44} - \frac{c\,m_{44}(c\,m_{44} - m_{34})}{\tau - m_{34} + c\,m_{44}} = 1. \tag{220}$$

Hence the v-sector is closed by (217)–(220).

The mixed sector. We finally derive the equations for $m_{13}, m_{14}, m_{23}, m_{24}$. At this stage the u-sector and v-sector are already closed, so the mixed equations will be linear in the unknown mixed variables.

From the basic identities (91), (92), (95), and (96), we have

$$-m_{11} + m_{13} - \beta_{14} = 0, \tag{221}$$
$$-v\,s_{13} + c_{14} = 0, \tag{222}$$
$$-m_{21} + m_{23} - \beta_{24} = 0, \tag{223}$$
$$-v\,s_{23} + c_{24} = 0. \tag{224}$$

For convenience, define

$$D_1 := \tau - u\,m_{21} + c\,m_{22}, \tag{225}$$
$$D_2 := \tau - m_{34} + c\,m_{44}. \tag{226}$$

Substituting (172) and (173) into (221) and (223), we obtain

$$m_{13} - m_{11} - \frac{\tau}{D_2}\left(m_{14} - c\,m_{24}\frac{m_{12} - u\,m_{11}}{D_1}\right) = 0, \tag{227}$$
$$m_{23} - m_{21} - \frac{\tau}{D_2}\left(m_{24} - c\,m_{24}\frac{m_{22} - u\,m_{21}}{D_1}\right) = 0. \tag{228}$$

Similarly, substituting the S-cavity relations (127) and (129), together with the C-cavity relations (193) and (194), into (222) and (224), we obtain

$$v\alpha\frac{m_{13}}{(1 - m_{12})(1 - v\,m_{43})} - \frac{\lambda}{2}m_{14} - \frac{c\tau}{D_2}\left(m_{14} - m_{24}\frac{c\,m_{12} - u\,m_{11}}{D_1}\right) = 0, \tag{229}$$

$$v\alpha\left(\frac{m_{23}}{1 - v\,m_{43}} + \frac{m_{22}m_{13}}{(1 - m_{12})(1 - v\,m_{43})}\right) - \frac{\lambda}{2}m_{24} - \frac{c\tau}{D_2}\left(m_{24} - m_{24}\frac{c\,m_{22} - u\,m_{21}}{D_1}\right) = 0. \tag{230}$$

Equations (227)–(230) close the mixed sector once the u- and v-sectors are known.

Summary of the closed system. Collecting the results above, the primary order parameters are determined by the following self-consistent system.

The u-sector satisfies

$$\begin{aligned}
&m_{11} = 1 + \alpha\,\frac{m_{12}}{1-m_{12}},\\
&m_{21} = \alpha\,\frac{m_{22}}{1-m_{12}},\\
&-u\,m_{11} + \left(c+\frac{\lambda}{2}\right) m_{12} + \frac{(u\,m_{21}-c\,m_{22})(c\,m_{12}-u\,m_{11})}{\tau - u\,m_{21} + c\,m_{22}} = 0,\\
&-u\,m_{21} + \left(c+\frac{\lambda}{2}\right) m_{22} + \frac{(u\,m_{21}-c\,m_{22})(c\,m_{22}-u\,m_{21})}{\tau - u\,m_{21} + c\,m_{22}} = 1,
\end{aligned} \tag{231}$$

where

$$c = 1 + \frac{1+\sigma^2}{\alpha}. \tag{232}$$

The v-sector satisfies

$$\begin{aligned}
&m_{33} = 1 + m_{34} - \frac{m_{34}(c\,m_{44}-m_{34})}{\tau - m_{34} + c\,m_{44}},\\
&-v\,\alpha\,\frac{m_{33}}{1-v\,m_{43}} + \left(c+\frac{\lambda}{2}\right) m_{34} - \frac{c\,m_{34}(c\,m_{44}-m_{34})}{\tau - m_{34} + c\,m_{44}} = 0,\\
&m_{43} = m_{44} - \frac{m_{44}(c\,m_{44}-m_{34})}{\tau - m_{34} + c\,m_{44}},\\
&-v\,\alpha\,\frac{m_{43}}{1-v\,m_{43}} + \left(c+\frac{\lambda}{2}\right) m_{44} - \frac{c\,m_{44}(c\,m_{44}-m_{34})}{\tau - m_{34} + c\,m_{44}} = 1.
\end{aligned} \tag{233}$$

Finally, the mixed sector satisfies

$$\begin{aligned}
&m_{13} - m_{11} - \frac{\tau}{D_2}\left(m_{14} - c\,m_{24}\frac{m_{12}-u\,m_{11}}{D_1}\right) = 0,\\
&m_{23} - m_{21} - \frac{\tau}{D_2}\left(m_{24} - c\,m_{24}\frac{m_{22}-u\,m_{21}}{D_1}\right) = 0,\\
&v\alpha\frac{m_{13}}{(1-m_{12})(1-v\,m_{43})} - \frac{\lambda}{2}m_{14} - \frac{c\tau}{D_2}\left(m_{14} - m_{24}\frac{c\,m_{12}-u\,m_{11}}{D_1}\right) = 0,\\
&v\alpha\left(\frac{m_{23}}{1-v\,m_{43}} + \frac{m_{22}m_{13}}{(1-m_{12})(1-v\,m_{43})}\right) - \frac{\lambda}{2}m_{24} - \frac{c\tau}{D_2}\left(m_{24} - m_{24}\frac{c\,m_{22}-u\,m_{21}}{D_1}\right) = 0,
\end{aligned} \tag{234}$$

with

$$D_1 = \tau - u\,m_{21} + c\,m_{22}, \qquad D_2 = \tau - m_{34} + c\,m_{44}. \tag{235}$$

This is the closed deterministic system announced above. Its structure is now transparent: first one solves (231) for the u-sector, then (233) for the v-sector, and finally (234) for $m_{13}, m_{14}, m_{23}, m_{24}$. In the next subsection, we simplify these equations in the limit $\lambda \to 0$ and derive explicit formulas for the order parameters, in particular for $m_{13}(u,v)$, which is the quantity needed for the evaluation of the generalization error.

C.10 Explicit solution in $\lambda = 0$

We now solve the closed system obtained in the previous subsection in the unregularized case $\lambda = 0$. In this regime, the self-consistent equations simplify substantially and admit an explicit solution. The main reason is that the u-sector and v-sector reduce to the same scalar equation, and the mixed sector then becomes a linear system with rational coefficients. As a result, all primary order parameters can be written in terms of a single scalar resolvent function.

Solution of the u-sector. We begin with the equations (231). A convenient way to solve them is to eliminate m_{11}, m_{21}, and m_{22} in favor of m_{12}, and then rewrite the resulting equation in terms of a single scalar function. To this end, define

$$g(u) := m_{11}(u). \tag{236}$$

Then, from (207),

$$g(u) = 1 + \alpha \, \frac{m_{12}(u)}{1 - m_{12}(u)}. \tag{237}$$

Solving for m_{12}, we obtain

$$m_{12}(u) = \frac{g(u) - 1}{\alpha + g(u) - 1}. \tag{238}$$

However, it is more convenient to parametrize the solution in the equivalent form

$$\frac{m_{12}(u)}{u} = \frac{g(u)}{c}. \tag{239}$$

Using this ansatz together with the remaining equations in the u-sector, one finds that

$$u \, m_{21}(u) = \frac{\tau}{\tau - 1}\big(g(u) - 1\big), \tag{240}$$

and

$$m_{22}(u) = \frac{\tau}{\tau - 1}\frac{g(u)}{c}. \tag{241}$$

Substituting these expressions back into (231), we find that $g(u)$ must satisfy the scalar equation

$$u \, g(u)^2 - \big(c + u(1 - \alpha)\big)g(u) + c = 0. \tag{242}$$

The solution that is regular at $u = 0$ is the resolvent function of the Wishart matrix, which is given by

$$g(u) = \frac{c + u(1 - \alpha) - \sqrt{\big(c + u(1 - \alpha)\big)^2 - 4cu}}{2u}. \tag{243}$$

Therefore the full u-sector is given by

$$m_{11}(u) = g(u), \tag{244}$$

$$\frac{m_{12}(u)}{u} = \frac{g(u)}{c}, \tag{245}$$

$$u \, m_{21}(u) = \frac{\tau}{\tau - 1}\big(g(u) - 1\big), \tag{246}$$

$$m_{22}(u) = \frac{\tau}{\tau - 1}\frac{g(u)}{c}. \tag{247}$$

Solution of the v-sector. The v-sector is completely analogous. Indeed, comparing (231) and (233), we see that the two systems are related by the substitutions

$$(m_{11}, m_{12}, u \, m_{21}, m_{22}) \longleftrightarrow (m_{33}, v \, m_{43}, m_{34}, m_{44}), \tag{248}$$

together with $u \leftrightarrow v$. More precisely, the explicit solution is

$$m_{33}(v) = g(v), \tag{249}$$

$$m_{43}(v) = \frac{g(v)}{c}, \tag{250}$$

$$m_{34}(v) = \frac{\tau}{\tau - 1}\big(g(v) - 1\big), \tag{251}$$

$$m_{44}(v) = \frac{\tau}{\tau - 1}\frac{g(v)}{c}. \tag{252}$$

Simplification of D_1 and D_2. Before solving the mixed sector, it is useful to simplify the denominators D_1 and D_2. Using (246) and (247), we obtain

$$\begin{aligned} D_1 &= \tau - u\, m_{21} + c\, m_{22} \\ &= \tau - \frac{\tau}{\tau-1}\big(g(u)-1\big) + \frac{\tau}{\tau-1} g(u) = \frac{\tau(\tau+\alpha+g(u)-1)}{\tau-1} \end{aligned} \tag{253}$$

in an intermediate parametrization. However, for the subsequent algebra it is more convenient to keep the compact symbolic notation

$$D_1 = \tau - u\, m_{21} + c\, m_{22}, \qquad D_2 = \tau - m_{34} + c\, m_{44}, \tag{254}$$

and substitute the explicit formulas only at the end. Doing so leads to cleaner rational expressions in $g(u)$ and $g(v)$.

Solution of the mixed sector. We now turn to $m_{13}, m_{14}, m_{23}, m_{24}$, which satisfy the linear system (234). Substituting the explicit u- and v-sector solutions (244)–(247) and (249)–(252) into (234), and solving the resulting linear equations, we obtain

$$m_{13}(u,v) = \frac{\alpha\, g(u)\, g(v)}{\Delta(u,v)}, \tag{255}$$

$$m_{14}(u,v) = \frac{\tau\, g(u)\big(g(v)-1\big)\big(\alpha+g(u)-1\big)}{(\tau-1)\,\Delta(u,v)}, \tag{256}$$

$$m_{23}(u,v) = \frac{\tau\, g(u)\, g(v)\big(\alpha+g(u)-1\big)\big(\alpha+g(v)-1\big)}{c(\tau-1)\,\Delta(u,v)}, \tag{257}$$

$$m_{24}(u,v) = \frac{\tau^3\, g(u)\big(g(v)-1\big)\big(\alpha+g(u)-1\big)\big(\alpha+g(u)+g(v)-1\big)}{c(\tau-1)^3\,\Delta(u,v)}, \tag{258}$$

where

$$\Delta(u,v) := \alpha - \big(g(u)-1\big)\big(g(v)-1\big)\left[1 + \frac{c-1}{\tau-1}\big(\alpha+g(u)+g(v)-1\big)\right]. \tag{259}$$

At this stage, the resolvent block relevant for the generalization error has been obtained explicitly, since

$$F(u,v) = m_{13}(u,v) \tag{260}$$

in the high-dimensional limit. Substituting (255) into (61) already yields an explicit representation of the asymptotic generalization error.

In conclusion, combining (61) with (255), we conclude that the asymptotic generalization error is given by

$$\mathcal{E}_t = \sum_{p=0}^{t}\sum_{q=0}^{t} \binom{t}{p}\binom{t}{q}\left(-\frac{1}{\alpha}\right)^{p+q} \frac{1}{p!\,q!}\partial_u^p \partial_v^q m_{13}(u,v)\Big|_{u=v=0}, \tag{261}$$

which is exactly the statement of Result 1.

This completes the derivation, up to the routine verification that the branch of $g(u)$ chosen in (243) is the one analytic at the origin and therefore compatible with the Taylor expansion used in (61).

D Numerical evaluation of the theoretical prediction

We numerically evaluate the theoretical formula in Result 1 through coefficient extraction from a bivariate generating function. A direct implementation of (261) is numerically unstable for moderate or large t, because it involves substantial cancellation among many terms. Instead, we compute the same quantity by introducing

$$B(x,y) := \frac{1}{(1-x)(1-y)}\, m_{13}\left(-\frac{x}{\alpha(1-x)}, -\frac{y}{\alpha(1-y)}\right), \tag{262}$$

for which

$$\mathcal{E}_t = [x^t y^t]\, B(x, y). \tag{263}$$

Therefore, the problem reduces to computing the diagonal coefficients of $B(x, y)$.

To do so, we first expand

$$h(x) := g\left(-\frac{x}{\alpha(1-x)}\right) = \sum_{n\geq 0} h_n x^n, \tag{264}$$

where $g(u)$ is the resolvent function appearing in Result 1 (see (243)). Substituting $u = -x/(\alpha(1-x))$ into the quadratic relation (242) for $g(u)$, i.e. into

$$ug(u)^2 - \big(c + (1-\alpha)u\big)g(u) + c = 0 \tag{265}$$

gives

$$xh(x)^2 + \big(\alpha c - \big(2+\sigma^2\big)x\big)h(x) - \alpha c(1-x) = 0, \tag{266}$$

with c as in (232). Writing $h(x) = \sum_{n\geq 0} h_n x^n$ and matching coefficients yields the recurrence

$$h_0 = 1, \tag{267}$$

$$h_n = \frac{\big(2+\sigma^2\big)h_{n-1} - \sum_{k=0}^{n-1} h_k h_{n-1-k} - \alpha c\, \mathbf{1}_{\{n=1\}}}{\alpha c}, \qquad n \geq 1. \tag{268}$$

This allows us to compute the coefficients of $h(x)$ stably up to any finite order.

For $\lambda = 0$, the closed form for $m_{13}(u, v)$ is (255), i.e. $\alpha\, g(u)g(v)/\Delta(u, v)$ with $\Delta(u, v)$ defined in (259). Equivalently, Result 1 records the same expression as

$$m_{13}(u, v) = \frac{\alpha\, g(u)\, g(v)}{\alpha - (g(u)-1)(g(v)-1)\left[1 + \frac{1+\sigma^2}{\alpha(\tau-1)}\big(\alpha + g(u) + g(v) - 1\big)\right]}. \tag{269}$$

After the change of variables above, $B(x, y)$ satisfies

$$D(x, y)\, B(x, y) = \frac{\alpha\, h(x)h(y)}{(1-x)(1-y)}, \tag{270}$$

where

$$D(x, y) = \alpha - (h(x)-1)(h(y)-1)\left[1 + \frac{c-1}{\tau-1}\big(\alpha + h(x) + h(y) - 1\big)\right]. \tag{271}$$

We then expand

$$B(x, y) = \sum_{m,n\geq 0} b_{m,n} x^m y^n, \qquad D(x, y) = \sum_{i,j\geq 0} d_{i,j} x^i y^j. \tag{272}$$

Since the constant term satisfies $d_{0,0} = \alpha \neq 0$, the coefficients $b_{m,n}$ are determined recursively from

$$\sum_{i=0}^{m}\sum_{j=0}^{n} d_{i,j}\, b_{m-i,n-j} = \alpha \left(\sum_{k=0}^{m} h_k\right)\left(\sum_{\ell=0}^{n} h_\ell\right). \tag{273}$$

Equivalently,

$$b_{m,n} = \frac{1}{\alpha}\left(\alpha \left(\sum_{k=0}^{m} h_k\right)\left(\sum_{\ell=0}^{n} h_\ell\right) - \sum_{\substack{0\leq i\leq m,\ 0\leq j\leq n \\ (i,j)\neq(0,0)}} d_{i,j}\, b_{m-i,n-j}\right). \tag{274}$$

We evaluate this recursion in increasing total degree $m + n$, so that each coefficient depends only on previously computed lower-degree coefficients.

Finally, the theoretical prediction is obtained from the diagonal coefficients in (263),

$$\mathcal{E}_t = b_{t,t}. \tag{275}$$

In practice, to compute $\mathcal{E}_t$ for $t = 0, \ldots, t_{\max}$, we generate the coefficients up to degree $t_{\max}$ in each variable and then read off the diagonal entries. Since the coefficients can vary substantially in magnitude for large t, we use high-precision arithmetic. This coefficient-based procedure is substantially more stable than directly evaluating the derivative formula (261) (Result 1).

E Theoretical and Experimental Validation

In this appendix, we empirically validate the theoretical prediction derived from random matrix theory for the weight prediction model with linear attention. Figure 4 compares the asymptotic theoretical curve under the ridgeless limit with numerical experiments performed at finite feature dimension D across several choices of (α, τ). These parameter settings are chosen to cover qualitatively different regimes, including the saturation, overthinking, polynomial decay, and exponential decay regimes.

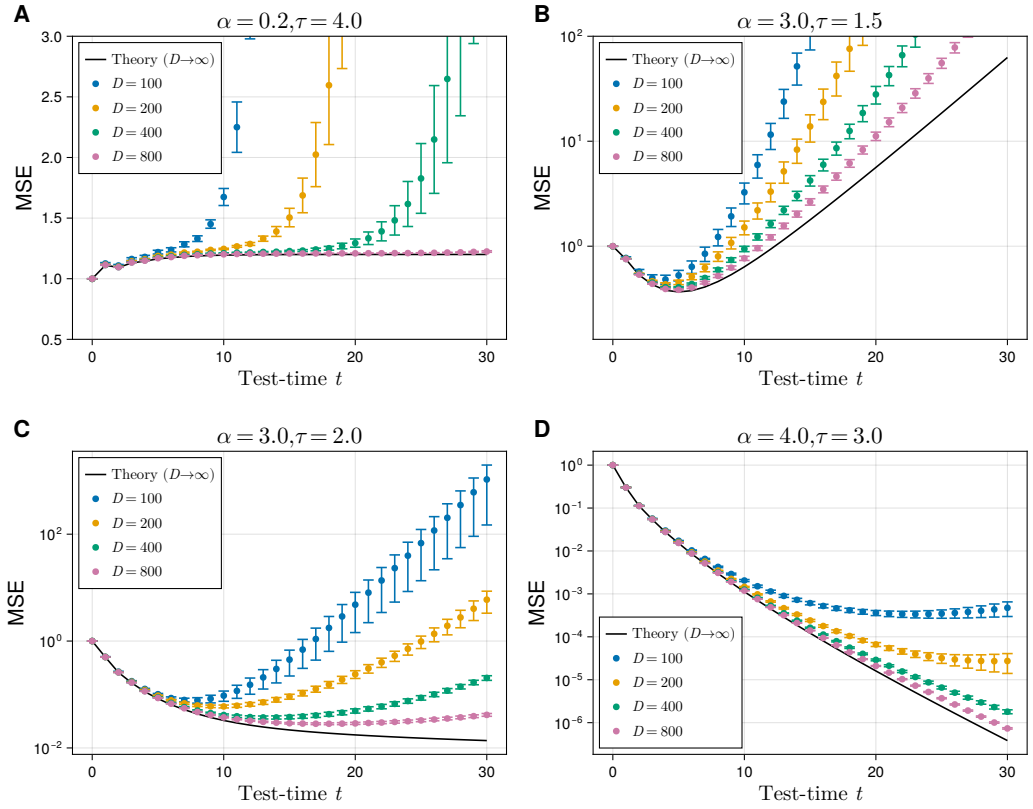

Figure 4: **Comparison between the theoretical prediction and numerical experiments.** The solid lines show the theoretical prediction in the $D \to \infty$ limit, while scatter points show numerical results at finite D. Parameters: (A-D) $\lambda = 10^{-5}$, $\sigma = 0.01$. Error bars represent the standard error of the mean over 5 trials per point.

Overall, the numerical results show clear agreement with the theoretical prediction. In all parameter settings we tested, the finite-D experimental curves approach the $D \to \infty$ theoretical prediction as D increases, indicating that the asymptotic theory accurately captures the large-dimensional behavior of the model even when compared against finite-size simulations.

The agreement is particularly close at small test-time depth t, where finite-size effects remain limited. As t becomes larger, discrepancies between theory and experiment gradually accumulate, and the empirical MSE tends to become larger than the theoretical prediction. This deviation is consistent with finite-size corrections that are not captured by the leading-order asymptotic analysis. Importantly, however, these discrepancies diminish as D increases, and the experiments continue to converge toward the theoretical curve, supporting the validity of the theoretical characterization.

F Derivation of Result 2

In this appendix, we derive the switching criterion stated in Result 2. The starting point is the diagonal coefficient representation of the quantity $\mathcal{E}_t^{\text{new}}$, which allows us to study its large-t behavior through

the singularities of a two-variable generating function. Our goal is to identify the singular point that controls the diagonal asymptotics and to determine when its modulus crosses the unit circle. This yields an explicit critical value $\tau_c(\alpha, \sigma^2)$ at which the exponential rate changes from growth to decay.

We begin by restating the result.

Result 6. *Assume $\tau > 1$. Asymptotically as $t \to \infty$, the test-time scaling law of the generalization error obeys*

$$\mathcal{E}_t - \mathcal{E}_\infty \sim K\, t^{-1/2}\, \Lambda(\alpha, \tau, \sigma^2)^t, \tag{18}$$

where K is a positive constant and the limiting value $\mathcal{E}_\infty$ is given by

$$\mathcal{E}_\infty = \begin{cases} \dfrac{(1-\alpha)(\tau-1)}{\tau - 2 - \sigma^2}, & 0 < \alpha < 1 \text{ and } \tau \geq \tau_c(\alpha, \sigma^2) \\ 0, & \text{otherwise}, \end{cases} \tag{19}$$

and the exponential rate satisfies

$$\Lambda(\alpha, \tau, \sigma^2) \begin{cases} > 1, & \tau < \tau_c(\alpha, \sigma^2), \\ = 1, & \tau = \tau_c(\alpha, \sigma^2), \\ < 1, & \tau > \tau_c(\alpha, \sigma^2). \end{cases} \tag{20}$$

Here, the critical value $\tau_c(\alpha, \sigma^2)$ is given by

$$\tau_c(\alpha, \sigma^2) = 1 + \frac{(1+\sigma^2)(\alpha + 1 + 2\sigma^2 - \sqrt{\Delta})(2\alpha + 2 + 2\sigma^2 - \sqrt{\Delta})}{2\alpha\sqrt{\Delta}}, \tag{21}$$

with $\Delta := (\alpha + 1 + 2\sigma^2)^2 - 4\alpha$.

We now derive this result in detail.

First, from the coefficient formula,

$$\mathcal{E}_t^{\text{new}} = [x^t y^t]\, B(x, y), \tag{276}$$

where

$$B(x, y) := \frac{1}{(1-x)(1-y)}\, m_{13}\left(-\frac{x}{\alpha(1-x)}, -\frac{y}{\alpha(1-y)} \right). \tag{277}$$

The function $m_{13}(u, v)$ is meromorphic away from the branch locus of the scalar Wishart resolvent g, and its singularities are determined by the vanishing of its denominator

$$D(u, v) := \alpha - (g(u) - 1)(g(v) - 1) \left[1 + \frac{1+\sigma^2}{\alpha(\tau-1)} \big(\alpha + g(u) + g(v) - 1 \big) \right]. \tag{278}$$

Accordingly, the singular variety relevant for the diagonal coefficient is

$$\mathcal{V} := \{(x, y) : D(u(x), v(y)) = 0\}, \tag{279}$$

where

$$u(x) := -\frac{x}{\alpha(1-x)}, \qquad v(y) := -\frac{y}{\alpha(1-y)}. \tag{280}$$

Because both the singular variety and the diagonal direction (t, t) are invariant under the exchange $x \leftrightarrow y$, the natural candidate for the dominant critical point is a symmetric point of the form

$$(x, y) = (\xi, \xi) \in \mathcal{V}. \tag{281}$$

Throughout this appendix, we assume that this symmetric point is indeed the unique dominant critical point and that no other singularity of smaller modulus contributes to the diagonal asymptotics. Under this assumption, the problem reduces to a one-parameter analysis along the symmetric branch, which we now describe.

On the symmetric branch, $u = v$ and hence $g(u) = g(v)$. The scalar resolvent $g(u)$ satisfies the quadratic equation

$$u\, g(u)^2 - \big(c + (1-\alpha)u \big) g(u) + c = 0, \tag{282}$$

where

$$c = 1 + \frac{1+\sigma^2}{\alpha}. \tag{283}$$

Solving this equation for u, we obtain

$$u = -\frac{c(1-g(u))}{g(u)(\alpha - 1 + g(u))}. \tag{284}$$

Combining this with

$$u(x) = -\frac{x}{\alpha(1-x)}, \tag{285}$$

we obtain

$$x = \frac{\alpha c(1-g(u))}{g(u)(\alpha-1+g(u)) + \alpha c(1-g(u))} = \frac{(\alpha+1+\sigma^2)(1-g(u))}{\alpha + \sigma^2(1-g(u)) + (1-g(u))^2}. \tag{286}$$

Next, we rewrite the singularity condition on the diagonal. The condition $D(u,u) = 0$ becomes

$$(1-g(u))^2 \left[1 + \frac{1+\sigma^2}{\alpha(\tau-1)}\big(\alpha - 1 + 2g(u)\big)\right] = \alpha. \tag{287}$$

Therefore, the symmetric singular points are determined by (287), with their location in the x-plane given by (286).

Under the dominant symmetric critical-point assumption, the diagonal coefficient is governed by a smooth critical point (ξ, ξ) on the singular variety. Standard smooth-point asymptotics in two variables then gives

$$[x^t y^t]\, B(x,y) \sim K\, t^{-1/2}\, (\xi^2)^{-t} \tag{288}$$

for some nonzero constant K. Equivalently,

$$\mathcal{E}_t^{\text{new}} \sim K\, t^{-1/2}\, \Lambda(\alpha,\tau,\sigma^2)^t, \qquad \Lambda(\alpha,\tau,\sigma^2) := |\xi|^{-2}. \tag{289}$$

Thus the long-time behavior is determined by the position of ξ relative to the unit circle:

$$|\xi| < 1 \Longrightarrow \Lambda > 1 \qquad \text{(exponential growth)}, \tag{290}$$
$$|\xi| = 1 \Longrightarrow \Lambda = 1 \qquad \text{(critical polynomial behavior)}, \tag{291}$$
$$|\xi| > 1 \Longrightarrow \Lambda < 1 \qquad \text{(exponential damping)}. \tag{292}$$

We now determine the parameter value at which this crossing occurs. Along the relevant symmetric branch, the transition takes place when the dominant point reaches the unit circle. The branch connected to the large-τ regime crosses at $\xi = -1$. Substituting $x = -1$ into (286), we obtain

$$(1-g(u))^2 + (\alpha + 1 + 2\sigma^2)(1-g(u)) + \alpha = 0. \tag{293}$$

Define

$$\Delta := (\alpha + 1 + 2\sigma^2)^2 - 4\alpha. \tag{294}$$

The relevant root on the dominant symmetric branch is

$$1 - g(u) = -\frac{\alpha + 1 + 2\sigma^2 - \sqrt{\Delta}}{2}. \tag{295}$$

Substituting this root into (287) and solving for τ, we obtain

$$\tau_c(\alpha,\sigma^2) = 1 + \frac{(1+\sigma^2)(\alpha+1+2\sigma^2-\sqrt{\Delta})(2\alpha+2+2\sigma^2-\sqrt{\Delta})}{2\alpha\sqrt{\Delta}}. \tag{296}$$

This is exactly the expression stated in Result 2.

F.1 Inference-time label noise

The phase transition above was derived with noiseless inference labels. We now add the same label noise used in pretraining and show that the stability phase diagram is unchanged. Let

$$\boldsymbol{y} = \mathbf{X}^\top \boldsymbol{w} + \boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_L), \tag{297}$$

where $\boldsymbol{\xi}$ is independent of $\boldsymbol{w}$, $\mathbf{X}$, and the pretraining data. Define $\mathbf{R} := \mathbf{I} - \alpha^{-1}\mathbf{A}\mathbf{S}$. The error recursion becomes

$$\boldsymbol{e}_{t+1} = \mathbf{R}\boldsymbol{e}_t - \frac{1}{\alpha}\mathbf{A}\mathbf{X}\boldsymbol{\xi}, \qquad \boldsymbol{e}_0 = \boldsymbol{w}, \tag{298}$$

and hence

$$\boldsymbol{e}_t = \mathbf{R}^t \boldsymbol{w} - (\mathbf{I} - \mathbf{R}^t)\mathbf{S}^\dagger \mathbf{X}\boldsymbol{\xi}, \tag{299}$$

where $\mathbf{S}^\dagger$ is the Moore–Penrose inverse of $\mathbf{S}$. Averaging over $\boldsymbol{w}$ and $\boldsymbol{\xi}$ gives

$$\mathcal{E}_t^{\text{noisy}} = \mathcal{E}_t + \frac{\sigma^2}{D}\mathbb{E}\,\mathrm{Tr}\left[(\mathbf{I} - \mathbf{R}^t)\mathbf{S}^\dagger(\mathbf{I} - \mathbf{R}^t)^\top\right]. \tag{300}$$

The inference noise changes only the additive term and leaves $\mathbf{R}$ unchanged. Therefore, the stability condition for $\mathbf{R}^t$ and the phase boundary $\tau_c(\alpha, \sigma^2)$ in Eq. (296) are unchanged. For $\alpha > 1$ in the stable regime, $\mathbf{R}^t \to \mathbf{0}$, and the additional term converges in the high-dimensional limit to

$$\frac{\sigma^2}{D}\mathbb{E}\,\mathrm{Tr}[\mathbf{S}^{-1}] \longrightarrow \frac{\sigma^2}{\alpha - 1}. \tag{301}$$

Thus, inference noise raises the error floor and can reduce the optimal reasoning depth, but it does not move the phase boundary.

G Derivation of Theorem 3

In this appendix, we derive the asymptotic scaling law stated in Theorem 3. We begin by restating the result.

Theorem 3. *Assume the population-risk regime $\tau \to \infty$, ridgeless learning $\lambda = 0$, and $\alpha > 1$. Then, as $t \to \infty$,*

$$\mathcal{E}_t \asymp t^{-3/2}\left(\frac{2\sqrt{\alpha} + \sigma^2}{\alpha + 1 + \sigma^2}\right)^{2t}. \tag{22}$$

Especially, in the context-rich limit $\alpha \gg \sigma^2 + 1$, $\mathcal{E}_t \asymp t^{-3/2}(4/\alpha)^t$.

Proof. We work in the population-risk regime $\tau \to \infty$ under ridgeless learning $\lambda = 0$, and assume $\alpha > 1$. In this regime, the test-time generalization error can be written as

$$\mathcal{E}_t = \int \left(1 - \frac{s}{\alpha c}\right)^{2t} \mu_\alpha(ds), \tag{302}$$

where $c = 1 + (1 + \sigma^2)/\alpha$ and μ_α is the Marchenko–Pastur law with aspect ratio α. Here, μ_α is the limiting empirical eigenvalue distribution of $\mathbf{S}$. Since $\alpha > 1$, its mass at zero is zero. Hence μ_α consists only of the following density:

$$\mu_\alpha(ds) = \frac{\sqrt{(s_+ - s)(s - s_-)}}{2\pi s}\mathbf{1}_{[s_-, s_+]}(s)\,ds, \qquad s_\pm = (1 \pm \sqrt{\alpha})^2. \tag{303}$$

Fix a small $\varepsilon > 0$, and split (302) as

$$\mathcal{E}_t = \int_{s_-}^{s_- + \varepsilon}\left(1 - \frac{s}{\alpha c}\right)^{2t}\mu_\alpha(ds) + \int_{s_- + \varepsilon}^{s_+}\left(1 - \frac{s}{\alpha c}\right)^{2t}\mu_\alpha(ds). \tag{304}$$

We begin with contribution of the first term in Equation (304). Write

$$s = s_- + x, \qquad x \in [0, \varepsilon]. \tag{305}$$

Since

$$s - s_- = x, \qquad s_+ - s = (s_+ - s_-) - x, \tag{306}$$

and $s_- = (\sqrt{\alpha} - 1)^2 > 0$, the density satisfies

$$\frac{\sqrt{(s_+ - s)(s - s_-)}}{2\pi s} \asymp x^{1/2} \qquad (x \downarrow 0). \tag{307}$$

On the other hand, we have

$$\left(1 - \frac{s}{\alpha c}\right)^{2t} = r_\alpha^{2t} \left(1 - \frac{x}{\alpha c\, r_\alpha}\right)^{2t}, \tag{308}$$

where

$$r_\alpha := 1 - \frac{s_-}{\alpha c} = \frac{2\sqrt{\alpha} + \sigma^2}{\alpha + 1 + \sigma^2} \in (0, 1). \tag{309}$$

For $x \in [0, \varepsilon]$ with ε sufficiently small, the factor in parentheses is bounded above and below by exponentials, namely

$$e^{-C_2 t x} \le \left(1 - \frac{x}{\alpha c\, r_\alpha}\right)^{2t} \le e^{-C_1 t x} \tag{310}$$

for some constants $C_1, C_2 > 0$ independent of t. Therefore,

$$\int_{s_-}^{s_- + \varepsilon} \left(1 - \frac{s}{\alpha c}\right)^{2t} \mu_\alpha(ds) \asymp r_\alpha^{2t} \int_0^\varepsilon e^{-Ctx} x^{1/2}\, dx \tag{311}$$

$$= r_\alpha^{2t} t^{-3/2} \int_0^{\varepsilon t} e^{-Cu} u^{1/2}\, du \tag{312}$$

$$\asymp r_\alpha^{2t} t^{-3/2}. \tag{313}$$

Next, we consider the contribution of the second term in Equation (304). For $\sigma^2 > 0$ and sufficiently small $\varepsilon > 0$, continuity and convexity give an $\eta > 0$ such that

$$\left|1 - \frac{s}{\alpha c}\right| \le r_\alpha - \eta \qquad (s \in [s_- + \varepsilon, s_+]). \tag{314}$$

Since μ_α is a finite measure, this yields

$$\int_{s_- + \varepsilon}^{s_+} \left(1 - \frac{s}{\alpha c}\right)^{2t} \mu_\alpha(ds) \le C(r_\alpha - \eta)^{2t}. \tag{315}$$

This term is negligible compared with (313).

If $\sigma^2 = 0$, the density in (303) is $\asymp (s_+ - s)^{1/2}$ near s_+. Writing $x = s_+ - s$, we have $1 - s/(\alpha c) = -r_\alpha + x/(\alpha c)$. Hence the same edge calculation gives

$$\int_{s_+ - \varepsilon}^{s_+} \left(1 - \frac{s}{\alpha c}\right)^{2t} \mu_\alpha(ds) \asymp t^{-3/2} r_\alpha^{2t}. \tag{316}$$

On $[s_- + \varepsilon, s_+ - \varepsilon]$, the absolute value of $1 - s/(\alpha c)$ is at most $r_\alpha - \eta$ for some $\eta > 0$. Its contribution is therefore exponentially smaller than either edge contribution.

Combining (304) and (313) with the bounds above, we conclude that

$$\mathcal{E}_t \asymp t^{-3/2} r_\alpha^{2t} = t^{-3/2} \left(\frac{2\sqrt{\alpha} + \sigma^2}{\alpha + 1 + \sigma^2}\right)^{2t}. \tag{317}$$

This proves the first claim.

Finally, when $\alpha \to \infty$ with $\sigma^2 = o(\sqrt{\alpha})$, the exponential base satisfies

$$\frac{2\sqrt{\alpha} + \sigma^2}{\alpha + 1 + \sigma^2} = \frac{2}{\sqrt{\alpha}}\big(1 + o(1)\big), \tag{318}$$

Hence, equivalently,

$$\mathcal{E}_t \asymp t^{-3/2} \left(\frac{4}{\alpha}\right)^t. \tag{319}$$

□

H Derivation of Result 4

In this appendix, we derive the optimal test-time depth and minimum error in the overthinking regime stated in Result 4.

We begin by restating the result.

Result 7. *Assume that (α, τ, σ^2) is in the overthinking regime. Let $t^* := \operatorname{argmin}_{t \in \mathbb{Z}_{\geq 0}} \mathcal{E}_t$. Then, as $\tau \to \tau_c(\alpha, \sigma^2)$ from below,*

$$t^* = C_t(\alpha, \sigma^2)\big(\tau_c(\alpha, \sigma^2) - \tau\big)^{-1}\big(1 + o(1)\big), \tag{23}$$

for some positive function $C_t(\alpha, \sigma^2)$. Moreover, the generalization error at t^* satisfies

$$\mathcal{E}_{t^*} = C_E(\alpha, \sigma^2)\,(t^*)^{-1/2}\big(1 + o(1)\big), \tag{24}$$

for some positive function $C_E(\alpha, \sigma^2)$.

We derive Result 4 from Result 2. Let

$$\delta := \tau_c(\alpha, \sigma^2) - \tau > 0. \tag{320}$$

In the overthinking regime, we have $\tau < \tau_c(\alpha, \sigma^2)$, and hence Result 2 gives

$$\mathcal{E}_t \sim K(\alpha, \tau, \sigma^2)\, t^{-1/2} \Lambda(\alpha, \tau, \sigma^2)^t \qquad (t \to \infty), \tag{321}$$

since $\mathcal{E}_\infty = 0$ in this regime.

Assume, as in the derivation of Result 2, that the dominant symmetric critical point remains unique near the transition and crosses the unit circle transversely at $\tau = \tau_c$. Then $\Lambda(\alpha, \tau, \sigma^2)$ is smooth near τ_c, and we have

$$\log \Lambda(\alpha, \tau, \sigma^2) = \kappa(\alpha, \sigma^2)\,(\tau_c(\alpha, \sigma^2) - \tau) + o(\tau_c - \tau), \tag{322}$$

for some $\kappa(\alpha, \sigma^2) > 0$. Moreover,

$$K(\alpha, \tau, \sigma^2) = K_c(\alpha, \sigma^2) + o(1), \tag{323}$$

with $K_c(\alpha, \sigma^2) > 0$. Substituting (320), (322), and (323) into (321), we obtain

$$\mathcal{E}_t \sim K_c(\alpha, \sigma^2)\, t^{-1/2} \exp\big(\kappa(\alpha, \sigma^2)\delta\, t\big). \tag{324}$$

We now minimize the continuous proxy suggested by (324),

$$\Phi_\delta(t) := t^{-1/2} e^{\kappa \delta t}. \tag{325}$$

One can check that the unique minimizer of (325) is

$$t_{\text{cont}} = \frac{1}{2\kappa(\alpha, \sigma^2)\delta}. \tag{326}$$

As $\delta \downarrow 0$, this tends to $+\infty$, so the large-t asymptotics is self-consistent. Passing back to the discrete minimizer $t^* \in \mathbb{Z}_{\geq 0}$, we obtain

$$t^* = \frac{C_t(\alpha, \sigma^2)}{\tau_c(\alpha, \sigma^2) - \tau}\big(1 + o(1)\big), \tag{327}$$

where

$$C_t(\alpha, \sigma^2) = \frac{1}{2\kappa(\alpha, \sigma^2)} > 0. \tag{328}$$

Finally, substituting (327) into (324) gives

$$\mathcal{E}_{t^*} \sim K_c(\alpha, \sigma^2)\,(t^*)^{-1/2} \exp\big(\kappa(\alpha, \sigma^2)\delta\, t^*\big) \tag{329}$$

$$= K_c(\alpha, \sigma^2) e^{1/2} (t^*)^{-1/2} (1 + o(1)). \tag{330}$$

Hence

$$\mathcal{E}_{t^*} = C_E(\alpha, \sigma^2)\,(t^*)^{-1/2}(1 + o(1)), \tag{331}$$

where

$$C_E(\alpha, \sigma^2) := K_c(\alpha, \sigma^2) e^{1/2} > 0. \tag{332}$$

Equation (331) is exactly the claimed form (24), which proves Result 4.

I Experimental details

In this appendix, we summarize the experimental details for the fully learned linear-attention and softmax-attention experiments in Section 6. All experiments used synthetic linear-regression tasks with ambient dimension $D = 50$. Each task was defined by a ground-truth parameter $\boldsymbol{w}_\mu \sim \mathcal{N}(\mathbf{0}, I_D)$, and each in-context example was generated as

$$\boldsymbol{x}_{\mu,l} \sim \mathcal{N}(\mathbf{0}, I_D), \qquad y_{\mu,l} = \boldsymbol{w}_\mu^\top \boldsymbol{x}_{\mu,l}.$$

Thus, the experiments were noiseless. We used the same prompt structure as in Section 3, but initialized the first parameter-estimate token with a random Gaussian vector.

I.1 Model architecture

Both models used a single residual self-attention layer acting on tokens in $\mathbb{R}^{2D+2}$. For $D = 50$, the token dimension was therefore $2D + 2 = 102$. To compute the next estimate from reasoning depth t, the prompt contained L example tokens followed by $t + 1$ parameter-estimate tokens, corresponding to

$$[\hat{\boldsymbol{w}}_0, \ldots, \hat{\boldsymbol{w}}_t].$$

The total number of tokens was therefore

$$T = L + t + 1.$$

The prediction was always read from the parameter block of the final token, which we denote by $\boldsymbol{z}_T$.

In both models, only the final token was updated according to

$$\boldsymbol{z}_T \mapsto \boldsymbol{z}_T + W_V \sum_{s=1}^{T} a_s \boldsymbol{z}_s, \tag{333}$$

where $W_K, W_Q, W_V \in \mathbb{R}^{(2D+2)\times(2D+2)}$ denote the key, query, and value matrices.

In the fully learned linear-attention model, we used

$$a_s = m_s^{(t)} \frac{1}{L} \left(W_K \boldsymbol{z}_s\right)^\top \left(W_Q \boldsymbol{z}_T\right), \tag{334}$$

where $m_s^{(t)}$ is a mask over source positions that determines which source tokens are visible to the final token.

In the softmax-attention model, we used the same bilinear score, but normalized it with a softmax over the visible source positions:

$$a_s = \frac{\exp\left(\frac{(W_K \boldsymbol{z}_s)^\top (W_Q \boldsymbol{z}_T)}{\sqrt{2D+2}} + b_s^{(t)}\right)}{\sum_{r=1}^{T} \exp\left(\frac{(W_K \boldsymbol{z}_r)^\top (W_Q \boldsymbol{z}_T)}{\sqrt{2D+2}} + b_r^{(t)}\right)}, \tag{335}$$

where $b_s^{(t)}$ is an additive attention mask.

For both the fully learned linear-attention and softmax-attention experiments, we used a full-history visibility pattern for the final parameter-estimate token. That is, the final token was allowed to attend to all L example tokens and all previously generated estimate tokens, but not to itself. For the fully learned linear-attention model, this mask was

$$m_s^{(t)} = \begin{cases} 1, & 1 \le s \le L + t, \\ 0, & s = T. \end{cases} \tag{336}$$

For the softmax-attention model, the corresponding additive mask was

$$b_s^{(t)} = \begin{cases} 0, & 1 \le s \le L + t, \\ -\infty, & s = T. \end{cases} \tag{337}$$

Thus, in both models, the visible source tokens consisted of the L example tokens and the previous estimate tokens $[\hat{\boldsymbol{w}}_0, \ldots, \hat{\boldsymbol{w}}_{t-1}]$, while the final query token $\hat{\boldsymbol{w}}_t$ was masked from attending to itself.

I.2 Learning protocol

For each pair (M, L), we first sampled a fixed training set of M tasks, each with L in-context examples. We then trained the model parameters on this fixed set by minimizing the MSE between the predicted parameter vector and the ground-truth task parameter, with ℓ_2 regularization on all trainable matrices.

Training used a one-step objective. For each sampled training task, we initialized the first parameter estimate $\hat{\boldsymbol{w}}_0$ as a random Gaussian vector, constructed a prompt containing the L example tokens and this single estimate token, and trained the model to predict the ground-truth parameter $\boldsymbol{w}$ from the updated final token.

Optimization used Adam with learning rate 10^{-3}, minibatch size 300, 3000 gradient steps, and regularization coefficient $\lambda = 10^{-5}$.

I.3 Inference protocol

After training, we evaluated each model on newly sampled test tasks from the same noiseless distribution. For each test task, we initialized the first parameter estimate $\hat{\boldsymbol{w}}_0$ as a random Gaussian vector and then iteratively reapplied the same learned attention layer to generate a trajectory $(\hat{\boldsymbol{w}}_t)_{t\geq 0}$.

At reasoning depth t, the prompt contained the L example tokens and the history of generated estimates

$$[\hat{\boldsymbol{w}}_0, \ldots, \hat{\boldsymbol{w}}_t].$$

The model updated the final estimate token to produce the next estimate $\hat{\boldsymbol{w}}_{t+1}$. Under the full-history mask, the final token could attend to the example tokens and the previous estimate tokens $[\hat{\boldsymbol{w}}_0, \ldots, \hat{\boldsymbol{w}}_{t-1}]$, but not to itself. Thus, both the fully learned linear-attention and softmax-attention models had access to the full generated history during inference, except for the current final token itself as an attention source.

We evaluated the MSE of the parameter estimate

$$\mathcal{E}_t = \frac{1}{D}\,\mathbb{E}\big[\|\boldsymbol{w} - \hat{\boldsymbol{w}}_t\|^2\big] \tag{338}$$

for reasoning depth t. Evaluation used 4096 independently sampled test tasks for each trained model.

I.4 Computational resources

All experiments were run in PyTorch. The experiments summarized here were run on CPU using an AMD EPYC 9654P 96-Core processor.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We state the main contributions in Section 1. These claims are matched by the results in Section 4 and Section 5, and their assumptions and scope are clarified in the corresponding statements.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss the limitations of the work in the Section 7.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: We discuss the assumptions and proofs in the Appendices F, G, and H.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: We provide detailed information to reproduce our experiments in Appendix I.

   Guidelines:

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

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

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

Answer: [Yes]

Justification: We provide the reproduction code in a GitHub repository in the last of Section 1.

Guidelines:

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

6. **Experimental setting/details**

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

Answer: [Yes]

Justification: We discuss the experimental setting in the Appendix I.

Guidelines:

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

7. **Experiment statistical significance**

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

Answer: [Yes]

Justification: We report error bars in the caption of the each figure.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: We report the compute resources in the Appendix I.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: We conform to the NeurIPS Code of Ethics.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: We provide a discussion of potential impacts in a Impact Statement subsection at the end of Section 1.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: We do not release any datasets or models that pose a significant risk of misuse, so no additional safeguards are required.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We have credited all third-party assets and fully complied with their license terms.

    Guidelines:

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

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [Yes]

    Justification: We provide the reproduction code in a GitHub repository including documentation.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: We do not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: We do not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: The core methodology of this research does not involve LLMs as an important, original, or non-standard component.

    Guidelines:

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