A SCALING THEORY OF MULTI-EPOCH PRETRAINING: LATE-TIME OVERFITTING AND ENSEMBLING

Anonymous authors
Paper under double-blind review

ABSTRACT

In the near future language model pretraining will become bound by total data rather than total compute. Such a regime requires revisiting scaling laws and scaling law theory. Motivated by prior works that have found utility in ensembling, we study how the number of training epochs and ensemble size jointly shape generalization when training data are fixed. In language-model experiments on FineWeb, logit averaging reduces late-training deterioration, delays optimal stopping, and can outperform wider or deeper single models at matched parameter-token budgets over a common training horizon. To analyze this interaction between data reuse and ensembling, we derive a dynamical mean-field theory (DMFT) for multi-epoch stochastic gradient descent in power-law random-feature ensembles, tracking error correlations across epochs and models. Ensembling reduces the prediction variance responsible for late-time growth in expected test error, changing the optimal stopping time. We use these stopping times to derive compute-allocation laws at fixed data and individual model sizes. Under this allocation, ensemble size grows with compute at least as fast as training time per member, while excess test loss above the limiting loss decreases as a power of compute.

1 INTRODUCTION

Scaling laws for one-pass training prescribe how model size and training data should grow with compute (Kaplan et al., 2020; Hoffmann et al., 2022). The supply of high-quality unique text, however, need not grow at the same rate as compute (Villalobos et al., 2024; Muennighoff et al., 2023; Kim et al., 2025). Once a corpus is reused, additional optimization has two opposing effects: it continues to fit the available data, but it can also amplify noise tied to that particular dataset, model, or training order. The resulting validation curve is often non-monotone even though training loss keeps falling.

Non-Monotone → Monotone Validation with Epoching. In this data-reuse regime, simple small scale language-model experiments can reproduce how validation loss evolves over repeated training epochs and how ensembling changes this trajectory. At fixed model size, across 10–100 million unique FineWeb tokens, validation loss reaches its minimum after roughly 9–14 epochs and then rises with further training (Figure 1). Averaging independently trained models' logits reduces this rise and can outperform a wider or deeper model at matched compute (Figure 2 and Appendix A.2). These observations motivate two questions: which component of late-training error can be reduced by averaging independently trained models, and how does this reduction change the loss-minimizing training time?

Theoretical Model. We address these questions using an ensemble of E random-feature models, each with N trainable parameters and an independently sampled feature matrix. The models train on the same P examples using fixed-order stochastic gradient descent (SGD), which reuses the same ordered minibatches in every epoch. We derive a dynamical mean-field theory (DMFT) for the training and test losses of individual models and their averaged predictions. The theory describes these losses through error correlations within and across epochs and between ensemble members, together with response functions that measure sensitivity to small perturbations and quantify overfitting dynamics that build up across epochs.

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

Relevant Prior Works Our closest theoretical starting point is the DMFT framework of Bordelon et al. (2024), which describes finite-data and finite-width learning dynamics, including overfitting and ensembling. We extend this framework to fixed-order, multi-epoch SGD and derive ensemble-dependent stopping and compute-allocation laws for noiseless gradient flow at $N = P$.

The analysis also draws on spectral learning curves (Bordelon et al., 2020; Maloney et al., 2022; Simon et al., 2021), high-dimensional SGD dynamics (Mignacco et al., 2020; Gerbelot et al., 2022; Paquette et al., 2021), and scaling-law analyses based on Volterra equations (Paquette et al., 2024) and linear-regression risk bounds (Lin et al., 2024; Yan et al., 2025). Our experiments are motivated by data-constrained scaling (Muennighoff et al., 2023) and fixed-data, high-compute pretraining (Kim et al., 2025).

Contributions. (i) We derive a multi-epoch SGD dynamical mean field theory (DMFT) and prove a finite-horizon loss-approximation bound for Gaussian features. Our theory quantifies how both train loss and validation loss evolve over both steps within epochs and across epochs. The theory requires computing correlation and response functions which characterize both optimization progress across steps and also overfitting effects across epoch timescales. (ii) We derive ensemble-dependent stopping laws for noiseless Gaussian gradient flow at $N = P$, and compute-allocation laws under power-law source/capacity conditions at fixed data and member sizes, with each ensemble trained to its own optimal stopping time. (iii) In language-model experiments on fixed FineWeb corpora, we show that logit averaging reduces late-training deterioration and delays optimal stopping. Ensembles can outperform wider or deeper single models at matched parameter-token budgets over a common training horizon.

2 Multi-Epoch Pretraining Phenomena

We measure validation cross-entropy for autoregressive Transformers trained repeatedly on fixed FineWeb subsets, varying model and corpus size and averaging logits from independently trained models. Appendix A gives training and evaluation details.

Repeated training eventually hurts validation performance. Validation loss follows a U-shaped trajectory while training loss continues to fall (Figure 1), consistent with earlier observations under repeated pretraining data (Muennighoff et al., 2023; Kim et al., 2025). At fixed model size, validation loss reaches its minimum after roughly 9–14 epochs across the tested corpus sizes; at fixed data size, larger models reach their minima earlier (Section 3.8).

Ensembling improves the use of a fixed training budget. Logit averaging lowers validation loss and reduces its late-time rise (Figure 2A). At matched parameter-token compute, an ensemble of four base models attains a validation loss 0.16 nats per token lower than a single model with twice the width (Figure 2B; Appendix A.2). Related work also finds gains from ensembling on fixed data (Kim et al., 2025) and allocating multi-epoch budgets to diverse predictors (Mandal et al., 2026).

These observations motivate a model of repeated data use that tracks both within-model errors and the errors shared across ensemble members.

3 A Dynamical Mean-Field Theory of Fixed-Data Reuse

3.1 Random-Feature Model and Multi-Epoch Dynamics

Model and observables. We study a random-feature predictor of width N, with N trainable parameters and P training examples. SGD uses minibatches of size B:

$$f(\boldsymbol{x}) = \frac{1}{N}\boldsymbol{w}^\top \boldsymbol{A}\boldsymbol{\psi}(\boldsymbol{x}), \qquad y(\boldsymbol{x}) = \boldsymbol{w}^\star \cdot \boldsymbol{\psi}(\boldsymbol{x}) + \sigma\epsilon, \tag{1}$$

where the training feature vectors are independent draws from $\mathcal{N}(\mathbf{0}, \mathbf{\Lambda})$, with $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots)$. The matrix $\boldsymbol{A} \in \mathbb{R}^{N\times M}$ has independent standard Gaussian entries, and $\epsilon \sim \mathcal{N}(0, 1)$. The population features, random-feature matrix, and label noises are mutually independent. Here σ^2 is the variance of additive label noise, drawn once per training example, and M is the population feature dimension. We initialize the trainable weights at $\boldsymbol{w}_0(0) = 0$; ensemble members use independent draws of $\boldsymbol{A}$ on the same dataset. The DMFT allows $\sigma \geq 0$. We first analyze late-time scaling for noiseless targets and then give the label-noise extension; the compute-allocation result uses the

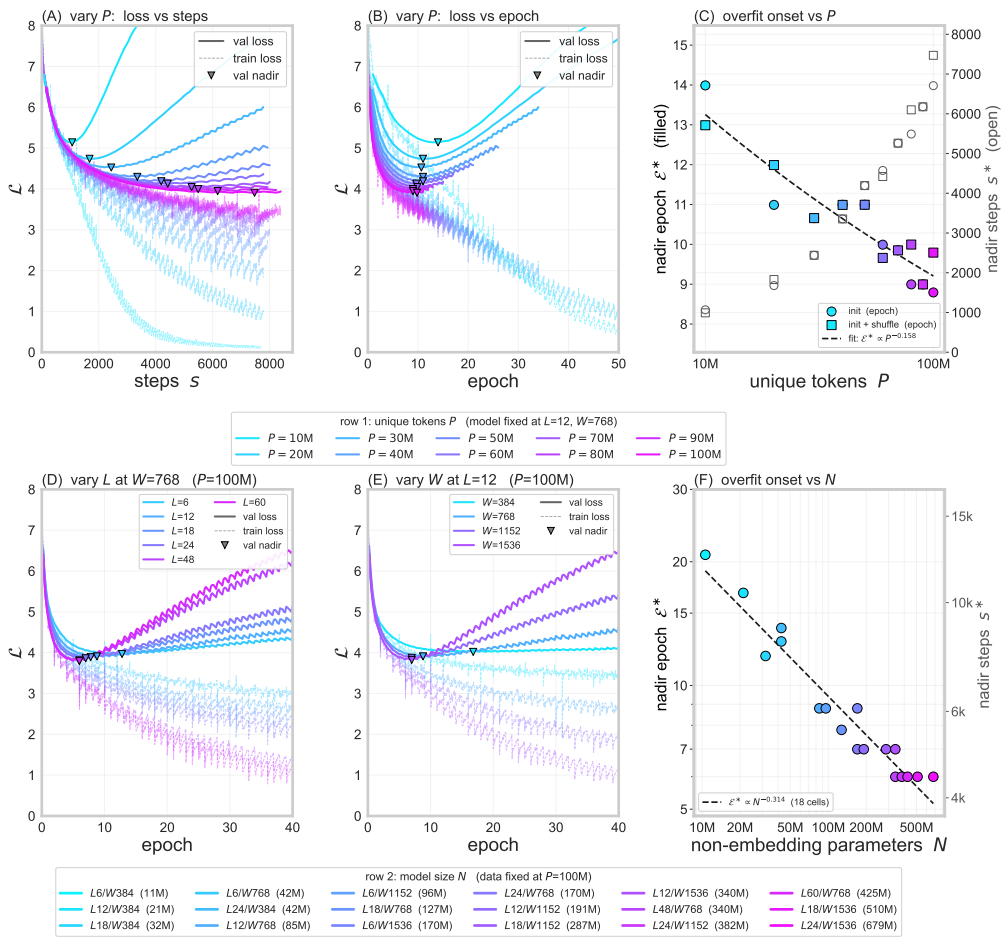

Figure 1: **Repeated training lowers training loss but eventually raises validation loss.** Top: corpus-size sweep at fixed model size; bottom: model-size sweep at fixed corpus size. Dashed/solid curves show training/validation loss; triangles mark minima. (C, F) Optimal stopping in epochs (filled) and steps (open). Larger models reach their optimum in fewer epochs; optimal epochs vary less across corpus sizes. Settings: Table 3.

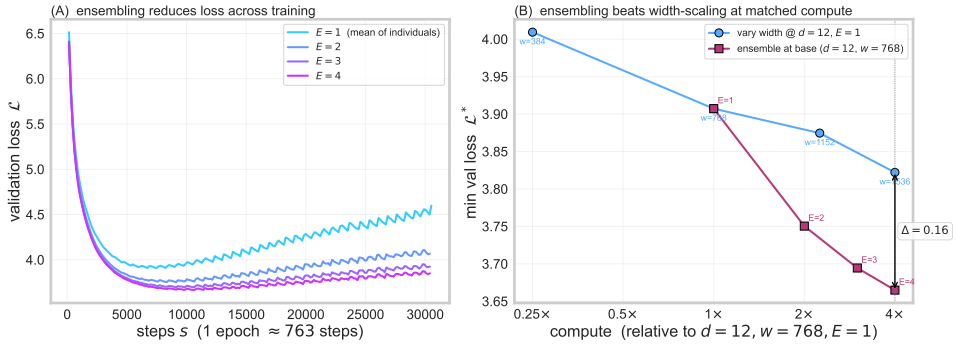

Figure 2: **Ensembling reduces late-time deterioration and improves loss at matched training budgets.** (A) Validation trajectories by ensemble size. (B) Four base models outperform one twice-wider model by 0.16 nats. Both panels use $P = 100$M; settings and evaluation coverage: Table 3.

noiseless scaling form. Setting $\sigma = 0$ removes label noise, while the dataset and random features remain random. Table 1 collects the size and training-horizon notation. For the scaling analysis, we allow infinitely many population features with summable eigenvalues, $\sum_{k=1}^{\infty} \lambda_k < \infty$, as in

dimension-free analyses of random-feature regression (Defilippis et al., 2024). Appendix B.1 separates this setting from an isotropic finite-dimensional proportional specialization.

Table 1: Size and training-horizon notation for the random-feature model.

Symbol	Meaning	Symbol	Meaning
N	Model width	M	Population feature dimension
P	Training samples	B	Samples per update
E	Ensemble members	$\mathcal{E}$	Training epochs
$T = P/B$	Updates per epoch	$S = T\mathcal{E}$	Total updates

The discrepancy between the target and learned feature coefficients is

$$\boldsymbol{v}_n^0(t) := \boldsymbol{w}^\star - \frac{1}{N}\boldsymbol{A}^\top \boldsymbol{w}_n(t).$$

For a fixed realization of the training data and random features, the test mean-squared error is

$$\mathcal{L}_{N,B}(n,t) = \boldsymbol{v}_n^0(t)^\top \boldsymbol{\Lambda} \boldsymbol{v}_n^0(t) + \sigma^2.$$

Epoch and minibatch indices. We use $n \in \{0, \dots, \mathcal{E}-1\}$ for the epoch index and $t \in \{0, \dots, T-1\}$ for the step within that epoch. The global update index and effective training time are

$$s = nT + t, \qquad \tau = \eta s.$$

Thus s/T counts the elapsed epochs. In the fixed-order setting, the same feature block $\boldsymbol{\Psi}(t) \in \mathbb{R}^{B\times M}$ and its fixed label-noise vector $\boldsymbol{\epsilon}(t)$ reappear at position t in every epoch. The exact residual recursion is

$$\begin{aligned} \boldsymbol{v}_n^0(t+1) &= \boldsymbol{v}_n^0(t) - \eta \left(\frac{1}{N}\boldsymbol{A}^\top \boldsymbol{A}\right) \frac{1}{B} \boldsymbol{\Psi}(t)^\top \left(\boldsymbol{\Psi}(t)\boldsymbol{v}_n^0(t) + \sigma\boldsymbol{\epsilon}(t)\right), \\ \boldsymbol{v}_{n+1}^0(0) &= \boldsymbol{v}_n^0(T). \end{aligned} \tag{2}$$

Here η is the N-rescaled learning rate; the corresponding learning rate on $\boldsymbol{w}$ is ηN.

Causality and fixed-data reuse. To simplify the equations, we use the combined epoch-step index $\alpha = (n,t)$, where n denotes the epoch and t the step within that epoch. The causal matrix $\boldsymbol{\Theta}$ and the projection $\mathcal{D}$ distinguish ordinary temporal dependence from repeated-data dependence:

$$\Theta_{\alpha,\beta} = \delta_{nn'}\mathbf{1}\{t > t'\} + \mathbf{1}\{n > n'\}, \quad \alpha = (n,t), \quad \beta = (n',t'), \quad [\mathcal{D}(\boldsymbol{Q})]_{\alpha,\beta} = Q_{(n,t),(n',t')}\delta_{tt'}.$$

The matrix $\boldsymbol{\Theta}$ sums updates that occur before a given state. The map $\mathcal{D}$ sets entries with $t \neq t'$ to zero, thereby retaining cross-epoch correlations only between repeated uses of the same minibatch.

3.2 High Level Findings from Theory

Overfitting Effects from Multi-Pass SGD and Cross-Epoch Response Functions. The theory describes how cross-epoch correlations build up as the same data are reused. These effects are captured by the DMFT response and correlation functions, which predict how validation loss and training loss separate as epoch count grows.

Regimes of Training in Terms of Compute. Our theory predicts that Chinchilla-style compute-optimal scaling breaks down around the compute scale $C \sim P^\chi$, where $\chi = b+1$ for the power-law spectrum studied below.

- **Effectively Online Regime**: At low compute, when data-reuse effects are small, the leading scaling follows the effectively online approximation used for one-pass training.
- **Possible Overfit Regime**: Once the model size selected by the online allocation becomes comparable to the dataset size, overfitting can affect the optimal use of compute. Optimal stopping and approaches such as ensembling then become important. With more compute, training time, model size, and ensemble count can be adjusted to further reduce test loss using the same data.

Optimal Stopping and Ensembling as a Function of Compute. For noiseless gradient flow at the interpolation threshold, the theory predicts growing expected validation loss at late times. This growth can be controlled through a combination of early stopping and ensembling. We characterize

how the optimal stopping time and ensemble count depend on compute at fixed data and member size (Section 3.6).

3.3 DMFT CHARACTERIZATION

Averaging over the random matrices reused in (2), DMFT gives a Gaussian single-site process (Mignacco et al., 2020; Gerbelot et al., 2022; Bordelon et al., 2024; Bordelon & Pehlevan, 2026). The response $R_{\alpha,\beta}$ measures the effect at α of a perturbation at β; $C_{\alpha,\beta}$ is a two-time correlation. Bold symbols denote matrices on the $S = T\mathcal{E}$ epoch-step indices. The response equations are

$$\boldsymbol{H}_k = \left[\boldsymbol{I} + \eta\lambda_k \boldsymbol{\Theta}\boldsymbol{R}_3\boldsymbol{R}_1\right]^{-1}, \qquad \bar{\boldsymbol{H}} = \sum_k \lambda_k \boldsymbol{H}_k,$$

$$\boldsymbol{R}_1 = \left[\boldsymbol{I} + \frac{\eta}{B}\mathcal{D}(\bar{\boldsymbol{H}}\boldsymbol{\Theta}\boldsymbol{R}_3)\right]^{-1}, \qquad \boldsymbol{R}_3 = \left[\boldsymbol{I} + \frac{\eta}{N}\boldsymbol{R}_1\bar{\boldsymbol{H}}\boldsymbol{\Theta}\right]^{-1}. \tag{3}$$

The transfer matrix $\boldsymbol{H}_k$ propagates spectral mode k. The responses $\boldsymbol{R}_1$ and $\boldsymbol{R}_3$ contain the finite-data and finite-width feedback, respectively.

Writing v^1_α for the residual on a current minibatch example, including label noise, the correlations and losses are

$$[\boldsymbol{C}_0]_{\alpha,\beta} = \sum_k \lambda_k \left\langle v^0_{k,\alpha} v^0_{k,\beta} \right\rangle, \qquad [\boldsymbol{C}_1]_{\alpha,\beta} = \left\langle v^1_\alpha v^1_\beta \right\rangle,$$

$$\mathcal{L}_{\text{test}}(\alpha) = [\boldsymbol{C}_0]_{\alpha\alpha} + \sigma^2, \qquad \widehat{\mathcal{L}}_{\text{train}}(\alpha) = [\boldsymbol{C}_1]_{\alpha\alpha}.$$

The brackets average over the effective process. The correlation equations close with (3); the minibatch-feature field has covariance proportional to $\mathcal{D}(\boldsymbol{C}_1)$, retaining correlations between repeated uses of the same minibatch. Appendices B.8 and B.9 derive the process, and Appendices B.3–B.7 give the full equations.

Figure 3 compares the closed DMFT equations with direct simulation of the microscopic random-feature model. The theory tracks changes in width N, batch size B, and ensemble size E, as well as the late-time growth at interpolation.

3.4 LATE-TIME GROWTH NEAR INTERPOLATION

We study the gradient-flow limit under power-law source and capacity conditions (Bordelon et al., 2024):

$$\lambda_k \sim k^{-b}, \qquad \lambda_k (w^\star_k)^2 \sim k^{-a}, \qquad a > 1, \quad b > 1.$$

Here b describes feature-spectrum decay and a describes target power; $b > 1$ makes the eigenvalue tail summable. Write $\mathcal{L}(\tau, N, P, E)$ for ensemble test loss at effective time τ. We first take $\sigma = 0$ and $1 < a < b + 1$, for which the full signal source scales as $P^{-(a-1)}$ (Appendix F).

At $N = P$, the response becomes marginal at low frequencies on the scale $\tau_c \asymp P^b$. Under the conditions of Theorem 2, the growing contribution $\mathcal{L}_{\text{grow}}(\tau)$ to a single model's test loss ($E = 1$) is

$$\mathcal{L}_{\text{grow}}(\tau) \asymp P^{-(a-1+b/2)} \tau^{1/2}, \qquad \tau \gtrsim \tau_c \asymp P^b. \tag{4}$$

On the regular GF branch at fixed $N/P \neq 1$, a positive low-frequency denominator cuts off this critical growth. Before finite data limit learning, the unlearned modes give a contribution of order $\tau^{-(a-1)/b}$ for $\tau \ll P^b$. To determine when to stop an ensemble, we also need the time dependence of the error that remains after averaging over models, which we analyze next.

Label noise adds a source to the correlation equations without changing the responses. Its conditional late-time growth and weak-noise stopping conditions are derived in Appendices F.3 and I.4; the stopping and allocation laws below concern noiseless targets.

3.5 ENSEMBLING AND OPTIMAL STOPPING

Consider E models with independent random projections trained on a shared dataset in a common fixed order. Write $C_{\text{self}}(\tau)$ for one member's expected squared error and $C_{\text{cross}}(\tau)$ for the expected

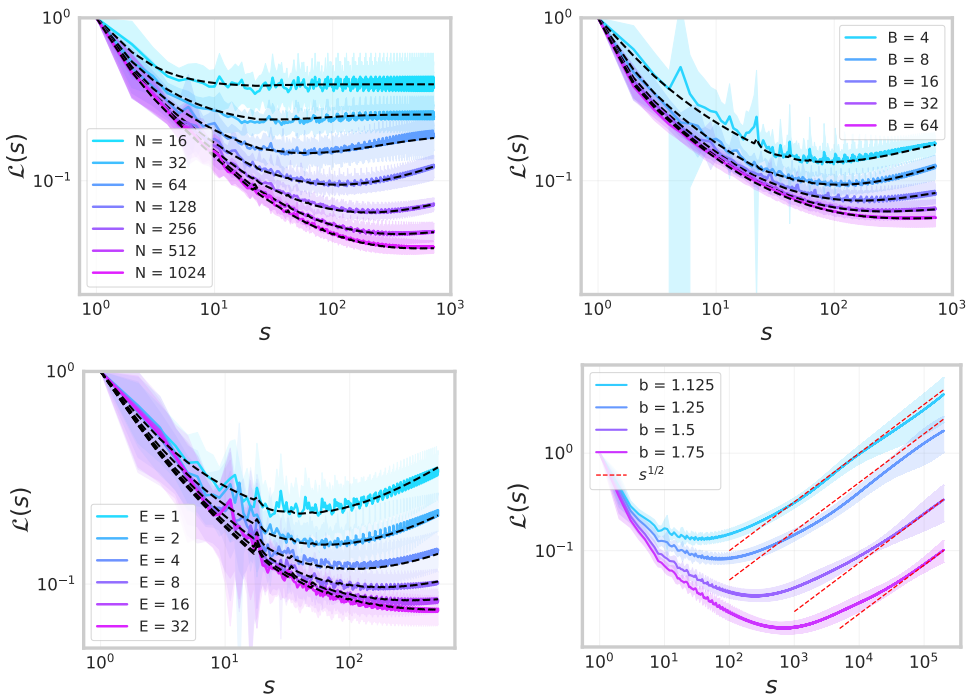

Figure 3: Microscopic random-feature simulations and DMFT predictions (dashed black) versus global step. Panels vary model width N, batch size B, ensemble size E, and capacity exponent b at interpolation.

product of errors from two distinct members on the same input. At $N = P$, the ensemble loss is

$$\mathcal{L}(\tau, P, P, E) = C_{\text{cross}}(\tau) + \frac{C_{\text{self}}(\tau) - C_{\text{cross}}(\tau)}{E}.$$

The difference $C_{\text{self}} - C_{\text{cross}}$ is the prediction variance across random projections, averaged over datasets. Ensembling reduces it by $1/E$ at fixed time, while cross includes both bias and error due to the shared data, as in previous analyses of ensembling (Adlam & Pennington, 2020; Geiger et al., 2020; Bordelon et al., 2024; Ruben & Pehlevan, 2023; Ruben et al., 2025). Appendix C derives the ensemble process.

For Gaussian data and random features, zero initialization, and the pure-power continuum spectrum $\lambda_k = k^{-b}$, $(w_k^\star)^2 = k^{b-a}$, with $b > 1$ and $1 < a < b + 1$, noiseless GF equations at $N = P$ give

$$C_{\text{cross}}(\tau) = C_\infty + J\tau^{-1/2} + D\tau^{-1} + O(\tau^{-3/2}),$$
$$C_{\text{self}}(\tau) = H\sqrt{\tau} + O(1), \qquad D, H > 0.$$

Here $C_\infty \asymp P^{-(a-1)}$, and J has the sign of $a - b$. Since cross stays bounded, only the variance grows at late times. At fixed P and large E, the loss-minimizing time $\tau_\star(P, E)$ balances this growth with the cross correction: $J\tau^{-1/2}$ for $a > b$, or D/τ when $a = b$ and $J = 0$. For $a < b$, cross approaches C_∞ from below and attains a finite-time minimum; stopping times approach its minimizer set. These cases give the stopping laws in Table 2 (Appendix H).

The $a = b$ cancellation follows from equal squared target coefficients, not merely equal tail exponents. For the finite noiseless Gaussian model with zero initialization, $N = P$, a positive spectrum, $M > P + 1$, and $(w_k^\star)^2 = q > 0$, Theorem 3 proves $\tau_\star \asymp E^{2/3}$ for the expected loss at fixed P, M as $E \to \infty$. The additional P^b scaling comes from the continuum calculation.

To connect this variance reduction with our language-model experiments, we consider logit averaging under cross-entropy. For a fixed dataset and evaluation input, a Taylor expansion of cross-entropy gives a leading $1/E$ correction under logit averaging at a fixed checkpoint (Appendix C.7).

3.6 Compute Allocation at Optimal Stopping

We study how to allocate additional compute between ensemble size and training time when dataset and individual model sizes are fixed. Under the noiseless continuum assumptions of Section 3.5, we fix P and $N = P$. For each ensemble size E, let $\tau_\star(P, E)$ be the loss-minimizing training time and $\mathcal{L}^*(P, E)$ the corresponding minimum test loss. At fixed batch size and learning rate, this training requires compute $EP\tau_\star(P, E)$, giving the allocation problem

$$\min_{E \in \mathbb{N}_{\geq 1}} \mathcal{L}^*(P, E) \quad \text{subject to} \quad EP\tau_\star(P, E) \leq C, \qquad P \text{ fixed.} \tag{5}$$

Ensembling reduces prediction variance, so the minimum loss is nonincreasing in E. An optimal choice $E_\star(C; P)$ is therefore the largest ensemble that can reach its loss minimum within budget.

Write $\mathcal{L}^*(C; P)$ for the resulting minimum loss. The excess loss in Table 2 is measured relative to $\mathcal{L}_\infty(P)$, the best test loss approached as ensemble size grows and training time is chosen optimally. This limit is of order $P^{-(a-1)}$ and reflects the error shared across models (Appendix I). For $a < b$, the excess-loss rate assumes positive minimum prediction variance over the times minimizing C_{cross}.

Table 2: Optimal stopping and compute allocation at fixed $N = P$. Entries give scalings up to positive constants as $E \to \infty$ for stopping and $C/P^{b+1} \to \infty$ for allocation and loss.

Condition	Optimal stopping $\tau_\star(P, E)/P^b$	Ensemble size $E_\star(C; P)$	Excess loss $P^{a-1}[\mathcal{L}^*(C; P) - \mathcal{L}_\infty(P)]$
$a > b$	E	$(C/P^{b+1})^{1/2}$	$(C/P^{b+1})^{-1/4}$
$a = b$	$E^{2/3}$	$(C/P^{b+1})^{3/5}$	$(C/P^{b+1})^{-2/5}$
$a < b$	1	C/P^{b+1}	$(C/P^{b+1})^{-1}$

At large budgets, ensemble size grows at least as fast as training time per model. For $a > b$, the two grow with the same power of compute; for equal squared target coefficients ($a = b$), ensemble size grows faster. In both cases, averaging suppresses the growing variance enough to make longer training worthwhile. For $a < b$, the optimal training horizon remains bounded, so further compute increases support more ensemble members.

Allowing any affordable checkpoint gives a broader optimization over ensemble size and training time. At $N = P$ with equal squared target coefficients, this compute frontier retains the $-2/5$ excess-loss exponent, with a different leading coefficient (Appendix I.1).

For finite-step SGD, batch size and learning rate also affect the effective training time available at fixed compute (Appendix D). Figure 4 illustrates these choices in the finite random-feature model.

3.7 Testing the Theoretical Predictions

We first give a finite-horizon guarantee for the loss approximation, then examine the stopping and compute-allocation predictions numerically.

Finite-horizon accuracy. For Gaussian population features, fixed feature dimension M, and a fixed horizon S, Theorem 1 bounds the absolute loss error by $\widetilde{\mathcal{O}}_{M,S}(N^{-1/2} + B^{-1/2})$, simultaneously for all $\alpha \leq S$, at a sufficiently small fixed learning rate. The same rate holds for relative loss error when the reference loss is bounded below on this horizon. Appendix J proves the bound by comparison with population gradient descent.

Compute allocation in the random-feature model. Figure 4 compares ways to use additional compute at fixed data size. In (A), the dashed curve and inset implement the stopped allocation in (5); the solid curve jointly selects ensemble size and an affordable checkpoint. Panel (B) starts below interpolation and compares increasing width with ensembling, while (C, D) show how learning-rate selection changes the batch-size comparison. Appendix K gives the full learning curves, self/cross correlations, stopping-law comparisons, and numerical checks, together with allocation results for widths above interpolation.

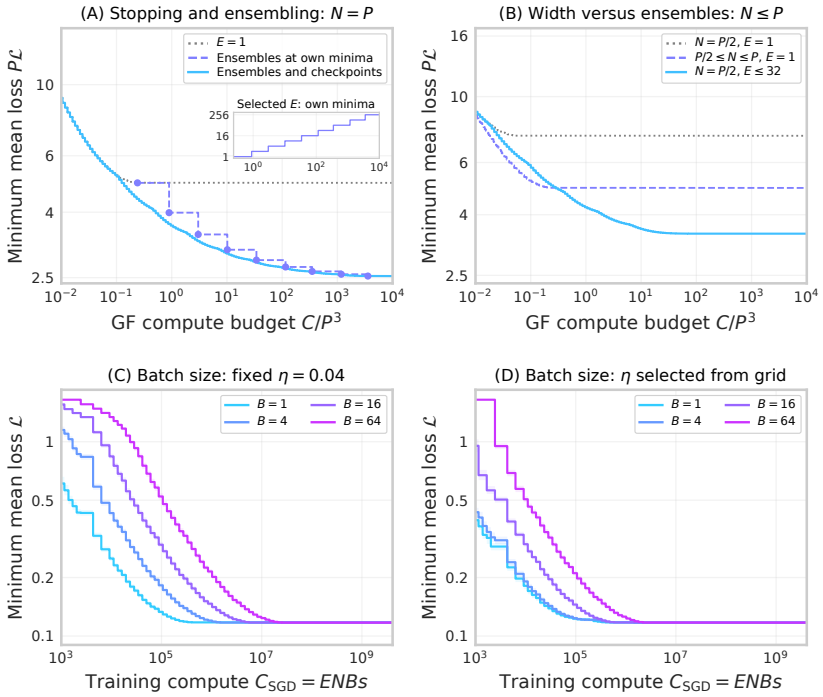

Figure 4: **Compute allocation at fixed data size.** Noiseless random-feature simulations with $a = b = 2$. (A) In GF at $N = P$, ensembling lowers loss beyond the single-model minimum. Inset: selected E when ensembles are trained to their own minima. (B) Starting from $N = P/2$, ensembling outperforms increasing width up to P at large budgets on this grid. (C, D) For cyclic SGD at $N = P/2$ and $E = 1$, selecting learning rates separately narrows the small-batch advantage. Shading indicates 95% dataset-bootstrap intervals. See Appendix K for settings and comparisons.

3.8 Language-Model Experiments

We now return to the FineWeb experiments to examine whether ensembling suppresses late-time deterioration and delays optimal stopping in language models, and to compare its benefits with those of increasing model size. The experiments use the FineWeb setup introduced in Section 2; Appendix A collects the configurations and fitting procedures.

Ensembling suppresses the late-time rise. The loss trajectories in Figure 2 show that averaging independently trained models reduces late-time deterioration. At fixed training steps, validation loss also decreases with ensemble size for both shared and independently shuffled data orders. Appendix A.2 shows that reducible loss decays approximately as E^{-1} for both, as the theory predicts.

Model size affects attainable loss and optimal stopping. We next vary model depth and width while holding the corpus at 100M tokens. Figure 5A shows that larger models generally attain lower validation loss. The separable relation $\mathcal{L}^*(L, W) - \mathcal{L}_\infty = c\, L^{-b_L} W^{-b_W}$ captures additional variation across architectures (Figure 5B,C). At approximately 170M parameters, a deeper, narrower model attains a loss 0.06 nats per token below a shallower, wider model. Model size therefore summarizes the main trend, with architecture accounting for some of the remaining variation.

Larger models also reach their minimum validation loss earlier: the optimal epoch count falls from 21 to 6 over the tested model-size range (Figure 1F). By comparison, it varies from 14 to 9 across the data-size sweep at fixed model size. Over the tested ranges, optimal epoch counts decrease more rapidly with model size than with corpus size. Appendix A.3 collects the fitted exponents and compares attainable loss along the model and data axes.

The ensemble advantage persists across model sizes. We compare ensembles of six model configurations with larger single models, holding the corpus and training horizon fixed. Matching effective size EN (total non-embedding parameters) therefore matches parameter-token training budgets. Four-member ensembles achieve validation losses 0.12–0.20 nats lower than the corre-

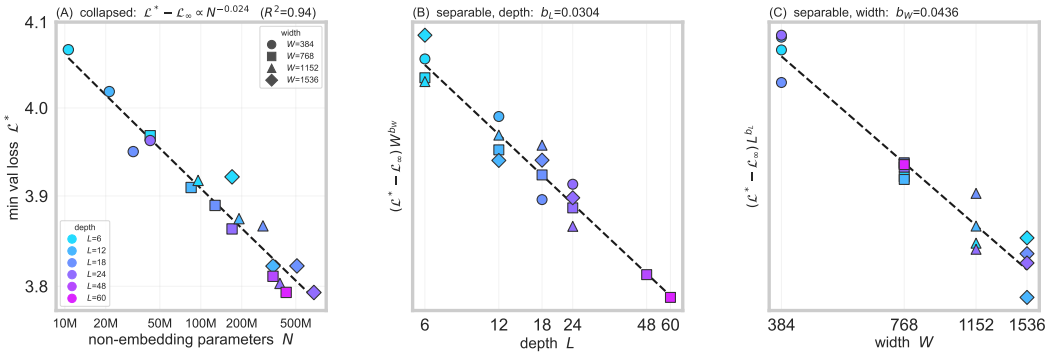

Figure 5: **Larger models attain lower loss, with additional depth and width effects.** At $P = 100$M: (A) minimum validation loss versus non-embedding parameter count; (B, C) depth and width dependence after removing the other variable's fitted contribution (joint $R^2 = 0.97$). Shape denotes width and colour denotes depth. Shared floor: Appendix A.3; fitted coefficients: Table 7.

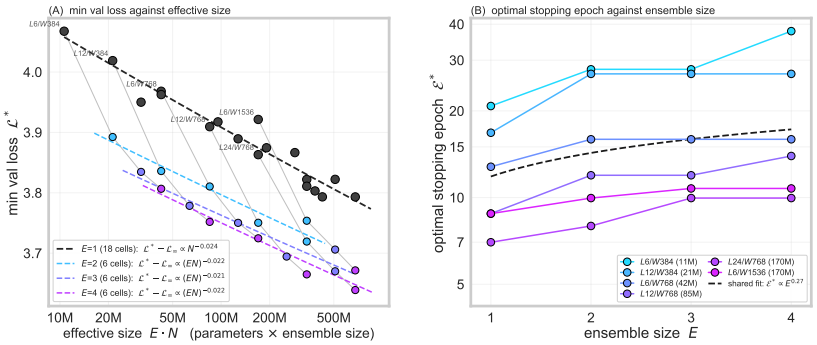

Figure 6: **Ensembles outperform larger single models and reach their minimum later.** At $P = 100$M: (A) minimum validation loss versus effective size EN; black circles show single models, coloured points ensembles of six configurations. (B) Optimal stopping epoch versus ensemble size. Dashed lines show fitted power laws, with a shared slope in (B); details: Appendix A.3.

sponding wider or deeper single models (Figure 6A). Ensembling also shifts the minimum to later training epochs (Figure 6B). This later stopping is qualitatively consistent with the RF mechanism in Section 3.5: averaging suppresses a growing error, allowing further training to remain beneficial.

4 Discussion and Conclusion

Our results show how repeated training epochs and ensembling jointly shape generalization on a fixed dataset. The multi-epoch DMFT tracks error correlations within and across epochs and between ensemble members. For noiseless Gaussian RF gradient flow at $N = P$, a component of the expected test error grows at late training times. Averaging the predictions of E models reduces this component by a factor of E, changing the balance that determines optimal stopping. The pure-power continuum analysis translates the resulting stopping laws into rules for allocating compute between ensemble size and training duration, with data and member sizes fixed and each ensemble trained to its own loss minimum.

Our language-model experiments show lower validation loss and later optimal stopping with ensembling, qualitatively consistent with this mechanism. They also show gains over wider or deeper single models at matched parameter-token training budgets. A quantitative connection to the RF theory requires accounting for feature learning and optimizer dynamics, and relating RF feature and sample counts to Transformer parameter and token counts.

Our label-noise analysis provides a starting point for extending these stopping and allocation laws to noisy targets. Reshuffling requires further analysis of how changing the data order affects correlations across epochs. Together, the theory and experiments motivate choosing the number of training epochs and ensemble members jointly when allocating compute to a fixed dataset.

AI USE STATEMENT

Generative AI tools assisted with drafting and revising sections of the manuscript, organizing the presentation, and polishing the writing. They also supported research ideation and execution, including feedback on experimental design, interpretation of results, and checks of notation, proof strategies, and scaling arguments. The authors reviewed the AI-assisted text, analyses, code, and figures and take responsibility for the final content.

REPRODUCIBILITY STATEMENT

The appendices provide the model assumptions, effective-process equations, and derivations underlying the theoretical results. Appendix A documents the language-model data and training configurations, evaluation and checkpoint-selection procedures, and fitting conventions. Appendix K describes the random-feature simulations, parameter settings, numerical checks, and uncertainty estimates.

REFERENCES

Ben Adlam and Jeffrey Pennington. Understanding double descent requires a fine-grained bias-variance decomposition. *Advances in neural information processing systems*, 33:11022–11032, 2020.

Blake Bordelon and Cengiz Pehlevan. Disordered dynamics in high dimensions: Connections to random matrices and machine learning. *arXiv preprint arXiv:2601.01010*, 2026.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pp. 1024–1034. PMLR, 2020.

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *International Conference on Machine Learning*, pp. 4345–4382. PMLR, 2024.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. *Nature communications*, 12(1):2914, 2021.

A Crisanti and H Sompolinsky. Path integral approach to random neural networks. *Physical Review E*, 98(6):062120, 2018.

Leonardo Defilippis, Bruno Loureiro, and Theodor Misiakiewicz. Dimension-free deterministic equivalents and scaling laws for random feature regression. *Advances in Neural Information Processing Systems*, 37:104630–104693, 2024.

Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stéphane d'Ascoli, Giulio Biroli, Clément Hongler, and Matthieu Wyart. Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401, 2020.

Cedric Gerbelot, Emanuele Troiani, Francesca Mignacco, Florent Krzakala, and Lenka Zdeborova. Rigorous dynamical mean field theory for stochastic gradient descent methods. *arXiv preprint arXiv:2210.06591*, 2022.

Moritz Helias and David Dahmen. *Statistical field theory for neural networks*, volume 970. Springer, 2020.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Konwoo Kim, Suhas Kotha, Percy Liang, and Tatsunori Hashimoto. Pre-training under infinite compute. *arXiv preprint arXiv:2509.14786*, 2025.

Licong Lin, Jingfeng Wu, Sham M Kakade, Peter L Bartlett, and Jason D Lee. Scaling laws in linear regression: Compute, parameters, and data. *arXiv preprint arXiv:2406.08466*, 2024.

Alexander Maloney, Daniel A Roberts, and James Sully. A solvable model of neural scaling laws. *arXiv preprint arXiv:2210.16859*, 2022.

Bishwas Mandal, Shmuel Berman, Akshay Vegesna, and Samip Dahal. q0: Primitives for hyper-epoch pretraining. *arXiv preprint arXiv:2606.03938*, 2026. doi: 10.48550/arXiv.2606.03938. URL https://arxiv.org/abs/2606.03938.

Francesca Mignacco, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification. *Advances in Neural Information Processing Systems*, 33:9540–9550, 2020.

Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.

Courtney Paquette, Kiwon Lee, Fabian Pedregosa, and Elliot Paquette. Sgd in the large: Average-case analysis, asymptotics, and stepsize criticality. In *Conference on Learning Theory*, pp. 3548–3626. PMLR, 2021.

Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+ 3 phases of compute-optimal neural scaling laws. *arXiv preprint arXiv:2405.15074*, 2024.

Ben Ruben and Cengiz Pehlevan. Learning curves for noisy heterogeneous feature-subsampled ridge ensembles. *Advances in Neural Information Processing Systems*, 36:50041–50093, 2023.

Benjamin Samuel Ruben, William Lingxiao Tong, Hamza Tahir Chaudhry, and Cengiz Pehlevan. No free lunch from random feature ensembles: Scaling laws and near-optimality conditions. In *International Conference on Machine Learning*, pp. 52198–52224. PMLR, 2025.

James B Simon, Madeline Dickens, Dhruva Karkada, and Michael R DeWeese. The eigenlearning framework: A conservation law perspective on kernel regression and wide neural networks. *arXiv preprint arXiv:2110.03922*, 2021.

Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Will we run out of data? limits of llm scaling based on human-generated data, 2024.

Tingkai Yan, Haodong Wen, Binghui Li, Kairong Luo, Wenguang Chen, and Kaifeng Lyu. Larger datasets can be repeated more: A theoretical analysis of multi-epoch scaling in linear regression. *arXiv preprint arXiv:2511.13421*, 2025.

A Experimental Protocol and Additional Fits

This appendix collects the experiment designs, training settings, measurement conventions, and fit results supporting Section 2 and the language-model experiments. Figures and numerical results retain their reported training recipes.

A.1 Experiment designs and training settings

Table 3: Language-model experiments. All listed sweeps use constant learning rates; λ is the AdamW weight-decay coefficient.

Experiment	Models and data	Horizon and replication
Data-size sweep Fig. 1, top	$L = 12$, $W = 768$ (85M); $P \in \{10, 20, \ldots, 100\}$M; $\lambda = 0$	About 8,000 updates (1.0–1.1B tokens), or 11–101 epochs; one run per strategy at each P
Model-size grid Figs. 1, bottom, 5, 8	$L \in \{6, 12, 18, 24\} \times W \in \{384, 768, 1152, 1536\}$, plus $L \in \{48, 60\}$ at $W = 768$; $P = 100$M; $\lambda = 0$	40 epochs; one run per cell; 18 cells spanning 11M–680M parameters
Matched-budget comparison Figs. 2, 9	Base $L = 12$, $W = 768$; ensembles versus wider/deeper single models; $P = 100$M; $\lambda = 0$	40-epoch training budget; ensembles in Fig. 2 evaluated from saved model states through epoch 31
Ensemble-size scaling Fig. 7	$L = 12$, $W = 768$; $P = 20$M; $\lambda = 0$; shared or independent data orders	40 epochs; 20 trained members per strategy; 10 random member orderings per strategy
Ensembles across model sizes Fig. 6	Six replicated cells (11M–170M); $P = 100$M; $\lambda = 0$; $E = 2, 3, 4$	40 epochs; at least four independent members per cell; up to 680M total counted parameters
Additional joint-fit data	Seven-cell model-size ladder at $P = 20$M; $\lambda = 0$	50 epochs; one run per configuration

Table 4: Language-model training configuration.

Data	FineWeb; GPT-2 BPE tokenizer (vocabulary 50,257); 10M held-out validation tokens
Architecture	Decoder-only Transformer, depth L, width W, head dimension 64 ($W/64$ heads); rotary embeddings, QK-norm, RMSNorm without weights, no biases, no dropout; SwiGLU MLP with hidden width $8W/3$; untied input and output embeddings; learned residual, input-injection, and U-Net skip gains; sliding-window attention on three of every four layers
Parameterization	CompleteP with base L=12, W=768; truncated-normal initialization, std 0.02 at base width
Sequence length	2048 tokens
Batch size	131,072 tokens (64 sequences) per update
Optimizer	AdamW, $\beta = (0.8, 0.95)$, $\epsilon = 10^{-10}$
Learning rates	matrices 0.01, embeddings 0.0375, output head 0.0005, scalar gains 0.025
Schedule	constant, no warmup (λ=0 runs); last 20% linear decay to zero (runs with $\lambda > 0$ and cooldown)
Precision, hardware	bfloat16 autocast; one NVIDIA H100-80GB per run
Seeds	one run per configuration; ensemble members differ in initialization and, where stated, data order

Table 5: Measurement and selection conventions.

Quantity	Definition or procedure
Validation loss	Cross-entropy in nats per token on a held-out FineWeb split.
Epoch and step	Each update processes 131,072 tokens; one epoch is $P/131{,}072$ updates (763 at $P = 100$M; 152 at $P = 20$M).
Evaluation cadence	Every 76 updates in the data sweep and 152 updates in the model grids, corresponding to about 10M and 19.9M training tokens.
Saved model state	A checkpoint stores the model parameters at a training step. The optimal checkpoint is the evaluated state with minimum validation loss.
Optimal stopping	s^* is the evaluated step with minimum validation loss. $\mathcal{E}^* = s^*/(P/131{,}072)$ is the corresponding epoch count.
Data-sweep aggregation	One minimum per strategy at each corpus size; the reported minimum loss and optimal epoch average the two strategies' minima.
Ensemble prediction	Average member logits before computing validation cross-entropy. Initialization-only members share the data order; initialization-plus-shuffle members vary both.

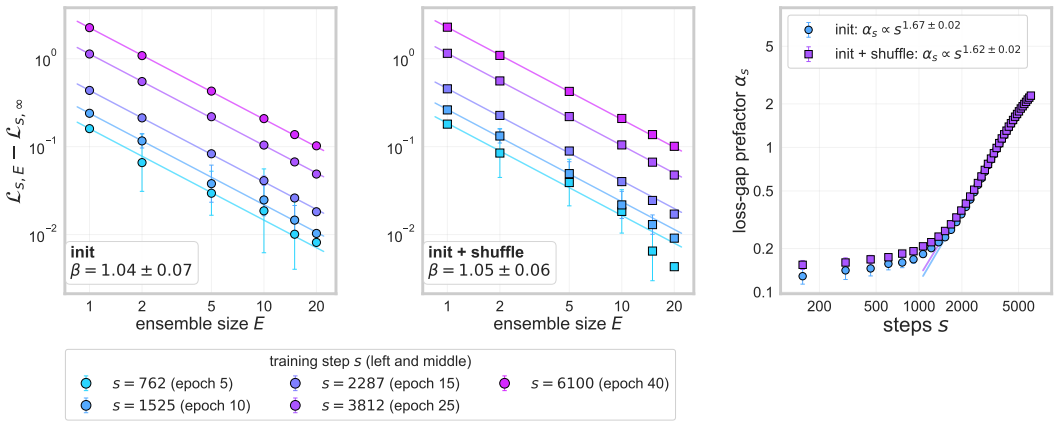

Figure 7: **The reducible ensemble loss decays approximately as E^{-1} under both sources of member diversity.** Left and middle: loss above the fitted infinite-ensemble asymptote against ensemble size, for members with shared or independently shuffled data orders; colour marks the training step and lines are fits of (6). Right: fitted amplitude α_s against step. $P = 20$M, 40 epochs.

A.2 Evaluation, ensembling, and compute accounting

Ensemble fits and resampling. At optimizer step s, each data-order strategy is fitted separately with

$$\mathcal{L}_{s,E} = \alpha_s E^{-\beta} + \mathcal{L}_{s,\infty}, \tag{6}$$

with one exponent β per strategy and an amplitude α_s and asymptote $\mathcal{L}_{s,\infty}$ at each step. The asymptote is strategy-specific because initialization-only members share one data order, which their infinite ensemble retains. Each of 10 resampling iterations draws a random ordering of the 20 trained members of each strategy; the ensemble of size E consists of the first E members, and (6) is refitted per iteration. Error bars show the spread over iterations.

The fitted ensemble-size exponents are close to one (Figure 7, left and middle), consistent with the E^{-1} dependence predicted by the fixed-checkpoint cross-entropy expansion (Appendix C.7). The amplitude α_s is nearly constant over the first 1,000 steps and grows about tenfold by epoch 40 (right).

Training-budget proxy. We count the attention and MLP matrices as $N = 12LW^2$: each layer has four $W \times W$ attention matrices and a gated MLP with hidden width $8W/3$ (three matrices totalling $8W^2$). Token embeddings, the output head, and per-layer scalar gains are excluded. With D training tokens per member, the compute proxy is $C = END$. For the matched-budget experiments, $D = 40P$: we charge the full 40-epoch run and report the best evaluated checkpoint within the available evaluation horizon. Thus four base models and one twice-wider model have the same proxy cost ($4 \times 85\text{M} = 340\text{M}$ counted parameters at the same D). The fixed-horizon LM comparison differs from the RF allocation, which charges training up to each ensemble's own optimal stopping time. The weight-decay coefficient λ is unrelated to the feature-spectrum eigenvalues λ_k in Section 3.

A.3 Loss laws and fitted exponents

Shared loss floor. Loss laws in N, L, W, and P use the shared floor $\mathcal{L}_\infty = 1.25$. Table 6 gives its joint-fit source and the observations used in each subsequent fit.

Table 6: Fitting conventions for attainable loss and optimal stopping.

Fit	Model and observations
Joint N, P law	$\mathcal{L}^* = \mathcal{L}_\infty + AN^{-\alpha} + BP^{-\alpha}$ on 35 observations: the 18-cell 100M grid, seven-cell 20M ladder, and ten data-sweep settings. Shared exponent $\alpha = 0.19$, floor 1.25, $R^2 = 0.996$.
Model-size laws	One minimum per cell on the 18-cell 100M grid. Fit $\mathcal{L}^* - \mathcal{L}_\infty$ against N, or against L, W with separate exponents. Optimal-epoch fits use the same cells.
Data-size laws	Ten corpus sizes at fixed $L = 12, W = 768$, using the strategy-averaged minima and optimal epochs defined in Table 5.
Ensemble/model-size laws	A separate power law in total counted parameters EN for each $E = 2, 3, 4$, using the six replicated cells and the same loss floor.
Ensemble stopping	Fit $\log \mathcal{E}_j^* = \log c_j + \gamma \log E$ over $E = 1, 2, 3, 4$ for six model configurations j, with a shared exponent $\gamma = 0.27$ and separate prefactors c_j. The dashed guide in Figure 6B uses the geometric mean of these prefactors.
Ensemble-size law	Equation (6) fits an infinite-ensemble asymptote for each strategy at each checkpoint; this is distinct from the shared N, P loss floor.

Table 7: Fitted exponents for optimal epochs $\mathcal{E}^*$ and minimum loss above the shared floor, $\Delta\mathcal{L}^* = \mathcal{L}^* - 1.25$. Table 6 specifies the observations and fitting conventions.

Quantity	Fitted dependence	Exponent(s)	R^2	Points
$\mathcal{E}^*$	$N^{-\alpha}$	0.314	0.938	18
$\mathcal{E}^*$	$L^{-a_L} W^{-a_W}$	$a_L = 0.336,\ a_W = 0.608$	0.940	18
$\mathcal{E}^*$	$P^{-\alpha}$	0.158	0.917	10
$\Delta\mathcal{L}^*$	$N^{-\beta_N}$	0.024	0.943	18
$\Delta\mathcal{L}^*$	$L^{-b_L} W^{-b_W}$	$b_L = 0.0304,\ b_W = 0.0436$	0.969	18
$\Delta\mathcal{L}^*$	$P^{-\beta_P}$	0.169	0.997	10

Figure 8 collects the model-size stopping curve and the attainable-loss fits along the model and data axes.

Table 8: Additional fit diagnostics and quantitative comparisons.

Comparison	Result on the measured configurations
Depth/width loss fit	$b_L = 0.0304$, $b_W = 0.0436$, each stable to ± 0.002 under leave-one-out. The fitted sensitivity ratio is $b_W/(2b_L) = 0.72$.
Matched-size example	At $N \approx 170$M, $L = 6, W = 1536$ attains loss 3.92, versus 3.86 for $L = 24, W = 768$.
Depth/width stopping fit	$a_L = 0.336$, $a_W = 0.608$, so $a_W/(2a_L) = 0.90$. Its rounded $R^2 = 0.94$ matches the fit using N alone. Dependence solely on $N \propto LW^2$ would give $a_W = 2a_L$.
Resource scaling	Optimal epochs fall by a fitted factor of 1.4 per decade of data scaling, versus 2.1 per decade of model scaling.
Ensemble/model-size fits	Exponents 0.022, 0.021, 0.022 for $E = 2, 3, 4$, versus 0.024 for single models. The $E = 4$ fitted law lies about 0.15 nats below the single-model law over the comparison range.

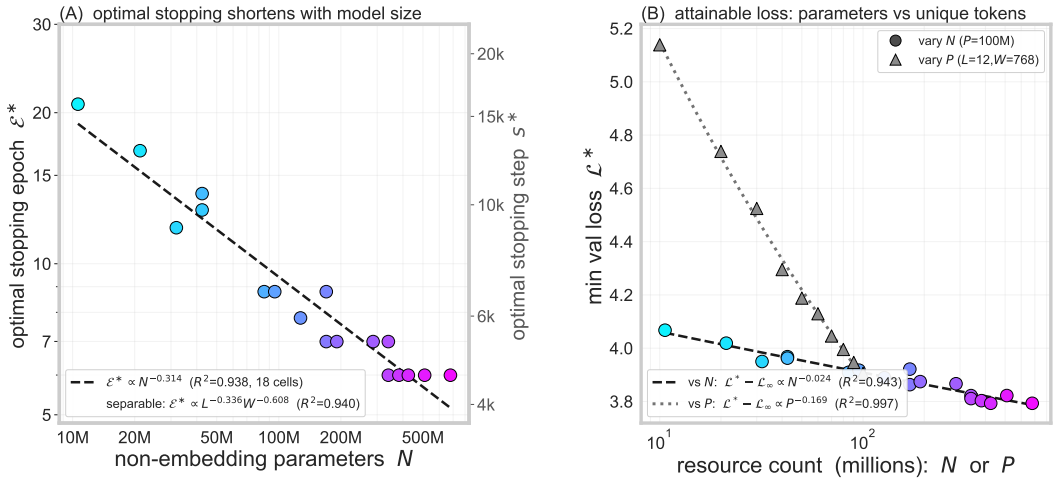

Figure 8: **Model size changes both the best stopping point and the attainable loss.** (A) Optimal stopping versus model size on the 100M-token corpus, shown in epochs and optimizer steps. (B) Minimum validation loss when varying model size at fixed data (circles) or data size at fixed model (triangles). Dashed curves show power-law fits over the respective measured ranges. The loss fits in (B) use the shared floor from Appendix A.3. Colour denotes model size.

A.4 Additional matched-budget comparisons

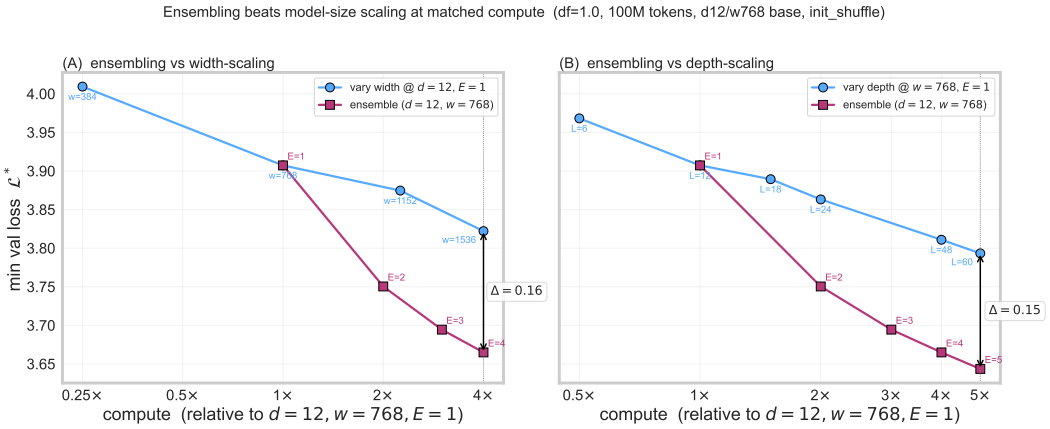

Figure 9: Ensembles compared with wider (left) or deeper (right) single models at matched parameter-token training budgets. Both comparisons use $P = 100$M, constant learning rates, zero weight decay, and a 40-epoch budget.

B Derivation of the Multi-Epoch DMFT Equations

We derive the dynamical mean-field theory (DMFT) for fixed-order multi-epoch training with independent Gaussian population features $\boldsymbol{\psi}(\boldsymbol{x}_\mu) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Lambda})$. The population features, Gaussian random-feature matrix, and label noises are mutually independent. Feature-space quantities are written in the eigenbasis of the population covariance,

$$\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots), \qquad \big\langle \psi_k(\boldsymbol{x}) \psi_\ell(\boldsymbol{x}) \big\rangle = \lambda_k \delta_{k\ell}.$$

The target has coefficients $w_k^\star$, so that

$$y(\boldsymbol{x}) = \boldsymbol{\psi}(\boldsymbol{x}) \cdot \boldsymbol{w}^\star + \sigma\epsilon.$$

The DMFT equations allow an arbitrary spectrum and target. We impose power laws only in Appendix F.

We first state the finite-horizon single-site equations and derive them by cavity and MSR methods. Later sections treat finite-size error, ensembles, continuum limits, and the low-frequency scaling near interpolation.

B.1 Asymptotic settings

We first derive the closure at finite feature dimension M and trajectory length $S = T\mathcal{E}$. The population features enter through $\boldsymbol{\Lambda}$ and the target coefficients. The closure retains finite-width and finite-minibatch contributions through its N^{-1} and B^{-1} terms. The loss approximation in Theorem 1 keeps M, S fixed as N, B grow. The spectral, small-learning-rate, and long-time limits used for scaling are specified separately below. We distinguish two settings for the population spectrum.

Dimension-free power-law setting. For the scaling analysis, we let $M \to \infty$ along a trace-class population spectrum and assume

$$\sum_{k=1}^{\infty} \lambda_k < \infty, \qquad \sum_{k=1}^{\infty} \lambda_k (w_k^\star)^2 < \infty.$$

This is the dimension-free setting in which the power laws $\lambda_k \sim k^{-b}$ and $\lambda_k (w_k^\star)^2 \sim k^{-a}$, with $a, b > 1$, are imposed in Appendix F.

Finite-dimensional isotropic specialization. For comparison, we also consider a proportional sequence with

$$M, N, P \to \infty, \qquad \frac{N}{M} \to \phi \in (0, \infty), \qquad \frac{P}{M} \to \gamma \in (0, \infty), \qquad \boldsymbol{\Lambda}_M = \boldsymbol{I}_M.$$

We choose a sequence of targets satisfying

$$\|\boldsymbol{w}_M^\star\|_2^2 \to \rho \in (0, \infty).$$

Equivalently, an isotropically oriented target may be generated with coordinates of order $M^{-1/2}$, for example $\boldsymbol{w}_M^\star = \sqrt{\rho/M}\,\boldsymbol{g}$ with $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_M)$. Since

$$\mathbb{E}_{\boldsymbol{x}}\left[(\boldsymbol{w}_M^\star \cdot \boldsymbol{\psi}(\boldsymbol{x}))^2\right] = \boldsymbol{w}_M^{\star\top} \boldsymbol{\Lambda}_M \boldsymbol{w}_M^\star = \|\boldsymbol{w}_M^\star\|_2^2,$$

this scaling keeps the target signal and loss of order one. This proportional isotropic sequence is not an instance of the infinite-dimensional trace-class limit, since $\mathrm{Tr}(\boldsymbol{\Lambda}_M) = M$.

We use a combined epoch-step index

$$\alpha = (n, t), \qquad n = 0, \dots, \mathcal{E} - 1, \qquad t = 0, \dots, T - 1,$$

and let $S = T\mathcal{E}$ be the total number of update times. Greek indices $\alpha, \beta, \gamma, \dots$ always denote epoch-step indices. We write sums over epoch-step indices explicitly unless matrix notation is used. The global update index is $s = nT + t$, and effective training time is $\tau = \eta s$, as in the main text. Whenever an inequality such as $\alpha \le S$ is used, the pairs are ordered causally through the one-based index $s + 1 \in \{1, \dots, S\}$.

Define the causal matrix $\boldsymbol{\Theta}$ by

$$\Theta_{\alpha\beta} = \Theta_{(n,t),(n',t')} = \delta_{nn'} \mathbf{1}\{t > t'\} + \mathbf{1}\{n > n'\}.$$

Thus $\Theta_{\alpha\beta} = 1$ when update β precedes state α. It sums the past gradient updates in the residual recursion.

For any $S \times S$ matrix $\boldsymbol{A}$, define $\mathcal{D}$ by

$$[\mathcal{D}(\boldsymbol{A})]_{(n,t),(n',t')} = A_{(n,t),(n',t)} \delta_{tt'}.$$

The map $\mathcal{D}$ retains pairs with the same minibatch label t and sets all other entries to zero. It appears in covariance and response terms generated by reusing $\boldsymbol{\Psi}(t)$ in each epoch.

B.2 Decomposing the dynamics

We begin from the feature-space residual

$$\boldsymbol{v}_n^0(t) = \boldsymbol{w}^\star - \frac{1}{N} \boldsymbol{A}^\top \boldsymbol{w}_n(t) \in \mathbb{R}^M,$$

where $\boldsymbol{A} \in \mathbb{R}^{N \times M}$ is the random feature matrix, $\boldsymbol{w}_n(t) \in \mathbb{R}^N$ is the trainable weight vector, and $\boldsymbol{w}^\star \in \mathbb{R}^M$ is the target coefficient vector. The minibatch feature matrix at within-epoch step t is

$$\boldsymbol{\Psi}(t) \in \mathbb{R}^{B \times M}.$$

In the fixed-order multi-epoch setting, the same $\boldsymbol{\Psi}(t)$ is reused at within-epoch step t of every epoch.

The feature-space residual update, including label noise, is

$$\boldsymbol{v}_n^0(t+1) = \boldsymbol{v}_n^0(t) - \eta \left(\frac{1}{N} \boldsymbol{A}^\top \boldsymbol{A}\right) \frac{1}{B} \boldsymbol{\Psi}(t)^\top \left(\boldsymbol{\Psi}(t) \boldsymbol{v}_n^0(t) + \sigma \boldsymbol{\epsilon}(t)\right), \tag{7}$$

with epoch boundary condition

$$\boldsymbol{v}_{n+1}^0(0) = \boldsymbol{v}_n^0(T).$$

Here $\boldsymbol{\epsilon}(t) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_B)$ is drawn once with minibatch t, reused at the same within-epoch step of every epoch, and is independent across different minibatch labels. Under the normalization $f(\boldsymbol{x}) = N^{-1} \boldsymbol{w}^\top \boldsymbol{A} \psi(\boldsymbol{x})$, the learning rate η in (7) denotes the N-rescaled learning rate; equivalently, the parameter-space learning rate for $\boldsymbol{w}$ is ηN. The noiseless case is recovered by setting $\sigma = 0$.

Following the standard random-feature construction Bordelon et al. (2024); Bordelon & Pehlevan (2026); Paquette et al. (2024), introduce four auxiliary fields. Each equation below contains only

one of the random matrices $\boldsymbol{\Psi}(t)$ and $\boldsymbol{A}$:

$$\boldsymbol{v}_n^1(t) = \boldsymbol{\Psi}(t)\boldsymbol{v}_n^0(t) + \sigma\boldsymbol{\epsilon}(t) \in \mathbb{R}^B, \tag{8}$$

$$\boldsymbol{v}_n^2(t) = \frac{1}{B}\boldsymbol{\Psi}(t)^\top \boldsymbol{v}_n^1(t) \in \mathbb{R}^M,$$

$$\boldsymbol{v}_n^3(t) = \boldsymbol{A}\boldsymbol{v}_n^2(t) \in \mathbb{R}^N,$$

$$\boldsymbol{v}_n^4(t) = \frac{1}{N}\boldsymbol{A}^\top \boldsymbol{v}_n^3(t) \in \mathbb{R}^M. \tag{9}$$

The entries of $\boldsymbol{v}_n^1(t)$ are the residuals on minibatch t, including label noise. Mapping them back to feature space gives

$$\boldsymbol{v}_n^2(t) = \frac{1}{B}\sum_{\mu\in\mathrm{batch}(t)} \psi(\boldsymbol{x}_\mu) v_{\mu,n}^1(t).$$

so $\boldsymbol{v}_n^2(t)$ is the empirical gradient with respect to the feature coefficient. The trained parameter is $\boldsymbol{w}_n(t) \in \mathbb{R}^N$, and $\boldsymbol{v}_n^3(t) = \boldsymbol{A}\boldsymbol{v}_n^2(t)$ is its update direction. The induced change in the feature coefficient is

$$\boldsymbol{v}_n^4(t) = \frac{1}{N}\boldsymbol{A}^\top \boldsymbol{v}_n^3(t) = \left(\frac{1}{N}\boldsymbol{A}^\top \boldsymbol{A}\right) \boldsymbol{v}_n^2(t)$$

Thus $\boldsymbol{v}_n^4(t)$ is $\boldsymbol{v}_n^2(t)$ filtered by the random-feature Gram matrix $\boldsymbol{A}^\top\boldsymbol{A}/N$, and it is the direction that enters the residual update.

Equations (8)–(9) reduce (7) to

$$\boldsymbol{v}_n^0(t+1) = \boldsymbol{v}_n^0(t) - \eta\boldsymbol{v}_n^4(t), \qquad \boldsymbol{v}_{n+1}^0(0) = \boldsymbol{v}_n^0(T).$$

The label-noise contribution is contained in $\boldsymbol{v}_n^4(t)$ through $\boldsymbol{v}_n^1(t)$.

Define the initial residual

$$\boldsymbol{v}_{\mathrm{init}}^0 = \boldsymbol{w}^\star - \frac{1}{N}\boldsymbol{A}^\top \boldsymbol{w}_0(0).$$

The full trajectory is

$$\boldsymbol{v}_\alpha^0 = \boldsymbol{v}_{\mathrm{init}}^0 - \eta\sum_\beta \Theta_{\alpha\beta}\boldsymbol{v}_\beta^4.$$

For zero initialization, $\boldsymbol{w}_0(0) = 0$, this reduces to $\boldsymbol{v}_{\mathrm{init}}^0 = \boldsymbol{w}^\star$.

Here $\boldsymbol{v}_\alpha^i = \boldsymbol{v}_n^i(t)$ for $\alpha = (n,t)$. Stacked notation collects the S vectors into one vector; for example,

$$\boldsymbol{v}^0 = \begin{pmatrix} \boldsymbol{v}_1^0 \\ \vdots \\ \boldsymbol{v}_S^0 \end{pmatrix} \in \mathbb{R}^{SM}.$$

In this notation,

$$\boldsymbol{v}^0 = \mathbf{1} \otimes \boldsymbol{v}_{\mathrm{init}}^0 - \eta\boldsymbol{\Theta}\boldsymbol{v}^4, \tag{10}$$

where $\otimes$ denotes the Kronecker product and $\mathbf{1} \otimes \boldsymbol{v}_{\mathrm{init}}^0$ repeats the initial residual across all S epoch-step indices. The matrix $\boldsymbol{\Theta}$ acts only on the epoch-step indices; equivalently, the last term is $(\boldsymbol{\Theta} \otimes \boldsymbol{I}_M)\boldsymbol{v}^4$.

B.3 Single-site stochastic processes

The DMFT fixed point is as follows. Appendices B.8 and B.9 give its cavity and MSR derivations. In the population eigenbasis, the effective variables are a representative minibatch coordinate v^1, a representative random-feature coordinate v^3, and a spectral mode (v_k^0, v_k^2, v_k^4). Each variable retains the epoch-step index $\alpha = (n,t)$.

The effective process is

$$v^0_{k,\alpha} = v^0_{\text{init},k} - \eta \sum_\beta \Theta_{\alpha\beta} v^4_{k,\beta}, \tag{11}$$

$$v^1_\alpha = u^1_\alpha + \sigma\epsilon_\alpha + \frac{1}{B}\sum_\beta [\mathcal{D}(\boldsymbol{R}_{0,2})]_{\alpha\beta} v^1_\beta, \tag{12}$$

$$v^2_{k,\alpha} = u^2_{k,\alpha} + \lambda_k \sum_\beta R_{1;\alpha\beta} v^0_{k,\beta}, \tag{13}$$

$$v^3_\alpha = u^3_\alpha + \frac{1}{N}\sum_\beta [\boldsymbol{R}_{2,4}]_{\alpha\beta} v^3_\beta, \tag{14}$$

$$v^4_{k,\alpha} = u^4_{k,\alpha} + \sum_\beta R_{3;\alpha\beta} v^2_{k,\beta}. \tag{15}$$

Here $v^0_{\text{init},k}$ is the k-th coordinate of the initial feature-space residual. For zero initialization, $v^0_{\text{init},k} = w^\star_k$. For $\alpha = (n,t)$, the scalar $\epsilon_\alpha \equiv \epsilon_t$ is the effective single-site label-noise variable inherited from $\boldsymbol{\epsilon}(t)$. It is therefore the same variable at the same within-epoch label t in every epoch.

The zero-mean Gaussian fields have covariances

$$\begin{aligned}
\left\langle u^1_\alpha u^1_\beta \right\rangle &= [\mathcal{D}(\boldsymbol{C}_0)]_{\alpha\beta}, \\
\left\langle u^2_{k,\alpha} u^2_{\ell,\beta} \right\rangle &= \delta_{k\ell}\frac{\lambda_k}{B}[\mathcal{D}(\boldsymbol{C}_1)]_{\alpha\beta}, \\
\left\langle u^3_\alpha u^3_\beta \right\rangle &= [\boldsymbol{C}_2]_{\alpha\beta}, \\
\left\langle u^4_{k,\alpha} u^4_{\ell,\beta} \right\rangle &= \delta_{k\ell}\frac{1}{N}[\boldsymbol{C}_3]_{\alpha\beta}.
\end{aligned}$$

All omitted cross-covariances among the Gaussian fields and the label-noise field vanish. The label-noise covariance is

$$\left\langle \epsilon_\alpha \epsilon_\beta \right\rangle = [\mathcal{D}(\mathbf{1}\mathbf{1}^\top)]_{\alpha\beta}.$$

Both u^1 and the label noise are correlated across epochs only when the minibatch labels agree.

The correlations $\boldsymbol{C}_i$ and responses $\boldsymbol{R}_{i,j}$ are determined self-consistently below.

B.4 Correlation functions

The Gaussian covariances depend on the following correlation functions:

$$\begin{aligned}
[\boldsymbol{C}_0]_{\alpha\beta} &= \sum_k \lambda_k \left\langle v^0_{k,\alpha} v^0_{k,\beta} \right\rangle, \\
[\boldsymbol{C}_1]_{\alpha\beta} &= \left\langle v^1_\alpha v^1_\beta \right\rangle, \\
[\boldsymbol{C}_2]_{\alpha\beta} &= \sum_k \left\langle v^2_{k,\alpha} v^2_{k,\beta} \right\rangle, \\
[\boldsymbol{C}_3]_{\alpha\beta} &= \left\langle v^3_\alpha v^3_\beta \right\rangle.
\end{aligned}$$

These are uncentered second moments. In particular, v^0_k generally has a nonzero mean, so $\boldsymbol{C}_0$ includes both bias and variance. The test and train losses are read from the diagonal entries

$$\mathcal{L}_{\text{test}}(\alpha) = [\boldsymbol{C}_0]_{\alpha\alpha} + \sigma^2, \qquad \widehat{\mathcal{L}}_{\text{train}}(\alpha) = [\boldsymbol{C}_1]_{\alpha\alpha}.$$

The additive σ^2 is the irreducible label-noise contribution to the test mean-squared error.

The v^1 and v^3 equations give

$$\begin{aligned}
\boldsymbol{C}_1 &= \boldsymbol{R}_1 \mathcal{D}\left(\boldsymbol{C}_0 + \sigma^2 \mathbf{1}\mathbf{1}^\top\right)\boldsymbol{R}_1^\top, \\
\boldsymbol{C}_3 &= \boldsymbol{R}_3 \boldsymbol{C}_2 \boldsymbol{R}_3^\top.
\end{aligned}$$

Here $\boldsymbol{C}_0$ is the full two-time residual correlation. The projection $\mathcal{D}$ enters when this residual is sampled through the fixed minibatches.

B.5 Response functions

The response functions use the standard random-feature DMFT convention. $R_{0,2}$ and $R_{2,4}$ measure the responses of v^0 to u^2 and of v^2 to u^4; R_1 and R_3 are the site responses to u^1 and u^3. Label noise changes correlations but not these linear responses.

For each mode k, define

$$R_{0,2;k;\alpha\beta} = \left\langle \frac{\delta v^0_{k,\alpha}}{\delta u^2_{k,\beta}} \right\rangle,$$

$$R_{2,4;k;\alpha\beta} = \left\langle \frac{\delta v^2_{k,\alpha}}{\delta u^4_{k,\beta}} \right\rangle.$$

The spectrum-averaged responses are

$$[\boldsymbol{R}_{0,2}]_{\alpha\beta} = \sum_k \lambda_k R_{0,2;k;\alpha\beta},$$

$$[\boldsymbol{R}_{2,4}]_{\alpha\beta} = \sum_k R_{2,4;k;\alpha\beta}.$$

The site responses are

$$R_{1;\alpha\beta} = \left\langle \frac{\delta v^1_{\alpha}}{\delta u^1_{\beta}} \right\rangle,$$

$$R_{3;\alpha\beta} = \left\langle \frac{\delta v^3_{\alpha}}{\delta u^3_{\beta}} \right\rangle.$$

From (12) and (14), these responses obey

$$\boldsymbol{R}_1 = \left[\boldsymbol{I} - \frac{1}{B} \mathcal{D}(\boldsymbol{R}_{0,2}) \right]^{-1}, \tag{16}$$

$$\boldsymbol{R}_3 = \left[\boldsymbol{I} - \frac{1}{N} \boldsymbol{R}_{2,4} \right]^{-1}. \tag{17}$$

These equations contain the feedback through the remaining sites.

B.6 Mode transfer functions

Combining (11), (13), and (15) gives the residual equation for each mode:

$$v^0_{k,\alpha} = v^0_{\mathrm{init},k} - \eta \sum_\beta \Theta_{\alpha\beta} \left[u^4_{k,\beta} + \sum_\gamma R_{3;\beta\gamma} u^2_{k,\gamma} + \lambda_k \sum_{\gamma,\rho} R_{3;\beta\gamma} R_{1;\gamma\rho} v^0_{k,\rho} \right].$$

Solving for $\boldsymbol{v}^0_k$ gives

$$\boldsymbol{v}^0_k = \boldsymbol{H}_k \left[v^0_{\mathrm{init},k} \mathbf{1} - \eta \boldsymbol{\Theta} \left(\boldsymbol{u}^4_k + \boldsymbol{R}_3 \boldsymbol{u}^2_k \right) \right],$$

where

$$\boldsymbol{H}_k = \left[\boldsymbol{I} + \eta \lambda_k \boldsymbol{\Theta} \boldsymbol{R}_3 \boldsymbol{R}_1 \right]^{-1}.$$

We also define the spectrum-averaged transfer operator

$$\bar{\boldsymbol{H}} = \sum_k \lambda_k \boldsymbol{H}_k.$$

The auxiliary responses can now be written in terms of the transfer functions:

$$\boldsymbol{R}_{0,2;k} = -\eta \boldsymbol{H}_k \boldsymbol{\Theta} \boldsymbol{R}_3, \qquad \boldsymbol{R}_{0,2} = -\eta \bar{\boldsymbol{H}} \boldsymbol{\Theta} \boldsymbol{R}_3,$$
$$\boldsymbol{R}_{2,4;k} = -\eta \lambda_k \boldsymbol{R}_1 \boldsymbol{H}_k \boldsymbol{\Theta}, \qquad \boldsymbol{R}_{2,4} = -\eta \boldsymbol{R}_1 \bar{\boldsymbol{H}} \boldsymbol{\Theta}.$$

Substituting these into (16)–(17) gives the closed response equations

$$\boldsymbol{R}_1 = \left[\boldsymbol{I} + \frac{\eta}{B} \mathcal{D}(\bar{\boldsymbol{H}} \boldsymbol{\Theta} \boldsymbol{R}_3) \right]^{-1},$$
$$\boldsymbol{R}_3 = \left[\boldsymbol{I} + \frac{\eta}{N} \boldsymbol{R}_1 \bar{\boldsymbol{H}} \boldsymbol{\Theta} \right]^{-1}.$$

B.7 Closed Matrix Equations for the Multi-Epoch DMFT

The equations used for numerical solution are collected here. Every matrix is indexed by the $S = T\mathcal{E}$ epoch-step pairs. Define

$$\boldsymbol{M}_k = \boldsymbol{I} - \eta \lambda_k \boldsymbol{R}_1 \boldsymbol{H}_k \boldsymbol{\Theta} \boldsymbol{R}_3. \tag{18}$$

This matrix propagates the Gaussian field $\boldsymbol{u}_k^2$, generated by the minibatch features, into $\boldsymbol{v}_k^2$.

The closed response equations are

$$\boldsymbol{H}_k = \left[\boldsymbol{I} + \eta \lambda_k \boldsymbol{\Theta} \boldsymbol{R}_3 \boldsymbol{R}_1 \right]^{-1}, \qquad \bar{\boldsymbol{H}} = \sum_k \lambda_k \boldsymbol{H}_k, \tag{19}$$

$$\boldsymbol{R}_1 = \left[\boldsymbol{I} + \frac{\eta}{B} \mathcal{D}(\bar{\boldsymbol{H}} \boldsymbol{\Theta} \boldsymbol{R}_3) \right]^{-1},$$
$$\boldsymbol{R}_3 = \left[\boldsymbol{I} + \frac{\eta}{N} \boldsymbol{R}_1 \bar{\boldsymbol{H}} \boldsymbol{\Theta} \right]^{-1}. \tag{20}$$

The closed correlation equations are

$$\boldsymbol{C}_1 = \boldsymbol{R}_1 \mathcal{D} \left(\boldsymbol{C}_0 + \sigma^2 \mathbf{1}\mathbf{1}^\top \right) \boldsymbol{R}_1^\top, \qquad \boldsymbol{C}_3 = \boldsymbol{R}_3 \boldsymbol{C}_2 \boldsymbol{R}_3^\top, \tag{21}$$

$$\boldsymbol{C}_0 = \sum_k \lambda_k \boldsymbol{H}_k \left[(v_{\mathrm{init},k}^0)^2 \mathbf{1}\mathbf{1}^\top + \frac{\eta^2}{N} \boldsymbol{\Theta} \boldsymbol{R}_3 \boldsymbol{C}_2 \boldsymbol{R}_3^\top \boldsymbol{\Theta}^\top + \frac{\eta^2 \lambda_k}{B} \boldsymbol{\Theta} \boldsymbol{R}_3 \mathcal{D}(\boldsymbol{C}_1) \boldsymbol{R}_3^\top \boldsymbol{\Theta}^\top \right] \boldsymbol{H}_k^\top, \tag{22}$$

$$\boldsymbol{C}_2 = \sum_k \lambda_k^2 \boldsymbol{R}_1 \boldsymbol{H}_k \left[(v_{\mathrm{init},k}^0)^2 \mathbf{1}\mathbf{1}^\top + \frac{\eta^2}{N} \boldsymbol{\Theta} \boldsymbol{R}_3 \boldsymbol{C}_2 \boldsymbol{R}_3^\top \boldsymbol{\Theta}^\top \right] \boldsymbol{H}_k^\top \boldsymbol{R}_1^\top + \frac{1}{B} \sum_k \lambda_k \boldsymbol{M}_k \mathcal{D}(\boldsymbol{C}_1) \boldsymbol{M}_k^\top. \tag{23}$$

For zero initialization, these equations reduce to the same expressions with $v_{\mathrm{init},k}^0 = w_k^\star$.

Finally, for $\alpha = (n, t)$, the test and train losses are

$$\mathcal{L}_{\mathrm{test}}(n,t) = [\boldsymbol{C}_0]_{\alpha\alpha} + \sigma^2, \qquad \widehat{\mathcal{L}}_{\mathrm{train}}(n,t) = [\boldsymbol{C}_1]_{\alpha\alpha}.$$

The additive σ^2 is the irreducible label-noise contribution to the test mean-squared error. In noiseless experiments, set $\sigma = 0$.

B.8 Cavity Derivation of the DMFT

We derive the single-site process in Appendix B.3 by adding one coordinate to the system and expanding the old trajectory to first order in its coupling to that coordinate. The fields $\boldsymbol{v}^0, \ldots, \boldsymbol{v}^4$ are defined in Appendix B.2.

Throughout, $\alpha = (n, t)$ and $\beta = (n', t')$. The operators $\boldsymbol{\Theta}$ and $\mathcal{D}$ enforce causality and fixed-minibatch reuse.

Adding a spectral mode. Add a population feature mode $k = 0$ with eigenvalue λ_0. The added mode has variables $v_{0,\alpha}^0, v_{0,\alpha}^2, v_{0,\alpha}^4$. It couples weakly to many data and random-feature sites, so its leading effect is given by linear response. Tildes denote the enlarged system. The data field responds

to $\psi_{b0}(t_\beta)v^0_{0,\beta}$ as

$$\widetilde{v}^1_{b,\alpha} = v^1_{b,\alpha} + \sum_\beta R_{1;\alpha\beta}\psi_{b0}(t_\beta)v^0_{0,\beta} + o(1).$$

Similarly, the old random-feature-site field responds to the perturbation $A_{i0}v^2_{0,\beta}$:

$$\widetilde{v}^3_{i,\alpha} = v^3_{i,\alpha} + \sum_\beta R_{3;\alpha\beta}A_{i0}v^2_{0,\beta} + o(1).$$

The added v^2 coordinate is

$$\begin{aligned} v^2_{0,\alpha} &= \frac{1}{B}\sum_{b=1}^{B}\psi_{b0}(t_\alpha)\widetilde{v}^1_{b,\alpha} \\ &= \underbrace{\frac{1}{B}\sum_{b=1}^{B}\psi_{b0}(t_\alpha)v^1_{b,\alpha}}_{u^2_{0,\alpha}} + \frac{1}{B}\sum_{b=1}^{B}\sum_\beta \psi_{b0}(t_\alpha)\psi_{b0}(t_\beta)R_{1;\alpha\beta}v^0_{0,\beta} + o(1). \end{aligned}$$

The first term is asymptotically Gaussian, with covariance

$$\left\langle u^2_{0,\alpha}u^2_{0,\beta}\right\rangle = \frac{\lambda_0}{B}[\mathcal{D}(\boldsymbol{C}_1)]_{\alpha\beta}.$$

Only equal minibatch labels share the same data, which gives $\mathcal{D}$. The second term self-averages to the Onsager term

$$v^2_{0,\alpha} = u^2_{0,\alpha} + \lambda_0\sum_\beta R_{1;\alpha\beta}v^0_{0,\beta}.$$

Likewise,

$$\begin{aligned} v^4_{0,\alpha} &= \frac{1}{N}\sum_{i=1}^{N}A_{i0}\widetilde{v}^3_{i,\alpha} \\ &= \underbrace{\frac{1}{N}\sum_{i=1}^{N}A_{i0}v^3_{i,\alpha}}_{u^4_{0,\alpha}} + \frac{1}{N}\sum_{i=1}^{N}\sum_\beta A^2_{i0}R_{3;\alpha\beta}v^2_{0,\beta} + o(1). \end{aligned}$$

The random term is Gaussian with covariance

$$\left\langle u^4_{0,\alpha}u^4_{0,\beta}\right\rangle = \frac{1}{N}[\boldsymbol{C}_3]_{\alpha\beta},$$

and the deterministic part self-averages to

$$v^4_{0,\alpha} = u^4_{0,\alpha} + \sum_\beta R_{3;\alpha\beta}v^2_{0,\beta}.$$

Together with

$$v^0_{0,\alpha} = v^0_{\mathrm{init},0} - \eta\sum_\beta \Theta_{\alpha\beta}v^4_{0,\beta},$$

this gives the mode part of the effective process.

Adding a data site. Add one data coordinate at a fixed within-epoch minibatch label. Its contribution to v^2 perturbs the residual according to

$$\widetilde{v}^0_{k,\alpha} = v^0_{k,\alpha} + \frac{1}{B}\sum_\beta R_{0,2;k;\alpha\beta}\psi_{0k}(t_\beta)v^1_{0,\beta} + o(1).$$

The new data-site field is

$$v^1_{0,\alpha} = \sum_k \psi_{0k}(t_\alpha)\widetilde{v}^0_{k,\alpha} + \sigma\epsilon_\alpha$$
$$= \underbrace{\sum_k \psi_{0k}(t_\alpha)v^0_{k,\alpha}}_{u^1_\alpha} + \sigma\epsilon_\alpha + \frac{1}{B}\sum_\beta \sum_k \psi_{0k}(t_\alpha)\psi_{0k}(t_\beta)R_{0,2;k;\alpha\beta}v^1_{0,\beta} + o(1).$$

The Gaussian source satisfies

$$\left\langle u^1_\alpha u^1_\beta \right\rangle = [\mathcal{D}(\boldsymbol{C}_0)]_{\alpha\beta}.$$

Rows from different minibatches are independent, while the row at a given label is reused across epochs. The second term therefore self-averages to

$$v^1_{0,\alpha} = u^1_\alpha + \sigma\epsilon_\alpha + \frac{1}{B}\sum_\beta [\mathcal{D}(\boldsymbol{R}_{0,2})]_{\alpha\beta}v^1_{0,\beta}.$$

The label-noise field has the same reuse pattern:

$$\left\langle \epsilon_\alpha \epsilon_\beta \right\rangle = [\mathcal{D}(\mathbf{1}\mathbf{1}^\top)]_{\alpha\beta}.$$

Adding a random-feature site. Add one random-feature coordinate $i = 0$. Its contribution to v^4 perturbs v^2 according to

$$\widetilde{v}^2_{k,\alpha} = v^2_{k,\alpha} + \frac{1}{N}\sum_\beta R_{2,4;k;\alpha\beta}A_{0k}v^3_{0,\beta} + o(1).$$

The newly added random-feature-site field is

$$v^3_{0,\alpha} = \sum_k A_{0k}\widetilde{v}^2_{k,\alpha}$$
$$= \underbrace{\sum_k A_{0k}v^2_{k,\alpha}}_{u^3_\alpha} + \frac{1}{N}\sum_\beta \sum_k A^2_{0k}R_{2,4;k;\alpha\beta}v^3_{0,\beta} + o(1).$$

The random part is Gaussian with covariance

$$\left\langle u^3_\alpha u^3_\beta \right\rangle = [\boldsymbol{C}_2]_{\alpha\beta},$$

and the deterministic part gives

$$v^3_{0,\alpha} = u^3_\alpha + \frac{1}{N}\sum_\beta [\boldsymbol{R}_{2,4}]_{\alpha\beta}v^3_{0,\beta}.$$

Closure. The three additions give the effective process (11)–(15). The response functions are the averaged linear sensitivities appearing in the cavity expansions:

$$R_{1;\alpha\beta} = \left\langle \frac{\delta v^1_\alpha}{\delta u^1_\beta} \right\rangle, \qquad R_{3;\alpha\beta} = \left\langle \frac{\delta v^3_\alpha}{\delta u^3_\beta} \right\rangle,$$
$$R_{0,2;k;\alpha\beta} = \left\langle \frac{\delta v^0_{k,\alpha}}{\delta u^2_{k,\beta}} \right\rangle, \qquad R_{2,4;k;\alpha\beta} = \left\langle \frac{\delta v^2_{k,\alpha}}{\delta u^4_{k,\beta}} \right\rangle.$$

Combining the equations for (v^0_k, v^2_k, v^4_k) gives

$$\boldsymbol{H}_k = \left[\boldsymbol{I} + \eta\lambda_k\boldsymbol{\Theta}\boldsymbol{R}_3\boldsymbol{R}_1\right]^{-1}.$$

Substitution gives the response closure (19)–(20). Taking second moments of the Gaussian effective process gives the correlation closure (21)–(23).

B.9 MSR Path Integral Derivation of the DMFT

We derive the same DMFT with the Martin–Siggia–Rose (MSR) generating functional. We extend the random-feature derivation of Bordelon et al. (2024) to multi-epoch SGD by retaining the epoch-step index $\alpha = (n, t)$. The operators $\mathbf{\Theta}$ and $\mathcal{D}$ encode causality and reuse of fixed minibatches, respectively.

We use the fields from Appendix B.2; their stacked residual equation is (10). Dot products below include the epoch-step and site indices.

Generating functional, normalization, and response convention. For a finite trajectory, we use the Fourier representation

$$\delta(g) = \int_{\mathbb{R}} \frac{d\hat{v}}{2\pi} \exp(i\hat{v}g).$$

Accordingly, $D\boldsymbol{v}$ denotes the product of Lebesgue measures over all ordinary fields, and $D\hat{\boldsymbol{v}}$ includes one factor $d\hat{v}/(2\pi)$ for every MSR response field. The discrete-time constraints are ordered causally and have unit coefficients on their ordinary fields. Their Jacobian is therefore block triangular with unit diagonal and contributes no trajectory-dependent determinant.

Introduce MSR response fields $\hat{\boldsymbol{v}}^0, \hat{\boldsymbol{v}}^1, \hat{\boldsymbol{v}}^2, \hat{\boldsymbol{v}}^3, \hat{\boldsymbol{v}}^4$ to enforce the constraints. We also add auxiliary kicks $\boldsymbol{h}^1, \boldsymbol{h}^2, \boldsymbol{h}^3, \boldsymbol{h}^4$ to the right-hand sides:

$$\begin{aligned}
\boldsymbol{v}^0 &= \mathbf{1} \otimes \boldsymbol{v}^0_{\text{init}} - \eta \mathbf{\Theta} \boldsymbol{v}^4, \\
\boldsymbol{v}^1_\alpha &= \mathbf{\Psi}(t_\alpha) \boldsymbol{v}^0_\alpha + \sigma \boldsymbol{\epsilon}(t_\alpha) + \boldsymbol{h}^1_\alpha, \\
\boldsymbol{v}^2_\alpha &= \frac{1}{B} \mathbf{\Psi}(t_\alpha)^\top \boldsymbol{v}^1_\alpha + \boldsymbol{h}^2_\alpha, \\
\boldsymbol{v}^3_\alpha &= \boldsymbol{A} \boldsymbol{v}^2_\alpha + \boldsymbol{h}^3_\alpha, \\
\boldsymbol{v}^4_\alpha &= \frac{1}{N} \boldsymbol{A}^\top \boldsymbol{v}^3_\alpha + \boldsymbol{h}^4_\alpha.
\end{aligned}$$

With a source J coupled to $\boldsymbol{v}^0$, the generating functional is

$$\begin{aligned}
Z[J, \boldsymbol{h}] = \Bigg\langle \int D\boldsymbol{v}\, D\hat{\boldsymbol{v}}\, \exp\Bigg\{ & i\hat{\boldsymbol{v}}^0 \cdot \Big(\boldsymbol{v}^0 - \mathbf{1} \otimes \boldsymbol{v}^0_{\text{init}} + \eta \mathbf{\Theta} \boldsymbol{v}^4 \Big) + \sum_\alpha J_\alpha \cdot \boldsymbol{v}^0_\alpha \\
& + i\hat{\boldsymbol{v}}^1 \cdot \Big(\boldsymbol{v}^1 - \mathbf{\Psi} \boldsymbol{v}^0 - \sigma \boldsymbol{\epsilon} - \boldsymbol{h}^1 \Big) + i\hat{\boldsymbol{v}}^2 \cdot \Big(\boldsymbol{v}^2 - \frac{1}{B} \mathbf{\Psi}^\top \boldsymbol{v}^1 - \boldsymbol{h}^2 \Big) \\
& + i\hat{\boldsymbol{v}}^3 \cdot \Big(\boldsymbol{v}^3 - \boldsymbol{A} \boldsymbol{v}^2 - \boldsymbol{h}^3 \Big) + i\hat{\boldsymbol{v}}^4 \cdot \Big(\boldsymbol{v}^4 - \frac{1}{N} \boldsymbol{A}^\top \boldsymbol{v}^3 - \boldsymbol{h}^4 \Big) \Bigg\} \Bigg\rangle_{\boldsymbol{A}, \mathbf{\Psi}, \boldsymbol{\epsilon}}.
\end{aligned}$$

For every fixed kick, normalization of the deterministic trajectory measure gives

$$Z[0, \boldsymbol{h}] = 1. \tag{24}$$

This identity fixes both the response sign and the pure-response correlators. Because a kick enters a constraint as $-\boldsymbol{h}^a$,

$$\left. \frac{\delta \langle v^a_\alpha \rangle}{\delta h^b_\beta} \right|_{\boldsymbol{h}=0} = -i \left\langle v^a_\alpha \hat{v}^b_\beta \right\rangle_c. \tag{25}$$

At zero source, $\langle \hat{v}^a_\alpha \rangle = 0$, so the connected correlator can be replaced by the ordinary one. Repeated derivatives of (24) with respect to the kicks also give $\langle \hat{v}^a_\alpha \hat{v}^b_\beta \rangle = 0$ (Crisanti & Sompolinsky, 2018; Helias & Dahmen, 2020). The additive kicks enter the same equations as the Gaussian fields u^a in the effective process; hence (25) agrees with the functional-derivative definitions in Appendix B.5.

Average over the random-feature matrix. The entries of $\boldsymbol{A}$ and the whitened minibatch features are independent Gaussians with the covariances specified below. The disorder averages in this step are therefore exact.

The terms depending on $\boldsymbol{A}$ are

$$-i \sum_{i,k,\alpha} A_{ik} \left[\hat{v}^3_{i,\alpha} v^2_{k,\alpha} + \frac{1}{N} \hat{v}^4_{k,\alpha} v^3_{i,\alpha} \right].$$

Define

$$Y_{ik} = \sum_{\alpha} \left[\hat{v}^3_{i,\alpha} v^2_{k,\alpha} + \frac{1}{N} \hat{v}^4_{k,\alpha} v^3_{i,\alpha} \right].$$

For Gaussian A_{ik} with unit variance,

$$\left\langle \exp\left(-i \sum_{i,k} A_{ik} Y_{ik} \right) \right\rangle_{\boldsymbol{A}} = \exp\left(-\frac{1}{2} \sum_{i,k} Y^2_{ik} \right).$$

Expanding the square gives three types of contractions:

$$-\frac{1}{2} \sum_{\alpha,\beta} \left(\sum_k v^2_{k,\alpha} v^2_{k,\beta} \right) \left(\sum_i \hat{v}^3_{i,\alpha} \hat{v}^3_{i,\beta} \right),$$

$$-\frac{1}{2N^2} \sum_{\alpha,\beta} \left(\sum_i v^3_{i,\alpha} v^3_{i,\beta} \right) \left(\sum_k \hat{v}^4_{k,\alpha} \hat{v}^4_{k,\beta} \right),$$

$$-\frac{1}{N} \sum_{\alpha,\beta} \left(\sum_k v^2_{k,\alpha} \hat{v}^4_{k,\beta} \right) \left(\sum_i \hat{v}^3_{i,\alpha} v^3_{i,\beta} \right).$$

The first two contractions generate the Gaussian fields u^3 and u^4_k, with covariances $\boldsymbol{C}_2$ and $N^{-1}\boldsymbol{C}_3$. The third contraction is the Onsager reaction term involving the responses $\boldsymbol{R}_{2,4}$ and $\boldsymbol{R}_3$.

Average over the fixed minibatch matrices. The terms depending on the data matrix at within-epoch label t are

$$-i \sum_{b,k} \Psi_{bk}(t) \sum_{\alpha : t_\alpha = t} \left[\hat{v}^1_{b,\alpha} v^0_{k,\alpha} + \frac{1}{B} \hat{v}^2_{k,\alpha} v^1_{b,\alpha} \right].$$

Define

$$X_{bk}(t) = \sum_{\alpha : t_\alpha = t} \left[\hat{v}^1_{b,\alpha} v^0_{k,\alpha} + \frac{1}{B} \hat{v}^2_{k,\alpha} v^1_{b,\alpha} \right].$$

Since $\langle \Psi_{bk}(t) \Psi_{b'\ell}(t') \rangle = \delta_{bb'} \delta_{tt'} \delta_{k\ell} \lambda_k$, the average over $\boldsymbol{\Psi}(t)$ gives

$$\left\langle \exp\left(-i \sum_{t,b,k} \Psi_{bk}(t) X_{bk}(t) \right) \right\rangle_{\boldsymbol{\Psi}} = \exp\left(-\frac{1}{2} \sum_{t,b,k} \lambda_k X_{bk}(t)^2 \right).$$

Expanding this square gives

$$-\frac{1}{2} \sum_{\alpha,\beta} \left(\sum_k \lambda_k v^0_{k,\alpha} v^0_{k,\beta} \right) \left(\sum_b \hat{v}^1_{b,\alpha} \hat{v}^1_{b,\beta} \right) \mathbf{1}\{t_\alpha = t_\beta\},$$

$$-\frac{1}{2B^2} \sum_{\alpha,\beta} \left(\sum_k \lambda_k \hat{v}^2_{k,\alpha} \hat{v}^2_{k,\beta} \right) \left(\sum_b v^1_{b,\alpha} v^1_{b,\beta} \right) \mathbf{1}\{t_\alpha = t_\beta\},$$

$$-\frac{1}{B} \sum_{\alpha,\beta} \left(\sum_k \lambda_k v^0_{k,\alpha} \hat{v}^2_{k,\beta} \right) \left(\sum_b \hat{v}^1_{b,\alpha} v^1_{b,\beta} \right) \mathbf{1}\{t_\alpha = t_\beta\}. \tag{26}$$

The first contraction generates the u^1 covariance $\mathcal{D}(\boldsymbol{C}_0)$. The second generates the u^2_k covariance

$$\delta_{k\ell} \frac{\lambda_k}{B} [\mathcal{D}(\boldsymbol{C}_1)]_{\alpha\beta}.$$

The third generates the Onsager term associated with the minibatch features, $\mathcal{D}(\boldsymbol{R}_{0,2})$ in the v^1 process and $\boldsymbol{R}_1$ in the v_k^2 process. The projection $\mathcal{D}$ appears because only epoch-step pairs with the same within-epoch minibatch label share the same data matrix.

Averaging over the fixed label-noise vectors similarly produces the quadratic term

$$-\frac{\sigma^2}{2}\sum_{\alpha,\beta}\left(\sum_b \hat{v}^1_{b,\alpha}\hat{v}^1_{b,\beta}\right)\mathbf{1}\{t_\alpha = t_\beta\}.$$

Equivalently, this term can be represented by the explicit Gaussian field $\sigma\epsilon_\alpha$ in the effective process, with

$$\left\langle \sigma\epsilon_\alpha\, \sigma\epsilon_\beta \right\rangle = \sigma^2[\mathcal{D}(\mathbf{1}\mathbf{1}^\top)]_{\alpha\beta}.$$

Order parameters and delta-functional insertions. Introduce the path-level order parameters below. In the generating functional, these are integration variables; their saddle-point values give the DMFT correlations and responses.

$$[\boldsymbol{C}_0]_{\alpha\beta} = \sum_k \lambda_k v^0_{k,\alpha} v^0_{k,\beta}, \qquad [\boldsymbol{C}_1]_{\alpha\beta} = \frac{1}{B}\sum_b v^1_{b,\alpha} v^1_{b,\beta},$$

$$[\boldsymbol{C}_2]_{\alpha\beta} = \sum_k v^2_{k,\alpha} v^2_{k,\beta}, \qquad [\boldsymbol{C}_3]_{\alpha\beta} = \frac{1}{N}\sum_i v^3_{i,\alpha} v^3_{i,\beta},$$

and the two site responses

$$R_{1;\alpha\beta} = -\frac{i}{B}\sum_b v^1_{b,\alpha}\hat{v}^1_{b,\beta}, \qquad R_{3;\alpha\beta} = -\frac{i}{N}\sum_i v^3_{i,\alpha}\hat{v}^3_{i,\beta}.$$

The factors $1/B$ and $1/N$ make these the empirical versions of the single-site responses. Their conjugate order parameters are denoted by $\hat{\boldsymbol{C}}_i, \hat{\boldsymbol{R}}_1, \hat{\boldsymbol{R}}_3$.

For matrices over epoch-step indices, define

$$\boldsymbol{X} : \boldsymbol{Y} := \sum_{\alpha,\beta} X_{\alpha\beta} Y_{\alpha\beta}.$$

Each definition is imposed by a Fourier representation of a delta functional. For example, with normalization constants chosen so that each identity equals one,

$$1 = \mathcal{N}_C \int D\boldsymbol{C}_1 D\hat{\boldsymbol{C}}_1 \exp\left\{\frac{B}{2}\hat{\boldsymbol{C}}_1 : \left(\boldsymbol{C}_1 - \frac{1}{B}\sum_b \boldsymbol{v}^1_b \boldsymbol{v}^{1\top}_b\right)\right\},$$

$$1 = \mathcal{N}_R \int D\boldsymbol{R}_1 D\hat{\boldsymbol{R}}_1 \exp\left\{-\boldsymbol{R}_1 : \hat{\boldsymbol{R}}_1^\top - \frac{i}{B}\sum_{b,\alpha,\beta} \hat{R}_{1;\alpha\beta}\hat{v}^1_{b,\alpha} v^1_{b,\beta}\right\}. \tag{27}$$

The conjugate integrations run along the corresponding Fourier contours. Analogous identities impose $\boldsymbol{C}_0, \boldsymbol{C}_2, \boldsymbol{C}_3$ and $\boldsymbol{R}_3$. The transpose in (27) is an index-ordering convention chosen so that the first index of every response is the responding time.

Explicit factorized action. Inserting these identities into the disorder-averaged generating functional factorizes the microscopic variables into a representative data-site path, a representative random-feature-site path, and one path for each spectral mode:

$$Z[J, \boldsymbol{h}] = \int D\mathcal{Q} D\hat{\mathcal{Q}} \exp\left\{\mathcal{S}_{\mathrm{op}}\right\} \mathcal{Z}_1^B \mathcal{Z}_3^N \prod_{k=1}^M \mathcal{Z}_{0,2,4;k}. \tag{28}$$

Here

$$\mathcal{Q} = \{\boldsymbol{C}_0, \boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{C}_3, \boldsymbol{R}_1, \boldsymbol{R}_3\},$$

and $\hat{\mathcal{Q}}$ contains the corresponding conjugates. The order-parameter part of the action is

$$\mathcal{S}_{\rm op} = \frac{1}{2}\left(\hat{\boldsymbol{C}}_0 : \boldsymbol{C}_0 + B\hat{\boldsymbol{C}}_1 : \boldsymbol{C}_1 + \hat{\boldsymbol{C}}_2 : \boldsymbol{C}_2 + N\hat{\boldsymbol{C}}_3 : \boldsymbol{C}_3\right) - \boldsymbol{R}_1 : \hat{\boldsymbol{R}}_1^\top - \boldsymbol{R}_3 : \hat{\boldsymbol{R}}_3^\top .$$

The three factors keep the B, N, and spectral-mode contributions separate.

The data-site factor is

$$\mathcal{Z}_1 = \mathbb{E}_\epsilon \int D\boldsymbol{v}^1 D\hat{\boldsymbol{v}}^1 \exp\left\{\mathcal{S}_1\right\},$$

where $\epsilon_{(n,t)} = \epsilon_t$, the T variables $\{\epsilon_t\}$ are independent standard Gaussians, and

$$\mathcal{S}_1 = i\sum_\alpha \hat{v}_\alpha^1 \left[v_\alpha^1 - h_\alpha^1 - \sigma\epsilon_\alpha - \frac{1}{B}\sum_\beta \hat{R}_{1;\alpha\beta} v_\beta^1\right] - \frac{1}{2}\hat{\boldsymbol{v}}^{1\top}\mathcal{D}(\boldsymbol{C}_0)\hat{\boldsymbol{v}}^1 - \frac{1}{2}\hat{\boldsymbol{C}}_1 : \left(\boldsymbol{v}^1\boldsymbol{v}^{1\top}\right).$$

The path in $\mathcal{Z}_1$ spans all epoch-step indices. The operator $\mathcal{D}$ makes its components at different within-epoch labels independent, while retaining reuse of the same label across epochs.

The random-feature-site factor is

$$\mathcal{Z}_3 = \int D\boldsymbol{v}^3 D\hat{\boldsymbol{v}}^3 \exp\left\{\mathcal{S}_3\right\},$$

with

$$\mathcal{S}_3 = i\sum_\alpha \hat{v}_\alpha^3 \left[v_\alpha^3 - h_\alpha^3 - \frac{1}{N}\sum_\beta \hat{R}_{3;\alpha\beta} v_\beta^3\right] - \frac{1}{2}\hat{\boldsymbol{v}}^{3\top}\boldsymbol{C}_2\hat{\boldsymbol{v}}^3 - \frac{1}{2}\hat{\boldsymbol{C}}_3 : \left(\boldsymbol{v}^3\boldsymbol{v}^{3\top}\right).$$

Finally, the factor for spectral mode k is

$$\mathcal{Z}_{0,2,4;k} = \int D\boldsymbol{v}_k^0 D\hat{\boldsymbol{v}}_k^0 D\boldsymbol{v}_k^2 D\hat{\boldsymbol{v}}_k^2 D\boldsymbol{v}_k^4 D\hat{\boldsymbol{v}}_k^4 \exp\left\{\mathcal{S}_{0,2,4;k}\right\},$$

where

$$\begin{aligned}\mathcal{S}_{0,2,4;k} = {} & i\hat{\boldsymbol{v}}_k^{0\top}\left(\boldsymbol{v}_k^0 - v_{{\rm init},k}^0\mathbf{1} + \eta\boldsymbol{\Theta}\boldsymbol{v}_k^4\right) + \boldsymbol{J}_k^\top\boldsymbol{v}_k^0 \\ & + i\hat{\boldsymbol{v}}_k^{2\top}\left[\boldsymbol{v}_k^2 - \boldsymbol{h}_k^2 - \lambda_k\mathcal{D}(\boldsymbol{R}_1)\boldsymbol{v}_k^0\right] \\ & + i\hat{\boldsymbol{v}}_k^{4\top}\left[\boldsymbol{v}_k^4 - \boldsymbol{h}_k^4 - \boldsymbol{R}_3\boldsymbol{v}_k^2\right] \\ & - \frac{\lambda_k}{2B}\hat{\boldsymbol{v}}_k^{2\top}\mathcal{D}(\boldsymbol{C}_1)\hat{\boldsymbol{v}}_k^2 - \frac{1}{2N}\hat{\boldsymbol{v}}_k^{4\top}\boldsymbol{C}_3\hat{\boldsymbol{v}}_k^4 \\ & - \frac{\lambda_k}{2}\hat{\boldsymbol{C}}_0 : \left(\boldsymbol{v}_k^0\boldsymbol{v}_k^{0\top}\right) - \frac{1}{2}\hat{\boldsymbol{C}}_2 : \left(\boldsymbol{v}_k^2\boldsymbol{v}_k^{2\top}\right).\end{aligned}$$

The projection $\mathcal{D}$ in the $\boldsymbol{v}_k^2$ response term follows directly from (26). At the physical saddle, $\boldsymbol{R}_1$ itself has this block support, so $\mathcal{D}(\boldsymbol{R}_1) = \boldsymbol{R}_1$.

Saddle-point equations. At fixed training horizon, we obtain the DMFT closure by evaluating the order-parameter integral in (28) in the saddle-point approximation. The preceding Gaussian disorder averages are exact; Theorem 1 establishes the loss approximation separately at fixed feature dimension and training horizon. Let $\langle\cdot\rangle_1$, $\langle\cdot\rangle_3$, and $\langle\cdot\rangle_k$ denote normalized averages under $\mathcal{Z}_1$, $\mathcal{Z}_3$,

and $\mathcal{Z}_{0,2,4;k}$, respectively. Variation with respect to the conjugate correlations gives

$$[\boldsymbol{C}_0]_{\alpha\beta} = \sum_k \lambda_k \left\langle v^0_{k,\alpha} v^0_{k,\beta} \right\rangle_k, \qquad [\boldsymbol{C}_1]_{\alpha\beta} = \left\langle v^1_\alpha v^1_\beta \right\rangle_1,$$
$$[\boldsymbol{C}_2]_{\alpha\beta} = \sum_k \left\langle v^2_{k,\alpha} v^2_{k,\beta} \right\rangle_k, \qquad [\boldsymbol{C}_3]_{\alpha\beta} = \left\langle v^3_\alpha v^3_\beta \right\rangle_3.$$

Variation with respect to the conjugate responses gives the response-sign relations derived independently in (25):

$$R_{1;\alpha\beta} = -i \left\langle v^1_\alpha \hat{v}^1_\beta \right\rangle_1, \qquad R_{3;\alpha\beta} = -i \left\langle v^3_\alpha \hat{v}^3_\beta \right\rangle_3.$$

The variations with respect to the physical site responses determine their conjugates. Define

$$R_{0,2;k;\alpha\beta} = -i \left\langle v^0_{k,\alpha} \hat{v}^2_{k,\beta} \right\rangle_k, \qquad R_{2,4;k;\alpha\beta} = -i \left\langle v^2_{k,\alpha} \hat{v}^4_{k,\beta} \right\rangle_k,$$
$$[\boldsymbol{R}_{0,2}]_{\alpha\beta} = \sum_k \lambda_k R_{0,2;k;\alpha\beta}, \qquad [\boldsymbol{R}_{2,4}]_{\alpha\beta} = \sum_k R_{2,4;k;\alpha\beta}.$$

The conjugate-response saddle equations are then

$$\hat{\boldsymbol{R}}_1 = \mathcal{D}(\boldsymbol{R}_{0,2}), \qquad \hat{\boldsymbol{R}}_3 = \boldsymbol{R}_{2,4}. \tag{29}$$

Thus the two response functions that appear as feedback on representative sites are the conjugates of the site responses.

Variation with respect to the physical correlations gives

$$\hat{\boldsymbol{C}}_0 = B\,\mathcal{D}\left(\left\langle \hat{\boldsymbol{v}}^1 \hat{\boldsymbol{v}}^{1\top} \right\rangle_1\right), \qquad \hat{\boldsymbol{C}}_1 = \frac{1}{B^2}\mathcal{D}\left(\sum_k \lambda_k \left\langle \hat{\boldsymbol{v}}^2_k \hat{\boldsymbol{v}}^{2\top}_k \right\rangle_k\right),$$
$$\hat{\boldsymbol{C}}_2 = N \left\langle \hat{\boldsymbol{v}}^3 \hat{\boldsymbol{v}}^{3\top} \right\rangle_3, \qquad \hat{\boldsymbol{C}}_3 = \frac{1}{N^2}\sum_k \left\langle \hat{\boldsymbol{v}}^4_k \hat{\boldsymbol{v}}^{4\top}_k \right\rangle_k.$$

All pure-response correlators on the right-hand side vanish by (24). Therefore the physical saddle has

$$\hat{\boldsymbol{C}}_0 = \hat{\boldsymbol{C}}_1 = \hat{\boldsymbol{C}}_2 = \hat{\boldsymbol{C}}_3 = \boldsymbol{0}. \tag{30}$$

Hubbard–Stratonovich reduction. For any positive semidefinite covariance $\boldsymbol{K}$, use

$$\exp\left(-\frac{1}{2}\hat{\boldsymbol{v}}^\top \boldsymbol{K} \hat{\boldsymbol{v}}\right) = \int d\mu_{\boldsymbol{K}}(\boldsymbol{u})\, \exp\left(-i\hat{\boldsymbol{v}}^\top \boldsymbol{u}\right), \tag{31}$$

where $d\mu_{\boldsymbol{K}}$ is the normalized centered Gaussian measure with covariance $\boldsymbol{K}$. Singular covariance matrices are understood through their Gaussian law on the support of $\boldsymbol{K}$.

Substituting (29) and (30) into $\mathcal{Z}_1$, and applying (31), gives

$$\mathcal{Z}_1 = \mathbb{E}_{\boldsymbol{u}^1,\epsilon} \int D\boldsymbol{v}^1 D\hat{\boldsymbol{v}}^1 \exp\left\{ i \sum_\alpha \hat{v}^1_\alpha \Big[v^1_\alpha - h^1_\alpha - u^1_\alpha - \sigma\epsilon_\alpha - \frac{1}{B}\sum_\beta [\mathcal{D}(\boldsymbol{R}_{0,2})]_{\alpha\beta} v^1_\beta \Big] \right\},$$

where

$$\boldsymbol{u}^1 \sim \mathcal{N}\left(\boldsymbol{0}, \mathcal{D}(\boldsymbol{C}_0)\right).$$

The $\hat{\boldsymbol{v}}^1$ integral is now a delta functional, yielding

$$v^1_\alpha = h^1_\alpha + u^1_\alpha + \sigma\epsilon_\alpha + \frac{1}{B}\sum_\beta [\mathcal{D}(\boldsymbol{R}_{0,2})]_{\alpha\beta} v^1_\beta. \tag{32}$$

The identical steps for the random-feature site give

$$v_\alpha^3 = h_\alpha^3 + u_\alpha^3 + \frac{1}{N}\sum_\beta [\boldsymbol{R}_{2,4}]_{\alpha\beta} v_\beta^3, \qquad \boldsymbol{u}^3 \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{C}_2).$$

For each spectral mode, the two quadratic response-field terms introduce independent Gaussian fields $\boldsymbol{u}_k^2, \boldsymbol{u}_k^4$. Integrating $\hat{\boldsymbol{v}}_k^2, \hat{\boldsymbol{v}}_k^4$, and $\hat{\boldsymbol{v}}_k^0$ gives

$$\begin{aligned} \boldsymbol{v}_k^0 &= v_{\mathrm{init},k}^0 \mathbf{1} - \eta \boldsymbol{\Theta} \boldsymbol{v}_k^4, \\ \boldsymbol{v}_k^2 &= \boldsymbol{h}_k^2 + \boldsymbol{u}_k^2 + \lambda_k \boldsymbol{R}_1 \boldsymbol{v}_k^0, \\ \boldsymbol{v}_k^4 &= \boldsymbol{h}_k^4 + \boldsymbol{u}_k^4 + \boldsymbol{R}_3 \boldsymbol{v}_k^2, \end{aligned} \tag{33}$$

where we used $\mathcal{D}(\boldsymbol{R}_1) = \boldsymbol{R}_1$. The Gaussian covariances are

$$\left\langle u_{k,\alpha}^2 u_{\ell,\beta}^2 \right\rangle = \delta_{k\ell} \frac{\lambda_k}{B} [\mathcal{D}(\boldsymbol{C}_1)]_{\alpha\beta}, \qquad \left\langle u_{k,\alpha}^4 u_{\ell,\beta}^4 \right\rangle = \delta_{k\ell} \frac{1}{N} [\boldsymbol{C}_3]_{\alpha\beta}.$$

The support statement $\mathcal{D}(\boldsymbol{R}_1) = \boldsymbol{R}_1$ follows directly from (32): its forcing and feedback connect only equal within-epoch labels, so the response to h_β^1 vanishes whenever $t_\alpha \neq t_\beta$.

Setting the kicks to zero in (32)–(33) recovers (11)–(15), including the fixed-order label-noise covariance. Differentiating these linear equations recovers the response closure, while taking second moments recovers the correlation closure in Appendix B.7.

B.10 Finite-Horizon Error Estimate

We prove a loss approximation for Gaussian population features at fixed feature dimension M and horizon $S = T\mathcal{E}$. The comparison uses population gradient descent as a common reference for the finite system and the closed DMFT equations.

Write $\boldsymbol{C}_0^{N,B}$ for the empirical residual correlation and $\boldsymbol{C}_0$ for the solution of Appendix B.7 at the same N, B, M:

$$[\boldsymbol{C}_0^{N,B}]_{\alpha\beta} := \sum_{k=1}^M \lambda_k v_{k,\alpha}^{0,N,B} v_{k,\beta}^{0,N,B}.$$

Their test losses are

$$\mathcal{L}_{N,B}(\alpha) = [\boldsymbol{C}_0^{N,B}]_{\alpha\alpha} + \sigma^2, \qquad \mathcal{L}_{\mathrm{DMFT}}(\alpha) = [\boldsymbol{C}_0]_{\alpha\alpha} + \sigma^2.$$

The population trajectory and loss are

$$\bar{\boldsymbol{v}}_s = (\boldsymbol{I} - \eta\boldsymbol{\Lambda})^s \boldsymbol{w}^\star, \qquad \mathcal{L}_{\mathrm{pop}}(s) = \bar{\boldsymbol{v}}_s^\top \boldsymbol{\Lambda} \bar{\boldsymbol{v}}_s + \sigma^2, \quad 0 \le s < S. \tag{34}$$

Here $s = nT + t$ corresponds to the one-based index $\alpha = s + 1$.

Theorem 1 (Fixed-horizon DMFT approximation)**.** *Fix $M, T, \mathcal{E}$, a finite spectrum $\lambda_k > 0$, a deterministic target $\boldsymbol{w}^\star$, and $\sigma \ge 0$. Let the entries of $\boldsymbol{A}$ be independent standard Gaussians, the training features be independent $\mathcal{N}(\boldsymbol{0}, \boldsymbol{\Lambda})$ vectors, and the label noises be independent standard Gaussians, with these three collections mutually independent. Use zero initialization, fixed-order minibatch reuse, and a fixed learning rate satisfying $0 < \eta\|\boldsymbol{\Lambda}\|_{\mathrm{op}} \le 1 - \rho$ for some $\rho \in (0,1)$. For $\delta \in (0,1)$, set*

$$h = \log \frac{8(T+1)}{\delta}, \qquad \varepsilon = \sqrt{\frac{M+h}{N}} + \sqrt{\frac{M+h}{B}}. \tag{35}$$

There are constants $K_{M,S} < \infty$ and $c_0 > 0$, depending on the fixed parameters but independent of N, B, δ, such that, whenever $\varepsilon \le c_0$, with probability at least $1 - \delta$,

$$\max_{\alpha \le S} \left| \mathcal{L}_{N,B}(\alpha) - \mathcal{L}_{\mathrm{DMFT}}(\alpha) \right| \le K_{M,S}\varepsilon. \tag{36}$$

If

$$\ell_S := \min_{0 \le s < S} \mathcal{L}_{\mathrm{pop}}(s) > 0,$$

the constants can be chosen so that, on the same event,

$$\max_{\alpha \le S} \frac{\left|\mathcal{L}_{N,B}(\alpha) - \mathcal{L}_{\mathrm{DMFT}}(\alpha)\right|}{\mathcal{L}_{\mathrm{DMFT}}(\alpha)} \le \frac{2K_{M,S}}{\ell_S}\varepsilon. \tag{37}$$

The condition $\ell_S > 0$ holds when $\sigma > 0$, or when $\boldsymbol{w}^\star \neq \mathbf{0}$ under the stated positive spectrum and learning rate. The constants depend on the fixed spectrum and horizon; extending the bounds to growing cutoffs or horizons requires control of this dependence.

Proof outline. Concentration of the fixed feature and minibatch covariance matrices gives a uniform comparison of the finite trajectory with population gradient descent. Causality makes the finite-horizon DMFT solution polynomial in $1/N$ and $1/B$; at $1/N = 1/B = 0$, it equals the same population trajectory. The triangle inequality gives the absolute loss bound. The relative bound uses only the positive minimum of the reference loss. Appendix J gives the proof.

Stability of the causal equations. Let $\mathfrak{Q} = (\boldsymbol{C}_0, \boldsymbol{C}_1, \boldsymbol{C}_2, \boldsymbol{C}_3, \boldsymbol{R}_1, \boldsymbol{R}_3)$, and define the absolute product norm by

$$\|\mathfrak{Q}\|_{\max} := \max\left\{ \max_{0\le j\le 3} \|\boldsymbol{C}_j\|_{\max}, \|\boldsymbol{R}_1\|_{\max}, \|\boldsymbol{R}_3\|_{\max} \right\}.$$

Each matrix norm on the right is the maximum absolute entry.

Lemma 1 (Causal closure at a fixed horizon)**.** *Fix M, S and the model parameters. Add residuals $\boldsymbol{r}$ to the six correlation and response equations in Appendix B.7, keeping $\boldsymbol{H}_k$ and $\boldsymbol{M}_k$ defined by* (19) *and* (18)*. Take the response residuals to be strictly lower triangular and the correlation residuals to be symmetric, with $r_{0,11} = 0$ for the fixed initial residual. The perturbed equations have a unique causal solution, polynomial in the residual entries and in $1/N, 1/B$. For sufficiently small residuals,*

$$\|\mathfrak{Q}(\boldsymbol{r}) - \mathfrak{Q}(\mathbf{0})\|_{\max} \le K_{M,S}\|\boldsymbol{r}\|_{\max}. \tag{38}$$

The local constant can be chosen uniformly for $N, B \ge 1$, with the remaining parameters fixed.

The lemma establishes algebraic stability of the closed equations at each fixed horizon through their triangular time dependence.

C DMFT FOR ENSEMBLING

Consider E models indexed by $e = 1, \dots, E$. Their random feature matrices $\boldsymbol{A}^{(e)}$ are independent, while the dataset and fixed minibatch order are shared.

For ensemble member e, the residual field is

$$\boldsymbol{v}_\alpha^{0,(e)} = \boldsymbol{w}^\star - \frac{1}{N}\boldsymbol{A}^{(e)\top}\boldsymbol{w}_\alpha^{(e)}.$$

The ensemble residual is the average across members,

$$\bar{\boldsymbol{v}}^0 = \frac{1}{E}\sum_{e=1}^{E} \boldsymbol{v}^{0,(e)}.$$

Its residual correlation is therefore

$$\bar{\boldsymbol{C}}_0 = \frac{1}{E^2}\sum_{e,e'} \boldsymbol{C}_0^{(e,e')}, \qquad [\boldsymbol{C}_0^{(e,e')}]_{\alpha\beta} = \sum_k \lambda_k \left\langle v_{k,\alpha}^{0,(e)} v_{k,\beta}^{0,(e')} \right\rangle.$$

C.1 WITHIN-MODEL AND CROSS-MODEL CORRELATIONS

The replicated generating functional contains E^2 correlation matrices. Independence of the random feature matrices makes the response diagonal in the ensemble indices, $R_{e,e'} = \delta_{e,e'}R$. For the disorder-averaged correlations, exchangeability of ensemble members gives the exact form

$$\boldsymbol{C}_0^{(e,e')} = \begin{cases} \boldsymbol{C}_0^{\mathrm{self}}, & e = e', \\ \boldsymbol{C}_0^{\mathrm{cross}}, & e \neq e', \end{cases}$$

where "self" denotes a within-model correlation and "cross" a correlation between two distinct models. Averaging over the ensemble indices gives

$$\bar{\boldsymbol{C}}_0 = \frac{1}{E}\boldsymbol{C}_0^{\text{self}} + \frac{E-1}{E}\boldsymbol{C}_0^{\text{cross}} = \boldsymbol{C}_0^{\text{cross}} + \frac{1}{E}\left(\boldsymbol{C}_0^{\text{self}} - \boldsymbol{C}_0^{\text{cross}}\right). \tag{39}$$

The difference between within-model and cross-model correlations is reduced by $1/E$. For models that are independent conditional on the dataset, this difference is the conditional covariance averaged over datasets. The cross-model term is the second moment of the conditional mean residual.

C.2 Effective Process for the Ensemble

Each ensemble member obeys the same effective single-site process as in Appendix B.7:

$$\boldsymbol{v}_k^{0,(e)} = \boldsymbol{H}_k\left[w_k^\star \mathbf{1} - \eta\boldsymbol{\Theta}\left(\boldsymbol{u}_k^{4,(e)} + \boldsymbol{R}_3\boldsymbol{u}_k^{2,(e)}\right)\right], \qquad \boldsymbol{H}_k = \left[\boldsymbol{I} + \eta\lambda_k\boldsymbol{\Theta}\boldsymbol{R}_3\boldsymbol{R}_1\right]^{-1}.$$

The response operators $\boldsymbol{H}_k, \boldsymbol{R}_1, \boldsymbol{R}_3$ are determined by the single-model dynamics and therefore do not depend on E.

The Gaussian fields generated by $\boldsymbol{A}^{(e)}$ are independent across models:

$$\left\langle \boldsymbol{u}_k^{4,(e)}\boldsymbol{u}_\ell^{4,(e')\top}\right\rangle = \delta_{k\ell}\delta_{ee'}\frac{1}{N}\boldsymbol{C}_3^{\text{self}}.$$

The fields generated by the minibatch features remain correlated because the data and their order are shared:

$$\left\langle \boldsymbol{u}_k^{2,(e)}\boldsymbol{u}_\ell^{2,(e')\top}\right\rangle = \delta_{k\ell}\frac{\lambda_k}{B}\mathcal{D}(\boldsymbol{C}_1^{(e,e')}).$$

Since label noise is also part of the shared dataset, it contributes equally to the within-model and cross-model correlations:

$$\boldsymbol{C}_1^{\text{self}} = \boldsymbol{R}_1\mathcal{D}\left(\boldsymbol{C}_0^{\text{self}} + \sigma^2\mathbf{1}\mathbf{1}^\top\right)\boldsymbol{R}_1^\top, \qquad \boldsymbol{C}_1^{\text{cross}} = \boldsymbol{R}_1\mathcal{D}\left(\boldsymbol{C}_0^{\text{cross}} + \sigma^2\mathbf{1}\mathbf{1}^\top\right)\boldsymbol{R}_1^\top.$$

If label noise were independently resampled across ensemble members, the σ^2-term would appear only in the within-model correlation.

C.3 Within-Model Correlations

The within-model equations are the single-model DMFT equations with $\boldsymbol{C}_i$ replaced by $\boldsymbol{C}_i^{\text{self}}$:

$$\begin{aligned}
\boldsymbol{C}_0^{\text{self}} &= \sum_k \lambda_k\boldsymbol{H}_k\left[(w_k^\star)^2\mathbf{1}\mathbf{1}^\top + \frac{\eta^2}{N}\boldsymbol{\Theta}\boldsymbol{R}_3\boldsymbol{C}_2^{\text{self}}\boldsymbol{R}_3^\top\boldsymbol{\Theta}^\top\right]\boldsymbol{H}_k^\top \\
&\quad + \frac{\eta^2}{B}\sum_k \lambda_k^2\boldsymbol{H}_k\boldsymbol{\Theta}\boldsymbol{R}_3\mathcal{D}(\boldsymbol{C}_1^{\text{self}})\boldsymbol{R}_3^\top\boldsymbol{\Theta}^\top\boldsymbol{H}_k^\top, \\
\boldsymbol{C}_2^{\text{self}} &= \sum_k \lambda_k^2\boldsymbol{R}_1\boldsymbol{H}_k(w_k^\star)^2\mathbf{1}\mathbf{1}^\top\boldsymbol{H}_k^\top\boldsymbol{R}_1^\top \\
&\quad + \frac{\eta^2}{N}\sum_k \lambda_k^2\boldsymbol{R}_1\boldsymbol{H}_k\boldsymbol{\Theta}\boldsymbol{R}_3\boldsymbol{C}_2^{\text{self}}\boldsymbol{R}_3^\top\boldsymbol{\Theta}^\top\boldsymbol{H}_k^\top\boldsymbol{R}_1^\top \\
&\quad + \frac{1}{B}\sum_k \lambda_k\boldsymbol{M}_k\mathcal{D}(\boldsymbol{C}_1^{\text{self}})\boldsymbol{M}_k^\top.
\end{aligned}$$

C.4 Cross-Model Correlations

For $e \neq e'$, the Gaussian fields generated by the random feature matrices have zero cross-model covariance:

$$\left\langle \boldsymbol{u}_k^{4,(e)}\boldsymbol{u}_\ell^{4,(e')\top}\right\rangle = 0.$$

Therefore the explicit $1/N$ variance term is absent from the cross-model correlation equation:

$$\boldsymbol{C}_0^{\text{cross}} = \sum_k \lambda_k \boldsymbol{H}_k \left[(w_k^\star)^2 \mathbf{1}\mathbf{1}^\top + \frac{\eta^2 \lambda_k}{B} \boldsymbol{\Theta} \boldsymbol{R}_3 \mathcal{D}(\boldsymbol{C}_1^{\text{cross}}) \boldsymbol{R}_3^\top \boldsymbol{\Theta}^\top \right] \boldsymbol{H}_k^\top. \tag{40}$$

The corresponding auxiliary cross-model correlation is

$$\boldsymbol{C}_2^{\text{cross}} = \sum_k \lambda_k^2 \boldsymbol{R}_1 \boldsymbol{H}_k (w_k^\star)^2 \mathbf{1}\mathbf{1}^\top \boldsymbol{H}_k^\top \boldsymbol{R}_1^\top + \frac{1}{B} \sum_k \lambda_k \boldsymbol{M}_k \mathcal{D}(\boldsymbol{C}_1^{\text{cross}}) \boldsymbol{M}_k^\top.$$

In the present closure, $\boldsymbol{C}_2^{\text{cross}}$ does not feed back into $\boldsymbol{C}_0^{\text{cross}}$, because fluctuations generated independently by the random feature matrices have zero cross-model covariance.

C.5 Variance Reduction by Ensembling

Define

$$\Delta \boldsymbol{C}_0 = \boldsymbol{C}_0^{\text{self}} - \boldsymbol{C}_0^{\text{cross}}.$$

Subtracting the cross equation from the self equation gives

$$\Delta \boldsymbol{C}_0 = \sum_k \lambda_k \boldsymbol{H}_k \left[\frac{\eta^2}{N} \boldsymbol{\Theta} \boldsymbol{R}_3 \boldsymbol{C}_2^{\text{self}} \boldsymbol{R}_3^\top \boldsymbol{\Theta}^\top + \frac{\eta^2 \lambda_k}{B} \boldsymbol{\Theta} \boldsymbol{R}_3 \mathcal{D}\Big(\boldsymbol{R}_1 \Delta \boldsymbol{C}_0 \boldsymbol{R}_1^\top \Big) \boldsymbol{R}_3^\top \boldsymbol{\Theta}^\top \right] \boldsymbol{H}_k^\top.$$

The explicit shared label-noise term cancels on subtraction. Random-feature fluctuations generate the conditional covariance $\Delta \boldsymbol{C}_0$, and minibatch feedback propagates it through training. Label noise can still affect this covariance through $\boldsymbol{C}_2^{\text{self}}$.

C.6 Ensemble loss

Using $\Delta \boldsymbol{C}_0$ in (39) gives

$$\bar{\boldsymbol{C}}_0 = \boldsymbol{C}_0^{\text{cross}} + \frac{1}{E} \Delta \boldsymbol{C}_0 = \frac{1}{E} \boldsymbol{C}_0^{\text{self}} + \frac{E-1}{E} \boldsymbol{C}_0^{\text{cross}}. \tag{41}$$

Thus the ensemble test loss at epoch-step $\alpha = (n, t)$ is

$$\mathcal{L}_{\text{ens}}(n, t) = [\bar{\boldsymbol{C}}_0]_{\alpha\alpha} + \sigma^2.$$

In effective time, this is $\mathcal{L}(\tau, N, P, E)$ with $\tau = \eta(nT + t)$.

For $E = 1$, $\bar{\boldsymbol{C}}_0 = \boldsymbol{C}_0^{\text{self}}$, recovering the single-model DMFT. As $E \to \infty$, $\bar{\boldsymbol{C}}_0 \to \boldsymbol{C}_0^{\text{cross}}$: variance across random-feature realizations vanishes. The remaining correlation includes the bias of the mean predictor and its dependence on the shared data and labels.

Equation (41) is exact at the symmetric saddle. Suppressing late-time growth by $1/E$ further requires that the growing contribution lie in $\Delta \boldsymbol{C}_0$, as specified in Appendix F.2.

Independent shuffles. The equations above correspond to fixed-order training with the same minibatch order for every ensemble member. If each member uses an independent data shuffle, $\mathcal{D}$ in (40) should be replaced by $\mathcal{D}_{\text{cross}}$, defined to measure the probability that two members see the same example at the same update time. For independent random shuffles, this overlap is heuristically $O(1/T)$, so the shared-minibatch feedback is weaker.

C.7 Cross-Entropy of Averaged Logits

Fix a training checkpoint and condition on the training data, any shared training randomness, and a validation example with label y. Let $\boldsymbol{z}_1, \dots, \boldsymbol{z}_E \in \mathbb{R}^K$ be conditionally independent, identically distributed member logits, with mean $\boldsymbol{m}$, covariance $\boldsymbol{\Sigma}$, and finite third centered moment. For $\bar{\boldsymbol{z}}_E = E^{-1} \sum_e \boldsymbol{z}_e$ and $\ell(\boldsymbol{z}, y) = \log \sum_{j=1}^K e^{z_j} - z_y$, Taylor expansion gives

$$\mathbb{E}\ell(\bar{\boldsymbol{z}}_E, y) = \ell(\boldsymbol{m}, y) + \frac{1}{2E} \operatorname{tr}\Big[(\operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top) \boldsymbol{\Sigma} \Big] + O(E^{-3/2}), \qquad \boldsymbol{p} = \operatorname{softmax}(\boldsymbol{m}).$$

Indeed, the centered average has mean zero and covariance $\boldsymbol{\Sigma}/E$. The third directional derivative of log-sum-exp obeys $|D^3 \ell(\boldsymbol{z}, y)[\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{u}]| \le 8 \|\boldsymbol{u}\|_2^3$, since it is the third centered moment of u_j

under the softmax probabilities. The finite third moment and independence give $\mathbb{E}\|\bar{\boldsymbol{z}}_E - \boldsymbol{m}\|_2^3 = O(E^{-3/2})$, which controls the Taylor remainder. Averaging over validation examples gives the same expansion when the conditional third moments and leading loss are integrable. The $1/E$ coefficient is nonnegative and measures the effect of member logit variation at a fixed checkpoint; it vanishes when that variation is only a common additive shift of all logits.

D Batch Size in Finite-Step SGD

Fix a realization of the dataset, random-feature matrix, and label noise, and let the integer batch size B divide P, with $T = P/B$. The same ordered partition is reused every epoch. The following identities hold at any width N. Define

$$\boldsymbol{W} = \frac{\boldsymbol{A}^\top \boldsymbol{A}}{N}, \quad \widehat{\boldsymbol{\Sigma}}_t = \frac{\boldsymbol{\Psi}(t)^\top \boldsymbol{\Psi}(t)}{B},$$
$$\boldsymbol{K}_t = \boldsymbol{W}\widehat{\boldsymbol{\Sigma}}_t, \quad \boldsymbol{g}_t = \frac{\boldsymbol{\Psi}(t)^\top \boldsymbol{\epsilon}(t)}{B}.$$

Thus one update is $\boldsymbol{v}_n^0(t+1) = (\boldsymbol{I} - \eta \boldsymbol{K}_t)\boldsymbol{v}_n^0(t) - \eta\sigma \boldsymbol{W}\boldsymbol{g}_t$.

Exact update over one epoch. Composing the T updates gives

$$\begin{aligned} \boldsymbol{v}_{n+1}^0(0) &= \boldsymbol{U}_\eta \boldsymbol{v}_n^0(0) + \boldsymbol{d}_\eta, \\ \boldsymbol{U}_\eta &= (\boldsymbol{I} - \eta \boldsymbol{K}_{T-1}) \cdots (\boldsymbol{I} - \eta \boldsymbol{K}_0), \\ \boldsymbol{d}_\eta &= -\eta\sigma \sum_{t=0}^{T-1} (\boldsymbol{I} - \eta \boldsymbol{K}_{T-1}) \cdots (\boldsymbol{I} - \eta \boldsymbol{K}_{t+1}) \boldsymbol{W}\boldsymbol{g}_t. \end{aligned} \tag{42}$$

The product in the last line is the identity when $t = T-1$. Since the matrices and noisy labels are fixed, the same affine update applies every epoch, and

$$\boldsymbol{v}_n^0(0) = \boldsymbol{U}_\eta^n \boldsymbol{w}^\star + \sum_{j=0}^{n-1} \boldsymbol{U}_\eta^j \boldsymbol{d}_\eta.$$

This expression also permits direct evaluation at late checkpoints by composing epoch updates. A checkpoint after n complete epochs and t further steps applies the first t batch updates to this state.

Comparison with gradient flow. For the noiseless compute analysis, set $\sigma = 0$. Write

$$h = \eta T, \qquad \bar{\boldsymbol{K}} = \frac{1}{T}\sum_{t=0}^{T-1} \boldsymbol{K}_t = \boldsymbol{W}\frac{\boldsymbol{\Psi}^\top \boldsymbol{\Psi}}{P}, \qquad L = \max_{t<T} \|\boldsymbol{K}_t\|_{\mathrm{op}},$$

where $\boldsymbol{\Psi}$ contains all P training examples and h is the effective duration of one epoch. The GF update over this duration is $e^{-h\bar{\boldsymbol{K}}}$. Expanding both updates to second order yields

$$\begin{aligned} \boldsymbol{U}_\eta - e^{-h\bar{\boldsymbol{K}}} &= \frac{\eta^2}{2}\left(\sum_{0 \le i < j < T} [\boldsymbol{K}_j, \boldsymbol{K}_i] - \sum_{t=0}^{T-1} \boldsymbol{K}_t^2 \right) + \boldsymbol{R}_\eta, \\ \|\boldsymbol{R}_\eta\|_{\mathrm{op}} &\le \frac{(hL)^3}{3} e^{hL}, \qquad [\boldsymbol{K}_j, \boldsymbol{K}_i] = \boldsymbol{K}_j\boldsymbol{K}_i - \boldsymbol{K}_i\boldsymbol{K}_j. \end{aligned} \tag{43}$$

Indeed, the product has quadratic term $\eta^2 \sum_{i<j} \boldsymbol{K}_j\boldsymbol{K}_i$, whereas the exponential has $\eta^2(\sum_t \boldsymbol{K}_t)^2/2$. Their difference gives the displayed commutators and squared matrices. In operator norm, each tail from degree three onward is bounded by $(hL)^3 e^{hL}/6$. This proves the remainder estimate without requiring $\boldsymbol{W}$ to be invertible or the $\boldsymbol{K}_t$ to be symmetric.

The squared-matrix term is the finite-step correction already present in full-batch gradient descent. The commutators record the order of distinct batch updates; they vanish when the $\boldsymbol{K}_t$ commute. At $B = P$, $T = 1$ and the leading difference is $-\eta^2\bar{\boldsymbol{K}}^2/2$. For a fixed dataset, changing its batch partition leaves $\bar{\boldsymbol{K}}$ unchanged but changes the two correction terms. Equation (43) is a small-step expansion when $hL \ll 1$.

The comparison also controls a fixed effective training horizon. The first-order expansions give $\|\boldsymbol{U}_\eta - e^{-h\bar{\boldsymbol{K}}}\|_{\mathrm{op}} \le (hL)^2 e^{hL}$, while both update norms are at most e^{hL}. Telescoping their nth powers therefore gives, at $\tau = nh$,

$$\|\boldsymbol{U}_\eta^n - e^{-\tau\bar{\boldsymbol{K}}}\|_{\mathrm{op}} \le n(hL)^2 e^{nhL} = \tau h L^2 e^{\tau L}.$$

Multiplying by $\boldsymbol{w}^\star$ bounds the residual difference. The corresponding loss difference is controlled by

$$\left| \|\boldsymbol{v}\|_{\boldsymbol{\Lambda}}^2 - \|\boldsymbol{u}\|_{\boldsymbol{\Lambda}}^2 \right| \le \|\boldsymbol{\Lambda}\|_{\mathrm{op}} (\|\boldsymbol{v}\|_2 + \|\boldsymbol{u}\|_2) \|\boldsymbol{v} - \boldsymbol{u}\|_2.$$

At fixed P, N, M, B, a fixed realization, and fixed τ, the noiseless SGD loss at epoch boundaries approaches GF with error $O(h)$ as $\eta \to 0$. The displayed bound specifies the dependence on the training horizon and batch matrices.

Batch size under a training budget. At s updates, the training cost of E members is proportional to $ENBs$, and the effective time is $\tau = \eta s$. At fixed N, E, B, η, spending an unnormalized budget C_{SGD} therefore gives $\tau = (\eta/B) C_{\mathrm{SGD}}/(EN)$, up to integer checkpoint rounding. At fixed epoch count, the same relation is expressed by $h = \eta P/B$. Holding η/B fixed matches the GF duration per epoch, while the finite-step corrections still depend on the partition and learning rate. A finite-step compute frontier evaluates (42) at all affordable checkpoints, with the learning rate specified or optimized over the allowed choices.

For noiseless targets, independent feature draws give the exact ensemble decomposition $\mathcal{L}_E = C_{\mathrm{cross}} + (C_{\mathrm{self}} - C_{\mathrm{cross}})/E$ at any common checkpoint. Here changing B changes the self and cross correlations through the epoch update, while averaging predictions reduces the conditional variance.

At fixed learning rate, a smaller batch therefore allows a longer effective training time at the same compute. Increasing the learning rate with batch size can compensate for this change in time, subject to the stability of the discrete updates. Appendix K.5 compares both learning-rate conventions. It evaluates the full epoch update, including the finite-step terms, over the tested batch sizes and learning rates.

E Formal Continuum Reductions: Noisy SGD and Gradient Flow

We use two small-learning-rate reductions. The first keeps the epoch index discrete and takes a continuum limit within each epoch. The second averages over infinitesimal passes and gives gradient flow on the empirical loss. These are different formal limits of the discrete equations. Appendix D gives a direct finite-step comparison with GF for each integer batch size and fixed dataset.

E.1 SGD limit: integer epochs, continuum within-epoch time

Consider the scaling

$$\eta = \frac{1}{T} \to 0, \qquad B = \frac{P}{T}, \qquad P = BT \text{ fixed}, \qquad n = 0, 1, \ldots.$$

Define the within-epoch time

$$u = \frac{t}{T} \in [0, 1].$$

The global effective time remains $\tau = \eta(nT + t) = \eta T(n + u)$. Under $\eta = 1/T$, this gives $\tau = n + u$; the coordinate u resets at each epoch. The relation $B = P/T$ is a continuum idealization, since minibatch size is integer in the original dynamics. One epoch still processes P examples. The drift is $O(1)$ per epoch and the data-noise covariance is $O(P^{-1})$, so this limit retains SGD noise.

We use the combined index

$$\alpha = (n, u), \qquad n = 0, 1, \ldots, \qquad u \in [0, 1].$$

The continuum causal operator is the integral operator

$$[\boldsymbol{\Theta} \boldsymbol{f}]_n(u) = \sum_{m<n} \int_0^1 du'\, f_m(u') + \int_0^u du'\, f_n(u').$$

Equivalently, its kernel is

$$\Theta_{nm}(u,u') = \delta_{nm} \mathbf{1}\{u > u'\} + \mathbf{1}\{n > m\}.$$

The fixed-order reuse of the same within-epoch minibatch label is encoded by

$$[\mathcal{D}(\boldsymbol{C})]_{nm}(u,u') = C_{nm}(u,u')\, \delta(u-u').$$

Thus $\mathcal{D}$ is local in u but can couple different epochs.

The corresponding single-site stochastic process is

$$\begin{aligned}
v^0_{k,n}(u) &= w^\star_k - [\boldsymbol{\Theta} \boldsymbol{v}^4_k]_n(u), \\
v^1_n(u) &= u^1_n(u) + \sigma \epsilon_n(u) + \frac{1}{P} [\mathcal{D}(\boldsymbol{R}_{0,2}) \boldsymbol{v}^1]_n(u), \\
v^2_{k,n}(u) &= u^2_{k,n}(u) + \lambda_k [\boldsymbol{R}_1 \boldsymbol{v}^0_k]_n(u), \\
v^3_n(u) &= u^3_n(u) + \frac{1}{N} [\boldsymbol{R}_{2,4} \boldsymbol{v}^3]_n(u), \\
v^4_{k,n}(u) &= u^4_{k,n}(u) + [\boldsymbol{R}_3 \boldsymbol{v}^2_k]_n(u).
\end{aligned}$$

Here all products denote integral-operator products over (n,u). The label field is reused across epochs, $\epsilon_n(u) \equiv \epsilon(u)$, and has covariance $\langle \epsilon_n(u) \epsilon_m(u') \rangle = \delta(u-u')$. The Gaussian fields have covariances

$$\begin{aligned}
\left\langle u^1_n(u) u^1_m(u') \right\rangle &= [\mathcal{D}(\boldsymbol{C}_0)]_{nm}(u,u'), \\
\left\langle u^2_{k,n}(u) u^2_{\ell,m}(u') \right\rangle &= \delta_{k\ell} \frac{\lambda_k}{P} [\mathcal{D}(\boldsymbol{C}_1)]_{nm}(u,u'), \\
\left\langle u^3_n(u) u^3_m(u') \right\rangle &= C_{2;nm}(u,u'), \\
\left\langle u^4_{k,n}(u) u^4_{\ell,m}(u') \right\rangle &= \delta_{k\ell} \frac{1}{N} C_{3;nm}(u,u').
\end{aligned}$$

Compared to the discrete-time equations, $\eta \boldsymbol{\Theta}$ has become the continuum integral operator $\boldsymbol{\Theta}$, and the minibatch-variance prefactor becomes $1/P$.

The closed response equations are

$$\begin{aligned}
\boldsymbol{H}_k &= [\boldsymbol{I} + \lambda_k \boldsymbol{\Theta} \boldsymbol{R}_3 \boldsymbol{R}_1]^{-1}, \qquad \bar{\boldsymbol{H}} = \sum_k \lambda_k \boldsymbol{H}_k, \\
\boldsymbol{R}_1 &= \left[\boldsymbol{I} + \frac{1}{P} \mathcal{D}(\bar{\boldsymbol{H}} \boldsymbol{\Theta} \boldsymbol{R}_3) \right]^{-1}, \\
\boldsymbol{R}_3 &= \left[\boldsymbol{I} + \frac{1}{N} \boldsymbol{R}_1 \bar{\boldsymbol{H}} \boldsymbol{\Theta} \right]^{-1}.
\end{aligned}$$

Define

$$\boldsymbol{M}_k = \boldsymbol{I} - \lambda_k \boldsymbol{R}_1 \boldsymbol{H}_k \boldsymbol{\Theta} \boldsymbol{R}_3.$$

Then the closed correlation equations are

$$\boldsymbol{C}_1 = \boldsymbol{R}_1 \mathcal{D}\left(\boldsymbol{C}_0 + \sigma^2 \mathbf{1}\mathbf{1}^\top\right) \boldsymbol{R}_1^\top, \qquad \boldsymbol{C}_3 = \boldsymbol{R}_3 \boldsymbol{C}_2 \boldsymbol{R}_3^\top,$$

$$\boldsymbol{C}_0 = \sum_k \lambda_k \boldsymbol{H}_k \Bigg[(w_k^\star)^2 \mathbf{1}\mathbf{1}^\top + \frac{1}{N} \boldsymbol{\Theta} \boldsymbol{R}_3 \boldsymbol{C}_2 \boldsymbol{R}_3^\top \boldsymbol{\Theta}^\top + \frac{\lambda_k}{P} \boldsymbol{\Theta} \boldsymbol{R}_3 \mathcal{D}(\boldsymbol{C}_1) \boldsymbol{R}_3^\top \boldsymbol{\Theta}^\top \Bigg] \boldsymbol{H}_k^\top,$$

$$\boldsymbol{C}_2 = \sum_k \lambda_k^2 \boldsymbol{R}_1 \boldsymbol{H}_k \left[(w_k^\star)^2 \mathbf{1}\mathbf{1}^\top + \frac{1}{N} \boldsymbol{\Theta} \boldsymbol{R}_3 \boldsymbol{C}_2 \boldsymbol{R}_3^\top \boldsymbol{\Theta}^\top \right] \boldsymbol{H}_k^\top \boldsymbol{R}_1^\top + \frac{1}{P} \sum_k \lambda_k \boldsymbol{M}_k \mathcal{D}(\boldsymbol{C}_1) \boldsymbol{M}_k^\top.$$

The epoch index remains explicit, so the response and correlation functions need not be time-translation invariant. The projection $\mathcal{D}$ retains the same within-epoch label across epochs.

E.2 Gradient-flow limit

For the gradient-flow limit, let many infinitesimal passes occur over an $O(1)$ interval of macroscopic time. The auxiliary parameter $\mathcal{N}$ counts passes per unit time and is distinct from the model width N; both N and P remain fixed in this limit:

$$\tau = \eta(nT + t), \qquad \eta = \frac{1}{T\mathcal{N}}, \qquad N, P \text{ fixed}, \qquad \mathcal{N} \to \infty.$$

One epoch has duration

$$\Delta\tau = \eta T = \frac{1}{\mathcal{N}}.$$

Over an $O(1)$ interval, the algorithm makes $\mathcal{N}$ infinitesimal passes. The accumulated minibatch variance is $O((P\mathcal{N})^{-1})$ and vanishes, while the label-noise vector remains fixed. The limit is gradient flow on the noisy empirical loss:

$$\begin{aligned} \partial_\tau \boldsymbol{v}^0(\tau) &= -\left(\frac{1}{N}\boldsymbol{A}^\top \boldsymbol{A}\right) \frac{1}{P} \boldsymbol{\Psi}^\top \left(\boldsymbol{\Psi} \boldsymbol{v}^0(\tau) + \sigma \boldsymbol{\epsilon}\right) \\ &= -\left(\frac{1}{N}\boldsymbol{A}^\top \boldsymbol{A}\right) \left[\frac{1}{P} \boldsymbol{\Psi}^\top \boldsymbol{\Psi} \boldsymbol{v}^0(\tau) + \frac{\sigma}{P} \boldsymbol{\Psi}^\top \boldsymbol{\epsilon}\right]. \end{aligned}$$

Here $\boldsymbol{\epsilon} \in \mathbb{R}^P$ collects the label-noise variables drawn once with the P training examples. It is fixed throughout gradient flow, whereas the stochasticity from minibatch sampling vanishes in the limit.

The auxiliary fields are now defined using the full dataset:

$$\begin{aligned} \boldsymbol{v}^1(\tau) &= \boldsymbol{\Psi} \boldsymbol{v}^0(\tau) + \sigma \boldsymbol{\epsilon}, \\ \boldsymbol{v}^2(\tau) &= \frac{1}{P} \boldsymbol{\Psi}^\top \boldsymbol{v}^1(\tau), \\ \boldsymbol{v}^3(\tau) &= \boldsymbol{A} \boldsymbol{v}^2(\tau), \\ \boldsymbol{v}^4(\tau) &= \frac{1}{N} \boldsymbol{A}^\top \boldsymbol{v}^3(\tau), \end{aligned}$$

with

$$\partial_\tau \boldsymbol{v}^0(\tau) = -\boldsymbol{v}^4(\tau).$$

The corresponding single-site stochastic process is

$$\begin{aligned}
\partial_\tau v_k^0(\tau) &= -v_k^4(\tau),\\
v^1(\tau) &= u^1(\tau) + \sigma\epsilon + \frac{1}{P}\int_0^\tau d\bar{\tau}\, R_{0,2}(\tau,\bar{\tau}) v^1(\bar{\tau}),\\
v_k^2(\tau) &= u_k^2(\tau) + \lambda_k \int_0^\tau d\bar{\tau}\, R_1(\tau,\bar{\tau}) v_k^0(\bar{\tau}),\\
v^3(\tau) &= u^3(\tau) + \frac{1}{N}\int_0^\tau d\bar{\tau}\, R_{2,4}(\tau,\bar{\tau}) v^3(\bar{\tau}),\\
v_k^4(\tau) &= u_k^4(\tau) + \int_0^\tau d\bar{\tau}\, R_3(\tau,\bar{\tau}) v_k^2(\bar{\tau}).
\end{aligned}$$

The Gaussian fields have covariances

$$\begin{aligned}
\left\langle u^1(\tau) u^1(\tau') \right\rangle &= C_0(\tau,\tau'),\\
\left\langle u_k^2(\tau) u_\ell^2(\tau') \right\rangle &= \delta_{k\ell} \frac{\lambda_k}{P} C_1(\tau,\tau'),\\
\left\langle u^3(\tau) u^3(\tau') \right\rangle &= C_2(\tau,\tau'),\\
\left\langle u_k^4(\tau) u_\ell^4(\tau') \right\rangle &= \delta_{k\ell} \frac{1}{N} C_3(\tau,\tau').
\end{aligned}$$

The projection $\mathcal{D}$ is absent because every infinitesimal update averages over the full dataset. The data term therefore contains the full two-time correlation $C_1(\tau,\tau')$.

Let

$$\Theta(\tau,\tau') = \mathbf{1}\{\tau > \tau'\}.$$

In operator notation, the response equations are

$$\begin{aligned}
\boldsymbol{H}_k &= \left[\boldsymbol{I} + \lambda_k \boldsymbol{\Theta} \boldsymbol{R}_3 \boldsymbol{R}_1\right]^{-1}, \qquad \bar{\boldsymbol{H}} = \sum_k \lambda_k \boldsymbol{H}_k,\\
\boldsymbol{R}_1 &= \left[\boldsymbol{I} + \frac{1}{P} \bar{\boldsymbol{H}} \boldsymbol{\Theta} \boldsymbol{R}_3\right]^{-1},\\
\boldsymbol{R}_3 &= \left[\boldsymbol{I} + \frac{1}{N} \boldsymbol{R}_1 \bar{\boldsymbol{H}} \boldsymbol{\Theta}\right]^{-1}.
\end{aligned}$$

The correlation equations are

$$\boldsymbol{C}_1 = \boldsymbol{R}_1 \left(\boldsymbol{C}_0 + \sigma^2 \mathbf{1}\mathbf{1}^\top\right) \boldsymbol{R}_1^\top, \qquad \boldsymbol{C}_3 = \boldsymbol{R}_3 \boldsymbol{C}_2 \boldsymbol{R}_3^\top, \tag{44}$$

$$\begin{aligned}
\boldsymbol{C}_0 = \sum_k \lambda_k \boldsymbol{H}_k \Bigg[&(w_k^\star)^2 \mathbf{1}\mathbf{1}^\top + \frac{1}{N} \boldsymbol{\Theta} \boldsymbol{R}_3 \boldsymbol{C}_2 \boldsymbol{R}_3^\top \boldsymbol{\Theta}^\top\\
&+ \frac{\lambda_k}{P} \boldsymbol{\Theta} \boldsymbol{R}_3 \boldsymbol{C}_1 \boldsymbol{R}_3^\top \boldsymbol{\Theta}^\top \Bigg] \boldsymbol{H}_k^\top,
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{C}_2 = &\sum_k \lambda_k^2 \boldsymbol{R}_1 \boldsymbol{H}_k \left[(w_k^\star)^2 \mathbf{1}\mathbf{1}^\top + \frac{1}{N} \boldsymbol{\Theta} \boldsymbol{R}_3 \boldsymbol{C}_2 \boldsymbol{R}_3^\top \boldsymbol{\Theta}^\top\right] \boldsymbol{H}_k^\top \boldsymbol{R}_1^\top\\
&+ \frac{1}{P} \sum_k \lambda_k \boldsymbol{M}_k \boldsymbol{C}_1 \boldsymbol{M}_k^\top,
\end{aligned} \tag{45}$$

where

$$\boldsymbol{M}_k = \boldsymbol{I} - \lambda_k \boldsymbol{R}_1 \boldsymbol{H}_k \boldsymbol{\Theta} \boldsymbol{R}_3.$$

For the stationary response solution,

$$R(\tau,\tau') = R(\tau - \tau').$$

Using the Fourier convention

$$R(\tau-\tau') = \int \frac{d\omega}{2\pi} e^{i\omega(\tau-\tau')} R(\omega),$$

we introduce the causal mode resolvent

$$H_k(\omega) := [\boldsymbol{\Theta} \boldsymbol{H}_k](\omega).$$

Thus $H_k(\omega)$ below is not the literal Fourier transform of the transfer operator $\boldsymbol{H}_k$; the extra causal integration contributes a factor $(i\omega)^{-1}$. With this convention, the response equations reduce to

$$H_k(\omega) = \frac{1}{i\omega + \lambda_k R_1(\omega) R_3(\omega)}, \tag{46}$$

$$R_1(\omega) = 1 - \frac{1}{P}\sum_k \frac{\lambda_k R_1(\omega) R_3(\omega)}{i\omega + \lambda_k R_1(\omega) R_3(\omega)}, \tag{47}$$

$$R_3(\omega) = 1 - \frac{1}{N}\sum_k \frac{\lambda_k R_1(\omega) R_3(\omega)}{i\omega + \lambda_k R_1(\omega) R_3(\omega)}. \tag{48}$$

The correlations still depend on both times through the initial signal. Define their two-frequency Fourier transform by

$$C(\omega,\omega') = \int d\tau\, d\tau'\, e^{-i\omega\tau - i\omega'\tau'} C(\tau,\tau').$$

For $\sigma = 0$, the Fourier-domain correlation equations are

$$C_0(\omega,\omega') = \sum_k \lambda_k H_k(\omega) H_k(\omega') \Bigg[(w_k^\star)^2 + \frac{1}{N} C_3(\omega,\omega') + \frac{\lambda_k}{P} R_3(\omega) R_3(\omega') C_1(\omega,\omega') \Bigg], \tag{49}$$

$$C_1(\omega,\omega') = R_1(\omega) R_1(\omega') C_0(\omega,\omega'),$$

$$C_2(\omega,\omega') = \frac{(i\omega)(i\omega')}{P} C_1(\omega,\omega') \sum_k \lambda_k H_k(\omega) H_k(\omega') + R_1(\omega) R_1(\omega') \sum_k \lambda_k^2 H_k(\omega) H_k(\omega') (w_k^\star)^2 + \frac{R_1(\omega) R_1(\omega')}{N} C_3(\omega,\omega') \sum_k \lambda_k^2 H_k(\omega) H_k(\omega'), \tag{50}$$

$$C_3(\omega,\omega') = R_3(\omega) R_3(\omega') C_2(\omega,\omega'). \tag{51}$$

These Fourier-domain equations use $\sigma = 0$. For $\sigma > 0$, the fixed label-noise vector adds a zero-frequency source; the noisy analysis uses the time-domain closure in (44)–(45).

The test and train losses are obtained from the diagonal:

$$\mathcal{L}_{\text{test}}(\tau) = C_0(\tau,\tau) + \sigma^2, \qquad \widehat{\mathcal{L}}_{\text{train}}(\tau) = C_1(\tau,\tau).$$

F Scaling Form for Interpolation-Induced Late-Time Growth

We analyze the low-frequency equations underlying (4). An equal-frequency calculation first identifies the relevant scales. The cross asymptotics and stopping laws are derived separately in Appendices G and H. Appendix F.2 then gives the equal-time result under joint double-Laplace assumptions. We set $\sigma = 0$ throughout this analysis, except in Appendix F.3, which derives the additional contribution from label noise.

At the interpolation threshold $N = P$, the finite-data and finite-width responses agree. Write

$$R_1(\omega) \simeq R_3(\omega) = R(\omega), \qquad H_k(\omega) = \frac{1}{i\omega + \lambda_k R(\omega)^2}.$$

The corresponding response equation is

$$R(\omega) = 1 - \frac{1}{P} R(\omega)^2 \sum_k \lambda_k H_k(\omega). \tag{52}$$

The frequency ω is conjugate to the global effective time τ defined in Appendix E.2. At fixed learning rate, replacing effective time by update count changes prefactors but leaves the exponents unchanged.

Fourier-domain correlation equations. Let $C_0(\omega, \omega')$ and $C_2(\omega, \omega')$ be the double Fourier transforms of the residual and feature-gradient correlations.

Their coupled equations are

$$\begin{aligned} C_0(\omega, \omega') &= B_0(\omega, \omega') + \frac{1}{P} T_2(\omega, \omega') R(\omega)^2 R(\omega')^2 C_0(\omega, \omega') \\ &\quad + \frac{1}{P} R(\omega) R(\omega') T_1(\omega, \omega') C_2(\omega, \omega'), \end{aligned} \tag{53}$$

$$\begin{aligned} C_2(\omega, \omega') &= B_2(\omega, \omega') + \frac{1}{P} (i\omega)(i\omega') T_1(\omega, \omega') R(\omega) R(\omega') C_0(\omega, \omega') \\ &\quad + \frac{1}{P} T_2(\omega, \omega') R(\omega)^2 R(\omega')^2 C_2(\omega, \omega'). \end{aligned} \tag{54}$$

Here

$$\begin{aligned} B_0(\omega, \omega') &= \sum_k \lambda_k (w_k^\star)^2 H_k(\omega) H_k(\omega'), \\ B_2(\omega, \omega') &= R(\omega) R(\omega') \sum_k \lambda_k^2 (w_k^\star)^2 H_k(\omega) H_k(\omega'), \end{aligned}$$

and

$$T_q(\omega, \omega') = \sum_k \lambda_k^q H_k(\omega) H_k(\omega'), \qquad q \in \{1, 2\}. \tag{55}$$

Here B_0, B_2 contain the target coefficients. The terms containing T_2 couple each correlation to itself, while those containing T_1 couple C_0 and C_2.

Equations (53)–(54) can be written as

$$\boldsymbol{G}(\omega, \omega') \begin{pmatrix} C_0(\omega, \omega') \\ C_2(\omega, \omega') \end{pmatrix} = \begin{pmatrix} B_0(\omega, \omega') \\ B_2(\omega, \omega') \end{pmatrix}, \tag{56}$$

where

$$\boldsymbol{G}(\omega, \omega') = \begin{pmatrix} 1 - \frac{1}{P} T_2 R(\omega)^2 R(\omega')^2 & -\frac{1}{P} R(\omega) R(\omega') T_1 \\ -\frac{1}{P} (i\omega)(i\omega') R(\omega) R(\omega') T_1 & 1 - \frac{1}{P} T_2 R(\omega)^2 R(\omega')^2 \end{pmatrix}. \tag{57}$$

For compactness, T_1, T_2 in (57) denote $T_1(\omega, \omega')$ and $T_2(\omega, \omega')$. The factor $(i\omega)(i\omega')$ appears only in the feedback from C_0 to C_2.

Solving (56) for C_0 introduces a denominator

$$1 - \Gamma(\omega, \omega') := \det \boldsymbol{G}(\omega, \omega'),$$

The low-frequency behavior of this determinant controls the variance amplification.

Equivalently, the residual correlation can be written as

$$C_0(\omega, \omega') = \frac{G_{22}(\omega, \omega') B_0(\omega, \omega') - G_{12}(\omega, \omega') B_2(\omega, \omega')}{1 - \Gamma(\omega, \omega')}.$$

The source terms contain the target spectrum, and $1 - \Gamma$ contains the finite-data and finite-width feedback.

F.1 Low-Frequency Scaling Heuristic

Low-frequency analysis at equal frequencies. We first examine

$$\omega \sim \omega' \to 0.$$

After deriving the two-frequency system, set $\omega' = \omega$ and define

$$T_q(\omega) := T_q(\omega, \omega),$$

and

$$\det \boldsymbol{G}(\omega) = \left(1 - \frac{1}{P} T_2(\omega) R(\omega)^4\right)^2 - \frac{1}{P^2} (i\omega)^2 R(\omega)^4 T_1(\omega)^2. \tag{58}$$

Using

$$H_k(\omega) = \frac{1}{i\omega + \lambda_k R(\omega)^2},$$

we have

$$\lambda_k R(\omega)^2 H_k(\omega) = 1 - i\omega H_k(\omega).$$

Therefore,

$$\begin{aligned}\frac{1}{P} R(\omega)^4 T_2(\omega) &= \frac{1}{P} R(\omega)^2 \sum_k \lambda_k H_k(\omega) \left[\lambda_k R(\omega)^2 H_k(\omega)\right] \\ &= \frac{1}{P} R(\omega)^2 \sum_k \lambda_k H_k(\omega) - \frac{1}{P} i\omega R(\omega)^2 \sum_k \lambda_k H_k(\omega)^2.\end{aligned}$$

By the response equation (52),

$$\frac{1}{P} R(\omega)^2 \sum_k \lambda_k H_k(\omega) = 1 - R(\omega).$$

Hence

$$\frac{1}{P} R(\omega)^4 T_2(\omega) = 1 - R(\omega) - \frac{1}{P} i\omega R(\omega)^2 T_1(\omega). \tag{59}$$

Substituting (59) into (58) gives

$$\begin{aligned}1 - \Gamma(\omega, \omega) &= \left[R(\omega) + \frac{1}{P} i\omega R(\omega)^2 T_1(\omega)\right]^2 - \frac{1}{P^2} (i\omega)^2 R(\omega)^4 T_1(\omega)^2 \\ &= R(\omega)^2 + \frac{2}{P} i\omega R(\omega)^3 T_1(\omega).\end{aligned} \tag{60}$$

F.1.1 Power-law bias and variance scaling

Power-law source and capacity estimates. The DMFT equations above hold for arbitrary spectra $\{\lambda_k\}$ and target coefficients $\{w_k^\star\}$. To obtain explicit exponents, we specialize to the standard power-law setting used in random-feature learning-curve analyses (Bordelon et al., 2020; Canatar et al., 2021; Simon et al., 2021; Bordelon et al., 2024):

$$\lambda_k \sim k^{-b}, \qquad \lambda_k (w_k^\star)^2 \sim k^{-a}, \qquad b > 1, \quad a > 1. \tag{61}$$

Here b controls spectral decay and a controls target power. The conditions $a, b > 1$ make the relevant tails summable. We use $\asymp$ when multiplicative constants are suppressed; frequency-domain order estimates below refer to magnitudes.

Scaling away from the interpolation threshold. Let

$$m_- := \min\{N, P\}, \qquad m_+ := \max\{N, P\}, \qquad Q(\omega) := R_1(\omega) R_3(\omega).$$

The regular low-frequency branch of (47)–(48) satisfies

$$Q(\omega) \sim i\omega r^2, \qquad 1 = \frac{1}{m_-} \sum_k \frac{\lambda_k r^2}{1 + \lambda_k r^2}. \tag{62}$$

Under the power-law assumption, the spectral sum is of order $(r^2)^{1/b}$, and hence

$$r^2 \asymp m_-^b.$$

For $a - 1 < 2b$, substitution into the asymptotic bias gives the residual-bias scale

$$\mathcal{B}_\infty(N, P) := \sum_k \frac{\lambda_k (w_k^\star)^2}{(1 + \lambda_k r^2)^2} \asymp m_-^{-(a-1)}.$$

For the off-threshold variance, write

$$H_k(\omega) = \frac{1}{i\omega + \lambda_k Q(\omega)}$$

and T_q defined in (55), the two-frequency feedback matrix is

$$\boldsymbol{G}_{N,P}(\omega, \omega') = \begin{pmatrix} 1 - \dfrac{Q(\omega)Q(\omega')}{P} T_2 & -\dfrac{R_3(\omega)R_3(\omega')}{N} T_1 \\ -\dfrac{(i\omega)(i\omega')R_1(\omega)R_1(\omega')}{P} T_1 & 1 - \dfrac{Q(\omega)Q(\omega')}{N} T_2 \end{pmatrix},$$

where $T_q = T_q(\omega, \omega')$. Thus

$$\begin{aligned} 1 - \Gamma_{N,P}(\omega, \omega') &:= \det \boldsymbol{G}_{N,P}(\omega, \omega') \\ &= \left(1 - \frac{Q(\omega)Q(\omega')}{P} T_2\right)\left(1 - \frac{Q(\omega)Q(\omega')}{N} T_2\right) \\ &\quad - \frac{(i\omega)(i\omega')Q(\omega)Q(\omega')}{NP} T_1^2. \end{aligned} \tag{63}$$

To evaluate the low-frequency limit, define

$$\chi_k := \frac{\lambda_k r^2}{1 + \lambda_k r^2}, \qquad A_2 := \sum_k \chi_k^2.$$

Equation (62) implies $\sum_k \chi_k = m_-$. Moreover,

$$Q(\omega)Q(\omega')T_2(\omega, \omega') \longrightarrow A_2,$$

and

$$(i\omega)(i\omega')Q(\omega)Q(\omega')T_1(\omega, \omega')^2 \longrightarrow (m_- - A_2)^2.$$

Substituting these limits into (63) gives

$$\lim_{\omega, \omega' \to 0} \left[1 - \Gamma_{N,P}(\omega, \omega')\right] = \frac{m_+ - m_-}{m_+} \left(1 - \frac{A_2}{m_-}\right). \tag{64}$$

For a nondegenerate spectrum, the right-hand side is positive when $N \neq P$. The determinant therefore vanishes at zero frequency only at $N = P$.

Low-frequency response at the interpolation threshold. At $N = P$, $R_1 = R_3 = R$, and the low-frequency solution has the form

$$R(\omega)^2 = i\omega r^2 + o(\omega), \qquad \omega \to 0, \tag{65}$$

where r is determined by

$$1 = \frac{1}{P} \sum_k \frac{\lambda_k r^2}{1 + \lambda_k r^2}. \tag{66}$$

Substitution into (52) gives (66) at leading order.

Let $k_\star$ be the cutoff satisfying

$$\lambda_{k_\star} r^2 \asymp 1.$$

Using $\lambda_k \sim k^{-b}$ from (61), this gives

$$k_\star \asymp r^{2/b}.$$

The sum in (66) is of order $k_\star$, so $k_\star \asymp P$. Therefore

$$r^2 \asymp P^b, \qquad r \asymp P^{b/2}. \tag{67}$$

Next,

$$T_1(\omega) = \sum_k \frac{\lambda_k}{\left(i\omega + \lambda_k R(\omega)^2\right)^2} = \frac{1}{(i\omega)^2} \sum_k \frac{\lambda_k}{(1+\lambda_k r^2)^2} = \frac{1}{(i\omega)^2} \mathcal{T}_1(P).$$

Using the same cutoff $k_\star \asymp P$,

$$\begin{aligned} \mathcal{T}_1(P) = \sum_k \frac{\lambda_k}{(1+\lambda_k r^2)^2} &\asymp r^{-4} \sum_{k \le P} \lambda_k^{-1} + \sum_{k>P} \lambda_k \\ &\asymp P^{-2b} \sum_{k \le P} k^b + \sum_{k>P} k^{-b} \\ &\asymp P^{-2b} \frac{P^{b+1}}{b+1} + \frac{P^{1-b}}{b-1} \asymp P^{1-b}, \qquad b > 1. \end{aligned}$$

Thus

$$T_1(\omega) \asymp (i\omega)^{-2} P^{1-b}. \tag{68}$$

Substituting (65), (67), and (68) into (60) gives

$$\begin{aligned} 1 - \Gamma(\omega, \omega) &= R(\omega)^2 + \frac{2}{P} i\omega R(\omega)^3 T_1(\omega) \\ &\asymp P^b (i\omega) + \frac{2}{P} (i\omega)(i\omega)^{3/2} P^{3b/2} (i\omega)^{-2} P^{1-b} \\ &\asymp P^b (i\omega) + 2P^{b/2} (i\omega)^{1/2}. \end{aligned}$$

The ratio of the analytic and square-root terms is

$$\frac{P^b |\omega|}{P^{b/2} |\omega|^{1/2}} = P^{b/2} |\omega|^{1/2}.$$

Thus the square-root term dominates when

$$P^{b/2} |\omega|^{1/2} \ll 1, \qquad \text{equivalently} \qquad |\omega| \ll P^{-b}.$$

Since $|\omega| \asymp \tau^{-1}$ at time τ, the square-root term appears for

$$\tau \gtrsim P^b.$$

For $\tau \ll P^b$, the analytic term $P^b(i\omega)$ dominates and gives the usual mode-wise relaxation.

Heuristic equal-time inversion. For a generic two-time correlation, the equal-time inverse transform is

$$C(\tau, \tau) = \int \frac{d\omega \, d\omega'}{(2\pi)^2} e^{i\tau(\omega+\omega')} \mathcal{C}(\omega, \omega').$$

Introduce the total and relative frequencies

$$\Omega = \omega + \omega', \qquad \nu = \frac{\omega - \omega'}{2}.$$

Then

$$C(\tau, \tau) = \int \frac{d\Omega}{2\pi} e^{i\tau\Omega} \overline{\mathcal{C}}(\Omega), \qquad \overline{\mathcal{C}}(\Omega) = \int \frac{d\nu}{2\pi} \mathcal{C}\left(\frac{\Omega}{2} + \nu, \frac{\Omega}{2} - \nu\right).$$

Large time selects small total frequency Ω, but does not set $\nu = 0$. The equal-frequency approximation therefore requires an additional assumption. Suppose the contours can be deformed to a unique dominant saddle of

$$\Phi_\tau(\omega, \omega') = i\tau(\omega + \omega') + \log \mathcal{C}(\omega, \omega').$$

Its saddle equations are

$$i\tau + \partial_\omega \log \mathcal{C}(\omega, \omega') = 0, \qquad i\tau + \partial_{\omega'} \log \mathcal{C}(\omega, \omega') = 0.$$

Exchange symmetry maps an off-diagonal saddle to a second saddle with the same value. Uniqueness therefore places the saddle at $\omega = \omega' = \omega_\star$. The diagonal exponent then determines the large-time exponent only if the rescaled ν-integral is finite and nonzero, including its Jacobian.

The joint inversion assumptions in Appendix F.2 provide the control away from equal frequencies that this heuristic requires.

Time bottleneck. For $\tau \ll P^b$, each mode relaxes independently to leading order. Since

$$\lambda_k \sim k^{-b},$$

the k-th mode relaxes on the timescale

$$\tau_k \asymp \lambda_k^{-1} \asymp k^b.$$

Hence, by time τ, only modes with

$$k \lesssim \tau^{1/b}$$

have relaxed appreciably. The higher modes retain most of their initial error.

Using the power-law task assumption

$$\lambda_k (w_k^\star)^2 \sim k^{-a},$$

the residual bias is dominated by these higher modes:

$$\mathcal{L}_{\mathrm{bias}}(\tau) \asymp \sum_{k \gtrsim \tau^{1/b}} \lambda_k (w_k^\star)^2 \asymp \sum_{k \gtrsim \tau^{1/b}} k^{-a}.$$

Approximating the sum by an integral gives

$$\mathcal{L}_{\mathrm{bias}}(\tau) \asymp \int_{\tau^{1/b}}^{\infty} k^{-a}\, dk \asymp \tau^{-(a-1)/b}.$$

For $\tau \gtrsim P^b$, the square-root term instead gives

$$1 - \Gamma(\omega, \omega) \asymp P^{b/2} (i\omega)^{1/2}. \tag{69}$$

Under the equal-frequency heuristic, the inverse determinant contributes $\tau^{1/2} P^{-b/2}$.

Signal source. The source B_0 alone has the low-frequency form

$$B_0(\omega, \omega) \sim \frac{\mathcal{B}_0(P)}{(i\omega)^2}, \qquad \mathcal{B}_0(P) := \sum_k \frac{\lambda_k (w_k^\star)^2}{(1 + \lambda_k r^2)^2}.$$

Splitting the sum at $k \asymp P$ gives

$$\mathcal{B}_0(P) \asymp P^{-2b} \sum_{k \le P} k^{2b-a} + \sum_{k > P} k^{-a}.$$

Consequently,

$$\mathcal{B}_0(P) \asymp P^{-(a-1)}, \qquad 1 < a < 2b + 1.$$

At $a = 2b + 1$, this estimate is $P^{-2b} \log P$; for larger a, it is P^{-2b}. These are estimates of B_0, not of the full numerator $G_{22} B_0 - G_{12} B_2$.

Write $r_P = r$ to display the dependence on P, and define

$$q_P := \frac{r_P^2}{P} \sum_k \frac{\lambda_k}{(1 + \lambda_k r_P^2)^2}, \qquad \mathcal{B}(P) := \sum_k \frac{\lambda_k (w_k^\star)^2}{1 + \lambda_k r_P^2}. \tag{70}$$

The response equation implies

$$q_P = 1 - \frac{1}{P} \sum_k \left(\frac{\lambda_k r_P^2}{1 + \lambda_k r_P^2} \right)^2,$$

and the power-law estimates give $q_P \asymp 1$. Under the low-frequency response expansion, $G_{22} \sim q_P$ and $-G_{12} \sim q_P/(i\omega)$ at equal frequencies. Thus

$$G_{22}B_0 - G_{12}B_2 \sim \frac{q_P}{(i\omega)^2} \sum_k \frac{\lambda_k (w_k^\star)^2 + r_P^2 \lambda_k^2 (w_k^\star)^2}{(1+\lambda_k r_P^2)^2} = \frac{q_P \mathcal{B}(P)}{(i\omega)^2}.$$

The two source terms add and leave a single power of the denominator. Its dependence on P is

$$\mathcal{B}(P) \asymp \begin{cases} P^{-(a-1)}, & 1 < a < b+1, \\ P^{-b}\log P, & a = b+1, \\ P^{-b}, & a > b+1. \end{cases} \tag{71}$$

Indeed, its low-index contribution is $P^{-b}\sum_{k\le P} k^{b-a}$, while its tail is $\sum_{k>P} k^{-a}$. The late-time and stopping laws below use the first branch, $1 < a < b+1$.

Resulting scaling form. Combining (69) and (71), for $N = P$, $1 < a < b+1$, and $\tau \gg P^b$, the equal-frequency estimate suggests a test-loss contribution of order

$$\frac{1}{E}\, P^{-b/2}\, \mathcal{B}(P)\, \tau^{1/2} \asymp \frac{1}{E}\, P^{-(a-1+b/2)}\, \tau^{1/2}.$$

The factor E^{-1} assumes that the growing contribution lies in $\boldsymbol{C}_0^{\mathrm{self}} - \boldsymbol{C}_0^{\mathrm{cross}}$, with the regularity condition on the cross-model correlation stated below.

The early-time contribution $\tau^{-(a-1)/b}$ above applies for $\tau \ll P^b$. To determine a late-time minimum, one instead needs the cross-model correlation and its derivative in that window. The exact noiseless ensemble loss is

$$\mathcal{L}(\tau, N, P, E) = C_{\mathrm{cross}}(\tau) + \frac{C_{\mathrm{self}}(\tau) - C_{\mathrm{cross}}(\tau)}{E}, \qquad N = P.$$

Appendix H derives the cross and self curves in the pure-power continuum model and uses their derivatives to determine optimal stopping. Its self-growth coefficient agrees with the calculation above.

With $T = P/B$ minibatches per epoch and $\mathcal{E}$ epochs, the available effective time is $\eta T \mathcal{E}$. Observing the late-time regime requires $\eta T \mathcal{E} \gtrsim P^b$; this kinematic condition is separate from the ensemble-dependent stopping balance.

F.2 Double-Laplace Analysis of Equal-Time Correlations

To recover the equal-time correlation, we analyze both frequency variables using a one-sided double Laplace transform. For $\Re z, \Re z' > 0$, define

$$\widehat{C}_0(z, z') = \int_0^\infty \int_0^\infty e^{-z\tau - z'\tau'} C_0(\tau, \tau')\, d\tau\, d\tau'.$$

The response and correlation equations are obtained from (46)–(51) by replacing $i\omega$ and $i\omega'$ with z and z', respectively. Let

$$D_P(z, z') := 1 - \Gamma_P(z, z') = \det \boldsymbol{G}_P(z, z')$$

and define the full source numerator

$$\mathcal{N}_P(z, z') := G_{22,P}(z, z') B_{0,P}(z, z') - G_{12,P}(z, z') B_{2,P}(z, z').$$

Then the exact transformed correlation is

$$\widehat{C}_0(z, z') = \frac{\mathcal{N}_P(z, z')}{D_P(z, z')}.$$

Both terms in the full numerator contribute at leading order. Their sum is proportional to $\mathcal{B}(P)$ in (70); its denominator has one power of $1 + \lambda_k r_P^2$.

Joint low-frequency estimate. The next lemma states the required joint expansion in z, z'.

Lemma 2 (Joint low-frequency expansion)**.** *Suppose $N = P$, $\sigma = 0$, $\lambda_k \sim c_\lambda k^{-b}$, and $\lambda_k (w_k^\star)^2 \sim c_w k^{-a}$, where $b > 1$ and $1 < a < b+1$. Let the square root denote the principal branch in $\Re z > 0$. In the joint low-frequency region relevant to the inverse Laplace contours, suppose the stationary DMFT solution satisfies, uniformly as $P \to \infty$,*

$$R_P(z)^2 = z r_P^2 \big(1 + o(1)\big), \qquad r_P^2 \asymp P^b,$$
$$D_P(z, z') = d_P \big(\sqrt{z} + \sqrt{z'}\big)\big(1 + o(1)\big), \qquad d_P \asymp P^{b/2}, \tag{72}$$
$$\mathcal{N}_P(z, z') = \frac{n_P}{z z'}\big(1 + o(1)\big), \qquad n_P := q_P \mathcal{B}(P) \asymp P^{-(a-1)}. \tag{73}$$

Assume in addition that the remainders in (72)–(73) *are dominated on the inversion contours by an integrable majorant and that the complementary-frequency contribution is $o(P^{-(a-1+b/2)} \tau^{1/2})$ in the late-time scaling window. Then*

$$\widehat{C}_{0,\mathrm{grow}}(z, z') = K_P \frac{1 + o(1)}{z z' (\sqrt{z} + \sqrt{z'})}, \qquad K_P := \frac{n_P}{d_P} \asymp P^{-(a-1+b/2)}. \tag{74}$$

To obtain the coefficient in (73), insert the response expansion into the spectral sums. In the same joint region,

$$G_{22,P} \sim q_P, \qquad -G_{12,P} \sim \frac{q_P}{\sqrt{z z'}}, \qquad B_{2,P} \sim \frac{r_P^2}{\sqrt{z z'}} \sum_k \frac{\lambda_k^2 (w_k^\star)^2}{(1 + \lambda_k r_P^2)^2}.$$

Combining these terms with $B_{0,P}$ gives $\mathcal{N}_P \sim q_P \mathcal{B}(P)/(z z')$. The spectral estimates therefore give $r_P^2 \asymp P^b$ and $n_P \asymp P^{-(a-1)}$ in the stated range. The uniform joint expansions, including the determinant expansion after its constant term cancels at $N = P$, and the inversion bounds remain assumptions of the lemma. Establishing them for the full DMFT requires control on both Bromwich contours beyond the equal-frequency calculation.

Theorem 2 (Conditional late-time critical growth)**.** *Under the conditions of Lemma 2, the singular contribution to the equal-time residual correlation satisfies*

$$C_{0,\mathrm{grow}}(\tau, \tau) = \frac{2(2 - \sqrt{2})}{\sqrt{\pi}} K_P \tau^{1/2} + o\Big(K_P \tau^{1/2}\Big). \tag{75}$$

Consequently,

$$C_{0,\mathrm{grow}}(\tau, \tau) \asymp P^{-(a-1+b/2)} \tau^{1/2}.$$

Proof. For $\Re z, \Re z' > 0$, use the integral representation

$$\frac{1}{\sqrt{z} + \sqrt{z'}} = \int_0^\infty e^{-u\sqrt{z}} e^{-u\sqrt{z'}}\, du.$$

The one-dimensional inverse Laplace identity

$$\mathcal{L}^{-1}_{z \to \tau}\left[\frac{e^{-u\sqrt{z}}}{z}\right] = \operatorname{erfc}\left(\frac{u}{2\sqrt{\tau}}\right)$$

therefore gives

$$\mathcal{L}^{-1}_{z, z' \to \tau, \tau'}\left[\frac{1}{z z' (\sqrt{z} + \sqrt{z'})}\right] = \int_0^\infty \operatorname{erfc}\left(\frac{u}{2\sqrt{\tau}}\right) \operatorname{erfc}\left(\frac{u}{2\sqrt{\tau'}}\right) du.$$

At equal times $\tau' = \tau$, the change of variables $x = u/(2\sqrt{\tau})$ yields

$$\mathcal{L}^{-1}_{z, z' \to \tau, \tau}\left[\frac{1}{z z' (\sqrt{z} + \sqrt{z'})}\right] = 2\sqrt{\tau} \int_0^\infty \operatorname{erfc}(x)^2\, dx = \frac{2(2 - \sqrt{2})}{\sqrt{\pi}} \sqrt{\tau}. \tag{76}$$

Applying this identity to the leading term in (74) gives the first term in (75). The uniform remainder and complementary-frequency bounds in Lemma 2 allow dominated contour inversion and give the stated $o(K_P\tau^{1/2})$ remainder. □

Ensemble consequence. For E ensemble members,

$$\bar{C}_0 = C_0^{\text{cross}} + \frac{1}{E}\left(C_0^{\text{self}} - C_0^{\text{cross}}\right).$$

Suppose the singular contribution in (74) belongs to $C_0^{\text{self}} - C_0^{\text{cross}}$, while the cross-model correlation has no growing equal-time contribution. Under the uniform remainder and complementary-frequency bounds for contour inversion in Lemma 2, Theorem 2 implies

$$\mathcal{L}^{(E)}_{\text{grow}}(\tau) \asymp \frac{1}{E} P^{-(a-1+b/2)}\tau^{1/2}.$$

The ensemble identity preserves the $1/E$ factor at a fixed time. Optimal stopping additionally depends on the decreasing cross contribution. Appendices G and H establish its boundedness and approach to the limit for the noiseless Gaussian specializations studied there and use the resulting derivatives to derive the stopping laws.

F.3 Effect of Label Noise

We now allow $\sigma > 0$ and average over the fixed training-label noise, which is shared by all ensemble members. The response equations are unchanged because the model is linear and the noise is independent of the features and target. The correlation equations acquire an additional source through (44). All statements below use the same gradient-flow limit as the preceding analysis.

Additional source in the single-model correlation. At $N = P$, let $\Delta\widehat{C}_i$ denote the difference between the correlations with noise variance σ^2 and those with $\sigma = 0$, with all other parameters fixed. The double Laplace transform of the constant noise covariance is $\sigma^2/(zz')$. Since $G_{11} = G_{22}$, the two correlation equations give

$$\boldsymbol{G}_P(z,z')\begin{pmatrix}\Delta\widehat{C}_0\\ \Delta\widehat{C}_2\end{pmatrix} = \frac{\sigma^2}{zz'}\begin{pmatrix}1-G_{22,P}\\ -G_{21,P}\end{pmatrix}.$$

The source vector is the first column of $\boldsymbol{I} - \boldsymbol{G}_P$, so inversion gives the exact identity within this closure

$$\Delta\widehat{C}_0(z,z') = \frac{\sigma^2}{zz'}\left[\frac{G_{22,P}(z,z')}{D_P(z,z')} - 1\right]. \tag{77}$$

Thus label noise changes the fitted residual as well as adding σ^2 to the test loss.

Conditional late-time growth. Assume the joint response and determinant expansions in Lemma 2. Assume also that they imply $G_{22,P}(z,z') = q_P(1+o(1))$ uniformly in the joint inversion region, and that the remainder and complementary-frequency bounds extend to the noise source, with error $o(\sigma^2 q_P d_P^{-1}\tau^{1/2})$ after inversion. Equation (77) then has the leading growing term

$$\Delta\widehat{C}_{0,\text{grow}}(z,z') = \frac{\sigma^2 q_P}{d_P}\frac{1+o(1)}{zz'(\sqrt{z}+\sqrt{z'})}.$$

The inverse transform in (76) gives

$$\Delta C_0(\tau,\tau) = \frac{2(2-\sqrt{2})}{\sqrt{\pi}}\frac{\sigma^2 q_P}{d_P}\sqrt{\tau} + o\left(\frac{\sigma^2 q_P}{d_P}\sqrt{\tau}\right), \qquad \tau/P^b \longrightarrow \infty. \tag{78}$$

The explicit $-\sigma^2/(zz')$ in (77) and any other terms bounded in time are lower order in this limit. Since $q_P \asymp 1$ and $d_P \asymp P^{b/2}$, the noise-induced growth has scale $\sigma^2 P^{-b/2}\sqrt{\tau}$. This part of the calculation depends on $b > 1$, but not on the target exponent a. Comparing it with the signal term in the range $1 < a < b+1$ gives the ratio

$$\frac{\sigma^2}{\mathcal{B}(P)} \asymp \sigma^2 P^{a-1}.$$

A fixed nonzero noise level therefore need not remain a small correction as the dataset size increases.

Label noise and ensembling. The cross-model closure retains the shared labels. In the gradient-flow limit of (40), its scalar feedback coefficient is $1 - G_{22,P}$; the additional $1/N$ feedback from independent random features is absent. Consequently,

$$\Delta\widehat{C}_0^{\mathrm{cross}}(z, z') = \frac{\sigma^2}{zz'}\left[\frac{1}{G_{22,P}(z, z')} - 1\right].$$

Under $G_{22,P} \to q_P > 0$, the leading low-frequency term is $\sigma^2(q_P^{-1} - 1)/(zz')$. If the inverse transforms of the cross-model remainders and complementary frequencies stay bounded in time, the contribution of label noise to the cross-model correlation is also bounded in time. The transform itself is not regular at zero: the factor $1/(zz')$ represents a constant equal-time contribution.

Under these conditions and those of (78), the ensemble identity in Appendix C.5 gives

$$\Delta\mathcal{L}_{\mathrm{grow}}^{(E)}(\tau) \sim \frac{1}{E}\frac{2(2-\sqrt{2})}{\sqrt{\pi}}\frac{\sigma^2 q_P}{d_P}\sqrt{\tau}.$$

This leading growth is reduced by $1/E$. The contribution of label noise to the cross-model correlation and the irreducible test-label variance σ^2 remain. For ensemble sizes that also grow in the limit, uniform control of the remainders is required before using this leading term.

Comparison of signal and noise sources. For $1 < a < b+1$, adding the signal and noise sources replaces $\mathcal{B}(P)$ by $\mathcal{B}(P) + \sigma^2$ in the leading growing coefficient, under the assumptions above. Thus $\sigma^2 \ll \mathcal{B}(P) \asymp P^{-(a-1)}$ makes the noise correction to this coefficient small. Optimal stopping also depends on the noise contribution to cross and its time derivative. Appendix I.4 describes the corresponding conditions for stability of the stopping time and compute allocation.

G Constant Squared Teacher Coefficients in the Finite Gaussian Model

We first establish the late-time result directly for a finite Gaussian model with an arbitrary positive spectrum, $N = P$, $\sigma = 0$, zero initialization, and gradient flow. Let $X \in \mathbb{R}^{P\times M}$ have independent $\mathcal{N}(0, \Lambda)$ rows, with $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_M) > 0$, and let $Q \in \mathbb{R}^{P\times M}$ have independent $\mathcal{N}(0, 1/P)$ entries. The predictor has coefficient vector $Q^\top\theta$, with gradient flow on $\|XQ^\top\theta - Xw^\star\|^2/(2P)$. This is the normalization of (1): $Q = A/\sqrt{N}$ and $\theta = w/\sqrt{N}$, with time τ defined using the N-rescaled learning rate. For $v_Q(\tau) = w^\star - Q^\top\theta(\tau)$, write

$$C_{\mathrm{self}}(\tau) = \mathbb{E}_{X,Q}[v_Q(\tau)^\top\Lambda v_Q(\tau)],$$
$$C_{\mathrm{cross}}(\tau) = \mathbb{E}_X[m_\tau(X)^\top\Lambda m_\tau(X)], \qquad m_\tau(X) = \mathbb{E}_Q[v_Q(\tau) \mid X].$$

Independent ensemble members share the same dataset. Consequently,

$$\mathcal{L}_E(\tau) = C_{\mathrm{cross}}(\tau) + \frac{C_{\mathrm{self}}(\tau) - C_{\mathrm{cross}}(\tau)}{E}, \qquad C_{\mathrm{self}}(\tau) - C_{\mathrm{cross}}(\tau) \geq 0. \tag{79}$$

G.1 Conditional mean and the cross correlation

Fix a full-row-rank dataset and define $G = XX^\top$, $U = X^\top G^{-1/2}$, and $\Pi = UU^\top$. Complete U to an orthogonal basis (U, V) of $\mathbb{R}^M$. The feature blocks QU and QV are independent. With $Z = XQ^\top$ and $K_X = ZZ^\top/P$, the fitted training predictions are $(I - e^{-\tau K_X})Xw^\star$. At every finite τ, the singular-value gain $(1 - e^{-\tau s^2/P})/s$ is bounded by $\sqrt{\tau/P}$, so the conditional feature averages exist. In particular, $\mathbb{E}[(QV)^\top\theta(\tau) \mid QU, X] = 0$. Set

$$F_\tau(G) = \mathbb{E}_Q[e^{-\tau K_X} \mid X], \qquad M_\tau = I - \Pi + UF_\tau(G)U^\top.$$

Gaussian sign symmetry in an eigenbasis of G implies that F_τ commutes with G. Hence $m_\tau = M_\tau w^\star$ and $0 \preceq M_\tau \preceq I$. This proves boundedness of the cross correlation for any fixed finite target. Moreover, $F_\tau \to 0$ by bounded convergence, so $m_\tau \to (I - \Pi)w^\star$.

Suppose now that $(w_k^\star)^2 = q > 0$ for every coordinate. The dataset distribution is invariant under coordinate sign changes, which make $\mathbb{E}_X[M_\tau \Lambda M_\tau]$ diagonal. Its quadratic form at $w^\star$ therefore equals q times its trace. Since $M_\tau^2 = I - \Pi + UF_\tau^2 U^\top$,

$$C_{\text{cross}}(\tau) = C_\infty + q\mathbb{E}_X \operatorname{tr}[\Lambda U F_\tau(G)^2 U^\top], \qquad C_\infty = q\mathbb{E}_X \operatorname{tr}[\Lambda(I-\Pi)]. \tag{80}$$

All $F_\tau(G)$ are simultaneously diagonal in an eigenbasis of G, and decrease in positive-semidefinite order with τ. Thus C_{cross} is strictly above C_∞ at every finite time, is nonincreasing, and is at most $q \operatorname{tr}\Lambda$. For a general target, the mixed term between $(I-\Pi)w^\star$ and $UF_\tau U^\top w^\star$ need not vanish. The cancellation in (80) uses constant squared target coefficients, not just equality of two asymptotic power-law exponents.

G.2 The spectral edge and integrable bounds

Write $K_X = G^{1/2} H^\top H G^{1/2}/P^2$ for a standard square Gaussian matrix H. At fixed $G > 0$, the density of K_X is proportional to

$$(\det G)^{-P/2} (\det K_X)^{-1/2} \exp[-(P^2/2)\operatorname{tr}(G^{-1}K_X)].$$

Let one eigenvalue s approach zero with eigenvector u. The Vandermonde factors involving s supply one determinant of the positive matrix on $u^\perp$. Integrating that matrix gives the factor

$$\det(U_\perp^\top G^{-1} U_\perp)^{-(P+1)/2} = [(\det G)^{-1} u^\top G u]^{-(P+1)/2}.$$

Thus the leading matrix spectral density is proportional to $s^{-1/2}\sqrt{\det G}\, uu^\top/(u^\top G u)^{(P+1)/2}$. The remaining eigenvalue integrals are dominated at zero by integrable negative half-powers times polynomials and at infinity by an exponential. The scalar coefficient for $H^\top H$ is $c_P = \Gamma((P+1)/2)/(\sqrt{2}\Gamma(P/2))$. It follows, for example, from the ratio $PZ_{P-1,1/2}/Z_{P,-1/2}$, where

$$Z_{d,\alpha} = 2^{d(\alpha+1)+d(d-1)/2} \prod_{j=1}^{d} \frac{\Gamma(1+j/2)\Gamma(\alpha+1+(j-1)/2)}{\Gamma(3/2)}.$$

Rescaling by P^2 and integrating $s^{-1/2}e^{-\tau s}$ gives

$$\sqrt{\tau} F_\tau(G) \longrightarrow W(G), \qquad \tau^{3/2}\partial_\tau F_\tau(G) \longrightarrow -\tfrac{1}{2}W(G), \tag{81}$$

$$W(G) = P\beta_P \sqrt{\det G} \int_{S^{P-1}} \frac{uu^\top}{(u^\top G u)^{(P+1)/2}}\, d\sigma(u), \qquad \beta_P = \frac{\sqrt{\pi}\Gamma((P+1)/2)}{\sqrt{2}\Gamma(P/2)}.$$

Here $d\sigma$ is normalized surface measure; $W(G) \succ 0$. The differentiated limit follows by including the factor $-s$ in the same integral. To control the subsequent data average, the density of the smallest eigenvalue of $H^\top H$ satisfies $f_{\min}(s) \le c_P s^{-1/2}$: shift the other eigenvalues by s in the joint density and bound their resulting negative half-powers. Eigenvalue comparison and $se^{-\tau s} \le 2e^{-\tau s/2}/(e\tau)$ then give

$$\|F_\tau(G)\| \le \frac{P\beta_P}{\sqrt{\tau \lambda_{\min}(G)}}, \qquad \|\partial_\tau F_\tau(G)\| \le \frac{2\sqrt{2}P\beta_P}{e\tau^{3/2}\sqrt{\lambda_{\min}(G)}}. \tag{82}$$

For $M > P+1$, the inverse-Wishart moment bound is

$$\mathbb{E}\lambda_{\min}(G)^{-1} \le \frac{P}{\lambda_{\min}(\Lambda)(M-P-1)} < \infty.$$

Dominated convergence in (80) now proves

$$C_{\text{cross}}(\tau) = C_\infty + D\tau^{-1} + o(\tau^{-1}), \qquad C'_{\text{cross}}(\tau) = -D\tau^{-2} + o(\tau^{-2}),$$
$$D = q\mathbb{E}_X \operatorname{tr}[\Lambda U W(G)^2 U^\top] > 0. \tag{83}$$

G.3 Self correlation and optimal stopping

Put $T_X = \operatorname{tr}[\Lambda(I-\Pi)]$, $\mathsf{E}_\tau = e^{-\tau K_X}$, and $\Psi_\tau(K_X) = (I - e^{-\tau K_X})^2 K_X^{-1}$. Conditional on QU, the independent block QV contributes $T_X\|\theta(\tau)\|^2/P$ to the risk. Applying the same data-coordinate

sign average as above gives

$$C_{\mathrm{self}}(\tau) = C_\infty + R_0(\tau) + V_0(\tau),$$
$$R_0(\tau) = q\mathbb{E}_{X,Q}\,\mathrm{tr}[\Lambda U G^{-1/2}\mathsf{E}_\tau G \mathsf{E}_\tau G^{-1/2}U^\top],$$
$$V_0(\tau) = \frac{q}{P^2}\mathbb{E}_X[T_X\,\mathrm{tr}(G\mathbb{E}_Q[\Psi_\tau(K_X)\mid X])].$$

The row-space term tends to zero and obeys

$$R_0(\tau) \le q\lambda_{\max}(\Lambda)\mathbb{E}[\mathrm{tr}\,G\,\mathrm{tr}\,G^{-1}], \qquad |R_0'(\tau)| \le \frac{2q\lambda_{\max}(\Lambda)}{e\tau}\mathbb{E}[\mathrm{tr}\,G\,\mathrm{tr}\,G^{-1}].$$

This moment is finite: for $W_0 = ZZ^\top$ with standard Gaussian $Z \in \mathbb{R}^{P\times M}$, radial independence gives $\mathbb{E}[\mathrm{tr}\,W_0\,\mathrm{tr}\,W_0^{-1}] = (PM-2)P/(M-P-1)$, and eigenvalue comparison transfers the bound to G. For the other term, the identity

$$\partial_\tau \mathbb{E}_Q[\Psi_\tau(K_X)\mid X] = 2(F_\tau(G) - F_{2\tau}(G))$$

and (81) yield limits $2(2-\sqrt{2})W(G)$ and $(2-\sqrt{2})W(G)$ after multiplying $\mathbb{E}_Q\Psi_\tau$ by $\tau^{-1/2}$ and its derivative by $\sqrt{\tau}$, respectively. The bounds in (82) dominate these quantities by $4P\beta_P/\sqrt{\lambda_{\min}(G)}$ and $2P\beta_P/\sqrt{\lambda_{\min}(G)}$. Since $T_X \le \mathrm{tr}\,\Lambda$, $\mathbb{E}(\mathrm{tr}\,G)^2 < \infty$, and the first inverse moment is finite, Cauchy–Schwarz permits the data averages. Therefore

$$C_{\mathrm{self}}(\tau) = H\sqrt{\tau} + o(\sqrt{\tau}), \qquad C_{\mathrm{self}}'(\tau) = \frac{H}{2\sqrt{\tau}} + o(\tau^{-1/2}),$$
$$H = \frac{2q(2-\sqrt{2})}{P^2}\mathbb{E}_X[T_X\,\mathrm{tr}(GW(G))] > 0. \tag{84}$$

These statements concern the expectation over data and random features. Every fixed full-rank realization has a finite limit. Nearly singular feature matrices prevent uniform integrability of the squared error.

Theorem 3 (Stopping with constant squared target coefficients)**.** *Fix $N = P$, $M > P+1$, $\Lambda > 0$, and $(w_k^\star)^2 = q > 0$, with the noiseless Gaussian model and initialization specified above. Every sequence of global minimizers $\tau_\star(P,E)$ of the expected loss satisfies, as $E \to \infty$,*

$$\tau_\star(P,E) \sim \left(\frac{2D}{H}E\right)^{2/3}, \qquad \min_\tau \mathcal{L}_E(\tau) - C_\infty \sim 3D^{1/3}(H/2)^{2/3}E^{-2/3}.$$

The positive coefficients D, H are given in (83) *and* (84)*.*

Proof. For finite E, continuity and self growth guarantee a minimizer. For every fixed time, the ensemble loss tends to the cross correlation as $E \to \infty$; the infimum therefore tends to C_∞. On each bounded time interval the cross correlation is strictly above C_∞, and (79) bounds the ensemble loss from below by cross. All minimizers consequently diverge and are eventually interior. At a minimizer, the differentiated tail estimates give $D(1-E^{-1})\tau_\star^{-2}(1+o(1)) = H(2E)^{-1}\tau_\star^{-1/2}(1+o(1))$. This proves the time formula; substitution proves the loss formula. □

The theorem gives the ensemble exponent at fixed P, M, with coefficient determined by D/H. Trace-class infinite spectra can also be treated at fixed P by using a fixed positive block with more than $P+3$ modes to dominate inverse moments before taking the feature cutoff to infinity. Uniform estimates for growing P or stopping times require control of these bounds in the joint limit.

H Continuum Cross Correlations and General Stopping Laws

We now evaluate the GF closure for the pure-power continuum spectrum $\lambda_k = k^{-b}$ and $(w_k^\star)^2 = k^{b-a}$, with $b > 1$ and $1 < a < b+1$. Mode sums are replaced by integrals over $k \in (0,\infty)$. This describes the scaling limit with a large feature cutoff relative to P; it differs from the ultimate finite-system spectral edge in Appendix G. The assumptions $N = P$, $\sigma = 0$, and zero initialization

remain in force. Define, locally in this section,

$$r = \frac{1}{b}, \quad p = \frac{a-1}{b}, \quad c_b = \frac{\pi}{b\sin(\pi/b)}, \quad K = \frac{\pi}{b\sin(\pi p)}, \quad \kappa = (c_b/P)^b, \quad g = \frac{2(2-\sqrt{2})}{\sqrt{\pi}}.$$

Both p and r lie in $(0,1)$.

H.1 Cross transform, expansion, and inversion

Set $h(z) = z/R(z)^2$, on the physical branch $h(z) > \kappa$ for $z > 0$. The response equation becomes

$$R(z) = 1 - (\kappa/h(z))^r, \qquad z = h(z)[1 - (\kappa/h(z))^r]^2.$$

Writing $h = h(z)$ and $h' = h(z')$, integration of the source and feedback spectra gives

$$U_p(h,h') = \int_0^\infty \frac{k^{-a}\,dk}{(k^{-b}+h)(k^{-b}+h')} = K\frac{h^{p-1} - (h')^{p-1}}{h'-h},$$
$$V(h,h') = \int_0^\infty \frac{k^{-2b}\,dk}{(k^{-b}+h)(k^{-b}+h')} = c_b\frac{(h')^{1-r} - h^{1-r}}{h'-h}.$$

Divided differences use their continuous value at coincident arguments. The cross equation is consequently

$$\widehat{C}_{\mathrm{cross}}(z,z') = \frac{F(h,h')}{zz'}, \qquad F(h,h') = \frac{hh'U_p(h,h')}{1 - V(h,h')/P} = Khh'\frac{h^{p-1} - (h')^{p-1}}{h'R(z') - hR(z)}. \tag{85}$$

At zero frequency the denominator $1 - V/P$ equals $r > 0$, and

$$C_\infty = F(\kappa,\kappa) = K(b-a+1)\kappa^p \asymp P^{-(a-1)}.$$

The response has expansion $h(z) = \kappa + (\sqrt{\kappa}/r)\sqrt{z} + O(z)$. Direct differentiation gives

$$\frac{F_h(\kappa,\kappa)}{C_\infty} = \frac{p+r-1}{2\kappa}, \qquad \frac{F_{hh'}(\kappa,\kappa)}{C_\infty} = \frac{p^2 - 2p + 3pr + 4 - 6r + 2r^2}{6\kappa^2}.$$

Inverting these terms, with the justification below, yields

$$C_{\mathrm{cross}}(\tau) = C_\infty + J\tau^{-1/2} + D\tau^{-1} + O(\tau^{-3/2}),$$
$$C'_{\mathrm{cross}}(\tau) = -\tfrac{1}{2}J\tau^{-3/2} - D\tau^{-2} + O(\tau^{-5/2}),$$
$$J = \frac{C_\infty(a-b)}{\sqrt{\pi\kappa}}, \qquad D = \frac{C_\infty[(a-b)^2 + (a-b) + 3b(b-1)]}{6\pi\kappa} > 0. \tag{86}$$

When $a = b$, $F(h,\kappa) = P\kappa$ for all h, so every pure-variable correction vanishes. The cross remainder then improves to $O(\tau^{-2})$, and $D = Pb(b-1)/(2\pi)$.

To justify the diagonal inversion, rescale $h = \kappa u$, $z = \kappa Z$. The upper boundary of the physical inverse of $Z = u(1-u^{-r})^2$ is

$$u = \rho(\vartheta)e^{i\vartheta}, \qquad \rho(\vartheta) = \left[\frac{\cos((r-1/2)\vartheta)}{\cos(\vartheta/2)}\right]^{1/r}, \qquad 0 < \vartheta < \pi.$$

It maps to the negative real Z axis; the other boundary $u \geq 1$ maps to the positive axis. The derivative $(1-u^{-r})[1 + (2r-1)u^{-r}]$ has no zero in the intervening upper domain, and $Z \sim u$ at infinity. The argument principle therefore gives a conformal inverse mapping the upper half-plane to itself. Thus h is a complete Bernstein function: it is positive on the positive axis, analytic off the negative axis, and maps the upper half-plane into itself. On the cut boundary, direct integration gives

$$\frac{V(h,\overline{h})}{P} = \frac{\sin((1-r)\vartheta)}{\sin(r\vartheta) + \sin((1-r)\vartheta)} \leq \max(1-r, 1/2) < 1.$$

The inequality follows from monotonicity of $x\cot x$ on $(0,\pi)$; the radial factor decreases this bound inside the domain. Cauchy–Schwarz in the spectral integral therefore proves the uniform gap

$$|1 - V(h(z), h(z'))/P| \geq \min(r, 1/2) > 0 \tag{87}$$

on all physical branches. There are no additional cross-denominator poles.

Subtract C_∞, the boundary term $F(h,\kappa)-C_\infty$, and its primed counterpart from F. The remainder vanishes if either argument is κ, and its transform divided by zz' has a double-cut inverse. The boundary terms have ordinary single-cut inverses. Their endpoint densities are respectively

$$\frac{F_h\sqrt{\kappa}}{\pi r}s^{-1/2}+O(s^{1/2}),\qquad \frac{\kappa F_{hh'}}{\pi^2 r^2}(ss')^{-1/2}+\text{higher endpoint powers.}$$

These follow from $h(-s\pm i0)-\kappa=\pm i(\sqrt{\kappa}/r)\sqrt{s}+O(s)$. Analytic integer powers have no cut discontinuity in that variable. The exact integrals and (87) bound the densities polynomially away from the endpoints, including regions where only one frequency is small. Exponential factors in the inverse transforms then justify termwise endpoint integration and time differentiation. Integrating $s^{-1/2}$ against $e^{-\tau s}$ proves (86), including its derivative.

H.2 A SCALAR INVERSE FOR SELF GROWTH

Solving the coupled self equations gives

$$\widehat{C}_{\mathrm{self}}(z,z')=Khh'\frac{R(z')h^{p-1}-R(z)(h')^{p-1}}{zz'(z'-z)}.$$

The determinant of the coupled system is $(z'-z)/(h'-h)$; the apparent singularity at $z=z'$ is removable. Define $A(z)=h(z)^p/z$ and $B(z)=\sqrt{h(z)/z}=1/R(z)$. Both are Stieltjes functions (positive on the positive axis and mapping the upper half-plane into the lower half-plane). Write $a_0(\tau)=\mathcal{L}^{-1}[A](\tau)$ and $\beta(\tau)=\mathcal{L}^{-1}[B-1](\tau)$, separating the constant part of B from its ordinary inverse. Then

$$\widehat{C}_{\mathrm{self}}(z,z')=K\frac{A(z)B(z')-B(z)A(z')}{z'-z}.$$

Its initial slice is $C_{\mathrm{self}}(\tau,0)=Ka_0(\tau)$. Inverting $(z'-z)\widehat{C}_{\mathrm{self}}$ and including the initial slice yields $(\partial_{\tau'}-\partial_\tau)C_{\mathrm{self}}=K[a_0(\tau)\beta(\tau')-\beta(\tau)a_0(\tau')]$. Integration along $(2\tau-v,v)$ therefore gives the exact identity

$$C_{\mathrm{self}}(\tau,\tau)=Ka_0(2\tau)+K\int_0^\tau[a_0(2\tau-v)\beta(v)-\beta(2\tau-v)a_0(v)]\,dv.$$

Single-cut inversion gives

$$a_0(\tau)=\kappa^p+O(\tau^{-1/2}),\qquad a_0'(\tau)=O(\tau^{-3/2}),$$
$$\beta(\tau)=\frac{\sqrt{\kappa}}{\sqrt{\pi\tau}}+O(\tau^{-3/2}),\qquad \beta'(\tau)=-\frac{\sqrt{\kappa}}{2\sqrt{\pi}\tau^{3/2}}+O(\tau^{-5/2}).$$

At zero, $a_0(\tau)\sim\tau^{-p}/\Gamma(1-p)$ and $\beta(\tau)\sim\kappa^r\tau^{r-1}/\Gamma(r)$, so both are locally integrable. Replacing a_0 by κ^p in the integral gives $gK\kappa^{p+1/2}\sqrt{\tau}+O(1)$. For each remaining term, the late argument lies in $[\tau,2\tau]$, and its $O(\tau^{-1/2})$ bound multiplies an $O(\sqrt{\tau})$ integral. The remainder is therefore $O(1)$. Differentiation has no moving-endpoint term, since the integrand vanishes at $v=\tau$; each differentiated remainder is $O(\tau^{-3/2})O(\sqrt{\tau})=O(\tau^{-1})$. Hence

$$C_{\mathrm{self}}(\tau)=H\sqrt{\tau}+O(1),\qquad C'_{\mathrm{self}}(\tau)=\frac{H}{2\sqrt{\tau}}+O(\tau^{-1}),\qquad H=gK\kappa^{p+1/2}\asymp P^{-(a-1+b/2)}.$$

H.3 GLOBAL MINIMA AND THE THREE STOPPING LAWS

A positive late cross correction alone would not exclude an earlier minimum below C_∞. Here this can be ruled out for $a\geq b$. Let $\alpha=b-a+1$ and $v(z)=(\kappa/h(z))^r$. The boundary transform is

$$F(h(z),\kappa)=K\kappa^p\frac{1-v(z)^\alpha}{1-v(z)},\qquad \frac{1-v^\alpha}{1-v}=\alpha\int_0^1(1-s+sv)^{\alpha-1}\,ds.$$

For $a>b$, $0<\alpha<1$, and v is Stieltjes. The expression in brackets after subtraction of α is a bounded complete Bernstein function vanishing at zero. Its quotient by z is Stieltjes, with a strictly positive inverse. Thus $f(\tau)=C_{\mathrm{cross}}(\tau,\infty)-C_\infty>0$ at every finite positive time. For $a=b$, f is identically zero.

To control the remaining term, expand $1/(1 - V/P)$ in (85). The scalar factors

$$a_\lambda(z) = \frac{h(z)}{z(\lambda + h(z))}, \qquad b_\lambda(z) = \frac{\lambda}{\lambda + h(z)}$$

have nonnegative time inverses. Each term in the expansion is an integral of outer products of a_λ convolved with the b_λ kernels. The mass per feedback step is $1 - r < 1$, so the expansion converges; one may first truncate the source and then remove the truncation by positivity and the positive-time inverse above. This Gram representation implies

$$C_{\mathrm{cross}}(\tau, \tau) - 2C_{\mathrm{cross}}(\tau, \infty) + C_\infty \geq 0.$$

It is strictly positive at finite positive time, since already its zeroth term contains $\int \lambda(w_\lambda^\star)^2[a_\lambda(\tau) - a_\lambda(\infty)]^2 \, d\rho(\lambda) > 0$. Consequently $C_{\mathrm{cross}}(\tau) > C_\infty$ for $a \geq b$. The continuum curve diverges at zero, as follows from its positive source term with short-time power τ^{-p}, while self growth controls infinity. Global ensemble minimizers exist and, for $a \geq b$, diverge as $E \to \infty$. Evaluating the differentiated expansions there proves

$$\begin{aligned} a > b: \quad & \tau_\star \sim \frac{J}{H} E = \frac{(b - a + 1)(a - b)}{2(2 - \sqrt{2})} \kappa^{-1} E, \\ & \mathcal{L}_E(\tau_\star) - C_\infty \sim 2\sqrt{JH}\, E^{-1/2}, \\ a = b: \quad & \tau_\star \sim \left(\frac{2D}{H} E\right)^{2/3} = \kappa^{-1} \left[\frac{b(b-1)}{\pi g} E\right]^{2/3}, \\ & \mathcal{L}_E(\tau_\star) - C_\infty \sim 3D^{1/3}(H/2)^{2/3} E^{-2/3}. \end{aligned} \tag{88}$$

For $a < b$, $J < 0$, so the cross curve has a value strictly below its long-time limit. Its minimizing set $\mathcal{S}(P) = \arg\min_{\tau > 0} C_{\mathrm{cross}}(\tau)$ is a compact subset of $(0, \infty)$. Since $\mathcal{L}_E \geq C_{\mathrm{cross}}$ and converges uniformly to cross on compact intervals, all large-E minimizers lie in one such interval, and every accumulation point belongs to $\mathcal{S}(P)$. With $m = \min_\tau C_{\mathrm{cross}}(\tau) < C_\infty$ and $V(\tau) = C_{\mathrm{self}}(\tau) - C_{\mathrm{cross}}(\tau)$, compactness and continuity also give

$$\min_\tau \mathcal{L}_E(\tau) = m + \frac{\min_{\tau \in \mathcal{S}(P)} V(\tau)}{E} + o(E^{-1}). \tag{89}$$

Indeed, evaluating at a minimizer of V on $\mathcal{S}(P)$ gives the upper bound; the lower bound follows by compactness of the ensemble minimizers and $C_{\mathrm{cross}} \geq m$. The E^{-1} scaling has a nonzero coefficient when $\min_{\mathcal{S}(P)} V > 0$. Thus the stopping times remain of order P^b, rather than growing as a power of E; their limiting location requires the complete cross curve.

Keeping both cross corrections gives a useful late-time balance. With $u_\star = \sqrt{\kappa \tau_\star}$, $j = (a - b)/\sqrt{\pi}$, and $d = [(a - b)^2 + (a - b) + 3b(b - 1)]/(6\pi)$,

$$\frac{g}{b - a + 1} u_\star^3 \simeq (E - 1)(j u_\star + 2d).$$

This requires a late stopping time and is not an exact finite-E formula. Near $a = b$, the two derivative terms become comparable at $E \asymp |a - b|^{-3}$. For $a < b$ close to this line it estimates $u_\star \simeq 2d/|j|$; far from the line this time need not be late.

H.4 Explicit curves for $a = b = 2$

Here $\kappa = (\pi/(2P))^2$ and $h(z) = (\sqrt{\kappa} + \sqrt{z})^2$. Substitution in (85) gives

$$\widehat{C}_{\mathrm{cross}}(z, z') = \frac{P\kappa}{zz'} + \frac{P\sqrt{\kappa}}{\sqrt{z}\sqrt{z'}(\sqrt{\kappa} + \sqrt{z} + \sqrt{z'})}.$$

Using $(\sqrt{\kappa} + \sqrt{z} + \sqrt{z'})^{-1} = \int_0^\infty e^{-s(\sqrt{\kappa} + \sqrt{z} + \sqrt{z'})}\, ds$ and $\mathcal{L}^{-1}[e^{-s\sqrt{z}}/\sqrt{z}](\tau) = e^{-s^2/(4\tau)}/\sqrt{\pi\tau}$ gives a positive integral inverse. The corresponding self transform is

$$\widehat{C}_{\mathrm{self}}(z, z') = \frac{P\kappa^{3/2}}{zz'(\sqrt{z} + \sqrt{z'})} + \frac{P\kappa}{zz'} + \frac{P\sqrt{\kappa}}{\sqrt{z}\sqrt{z'}(\sqrt{z} + \sqrt{z'})}.$$

Consequently, writing $x = \tau/P^2$, $\operatorname{erfcx}(s) = e^{s^2}\operatorname{erfc}(s)$, and $A = (\pi^3/8)g$,

$$PC_{\text{cross}}(P^2x) = \frac{\pi^2}{4} + \sqrt{\frac{\pi}{8}}x^{-1/2}\operatorname{erfcx}\left(\frac{\pi\sqrt{x}}{2\sqrt{2}}\right),$$

$$PC_{\text{self}}(P^2x) = \frac{\pi^2}{4} + \sqrt{\frac{\pi}{8}}x^{-1/2} + A\sqrt{x}, \tag{90}$$

$$P\mathcal{L}_E(P^2x) = \frac{\pi^2}{4} + \sqrt{\frac{\pi}{8}}x^{-1/2}\left[\frac{1}{E} + \left(1 - \frac{1}{E}\right)\operatorname{erfcx}\left(\frac{\pi\sqrt{x}}{2\sqrt{2}}\right)\right] + \frac{A\sqrt{x}}{E}. \tag{91}$$

Expanding the cross transient gives $PC_{\text{cross}} - \pi^2/4 = 1/(\pi x) - 4/(\pi^3 x^2) + O(x^{-3})$. The remaining $x^{-1/2}$ term in the ensemble loss carries $1/E$. Differentiating (91), the stopping equation is

$$Ax^{3/2} - \sqrt{\pi/8}\,x^{1/2} - \frac{2(E-1)}{\pi} + \frac{16(E-1)}{\pi^3 x} + O(Ex^{-2}) = 0.$$

Expanding around $x = [2E/(\pi A)]^{2/3}$ gives

$$\frac{\tau_\star}{P^2} = \left(\frac{2E}{\pi A}\right)^{2/3} + \frac{2\sqrt{\pi/8}}{3A} - \frac{16}{3\pi^2} + O(E^{-1/3}).$$

The constant correction is negative, so the large-E log–log slope approaches $2/3$ from above. These are complete positive-time curves of the noiseless continuum closure with zero initialization. The discrete model has a finite initial loss; comparisons as P, the cutoff, and E grow together require uniform control of the continuum approximation over the stopping window. Appendix I.1 develops the corresponding width comparison below interpolation.

I Compute Allocation with Fixed Data

We use $\mathcal{L}(\tau, N, P, E)$ for the ensemble test loss, with training time first. The main allocation problem holds the dataset and member size fixed at $N = P$. Each ensemble is trained to its own optimal stopping time, and only the number of members is chosen under the compute budget. We first optimize width and training time for a single model with abundant data, then turn to ensemble allocation at fixed $N = P$.

The stopping results below concern noiseless GF with zero initialization. The explicit dependence on a, b, P uses the pure-power continuum spectrum of Appendix H, with $b > 1$ and $1 < a < b + 1$. The cancellation at $a = b$ corresponds to constant squared teacher coefficients. Appendix G gives the corresponding finite Gaussian-model argument for that target.

Time and compute conventions. With s updates and $\mathcal{E}$ epochs, an ensemble of E models uses a number of random-feature multiply-adds proportional to

$$ENBs = EN\frac{B}{\eta}\tau = ENP\mathcal{E}, \qquad \tau = \eta s, \quad Bs = P\mathcal{E}.$$

We fix B and η, absorb the common factor B/η into the compute units, and write

$$C = EN\tau. \tag{92}$$

This normalization also permits comparisons with varying B and η if B/η is held fixed. Otherwise the factor must be restored. At fixed N, P, τ, the GF equations do not depend separately on batch size. The finite-step comparison in Appendix D shows how changing the batch partition and learning rate changes the epoch update. Batch-size allocation uses these discrete dynamics and the unnormalized cost $ENBs$.

Single-model allocation with abundant data. For $E = 1$, when the leading loss is insensitive to data reuse, the time and width terms give

$$\mathcal{L}(\tau, N) \asymp A_\tau \tau^{-(a-1)/b} + A_N N^{-(a-1)}. \tag{93}$$

This approximation underlies the usual compute-constrained calculation (Hoffmann et al., 2022; Bordelon et al., 2024). Substituting $\tau = C/N$ and balancing the two terms gives

$$N_\star(C) \asymp C^{1/(b+1)}, \qquad \tau_\star(C) \asymp C^{b/(b+1)},$$

and hence

$$\mathcal{L}^*(C) \asymp C^{-(a-1)/(b+1)}.$$

Loss at optimal stopping with fixed data. Hold $N = P$ fixed and abbreviate $\mathcal{L}(\tau, P, E) = \mathcal{L}(\tau, P, P, E)$. For independent members sharing one dataset,

$$\mathcal{L}(\tau, P, E) = \left(1 - \frac{1}{E}\right) C_{\text{cross}}(\tau) + \frac{1}{E} C_{\text{self}}(\tau).$$

Appendix H derives both the late-time expansions and the derivative estimates needed to optimize this loss:

$$C_{\text{cross}}(\tau) = C_\infty + \frac{J}{\sqrt{\tau}} + \frac{D}{\tau} + O(\tau^{-3/2}),$$
$$C_{\text{self}}(\tau) = H\sqrt{\tau} + O(1),$$

where $D, H > 0$. Put

$$c_b = \frac{\pi}{b \sin(\pi/b)}, \qquad \kappa = (c_b/P)^b, \qquad g = \frac{2(2-\sqrt{2})}{\sqrt{\pi}}.$$

In the continuum model, $C_\infty \asymp P^{-(a-1)}$ and

$$J = \frac{C_\infty j}{\sqrt{\kappa}}, \qquad D = \frac{C_\infty d}{\kappa}, \qquad H = \frac{g C_\infty \sqrt{\kappa}}{b - a + 1},$$
$$j = \frac{a-b}{\sqrt{\pi}}, \quad d = \frac{(a-b)^2 + (a-b) + 3b(b-1)}{6\pi} > 0.$$

The time-dependent terms at an interior late minimum obey

$$H\tau_\star^{3/2} \simeq (E-1)\big(J\sqrt{\tau_\star} + 2D\big). \tag{94}$$

The bounded and decreasing terms from self carry a factor $1/E$. Their derivatives are lower order on the diverging stopping scales below. Equation (94) is an asymptotic balance; small ensembles, including $E = 1$, require the full loss curve.

For $a > b$, $J > 0$, and the cross curve lies above C_∞ at every finite time. For $a = b$, $J = 0$, and the positive D/τ term controls the approach to C_∞. The global-minimum argument in Appendix H then gives

$$\begin{aligned} a > b: \quad & \tau_\star(P, E) \sim \frac{J}{H} E \asymp P^b E, \\ & \mathcal{L}^*(P, E) - C_\infty \sim 2\sqrt{JH}\, E^{-1/2}, \\ a = b: \quad & \tau_\star(P, E) \sim \left(\frac{2D}{H} E\right)^{2/3} \asymp P^b E^{2/3}, \\ & \mathcal{L}^*(P, E) - C_\infty \sim 3 D^{1/3} (H/2)^{2/3} E^{-2/3}, \end{aligned} \tag{95}$$

where $\mathcal{L}^*(P, E) = \min_{\tau > 0} \mathcal{L}(\tau, P, E)$. At $a = b = 2$, the decreasing term has size P/τ and the growing term has size $\sqrt{\tau}/(EP^2)$. Their balance gives $\tau_\star \asymp P^2 E^{2/3}$.

For $a < b$, $J < 0$, so cross is below its limiting value at some finite times. Define

$$m(P) = \min_{\tau > 0} C_{\text{cross}}(\tau), \qquad \mathcal{S}(P) = \arg\min_{\tau > 0} C_{\text{cross}}(\tau), \qquad V(\tau) = C_{\text{self}}(\tau) - C_{\text{cross}}(\tau).$$

Then $m(P) < C_\infty$, and optimal times remain in a compact interval of positive multiples of P^b, approaching $\mathcal{S}(P)$ as $E \to \infty$. The full cross curve is needed to locate this interval. Moreover,

$$\mathcal{L}^*(P, E) = m(P) + \frac{\min_{\tau \in \mathcal{S}(P)} V(\tau)}{E} + o(E^{-1}), \qquad a < b. \tag{96}$$

For the E^{-1} excess-loss rate below, we additionally assume $\min_{\tau\in\mathcal{S}(P)} V(\tau) > 0$; its coefficient then scales as $P^{-(a-1)}$ in the continuum description. If the cross minimum is unique, $\tau_\star(P,E)$ converges to its time. Thus the large-E excess in this branch is measured above $m(P)$, not C_∞. If the minimum variance vanishes, (96) only gives an $o(E^{-1})$ excess, and its leading rate requires a higher-order calculation.

Allocation over ensemble size. Let $\tau_\star(P,E)$ be the earliest global minimizer when there is a tie. The exact cost and feasible set for the allocation considered here are

$$C_E = EP\tau_\star(P,E), \qquad \mathcal{A}(C;P) = \{E \in \mathbb{N}_{\geq 1} : C_E \leq C\}. \tag{97}$$

The ensemble identity gives $V(\tau) \geq 0$, so $\mathcal{L}^*(P,E)$ is nonincreasing in E. When $\mathcal{A}(C;P)$ is nonempty, a largest affordable ensemble therefore attains the smallest loss in this set. At small E, feasibility is evaluated using stopping times obtained from the full loss curve. If the set is empty, no ensemble can reach its unconstrained minimum under this budget.

At large E, the costs implied by (95)–(96) are

$$C_E \asymp \begin{cases} P^{b+1}E^2, & a > b, \\ P^{b+1}E^{5/3}, & a = b, \\ P^{b+1}E, & a < b. \end{cases}$$

Consequently, as $C/P^{b+1} \to \infty$,

$$E_\star(C;P) \asymp \begin{cases} (C/P^{b+1})^{1/2}, & a > b, \\ (C/P^{b+1})^{3/5}, & a = b, \\ C/P^{b+1}, & a < b. \end{cases}$$

Integer rounding affects subleading terms in these large-budget laws. For the loss over the feasible set, write

$$\mathcal{L}^*(C;P) = \min_{E\in\mathcal{A}(C;P)} \mathcal{L}^*(P,E), \qquad \mathcal{L}_\infty(P) = \begin{cases} C_\infty, & a \geq b, \\ m(P), & a < b. \end{cases}$$

Then

$$\mathcal{L}^*(C;P) - \mathcal{L}_\infty(P) \asymp \begin{cases} P^{-(a-1)}(C/P^{b+1})^{-1/4}, & a > b, \\ P^{-(a-1)}(C/P^{b+1})^{-2/5}, & a = b, \\ P^{-(a-1)}(C/P^{b+1})^{-1}, & a < b. \end{cases} \tag{98}$$

In particular, constant squared teacher coefficients give an excess-loss compute exponent of $-2/5$ at fixed data. The $a < b$ rate assumes the positive minimum variance specified above.

Scope of the allocation. The large-E conclusions for $a \geq b$ use the differentiated late-time estimates along diverging stopping times. The branch $a < b$ uses convergence to the finite minimum of the full cross curve. Appendix H establishes these properties within the pure-power continuum GF equations. For a general discrete spectrum, the leading cross transient depends on the target coefficients throughout the spectrum. The exact cancellation giving the D/τ transient follows from constant squared teacher coefficients; matching the tail exponents alone is insufficient.

The dataset and member size remain fixed across all configurations in (5). This problem compares ensembles at their own minima. The broader problem

$$\min_{\tau\geq 0,\ E\in\mathbb{N}_{\geq 1}:\, EP\tau\leq C} \mathcal{L}(\tau,P,E), \qquad N = P \text{ fixed},$$

also permits stopping before an ensemble's unconstrained minimum. Appendix I.1 reduces this broader checkpoint optimization to a scalar problem at fixed width and solves its large-budget limit for constant squared teacher coefficients at $N = P$. It also gives a local comparison with changing width.

I.1 Width and Ensembling at Fixed Data

Fix a dataset size P and an initial member width $N_0 < P$. Two uses of an additional compute budget are to increase the width of one model, or to train more models of width N_0. Allowing each configuration to stop at any affordable time gives the comparison

$$\inf_{\substack{N\in\mathbb{N}\cap[N_0,P],\ \tau\geq 0\\ N\tau\leq C}} \mathcal{L}(\tau, N, P, 1) \qquad \text{and} \qquad \inf_{\substack{E\in\mathbb{N}_{\geq 1},\ \tau\geq 0\\ EN_0\tau\leq C}} \mathcal{L}(\tau, N_0, P, E). \tag{99}$$

Here the stopping time is optimized within the budget, whereas (97) compares configurations at their unconstrained loss minima.

Allocation at a finite compute budget. At fixed P, write $c(\tau, N) = C_{\text{cross}}(\tau, N)$ and $V(\tau, N) = C_{\text{self}}(\tau, N) - c(\tau, N) \geq 0$. For each width and checkpoint, loss decreases with ensemble size. The optimal integer ensemble at that checkpoint is therefore $\lfloor C/(N\tau)\rfloor$, giving the exact reduction

$$\inf_{\substack{E\in\mathbb{N}_{\geq 1},\ \tau>0\\ EN\tau\leq C}} \mathcal{L}(\tau, N, P, E) = \inf_{0<\tau\leq C/N} \left[c(\tau, N) + \frac{V(\tau, N)}{\lfloor C/(N\tau)\rfloor}\right]. \tag{100}$$

The initial loss is included through the limit $\tau \downarrow 0$. Allowing real $E \geq 1$ replaces the objective on the right by $c(\tau, N) + N\tau V(\tau, N)/C$. At an interior optimum,

$$\partial_\tau c + \frac{N}{C}(V + \tau\partial_\tau V) = 0, \qquad \partial_\tau \mathcal{L} = -\frac{V}{E\tau}.$$

Thus, when $V > 0$, the optimal checkpoint lies on a decreasing part of the fixed-E loss curve. This differs from training each ensemble to its own loss minimum. Equation (100) evaluated at $N = N_0$ gives the ensemble strategy in (99); the width strategy minimizes $c + V$ over the displayed feasible widths and checkpoints.

A local comparison of width and ensemble count. For a differentiable continuum width, compare $(N, E, \tau) \mapsto (fN, E, \tau/f)$ and $(N, E, \tau) \mapsto (N, fE, \tau/f)$ at $f = 1$. Both changes preserve $EN\tau = C$. Their loss derivatives are respectively $N\partial_N\mathcal{L} - \tau\partial_\tau\mathcal{L}$ and $-V/E - \tau\partial_\tau\mathcal{L}$. The width change has the smaller derivative precisely when

$$N\partial_N c + \frac{V + N\partial_N V}{E} < 0. \tag{101}$$

The reduction of training time is the same for both changes and cancels in this comparison. Width changes the cross correlation and the variance; ensembling changes only the variance at a fixed checkpoint. For integer widths and ensembles, the corresponding finite changes are compared directly using (99).

Bordelon et al. (2024, Appendix H.4) show that increasing width is at least as effective as ensembling in the infinite-data limit, under monotonicity conditions on their bias and variance terms. Equation (101) gives a local comparison at finite P, without imposing these monotonicity conditions.

Comparison with stopping each ensemble at its minimum. At $N = P$ with constant squared teacher coefficients, the late-time expansions give $c - C_\infty \sim D/\tau$ and $V \sim H\sqrt{\tau}$, with $D, H > 0$. Since $c(\tau) > C_\infty$ at each finite time, an optimal sequence approaching C_∞ has $\tau \to \infty$ and $E \to \infty$. The leading budget objective is therefore $D/\tau + (PH/C)\tau^{3/2}$. Its minimum satisfies

$$\begin{aligned} \tau_C &\sim \left(\frac{2DC}{3PH}\right)^{2/5}, \qquad E_C \sim \frac{C}{P\tau_C}, \\ \inf_{EP\tau\leq C} \mathcal{L} - C_\infty &\sim \frac{5D}{3}\left(\frac{3PH}{2DC}\right)^{2/5}. \end{aligned} \tag{102}$$

Here the infimum is over integer $E \geq 1$ and $\tau > 0$. The tail estimates bound the exact objective above and below by arbitrarily close multiples of the leading objective along minimizing sequences. Its unique minimum gives (102); rounding E_C changes the variance term by a relative $O(E_C^{-1})$. The compute exponent remains $-2/5$, while the optimal checkpoint and leading loss coefficient differ from the allocation in (98).

The following results determine the width dependence entering this comparison.

Cross correlations at arbitrary width. Use the noiseless finite Gaussian model of Appendix G, now with arbitrary N and $Q \in \mathbb{R}^{N\times M}$ having independent $\mathcal{N}(0, 1/N)$ entries. Keep $M > P$ and zero initialization. For a fixed full-row-rank dataset, write $G = XX^\top$, $U = X^\top G^{-1/2}$, and $\Pi = UU^\top$. Let $U_\perp$ be an orthonormal complement of U. The independent Gaussian blocks QU and $QU_\perp$ give

$$\mathbb{E}_Q[v_Q(\tau) \mid X] = \left[I - \Pi + UF_{N,\tau}(G)U^\top\right]w^\star,$$
$$F_{N,\tau}(G) = \mathbb{E}_Q e^{-\tau G^{1/2}(QU)^\top(QU)G^{1/2}/P}. \tag{103}$$

Indeed, $QU_\perp$ has zero conditional mean and is independent of the training predictions. Gaussian symmetry makes $F_{N,\tau}$ commute with G, giving the displayed row-space term. Since $0 \preceq F_{N,\tau} \preceq I$, the conditional mean is a Euclidean contraction and $C_{\text{cross}}(\tau) \le \lambda_{\max}(\Lambda)\|w^\star\|^2$ for every N and τ. For $(w_k^\star)^2 = q > 0$, coordinate-sign symmetry further gives

$$C_{\text{cross}}(\tau) = C_\infty + q\mathbb{E}_X \operatorname{tr}[\Lambda U F_{N,\tau}^2 U^\top], \qquad C_\infty = q\mathbb{E}_X \operatorname{tr}[\Lambda(I - \Pi)].$$

The matrices $F_{N,\tau}$ are simultaneously diagonal in an eigenbasis of G and decrease with τ, so cross is nonincreasing at every width. For $N \ge P$, $F_{N,\tau} \to 0$; for $N < P$, its limit is the mean projection onto the training kernel's null space. Thus finite width can also change the limiting cross correlation below interpolation.

Fully trained models above interpolation. For an arbitrary deterministic teacher and $N > P+1$, the exact finite-model loss is

$$\mathcal{L}(\infty, N, P, E) = C_\infty + \frac{K(P)}{E(N - P - 1)},$$
$$C_\infty = \mathbb{E}_X\|(I - \Pi)w^\star\|_\Lambda^2, \tag{104}$$
$$K(P) = \mathbb{E}_X\left[\operatorname{tr}[\Lambda(I - \Pi)]\|\Pi w^\star\|^2\right].$$

To see this, put $A = QU$, $B = QU_\perp$, and $u = U^\top w^\star$. The trained weights are $\theta_\infty = A(A^\top A)^{-1}u$, and $v_Q(\infty) = (I-\Pi)w^\star - U_\perp B^\top \theta_\infty$. Averaging over B gives the variance $\operatorname{tr}[\Lambda(I-\Pi)]\|\theta_\infty\|^2/N$. The Wishart inverse mean $\mathbb{E}_A(A^\top A)^{-1} = NI_P/(N - P - 1)$ and the ensemble identity then yield (104). If $(w_k^\star)^2 = q$, sign symmetry gives $K(P) = PC_\infty$. At $N \gg P$, the converged variance is consequently $K(P)/(NE)\,[1 + O((P + 1)/N)]$.

Finite training time. The cross correlation also depends on width before convergence. Fix $P, M, \tau > 0$ and put $K_0 = G/P$. Expanding the exponential in (103) in the centered Wishart fluctuation gives

$$F_{N,\tau} = e^{-\tau K_0} + N^{-1}F_1(\tau) + o(N^{-1}),$$
$$F_1(\tau) = \int_0^\tau (\tau - v)e^{-(\tau-v)K_0}\left[K_0^2 e^{-vK_0} + K_0 \operatorname{tr}(K_0 e^{-vK_0})\right] dv.$$

The two terms follow from $N\mathbb{E}[\Delta_{ij}\Delta_{kl}] = (K_0)_{ik}(K_0)_{jl} + (K_0)_{il}(K_0)_{jk}$ for $\Delta = G^{1/2}A^\top AG^{1/2}/P - K_0$. The third-order remainder is bounded by a fixed-time constant times $\mathbb{E}\|\Delta\|^3 = O(N^{-3/2})$; finite Gaussian data moments also permit averaging over X. Every eigenvalue of $F_1(\tau)$ is positive. For constant squared teacher coefficients,

$$C_{\text{cross}}(\tau, N) = C_{\text{cross}}(\tau, \infty) + \frac{A(\tau)}{N} + o(N^{-1}), \qquad A(\tau) = 2q\mathbb{E}_X \operatorname{tr}[\Lambda U e^{-\tau K_0}F_1(\tau)U^\top] > 0.$$

This correction remains when $E \to \infty$. Thus the equivalence of width and ensemble count in the converged variance does not extend directly to the finite-time cross correlation.

Stationary correlations in the GF closure. For positive Laplace variables z, z', set $h = z/(R_1R_3)$ and $h' = h(z')$. With $S(h) = \sum_k \lambda_k/(\lambda_k + h)$, the general-width response equations read

$$R_1 = 1 - \frac{S(h)}{P}, \qquad R_3 = 1 - \frac{S(h)}{N}, \qquad z = h\left(1 - \frac{S(h)}{P}\right)\left(1 - \frac{S(h)}{N}\right). \tag{105}$$

Define the spectral functions

$$(I_1, I_2, U, W) = \sum_k \frac{\left(\lambda_k, \lambda_k^2, \lambda_k (w_k^\star)^2, \lambda_k^2 (w_k^\star)^2\right)}{(\lambda_k + h)(\lambda_k + h')}.$$

Here U denotes a scalar spectral function. Solving the self system in (49) and (50), together with the GF limit of (40), gives

$$\begin{aligned} \widehat{C}_{\mathrm{cross}}(z, z') &= \frac{hh'}{zz'} \frac{U}{1 - I_2/P}, \\ \widehat{C}_{\mathrm{self}}(z, z') &= \frac{hh'}{zz'} \frac{(1 - I_2/N)U + I_1 W/N}{(1 - I_2/P)(1 - I_2/N) - hh' I_1^2/(NP)}. \end{aligned} \tag{106}$$

For the stationary branch at $N/P \neq 1$, $h \to h_0 > 0$ with $S(h_0) = m = \min(N, P)$. Writing $d = m - I_2(h_0, h_0) = h_0 I_1(h_0, h_0) > 0$, the self denominator tends to $(1 - m/\max(N, P))d/m$. The stationary values follow from the $1/(zz')$ residues, assuming the final-value limit for this branch.

For the pure-power continuum spectrum of Appendix H, put $r = 1/b$, $p = (a-1)/b$, $c_b = \pi/[b \sin(\pi/b)]$, and $K = \pi/[b \sin(\pi p)]$. Direct integration gives

$$S(h) = c_b h^{-r}, \quad h_0 = (c_b/m)^b, \quad I_2 = (1 - r)m, \quad U = K(1 - p)h_0^{p-2}, \quad W = Kph_0^{p-1}.$$

Substitution in (106) yields

$$\begin{aligned} N < P: \qquad C_{\mathrm{self}}(\infty) &= \frac{K(c_b/N)^{a-1}}{1 - N/P}, \\ C_{\mathrm{cross}}(\infty) &= \frac{K(1-p)(c_b/N)^{a-1}}{1 - (1-r)N/P}, \\ N > P: \qquad C_{\mathrm{cross}}(\infty) &= \frac{K(1-p)}{r}(c_b/P)^{a-1}, \\ C_{\mathrm{self}}(\infty) - C_{\mathrm{cross}}(\infty) &= K(c_b/P)^{a-1} \frac{P}{N - P}. \end{aligned} \tag{107}$$

The denominator $N - P$ in the continuum variance agrees to leading order with $N - P - 1$ in the exact finite-model expression at fixed $N/P > 1$. For $a = b = 2$ and $\nu = N/P < 1$, the ensemble plateau simplifies to

$$\mathcal{L}(\infty, N, P, E) = \frac{\pi^2}{4P}\left[\frac{1}{\nu(2 - \nu)} + \frac{1}{E\nu(1 - \nu)(2 - \nu)}\right]. \tag{108}$$

Increasing width changes both terms, while increasing E reduces the second. The finite-budget comparison in (99) uses the full trajectories to account for early stopping near interpolation.

Full training curves below interpolation. For noiseless GF with zero initialization, take $a = b = 2$, $\lambda_k = k^{-2}$, and $(w_k^\star)^2 = 1$, with spectral sums replaced by integrals over $k \in (0, \infty)$. Define

$$x = \tau/P^2, \qquad \nu = N/P \in (0, 1], \qquad c_2 = \pi/2, \qquad d = \frac{c_2}{2}(\nu^{-1} - 1), \qquad A = c_2 + d.$$

Let $S_\nu(x, x') = PC_{\mathrm{self}}(P^2 x, P^2 x')$ and $X_\nu(x, x') = PC_{\mathrm{cross}}(P^2 x, P^2 x')$ at $N = \nu P$, with $S_\nu(x) = S_\nu(x, x)$ and $X_\nu(x) = X_\nu(x, x)$. The normalized ensemble loss is $P\mathcal{L}(P^2 x, \nu P, P, E) = X_\nu(x) + [S_\nu(x) - X_\nu(x)]/E$. In the response equation, $w = P\sqrt{h}$ and $\zeta = P^2 z$ satisfy

$$\zeta = (w - c_2)(w - c_2/\nu), \qquad w = A + u, \qquad u = \sqrt{\zeta + d^2}.$$

Equation (106) then gives the two-time Laplace transforms of the normalized correlations as

$$\widehat{S}_\nu(\zeta, \zeta') = \frac{c_2 w w'}{\zeta\zeta'(u + u')}, \qquad \widehat{X}_\nu(\zeta, \zeta') = \frac{c_2 w w'}{\zeta\zeta'(u + u' + c_2/\nu)}. \tag{109}$$

Here the two time variables are both scaled by P^2. Define

$$\begin{aligned} F(r, x) &= \frac{1}{2}\left[e^{-dr}\,\mathrm{erfc}\left(\frac{r}{2\sqrt{x}} - d\sqrt{x}\right) + e^{dr}\,\mathrm{erfc}\left(\frac{r}{2\sqrt{x}} + d\sqrt{x}\right)\right], \\ f(r, x) &= AF(r, x) - \partial_r F(r, x). \end{aligned} \tag{110}$$

Since $\widehat{F}(r,\zeta) = e^{-r\sqrt{\zeta+d^2}}/\zeta$, factorizing the denominators in (109) yields the positive integrals

$$S_\nu(x) = c_2 \int_0^\infty f(r,x)^2\, dr, \qquad X_\nu(x) = c_2 \int_0^\infty e^{-c_2 r/\nu} f(r,x)^2\, dr. \tag{111}$$

These formulas hold for every $x > 0$ in the continuum closure, including $\nu = 1$. For $\nu < 1$, $f(r,x) \to (c_2/\nu)e^{-dr}$ as $x \to \infty$, recovering (108).

Theorem 4 (Ensembling below interpolation)**.** *Under the preceding continuum assumptions, every single model with $0 < N \le P$ satisfies, at every positive training time,*

$$\mathcal{L}(\tau, N, P, 1) \ge \frac{c_2^2}{P}(1 + P/N) \ge \frac{\pi^2}{2P}.$$

Fix $N_0 = \nu_0 P$ with $1 - 1/\sqrt{2} < \nu_0 < 1$. Any integer ensemble count satisfying

$$E > \frac{1}{(1-\nu_0)[2\nu_0(2-\nu_0) - 1]} \tag{112}$$

achieves strictly lower loss at a finite training time than every single model with $N \in [N_0, P]$, allowing each single model to choose its stopping time. Consequently, the ensemble strategy in (99) *has strictly smaller loss for all sufficiently large compute budgets. The sufficient counts are $E \ge 5$ at $N_0 = P/2$ and $E \ge 14$ at $N_0 = P/3$.*

Proof. At fixed $x > 0$, $F(0,x) = 1$ and $F(\infty,x) = 0$. With $J = \int_0^\infty F^2\, dr$ and $K = \int_0^\infty (\partial_r F)^2\, dr$, integration by parts gives $\int F \partial_r F\, dr = -1/2$, hence $JK \ge 1/4$. Therefore

$$S_\nu(x) = c_2(A + A^2 J + K) \ge 2c_2 A = c_2^2(1 + \nu^{-1}).$$

This proves the bound uniformly over training time and width. At the fixed starting width, the ensemble plateau is

$$\lim_{\tau\to\infty} P\mathcal{L}(\tau, N_0, P, E) = \frac{c_2^2}{\nu_0(2-\nu_0)}\left[1 + \frac{1}{E(1-\nu_0)}\right].$$

Condition (112) places this limit strictly below $2c_2^2$. Convergence therefore gives a finite τ_0 at which the ensemble beats the single-model bound. The same checkpoint is feasible for every $C \ge EN_0\tau_0$. □

Large budgets at a fixed width below interpolation. Keep $\nu \in (0,1)$ fixed and set $Q = C/P^3$. To optimize the stopping time within the budget, define the normalized loss

$$J_\nu(Q) = \inf_{\substack{E\in\mathbb{N}_{\ge 1},\ x>0 \\ E\nu x \le Q}} \left\{ X_\nu(x) + \frac{S_\nu(x) - X_\nu(x)}{E} \right\}.$$

Thus $J_\nu(Q)/P$ uses any affordable checkpoint, as in (99). Its large-budget asymptotics are

$$x_\star(Q) = \frac{1}{d^2}\left[\log Q - \frac{3}{2}\log\log Q + O(1)\right], \qquad E_\star(Q) \sim \frac{d^2 Q}{\nu \log Q},$$
$$J_\nu(Q) - \frac{c_2^2}{\nu(2-\nu)} \sim \frac{4\nu^2}{(1-\nu)^3(2-\nu)}\frac{\log Q}{Q}. \tag{113}$$

The floor $c_2^2/[P\nu(2-\nu)]$ exceeds the critical-width floor c_2^2/P. These asymptotics hold at fixed $\nu < 1$; the gap d^2 vanishes as $\nu \to 1$.

To obtain (113), first note that $X_\nu(x)$ strictly decreases to its plateau. Indeed, $\partial_x F > 0$ for $r > 0$, $\partial_x F = \partial_r^2 F - d^2 F$, and integration by parts gives

$$X_\nu'(x) = 2c_2 \int_0^\infty e^{-c_2 r/\nu}(\partial_r f - df)\partial_x F\, dr < 0.$$

The sign follows from $\partial_r f - df = (c_2/\nu)(\partial_r F - dF) - \partial_x F$ and $\partial_r F - dF < 0$, directly from (110). Expansion of $\widehat{f} = (A + \sqrt{\zeta + d^2})e^{-r\sqrt{\zeta+d^2}}/\zeta$ at $\zeta = -d^2$ gives

$$f(r,x) - \frac{c_2}{\nu}e^{-dr} \sim \frac{1 - Ar}{2d^2\sqrt{\pi}}\, x^{-3/2} e^{-d^2 x}.$$

Integrating the leading term in (111), and differentiating the same inverse transform, yields

$$X_\nu(x) - \frac{c_2^2}{\nu(2-\nu)} \sim K_\nu x^{-3/2} e^{-d^2 x}, \qquad X'_\nu(x) \sim -d^2 K_\nu x^{-3/2} e^{-d^2 x},$$
$$K_\nu = \frac{2c_2^2}{\nu d\sqrt{\pi}(d + c_2/\nu)^2} > 0, \qquad V_\nu(\infty) = \frac{c_2^2}{\nu(1-\nu)(2-\nu)},$$

Here $V_\nu = S_\nu - X_\nu$; the same expansion gives $V'_\nu(x) = O(x^{-3/2} e^{-d^2 x})$, hence $xV'_\nu(x) \to 0$. Taking x of order $\log Q$ and E of order $Q/\log Q$ approaches the cross plateau. Strict decrease of X_ν and positivity of V_ν then force any globally optimal sequence to satisfy $x_\star \to \infty$ and $E_\star \to \infty$. For real E, eliminating it gives $X_\nu(x) + \nu x V_\nu(x)/Q$. Its stationary equation becomes

$$d^2 K_\nu x^{-3/2} e^{-d^2 x} \sim \frac{\nu V_\nu(\infty)}{Q},$$

which gives the first line of (113). The variance term dominates the excess loss at this optimum, with coefficient $\nu V_\nu(\infty)/d^2 = 4\nu^2/[(1-\nu)^3(2-\nu)]$. At late times the fixed-E loss is decreasing near the optimum, so the budget is saturated. Adjacent integer choices of E give checkpoints $x = Q/(\nu E)$ separated by $O((\log Q)^2/Q) = o(1)$ there, preserving these asymptotics.

I.2 Discrete Spectra at Interpolation

The critical coefficient can also be evaluated without replacing the spectral sums by integrals. Set $N = P$, $\sigma = 0$, and $(w_k^\star)^2 = q > 0$. Take a finite positive spectrum with $M > P$, or a summable positive infinite spectrum. Let h_0 be the unique positive solution of

$$S(h_0) = \sum_k \frac{\lambda_k}{\lambda_k + h_0} = P.$$

In (106), $U = qI_1$ and $W = qI_2$, so the self numerator reduces to qI_1. The denominator factors as

$$\frac{(P - I_2 - \sqrt{hh'}I_1)(P - I_2 + \sqrt{hh'}I_1)}{P^2}.$$

The identities $S(h) = I_2 + h'I_1$ and $S(h') = I_2 + hI_1$ give

$$P - I_2 - \sqrt{hh'}I_1 = \frac{P}{2}(R(z) + R(z')) + \frac{1}{2}(\sqrt{h} - \sqrt{h'})^2 I_1.$$

Since $R(z) = \sqrt{z/h(z)}$ and $h(z) = h_0 + O(\sqrt{z})$, the joint low-frequency expansion is

$$\widehat{C}_{\text{self}}(z, z') = \frac{qPh_0^{3/2}}{zz'(\sqrt{z} + \sqrt{z'})}\big[1 + O(\sqrt{z} + \sqrt{z'})\big].$$

Under the same contour-inversion bounds used in Lemma 2, (76) therefore gives the discrete-spectrum GF coefficient

$$C_{\text{self}}(\tau) = H\sqrt{\tau} + o(\sqrt{\tau}), \qquad H = \frac{2(2-\sqrt{2})}{\sqrt{\pi}}\, qPh_0^{3/2}.$$

For $\lambda_k = k^{-2}$ with infinite cutoff, the sum is explicit:

$$S(h) = \frac{\pi}{2\sqrt{h}} \coth\!\left(\frac{\pi}{\sqrt{h}}\right) - \frac{1}{2}.$$

Consequently, relative to the continuum coefficient $H_{\text{cont}} = gqP[\pi/(2P)]^3$, the coefficient with cutoff M satisfies, as $P \to \infty$ and $M/P \to \infty$,

$$\frac{H_M}{H_{\text{cont}}} = \left(\frac{P}{P + 1/2}\right)^3 \left[1 - \frac{12(P + 1/2)}{\pi^2 M} + O\!\left(\frac{P^2}{M^2} + e^{-4P}\right)\right].$$

Here $g = 2(2-\sqrt{2})/\sqrt{\pi}$. The prefactor accounts for the discrete spectrum, and the M^{-1} term accounts for the feature cutoff. To obtain the latter, subtract $\sum_{k>M}(1 + hk^2)^{-1} = 1/(hM) + O(1/(hM^2) + 1/(h^2M^3))$ from the infinite sum and expand its root. These are coefficients of the GF

closure with discrete spectral sums. The exact finite Gaussian-model averages are given separately in (83) and (84). Comparisons over the stopping window use the time-dependent correlations.

Cross coefficient for the same spectrum. Writing $F(h,h') = zz'\widehat{C}_{\mathrm{cross}}(z,z')$, the identities above give

$$F(h,h_0) = F(h_0,h) = qPh_0, \qquad F_{hh'}(h_0,h_0) = \frac{qP}{h_0}\left(1 + \frac{h_0 S''(h_0)}{2S'(h_0)}\right).$$

Using $h(z) - h_0 = -P\sqrt{z}/[\sqrt{h_0}S'(h_0)] + O(z)$, the mixed term of F is $\pi D\sqrt{zz'}$, with

$$D = \frac{qP^3}{\pi h_0^2[S'(h_0)]^2}\left(1 + \frac{h_0 S''(h_0)}{2S'(h_0)}\right) > 0.$$

The bracket equals $\sum_k \lambda_k^2/(\lambda_k + h_0)^3 \big/ \sum_k \lambda_k/(\lambda_k + h_0)^2 > 0$. With the corresponding contour and differentiated remainder bounds, this is the coefficient of τ^{-1} in $C_{\mathrm{cross}}(\tau) - qPh_0$ and of $-\tau^{-2}$ in its derivative. Discreteness therefore changes both D and H in the late-time balance $\tau_\star \sim [2D(E-1)/H]^{2/3}$.

Time-dependent corrections for $a = b = 2$. The same spectral sums determine the correlations throughout the stopping window. Retain $\lambda_k = k^{-2}$ and $(w_k^\star)^2 = q$, and write

$$x = \frac{\tau}{P^2}, \qquad \zeta = P^2 z, \qquad w = P\sqrt{h(z)}, \qquad c_2 = \frac{\pi}{2}, \qquad \alpha = 1 + \frac{1}{2P}.$$

Here $w' = P\sqrt{h(z')}$ denotes the response at the second transform argument. The exact infinite-spectrum sum above becomes

$$\frac{S(h)}{P} = \frac{c_2}{w}\coth\!\left(\frac{\pi P}{w}\right) - \frac{1}{2P}, \qquad \sqrt{\zeta} = w\left(1 - \frac{S(h)}{P}\right).$$

To retain the leading discrete-spectrum correction, replace the hyperbolic cotangent by one. We use this expansion in stopping windows where the omitted term is small on the inverse Laplace contours. The response then satisfies

$$w \simeq \frac{c_2 + \sqrt{\zeta}}{\alpha}.$$

All square roots use the physical branch continued from positive real z. The identities $I_1 = [S(h) - S(h')]/(h' - h)$ and $I_2 = S(h) - h'I_1$ give, within the same approximation,

$$\frac{I_1}{P^3} \simeq \frac{c_2}{ww'(w+w')}, \qquad \frac{I_2}{P} \simeq \frac{c_2}{w+w'} - \frac{1}{2P}.$$

The coincident-argument values follow by continuity. Substituting these expressions into (106) and inverting gives

$$\begin{aligned} C_{\mathrm{self}}^{\mathrm{disc}}(P^2x) &\simeq \alpha^{-3} C_{\mathrm{self}}^{\mathrm{cont}}(P^2x), \\ C_{\mathrm{cross}}^{\mathrm{disc}}(P^2x) &\simeq \alpha^{-2} C_{\mathrm{cross}}^{\mathrm{cont}}(P^2x). \end{aligned} \tag{114}$$

The continuum curves are those in (90), multiplied by q. Thus the leading correction changes the self and cross amplitudes by different factors, without rescaling x.

For the ensemble loss, (114) is equivalent to

$$\mathcal{L}_E^{\mathrm{disc}}(P^2x) \simeq \alpha^{-3}\frac{E_{\mathrm{eff}}}{E}\mathcal{L}_{E_{\mathrm{eff}}}^{\mathrm{cont}}(P^2x), \qquad E_{\mathrm{eff}} = 1 + \alpha(E-1).$$

Here the continuum ensemble formula is evaluated at the real argument E_{eff}; the actual member count remains the integer E. The prefactor is independent of time, so the stopping time follows from the continuum curve at this effective count. In particular, the large-E exponent remains $2/3$, with a modified coefficient.

Finite spectral cutoff. For $M/P \gg 1$, let $\delta = P/M$. Subtracting the leading omitted tail from the same infinite sum gives

$$\frac{S_M(h)}{P} \simeq \frac{c_2}{w} - \frac{1}{2P} - \frac{\delta}{w^2}, \qquad \alpha w^2 - (c_2 + \sqrt{\zeta})w + \delta \simeq 0. \tag{115}$$

The physical solution is the larger root on the positive real axis, continued analytically from $\delta = 0$:

$$w \simeq \frac{c_2 + \sqrt{\zeta} + \sqrt{(c_2 + \sqrt{\zeta})^2 - 4\alpha\delta}}{2\alpha}.$$

The associated spectral functions must be corrected together:

$$\frac{I_1}{P^3} \simeq \frac{c_2}{ww'(w+w')} - \frac{\delta}{w^2 w'^2}, \qquad \frac{I_2}{P} \simeq \frac{c_2}{w+w'} - \frac{1}{2P}. \tag{116}$$

The leading cutoff term cancels explicitly from I_2, while its response arguments still depend on δ. Equations (115)–(116) can therefore be inserted into the full self and cross transforms to approximate the stopping window. Unlike the infinite-cutoff correction in (114), this correction is not a constant rescaling of the continuum curves. Its low-frequency limit recovers the leading cutoff corrections to H and D above.

I.3 Scaling with Model and Data Size

Consider the noiseless pure-power continuum spectrum $\lambda_k = k^{-b}$, $(w_k^\star)^2 = k^{b-a}$, integrated over $k \in (0, \infty)$, with $b > 1$ and $1 < a < b + 1$. The general GF equations give a common scaling form at each fixed $\nu = N/P > 0$. To derive it, set

$$k = Pu, \qquad h = P^{-b}y, \qquad z = P^{-b}\zeta.$$

Then $S(h)/P = c_b y^{-1/b}$, and the response equation (105) becomes $\zeta = y(1 - c_b y^{-1/b})(1 - c_b y^{-1/b}/\nu)$. Its physical branch is independent of P. The spectral functions in (106) scale as

$$I_2 = P\widetilde{I}_2, \qquad I_1 = P^{b+1}\widetilde{I}_1, \qquad U = P^{2b-a+1}\widetilde{U}, \qquad W = P^{b-a+1}\widetilde{W},$$

where the functions with tildes depend only on y, y' and the exponents. Both feedback denominators are dimensionless, while the numerator after factoring out $1/(zz')$ scales exactly as P^{1-a}. Thus either correlation transform has the exact scaling $\widehat{C}(z, z'; P, \nu) = P^{2b+1-a}\widehat{f}(\zeta, \zeta'; \nu)$. On the physical branch with well-defined inverse transforms, rescaling the two Laplace contours and applying the ensemble identity gives, at positive times,

$$\mathcal{L}(\tau, N, P, E) = P^{-(a-1)} F\left(\frac{\tau}{P^b}, \frac{N}{P}, E\right).$$

This identity holds within the pure-power continuum closure. For the discrete model, the continuum approximation predicts collapse of $P^{a-1}\mathcal{L}$ against τ/P^b at fixed $N/P, E$; its accuracy depends on the discrete spectrum and feature cutoff. The associated compute variable is

$$\frac{C}{P^{b+1}} = E\nu\frac{\tau}{P^b}.$$

The scaling form fixes these units; determining the optimal allocation across ratios requires the time dependence of F.

At $N = P$, the pure-power continuum calculation does supply a specialization. Put $x = \tau/P^b$, $u = c_b^{b/2}\sqrt{x}$, and $f_\infty = P^{a-1}C_\infty$, which is independent of P. Then

$$F(x, 1, E) = \left(1 - \frac{1}{E}\right) f_\infty \left[1 + \frac{j}{u} + \frac{d}{u^2} + O(u^{-3})\right] + \frac{f_\infty}{E}\left[\frac{g}{b-a+1}u + O(1)\right], \qquad u \to \infty.$$

The u^{-1} term vanishes for constant squared teacher coefficients, $a = b$. For $a < b$, this late expansion does not locate the finite cross minimum unless that minimum itself lies at large u.

At fixed $N/P \neq 1$, the positive low-frequency determinant in (64) cuts off the critical growth on the regular stationary branch, giving the plateaus in (107). The limit $x \to \infty$ therefore depends on the ratio: the critical $\sqrt{x}/E$ term is specific to $\nu = 1$. For $a = b = 2$ and $0 < \nu \leq 1$, the full continuum curves in (111) give the joint time and aspect-ratio dependence of F at every $x > 0$, including $\nu = 1$. A uniform near-interpolation expansion for general exponents or for the approach from $\nu > 1$ requires the corresponding joint analysis.

Matching the abundant-data and interpolation scales. Within the single-model approximation (93), the optimal width reaches order P at the same compute scale at which its training time reaches order P^b:

$$N_\star(C) \asymp P, \qquad \tau_\star(C) \asymp P^b.$$

Either matching condition gives

$$C \asymp P^{b+1}.$$

For $C \ll P^{b+1}$, the same abundant-data approximation gives $N_\star \ll P$ and $\tau_\star \ll P^b$. The matching estimate uses (93); locating transitions in the joint optimum over N, E, τ requires the full width- and time-dependent loss.

In unnormalized compute units, the matching cost is of order $(B/\eta)P^{b+1}$, after order P^b/η updates. One pass through the data instead takes P/B updates, or effective time $\eta P/B$. The abundant-data approximation treats the effect of reuse on the leading loss as negligible over the training times considered.

The limit of large ensembles at fixed data. For the continuum GF loss at fixed $N = P$,

$$\inf_{E \in \mathbb{N}_{\geq 1}} \mathcal{L}^*(P, E) = \inf_{\tau > 0} C_{\text{cross}}(\tau) = \mathcal{L}_\infty(P) \asymp P^{-(a-1)}.$$

The first equality follows from $V \geq 0$ and convergence of $\mathcal{L}(\tau, P, E)$ to cross at every fixed positive time. For $a \geq b$, the infimum of cross is its long-time limit C_∞; for $a < b$, it is the smaller value $m(P)$ attained at finite time.

To quantify the approach to this limit, consider a loss within $\epsilon P^{-(a-1)}$ of $\mathcal{L}_\infty(P)$, where $0 < \epsilon \ll 1$. Equation (98) gives

$$C_\epsilon \asymp \begin{cases} P^{b+1}\epsilon^{-4}, & a > b, \\ P^{b+1}\epsilon^{-5/2}, & a = b, \\ P^{b+1}\epsilon^{-1}, & a < b. \end{cases}$$

Because $\mathcal{L}_\infty(P) \asymp P^{-(a-1)}$, requiring a relative excess of at most ϵ gives the same exponents. The $a < b$ tolerance cost assumes $\min_{\tau \in \mathcal{S}(P)} V(\tau) > 0$. The compute needed to approach the fixed-width limit thus depends on the desired accuracy.

I.4 Effect of Weak Label Noise

Under the conditions of Appendix F.3, the signal source has order $P^{-(a-1)}$, whereas the additional source from labels has order σ^2. Their ratio is therefore of order

$$\sigma^2 P^{a-1}.$$

In particular,

$$\sigma^2 P^{a-1} \ll 1$$

means that the label-noise source is small relative to the signal source.

Stability of the noiseless stopping time additionally requires an isolated minimum and control of the noise-induced change in the loss derivative near that minimum, including the derivative of the cross contribution. A noisy allocation law requires these controls uniformly along the ensemble sizes and stopping times being compared. For constant squared teacher coefficients, noise can change the leading cross transient because the noiseless $\tau^{-1/2}$ coefficient cancels. The additive irreducible test noise σ^2 has zero time derivative, so the stopping condition depends on the noise-induced changes in the learned predictor.

J Proof of the Finite-Horizon DMFT Error Estimate

We first prove Lemma 1 and identify the population limit of the closure. We then compare the finite microscopic trajectory with that limit to prove Theorem 1.

J.1 CAUSAL RECURSION AND THE POPULATION LIMIT

Use the ordered indices $i, j \in \{1, \dots, S\}$. The response matrices $\boldsymbol{H}_k, \boldsymbol{R}_1, \boldsymbol{R}_3$ are lower triangular with unit diagonal. Every matrix to be inverted in (19)–(20) has the form $\boldsymbol{I} + \boldsymbol{K}$, where $\boldsymbol{K}$ is strictly lower triangular. Hence

$$(\boldsymbol{I} + \boldsymbol{K})^{-1} = \sum_{m=0}^{S-1} (-\boldsymbol{K})^m. \tag{117}$$

At lag $i - j = d > 0$, each feedback term contains a factor $\boldsymbol{\Theta}$ of strictly positive lag. Its other response entries therefore have lag less than d. The response equations can be solved successively in d, including strictly lower-triangular additive residuals.

Once the responses are determined, order the correlation entries by $i + j$. At a fixed pair (i, j), solve them in the order C_0, C_1, C_2, C_3. Every feedback term in the C_0 equation is sandwiched between $\boldsymbol{H}_k \boldsymbol{\Theta} \boldsymbol{R}_3$ and its transpose, so it uses only indices $p < i, q < j$. The C_1 equation can then use $C_{0,ij}$. The self dependence in C_2 again has a strictly lower triangular factor on both sides, while its same-pair dependence on $C_{1,ij}$ is already known. Finally, C_3 follows from C_2. Symmetric residuals preserve symmetry throughout this recursion.

This construction proves uniqueness of the causal solution. Every step uses finite sums and products, including the inverse (117), so the solution is polynomial in the residual entries and $1/N, 1/B$. Its derivatives are bounded on a compact neighborhood of zero residual and $(1/N, 1/B) \in [0, 1]^2$. The mean value inequality gives (38), proving the lemma.

Set $1/N = 1/B = 0$. The response solution becomes

$$\boldsymbol{R}_1 = \boldsymbol{R}_3 = \boldsymbol{I}, \qquad \boldsymbol{H}_k = (\boldsymbol{I} + \eta \lambda_k \boldsymbol{\Theta})^{-1}, \qquad (\boldsymbol{H}_k \mathbf{1})_{s+1} = (1 - \eta \lambda_k)^s.$$

Substitution into (22) gives the correlation of $\bar{\boldsymbol{v}}_s$ in (34). Polynomial dependence on $1/N, 1/B$ therefore implies

$$\max_{0 \le s < S} |\mathcal{L}_{\mathrm{DMFT}}(s+1) - \mathcal{L}_{\mathrm{pop}}(s)| \le K_{M,S} \left(\frac{1}{N} + \frac{1}{B} \right). \tag{118}$$

Here and below, $K_{M,S}$ denotes a finite constant depending on the fixed parameters, whose value may change between inequalities.

J.2 CONCENTRATION OF THE FIXED TRAINING MATRICES

Write

$$\boldsymbol{W} = \frac{\boldsymbol{A}^\top \boldsymbol{A}}{N}, \qquad \widehat{\boldsymbol{\Sigma}}_t = \frac{\boldsymbol{\Psi}(t)^\top \boldsymbol{\Psi}(t)}{B}, \qquad \boldsymbol{g}_t = \frac{\boldsymbol{\Psi}(t)^\top \boldsymbol{\epsilon}(t)}{B}.$$

Let $u_N = \sqrt{(M+h)/N}$ and $u_B = \sqrt{(M+h)/B}$, with h as in (35). Gaussian sample-covariance bounds and a union bound over the T distinct minibatches give, with probability at least $1 - \delta$,

$$\begin{aligned} d_N &:= \|\boldsymbol{W} - \boldsymbol{I}\|_{\mathrm{op}} \le c(u_N + u_N^2), \\ d_B &:= \max_{t<T} \|\widehat{\boldsymbol{\Sigma}}_t - \boldsymbol{\Lambda}\|_{\mathrm{op}} \le c\|\boldsymbol{\Lambda}\|_{\mathrm{op}}(u_B + u_B^2), \\ g &:= \max_{t<T} \|\boldsymbol{g}_t\|_2 \le \sqrt{\frac{\|\boldsymbol{\Lambda}\|_{\mathrm{op}} + d_B}{B}} \left(\sqrt{M} + \sqrt{2h} \right). \end{aligned} \tag{119}$$

For the last inequality, condition on the feature matrices and use $\boldsymbol{g}_t \sim \mathcal{N}(\mathbf{0}, \widehat{\boldsymbol{\Sigma}}_t / B)$. The universal constant c absorbs the Gaussian tail constants. All subsequent estimates hold on this event. Repeated uses of a minibatch involve the same matrices, so the event controls the whole trajectory.

Choose c_0 in the theorem small enough that $\varepsilon = u_N + u_B \le c_0$ implies

$$d_N < \frac{1}{2}, \qquad \eta(1 + d_N)(\|\boldsymbol{\Lambda}\|_{\mathrm{op}} + d_B) \le 1. \tag{120}$$

The strict margin ρ ensures this choice is possible. Also set

$$\Delta := \max_{t<T} \|\boldsymbol{W} \widehat{\boldsymbol{\Sigma}}_t - \boldsymbol{\Lambda}\|_{\mathrm{op}} \le (1 + d_N) d_B + d_N \|\boldsymbol{\Lambda}\|_{\mathrm{op}}. \tag{121}$$

J.3 TRAJECTORY AND LOSS COMPARISON

Flatten the microscopic residual to $\boldsymbol{v}_s = \boldsymbol{v}_n^0(t)$, where $s = nT + t$ and $t_s = s \bmod T$. Its exact recursion is

$$\boldsymbol{v}_{s+1} = (\boldsymbol{I} - \eta \boldsymbol{W} \widehat{\boldsymbol{\Sigma}}_{t_s}) \boldsymbol{v}_s - \eta\sigma \boldsymbol{W} \boldsymbol{g}_{t_s}, \qquad \boldsymbol{v}_0 = \boldsymbol{w}^\star.$$

Since $d_N < 1/2$, $\boldsymbol{W}$ is positive definite. In coordinates $\boldsymbol{z}_s = \boldsymbol{W}^{-1/2}\boldsymbol{v}_s$, the homogeneous update is $\boldsymbol{I} - \eta \boldsymbol{W}^{1/2} \widehat{\boldsymbol{\Sigma}}_{t_s} \boldsymbol{W}^{1/2}$, a positive semidefinite contraction by (120). It follows that

$$\|\boldsymbol{v}_s\|_2 \le V_0 + \eta\sigma s(1 + d_N)g, \qquad V_0 = \sqrt{\frac{1 + d_N}{1 - d_N}} \|\boldsymbol{w}^\star\|_2. \tag{122}$$

Let $\boldsymbol{e}_s = \boldsymbol{v}_s - \bar{\boldsymbol{v}}_s$. Subtracting the population recursion gives

$$\boldsymbol{e}_{s+1} = (\boldsymbol{I} - \eta \boldsymbol{\Lambda}) \boldsymbol{e}_s - \eta (\boldsymbol{W} \widehat{\boldsymbol{\Sigma}}_{t_s} - \boldsymbol{\Lambda}) \boldsymbol{v}_s - \eta\sigma \boldsymbol{W} \boldsymbol{g}_{t_s}.$$

The population propagator also has norm at most one. Summing the forcing terms and using (122) yields

$$\max_{0 \le s < S} \|\boldsymbol{e}_s\|_2 \le Q_S, \qquad Q_S = \eta S \Delta V_0 + \eta\sigma S(1 + d_N)g \left(1 + \frac{\eta S \Delta}{2}\right).$$

Equations (119)–(121) give $Q_S \le K_{M,S}\varepsilon$ when $\varepsilon \le c_0$. The test-loss difference therefore satisfies

$$\begin{aligned} \max_{0 \le s < S} |\mathcal{L}_{N,B}(s+1) - \mathcal{L}_{\mathrm{pop}}(s)| &\le \|\boldsymbol{\Lambda}\|_{\mathrm{op}} \left(2\|\boldsymbol{w}^\star\|_2 Q_S + Q_S^2\right) \\ &\le K_{M,S}\varepsilon. \end{aligned} \tag{123}$$

The additive test-noise variance cancels in this difference. Combining (123) with (118) proves (36), since $N^{-1} + B^{-1}$ is bounded by a constant times ε in the stated range.

If $\ell_S > 0$, reduce c_0 further so that (118) is at most $\ell_S/2$. Then $\mathcal{L}_{\mathrm{DMFT}}(\alpha) \ge \ell_S/2$ for every $\alpha \le S$. Dividing the absolute error bound by this lower bound gives (37) and completes the proof.

The argument bounds absolute matrix and trajectory errors and normalizes only the final loss. The constants depend on the fixed spectrum and horizon.

K NUMERICAL TESTS OF THE RANDOM-FEATURE THEORY

We test the learning curves, stopping laws, and compute allocations in the Gaussian random-feature model. The experiments use direct simulations of the finite model and compare them with the continuum predictions derived above. Figure 3 gives the separate comparison with the discrete-time DMFT equations.

The loss curves in Figure 4A, B, and C–D are also shown in Figures 13(b), 14(a), and 15(a–b), respectively. Panels A–B use $P = 256$, and C–D use $P = 64$, with data size fixed within each comparison. The inset in Figure 4A shows the ensemble size selected by the stopped allocation described in Appendix K.4.

K.1 EXPERIMENTAL SETUP

We use population eigenvalues $\lambda_k = k^{-b}$ and target coefficients $(w_k^\star)^2 = k^{b-a}$, for $k = 1, \dots, M$. Thus the target power in mode k is $\lambda_k (w_k^\star)^2 = k^{-a}$. All experiments in this section have $b = 2$, noiseless labels ($\sigma = 0$), and zero initialization. Unless we vary the target exponent explicitly, $a = b = 2$, so the squared teacher coefficients are constant. Each configuration uses 20 datasets. Ensemble members share the training data and use independent Gaussian random-feature matrices.

For GF, we evaluate the finite-model spectral solution at the saved times. For SGD, we reuse the same ordered minibatches each epoch and evaluate the exact affine epoch update in (42). This permits late checkpoints to be evaluated by matrix composition. The loss is the population mean-squared prediction error, evaluated analytically in the population eigenbasis.

We estimate C_{self} by averaging squared errors across a pool of models and C_{cross} by averaging error inner products over distinct pairs in that pool. Equation (41) then gives the exact average loss over all size-E subsets of the sampled pool. The dependence on E in this construction follows from the ensemble identity; the numerical tests examine the time dependence of the two correlations and the resulting stopping and allocation rules. Pool sizes are specified in the captions.

Stopping minimizes the dataset-mean population loss over saved checkpoints. We use $\mathcal{L}^*(P, E)$ for this minimum when $N = P$. Uncertainty bands are pointwise 95% percentile intervals from 1,000 whole-dataset bootstrap resamples, conditional on the sampled model pools. When a quantity involves selection, the checkpoint and other allowed choices are reselected within each resample. The curves therefore describe oracle selection using population risk.

GF compute is $C = EN\tau$, with the common factor B/η absorbed into its units as in (92). When batch size or learning rate varies, we instead use

$$C_{\mathrm{SGD}} = ENBs, \qquad \tau = \eta s, \qquad Bs = P\mathcal{E}.$$

Each compute comparison holds P fixed. The cost counts the training updates represented by a trajectory, rather than the runtime of the spectral or epoch-update evaluator.

K.2 LEARNING CURVES AND ENSEMBLE CORRELATIONS

Figure 10 tests the time and data dependence at $a = b = 2$. Along $N = P$, the early-time loss approaches the predicted $\tau^{-1/2}$ decay as P increases. The minimum loss is close to P^{-1} at each fixed ensemble size. Because width grows with data size in this comparison, this tests the joint scaling along $N = P$. At late times, $P(C_{\mathrm{self}} - C_{\mathrm{cross}})$ grows approximately as $(\tau/P^2)^{1/2}$. The lower row shows how the learning curves change below, at, and above interpolation. Ensembling reduces late-time loss in all three examples, while the strong growth at $N = P$ is specific to the critical ratio.

Figure 11 resolves the contributions that determine stopping at $N = P$. For constant squared teacher coefficients, the continuum cross correlation approaches its limit as $1/\tau$, while the self correlation grows as $\sqrt{\tau}$. At an interior minimum,

$$-\partial_\tau C_{\mathrm{cross}} = \frac{1}{E} \partial_\tau (C_{\mathrm{self}} - C_{\mathrm{cross}}).$$

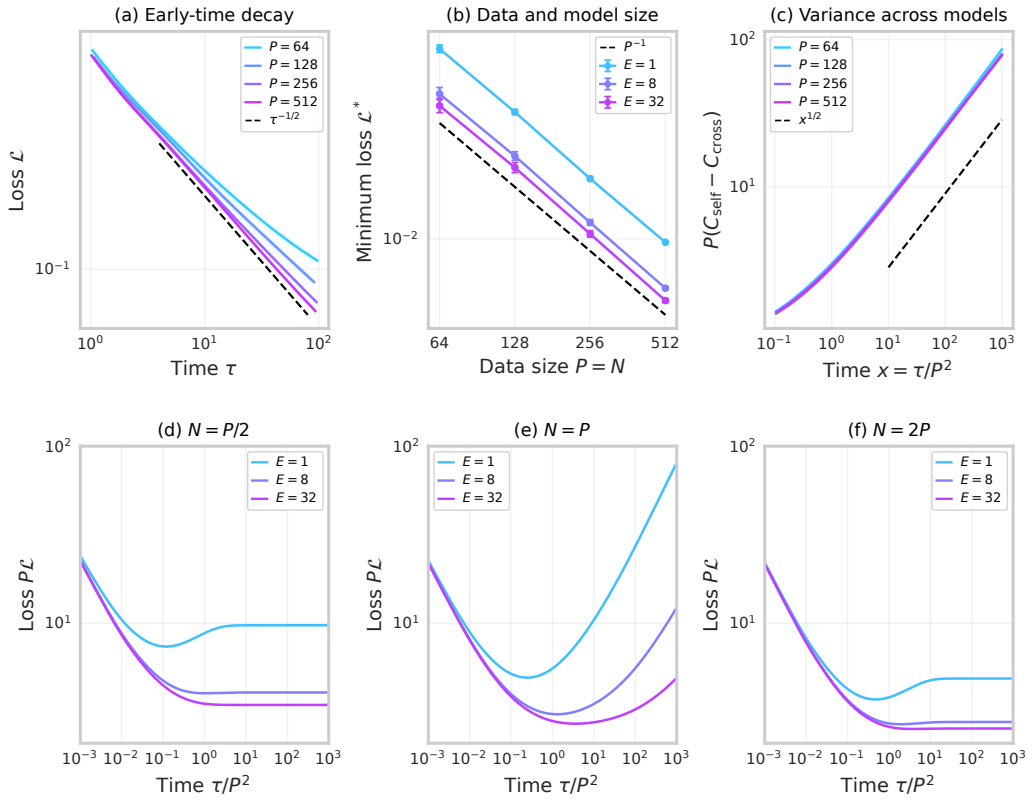

Figure 10: Learning curves in the noiseless random-feature model with $a = b = 2$ and $M = 16384$. (a) Single-model loss at early times for $N = P$. (b) Minimum loss versus P, also along $N = P$, at fixed ensemble size. (c) The self-minus-cross correlation, multiplied by P, versus scaled time. Dashed lines in (a)–(c) are slope guides with arbitrary amplitudes. Bands and bars are dataset-bootstrap intervals. (d)–(f) Ensemble learning curves at $P = 256$ and three width-to-data ratios. All panels use 20 datasets and 32 models per dataset.

The derivative comparison uses the full curves in (90), including their corrections at finite time. This balance gives the $E^{2/3}$ stopping law while retaining $1/E$ variance reduction at each fixed checkpoint.

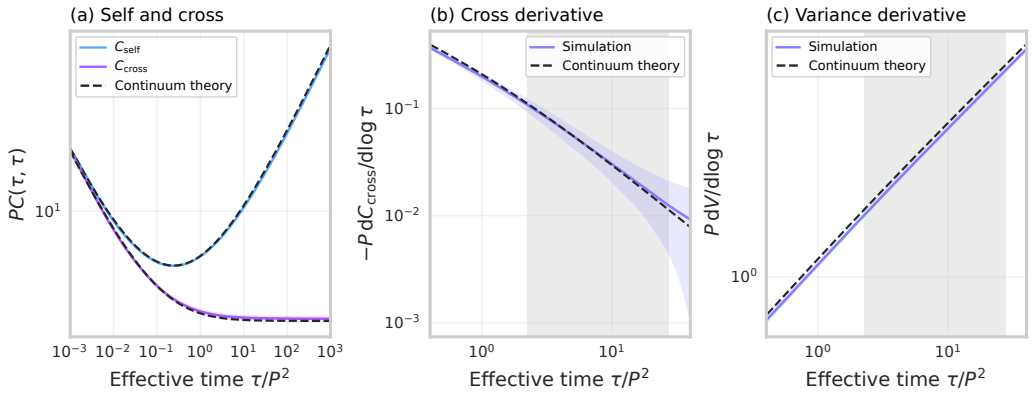

Figure 11: Self and cross correlations at interpolation, with $P = N = 256$, $M = 32768$, and $a = b = 2$. The simulations average 20 datasets and 512 models per dataset. (a) Equal-time correlations. (b) The negative cross derivative. (c) The derivative of $V = C_{\text{self}} - C_{\text{cross}}$. Gray shading in (b)–(c) spans the stopping times for $E = 16$ through 512; colored bands are dataset-bootstrap intervals. The continuum curves use the normalization of (90), with no fitted amplitudes. Derivatives are with respect to $\log x$, where $x = \tau/P^2$; multiplying the stopping equation by τ leaves its balance unchanged.

K.3 Optimal stopping for different teacher spectra

We next vary a at fixed $b = 2$. For each of $a \in \{1.5, 2, 2.5\}$, we use $N = P \in \{64, 128\}$, $M = 8192$, and pools of 256 models. A paired calculation at $P = 128$ increases the cutoff to $M = 16384$, retaining the same underlying Gaussian draws. These settings test the three branches of (88) and (89).

For each ensemble size, we minimize the complete continuum loss $(1 - 1/E)C_{\text{cross}} + C_{\text{self}}/E$ to obtain its theoretical stopping point. The comparison in Figure 12 uses these finite-E predictions. We also measure

$$\Delta_E = \mathcal{L}^*(P, E) - \mathcal{L}^*(P, 2E), \qquad C_E = EP\tau_\star(P, E).$$

This difference removes the common limiting loss without estimating it. The horizontal coordinate C_E is the stopping cost of the smaller ensemble. Each loss in the difference is evaluated at its own minimum.

The stopping time flattens for $a < b$, whereas it continues to grow for $a = b$ and $a > b$. Table 9 compares fits on the same ensemble range. The finite-range continuum predictions fall within the reported intervals at $P = 128, M = 16384$. In particular, for $a > b$, the continuum slope in this range is about 0.86, before it reaches its limiting value of one. The $P = 64$ curves show larger deviations, which remain visible in the figure.

Table 9: Stopping-time exponents at $b = 2$, fitted over $E = 16, 32, 64, 128, 256$. The simulation column uses $P = N = 128$, $M = 16384$, and a 256-member pool. Brackets give percentile dataset-bootstrap 95% intervals for the fitted exponent. The continuum fit uses the same five ensemble sizes.

a	Asymptotic exponent	Continuum fit	Simulation fit
1.5	0	0.167	0.172 [0.117, 0.234]
2	2/3	0.704	0.633 [0.535, 0.727]
2.5	1	0.860	0.776 [0.705, 0.865]

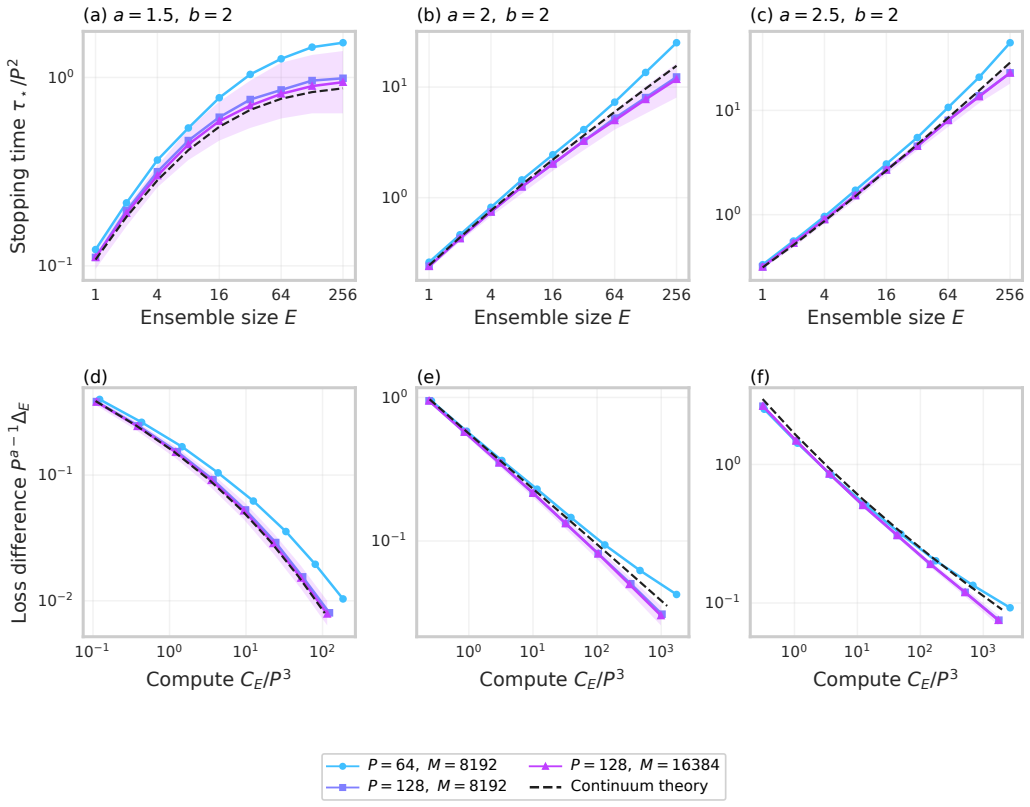

Figure 12: Teacher dependence of stopping and compute scaling at $N = P$ and $b = 2$. Columns vary a. Top: normalized stopping time versus ensemble size. Bottom: normalized adjacent-ensemble loss difference versus the smaller ensemble's stopping cost. Colors distinguish the simulated data sizes and cutoffs; shaded intervals are shown for $P = 128, M = 16384$. In the lower row, the loss intervals are plotted at the central estimate of compute. Dashed curves are the full continuum predictions, with no fitted amplitudes. All simulations use 20 datasets and 256 models per dataset.

K.4 Compute allocation with fixed data

Figure 13 compares the stopped allocation in (5) with the budget frontier in (100). The first chooses among ensembles that can afford to reach their own loss minimum. The second also allows earlier checkpoints. Its feasible set is larger, giving lower loss over part of the compute range. This distinction is visible even with fixed $N = P$, and it is separate from the loss-difference scaling in Figure 12.

For the inset in Figure 4A, we first find each ensemble's minimum dataset-mean population loss over the saved checkpoints. At each budget, we select the lowest-loss configuration among these minima whose cost $EN\tau$ is affordable, and plot its ensemble size. This is the allocation used by the dashed curve. The ensemble grid is $E \in \{1, 2, 4, 8, 16, 32, 64, 128, 256\}$, so the inset ends at the tested maximum of 256 members.

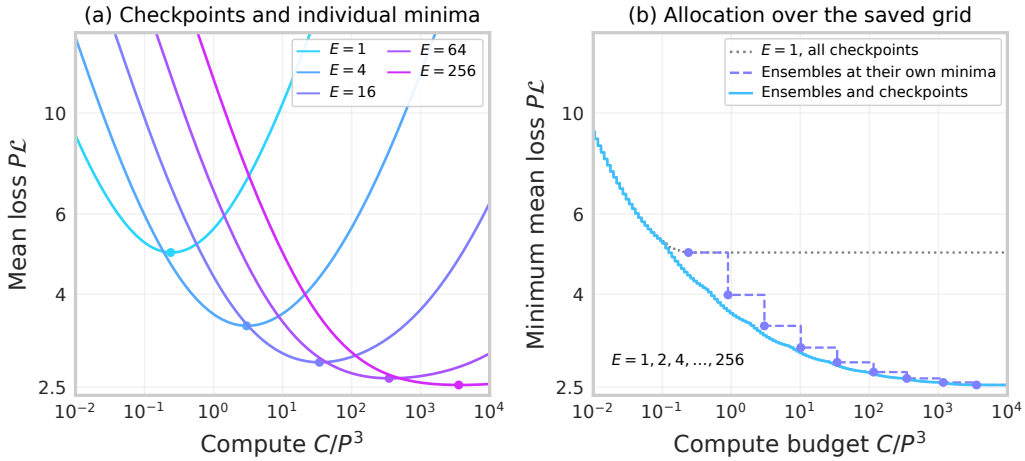

Figure 13: Compute allocation at fixed $P = N = 256$, with $M = 16384$, $a = b = 2$, 20 datasets, and a 256-member pool. Learning curves are plotted against $C = EN\tau$, and their minima define the stopped configurations in (a). The budget comparison in (b) selects either an affordable stopped configuration or any affordable saved checkpoint. Ensemble sizes are $E \in \{1, 2, 4, 8, 16, 32, 64, 128, 256\}$. The single-model checkpoint frontier is included for reference. Curves show means, and unused budget is allowed.

The width comparison starts from $N_0 = P/2$, as in (99). At fixed $P = 256$, we compare increasing the width of one model toward P with training an ensemble of models of width N_0. Both strategies choose their best affordable checkpoint. Figure 14 also extends the width range above interpolation. From $N_0 = P/2$, increasing width is more effective at low budgets, while ensembling achieves lower loss at larger budgets than the tested single models with $N \le P$. Allowing widths through $16P$ changes the comparison: with ensembles capped at 32 members, the widest single model achieves a lower loss than the $N = P$ ensembles at large budgets. Ensembling at $N = 16P$ lowers the loss further. Thus the comparison depends on the permitted widths and ensemble sizes.

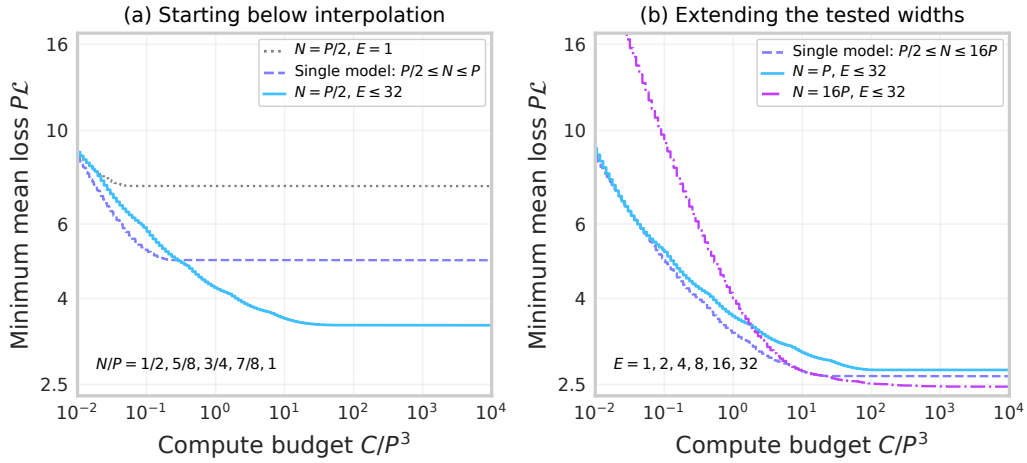

Figure 14: Width and ensemble allocation at fixed $P = 256$, with $M = 16384$ and $a = b = 2$. (a) Starting from $N_0 = P/2$, compare increasing single-model width through $N/P \in \{1/2, 5/8, 3/4, 7/8, 1\}$ with increasing ensemble size at fixed $N = N_0$. (b) Single-model widths through $16P$, compared with ensembles at $N = P$ and $N = 16P$. The full width grid is $N/P \in \{1/2, 5/8, 3/4, 7/8, 1, 5/4, 3/2, 2, 4, 8, 16\}$. Each curve minimizes dataset-mean loss over its allowed configurations and affordable saved checkpoints. The figure uses 20 datasets and 32 models per dataset at each width; $E \in \{1, 2, 4, 8, 16, 32\}$. Loss is averaged over datasets before selecting configurations.

K.5 Batch size and learning rate

We hold $P = 64$ fixed and study $N = 32$ and $N = 64$, with $M = 4096$, $a = b = 2$, and 128 models per dataset. Batch sizes are $B \in \{1, 4, 16, 64\}$. We use the same initial learning-rate grid $\eta \in \{0.01, 0.04, 0.12, 0.4, 1.2\}$ for every batch size. A candidate is retained when all its simulated members pass the fixed-order epoch stability check. This retains 30 of the 40 (N, B, η) settings: $\eta = 1.2$ is excluded throughout, and $\eta = 0.4$ is additionally excluded at $B = 1$. Comparisons use the remaining grid, with the same 20 datasets for every retained candidate. After the first epoch, checkpoints are saved at whole-epoch boundaries.

Figure 15 first fixes $\eta = 0.04$, then permits each batch size to select its learning rate under the compute budget. With fixed η, smaller batches make substantially more progress per unit compute. Learning-rate selection narrows the gap. At $C_{\mathrm{SGD}} = 10^4$ and $E = 1$, the loss reduction from $B = 4$ to $B = 1$ falls from 41.6% to 5.1% for $N = P/2$, and from 45.9% to 6.8% for $N = P$. This agrees with the time conversion in Appendix D: at fixed compute, the accessible effective time depends on η/B. The discrete epoch update determines the remaining dependence on batch size and learning rate.

Figure 16 then allocates compute over ensemble size. At low budgets, one model can achieve a lower loss than four models because it receives more updates. At $B = 1$ and $C_{\mathrm{SGD}} = 10^6$, four models instead reduce loss by 33.7% for $N = P/2$ and 17.6% for $N = P$, after learning-rate selection. The selected ensemble size increases with budget and eventually reaches the tested cap of 64. At fixed $E = 1$ or $E = 4$, the minimum losses across batch sizes are nearly equal at the largest budgets. In this experiment, the main batch-size effect is the compute needed to approach those losses.

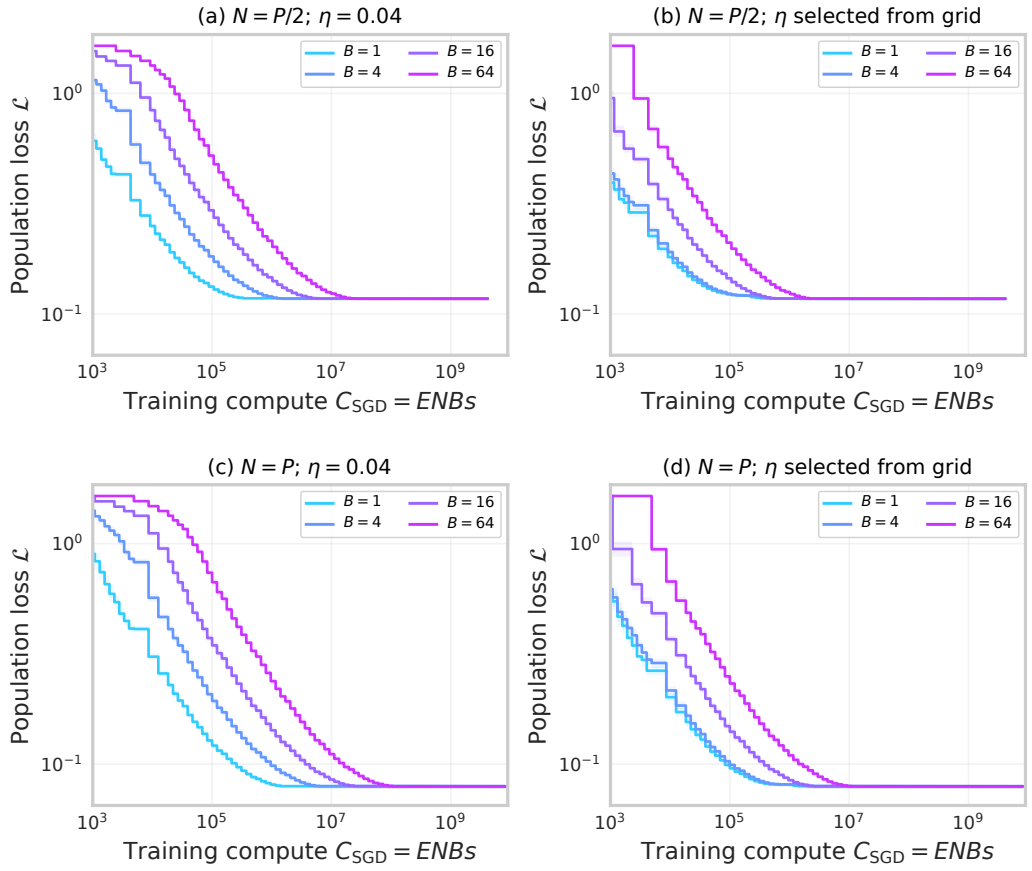

Figure 15: Batch size at matched training compute $C_{\mathrm{SGD}} = ENBs$, with $P = 64$, $M = 4096$, and $E = 1$. Rows show $N = P/2$ and $N = P$. Left: fixed learning rate $\eta = 0.04$. Right: learning rate selected from the retained grid separately for each batch size and budget. All curves select an affordable checkpoint. Bands are pointwise dataset-bootstrap intervals, with selection repeated in each resample. Each setting uses 20 datasets and 128 models per dataset.

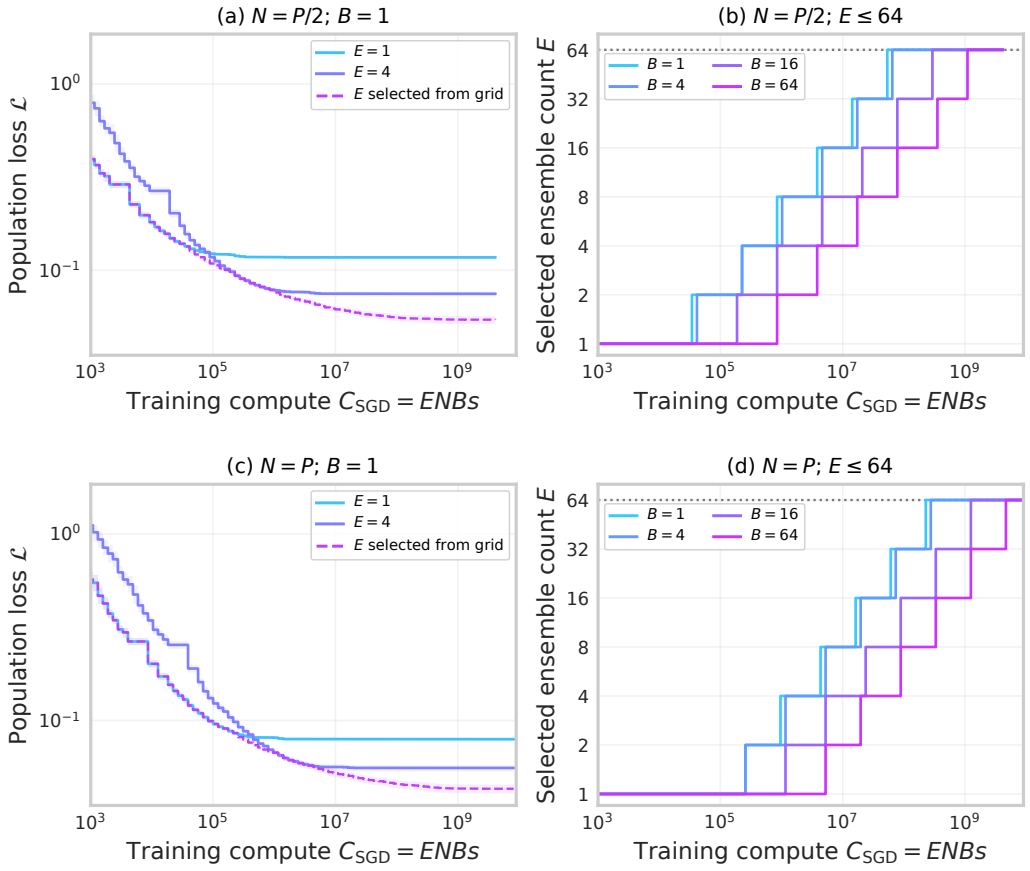

Figure 16: Ensemble allocation for cyclic SGD with the settings of Figure 15. Left: at $B = 1$, compare fixed $E = 1$, fixed $E = 4$, and selection over $E \in \{1, 2, 4, 8, 16, 32, 64\}$, allowing learning-rate and checkpoint selection under each budget. Right: selected ensemble size for each batch size. Rows show $N = P/2$ and $N = P$. The upper ensemble limit is 64. At budgets permitting only the zero predictor, ties are assigned to $E = 1$.

K.6 NUMERICAL CHECKS

Figure 17 compares spectral cutoffs, feature pools, and discrete updates. At $P = N = 256$, doubling M from 16384 to 32768 changes the stopping times only slightly. Increasing the number of sampled models gives a separate check on the self/cross estimates. With 512 models, the fitted stopping exponent over $E = 32, 64, 128, 256, 512$ is 0.716, with a 95% interval $[0.589, 0.874]$. This larger-size calculation is consistent with the constant-teacher stopping law and its finite-E corrections.

The paired teacher-cutoff comparison in Figure 12 gives a similar check for all three target spectra. Doubling M changes the fitted stopping exponents from $0.171, 0.652, 0.778$ to $0.172, 0.633, 0.776$ for $a = 1.5, 2, 2.5$, respectively. Absolute losses are more sensitive to the cutoff for $a = 1.5$, whose target power decays more slowly. Using the saved analytic derivatives to interpolate stationary points changes the largest-pool stopping exponents by less than 0.006 across these nine teacher configurations.

For the small-step comparison, GF and cyclic SGD use identical datasets and feature matrices at $P = N = 32$, $M = 4096$, and $B = 4$. At matched effective time, the maximum relative loss difference over $0.1 \le \tau/P^2 \le 20$ and $E \in \{1, 2, 4, 8, 16, 32\}$ is 0.127% at $\eta = 0.02$ and 0.064% at $\eta = 0.01$. The mean-squared difference between the paired SGD and GF predictions decreases by approximately a factor of four when the step size is halved at the final checkpoint. This is consistent with first-order prediction error in the small-step limit established in Appendix D.

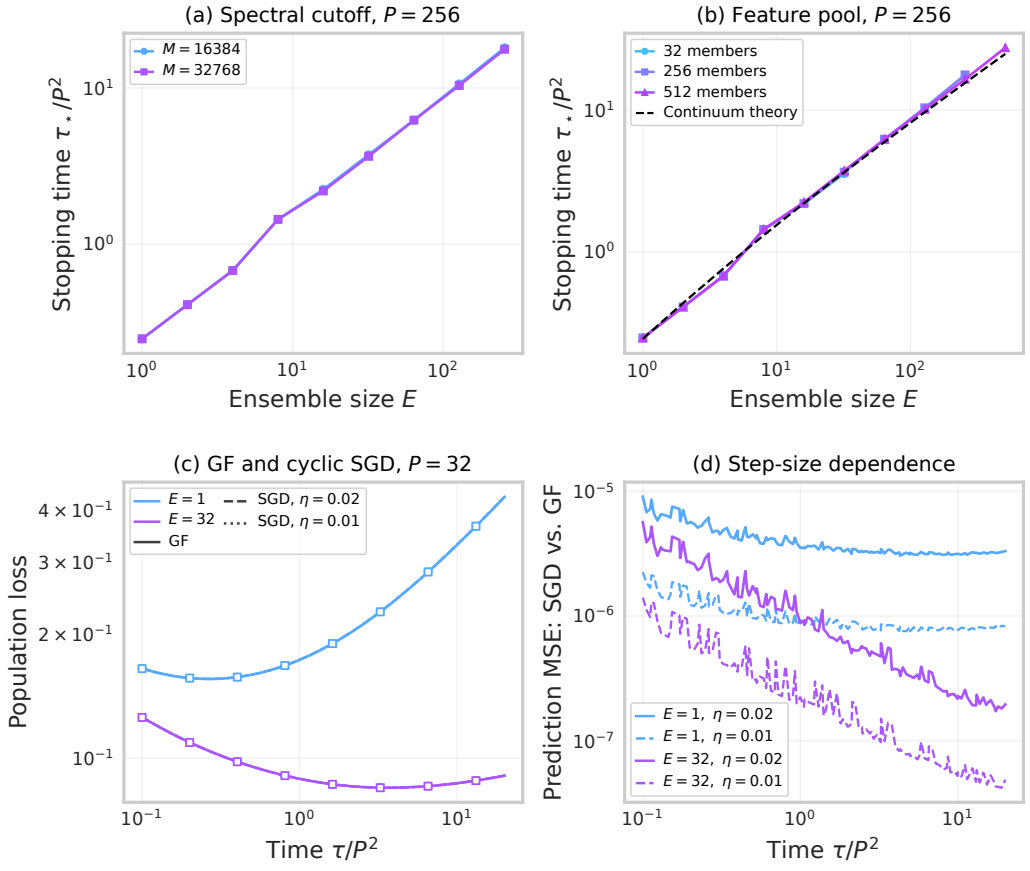

Figure 17: Numerical checks for $a = b = 2$. (a) Stopping times at $P = N = 256$ for paired cutoffs $M = 16384$ and 32768, using 256 models per dataset. (b) Nested pools of 32, 256, and 512 models at $P = N = 256, M = 32768$, together with the continuum prediction. (c) GF and fixed-order cyclic SGD at $P = N = 32, M = 4096, B = 4$, using 32 models per dataset. (d) Population mean-squared difference between the paired SGD and GF predictions. Panels (c)–(d) compare matched effective time $\tau = \eta s$. Every configuration uses 20 datasets.