# Scaling Theory for SlowRunning: Model size, Ensembling, and Training Horizon in the Multi-Epoch Regime

**author names withheld**

**Under Review for the Workshop on High-dimensional Learning Dynamics, 2026**


## Abstract

We study the learning dynamics of multi-epoch training both empirically and theoretically. Consistent with empirical works on language model training, naive scaling of training epochs and model size fail to deliver monotonic improvements in performance under multiple passes. However, denoising methods such as ensembling, regularization, and averaging over data shuffles can improve performance in the multi-epoch regime. We theoretically analyze multi-epoch training in a solvable powerlaw random feature model using dynamical mean field theory. This theory predicts how train and test loss evolve over iterations of SGD within-epoch and across epochs. We show that SGD noise adds variance across steps within epochs while systematic overfitting effects arise from the *cross-epoch* correlations in gradients which build up as response functions in the theory. Using the model, we provide an analysis of ensembling, model size, and SGD noise. We then conduct experiments in language model pretraining where we show that, in regimes where learning curves are non-monotonic, increasing ensembles can be preferable to increasing width at fixed compute. Using our model, we provide a theoretical argument for this account.


## 1. Introduction

Scaling laws for one-pass training are well established [14, 15], but compute is now growing faster than the supply of unique web text [21, 29], making multi-epoch training unavoidable. Empirically, the consequences are subtle: Muennighoff et al. [21] found $\sim 4$ epochs are essentially free but additional repetition causes naive parameter scaling to overfit, while Kim et al. [16] showed that under fixed data and unbounded compute, heavy weight decay and ensembling over independent runs together yield $\sim 5\times$ data efficiency over the data-constrained Chinchilla extrapolation. The NanoGPT Slowrun benchmark [24] — 100M FineWeb tokens with no compute cap — has crystallized this regime, with leading entries combining many epochs, heavy regularization, and ensembling.

**Theory for Multi-Epoch SGD** Theoretically, the dominant framework for modeling scaling laws by analyzing SGD in power-law random feature models either focuses on the online training regime or operates at small enough learning rates for epoch and SGD step timescales to be indistinguishable [6, 17, 23]. Further, these works do not focus on overfitting phenomenology, where training loss and test loss rapidly diverge over multiple passes. To address the theoretical gap in a systematic way, we develop a **Dynamical Mean Field Theory (DMFT)** which computes the typical loss evolution under a multi-pass SGD setting. Our theory predicts how the loss depends on batch $B$, steps per epoch $T$ (total data is $P = BT$), total epochs $\mathcal{E}$, model size $N$ and the structure of the limiting features through source and capacity exponents for the eigenvalues $\lambda_k$ and target coefficients $w_k^\star$.

**Related Work.** Empirically, Muennighoff et al. [21] fit a data-constrained scaling law over >400 runs up to 9B params and 1500 epochs, finding ∼4 free epochs before exponentially decaying returns; Kim et al. [16] remove the compute cap and report optimal weight decay ∼30× standard practice plus large ensembling gains, and the Slowrun benchmark [24] provides our experimental setting. Theoretically, scaling laws for one-pass SGD on power-law random features have been derived via DMFT [6], Volterra equations identifying 4+3 phases [23], and sketched linear regression [17], with extensions to feature learning [7], momentum/DANA [12], and optimal schedules [3]; static analyses include [1, 2, 5, 18, 27, 28]. Multi-pass SGD has been analyzed at proportional batch size via DMFT/state evolution [9, 13, 20], via Volterra equations for linear regression [22], and for sub-linear-batch multi-index models [11]; closest to us, Yan et al. [30] characterize the effective reuse rate $E(K, N)$ for multi-epoch SGD in linear regression, but we use a random feature model (giving access to model size for compute-matched ensemble-vs-width), operate near $N \approx P$ where noise amplification dominates, and track ensemble-independent randomness to obtain the $1/E$ variance reduction. We utilize bias-variance decompositions for ensembling theory [25, 26, 28].

**Notation Summary for this Work** $P$ unique data, $E$ ensembles, model size $N$, batch size $B$, $T = P/B$ steps per epoch, $\mathcal{E}$ epochs, global step $s = Tn + t$. Source/capacity exponents for the features are denoted by $a, b > 1$ with eigenvalues $\lambda_k \sim k^{-b}$ and target decay $\lambda_k (w_k^\star)^2 \sim k^{-a}$.

## 2. Pretraining Phenomena in Multi-Epoch Regime

**Optimal Stopping Under Multi-Pass Training** We start by experimentally probing the dynamics of training under multiple passes. We train ∼ 100M models on a subset of $P$ fineweb tokens. First we demonstrate that performing multiple passes induces non-monotonic validation losses in Figure 1. The optimal early stopping time $s_\star \sim \Theta(P)$ scales roughly linearly with total data $P$ and the optimal epochs $\mathcal{E} \sim \Theta(1)$ is approximately constant.

**Ensembling** Averaging logits over an ensemble of $E$ models with distinct random initializations and/or SGD data order can eliminate the non-monotonicity in the learning curve as we show in Figure 2. As $E \to \infty$ the ensembled predictor no longer inherits noise from the random model initialization, suggesting that amplification of this noise is generating the non-monotonic loss curve.

## 3. Theory for Multi-Epoch SGD in a Random Feature Model

We now aim to derive a theoretical model that captures these key phenomena in multi-epoch SGD.

**Model Definition** We study the powerlaw random feature model of [6, 23]

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

where the feature map is $\boldsymbol{\psi}(\boldsymbol{x}) \in \mathbb{R}^M$ the matrix $\boldsymbol{A} \in \mathbb{R}^{N \times M}$ is random and the features $\boldsymbol{\psi}(\boldsymbol{x}) \in \mathbb{R}^M$ have covariance $\left\langle \psi_k(\boldsymbol{x}) \psi_\ell(\boldsymbol{x}) \right\rangle = \delta_{k\ell} \lambda_k$. We will often take $M \to \infty$ with $\sum_{k=1}^{\infty} \lambda_k < \infty$. We will optimize the empirical MSE loss $\hat{\mathcal{L}} = \frac{1}{P} \sum_{\mu=1}^{P} \left( f(\boldsymbol{x}_\mu) - y(\boldsymbol{x}_\mu) \right)^2$ on a randomly sampled collection of $P$ data points using SGD, but we will ultimately care about the *test loss* $\mathcal{L} = \left\langle \left( f(\boldsymbol{x}) - y(\boldsymbol{x}) \right)^2 \right\rangle$, which measures errors over the full distribution.

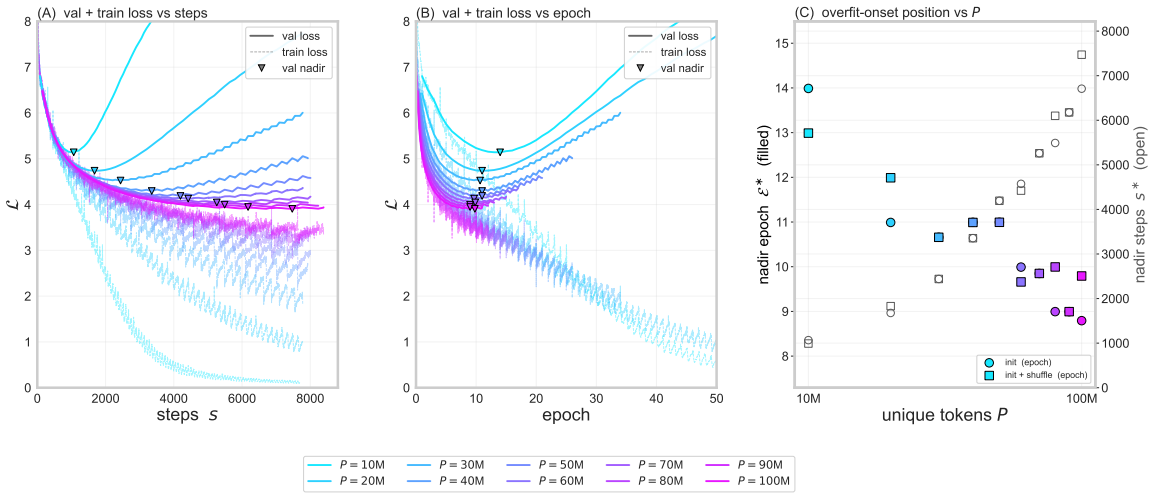


Figure 1: Multi-epoch training induces classical overfitting in the data-limited regime. (A) Validation loss (solid) and training loss (dashed) vs steps for ten unique-data sizes $P \in \{10, \ldots, 100\}$ M from FineWeb); ▼ marks each run's optimal early-stopping point. Training loss falls toward zero while validation loss U-curves, so the generalization gap grows by an order of magnitude as $P$ shrinks. (B) The same curves as a function of epochs: min val loss collapse to $\sim$ 9–13 epochs across the $10\times$ range of $P$. (C) Optimal stopping time.

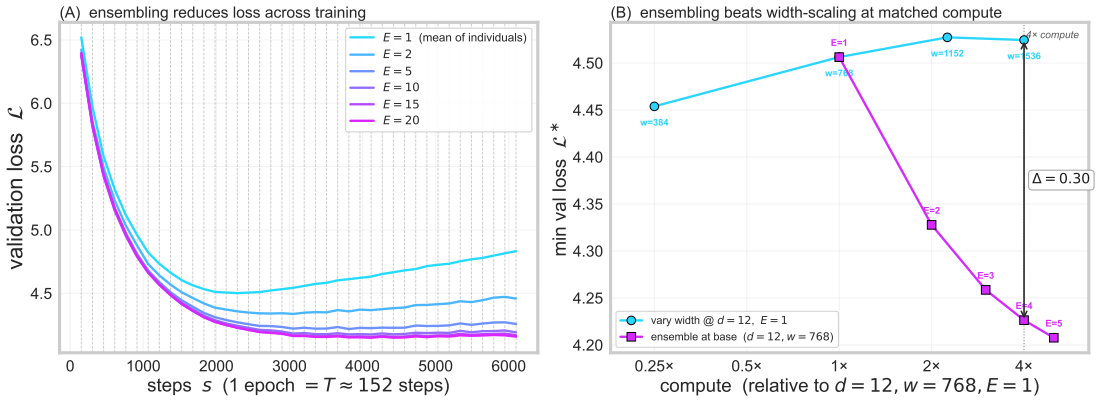


Figure 2: Ensembling reduces validation loss across training and beats single-model scaling at matched compute. (A) Validation loss $\mathcal{L}$ vs steps $s$ for ensemble sizes $E$'s; larger $E$ lowers the asymptote and dampens the post-overfit climb. (B) At each compute budget, two strategies are compared: cyan = train one wider model, magenta = train $E$ base-size models and ensemble them. At 4× compute, the ensemble of $E$=4 base models reaches lower val loss than the 2× wider (hence 4x compute) single model by $\Delta = 0.30$.

**Discrete Time Multi-Epoch SGD** We consider a multi-epoch training setting where each epoch contains $T$ timesteps, each of which has a minibatch of $B$ data. We will analyze the loss over epochs $n \in [\mathcal{E}]$ and time-steps within-epoch time $t \in [T]$. Let the residual $\boldsymbol{v}_n^0(t) \equiv \boldsymbol{w}^\star - \frac{1}{N}\boldsymbol{A}^\top \boldsymbol{w}_n(t)$. To simplify our initial analysis, we will focus on fixed-order multiepoch SGD. Each gradient step is

generated by a minibatch of features $\boldsymbol{\Psi}(t) \in \mathbb{R}^{B\times M}$. We have $T$ of these matrices so that the total data is $P \equiv BT$. The update rule for the error vectors is

$$\boldsymbol{v}_n^0(t+1) - \boldsymbol{v}_n^0(t) = -\eta \left(\frac{1}{N}\boldsymbol{A}^\top \boldsymbol{A}\right)\left(\frac{1}{B}\boldsymbol{\Psi}(t)^\top \boldsymbol{\Psi}(t)\right)\boldsymbol{v}_n^0(t) \ , \ \boldsymbol{v}_{n+1}^0(0) = \boldsymbol{v}_n^0(T)$$

The test loss dynamics are determined by the evolution of this vector.

$$\mathcal{L}_n(t) = \left\langle (f(\boldsymbol{x}) - y(\boldsymbol{x}))^2 \right\rangle_x = \left\langle \boldsymbol{v}_n^0(t)^\top \boldsymbol{\Lambda} \boldsymbol{v}_n^0(t) \right\rangle = \sum_k \lambda_k \left\langle v_{k,n}^0(t)^2 \right\rangle$$

**DMFT Description of the Dynamics** We now average these dynamics over random draws of matrices $\boldsymbol{A}$ and the minibatch SGD iterates $\boldsymbol{\Psi}(t)$. To simplify the equations, we introduce

$$\text{Multi-epoch integration tensor}: \quad \Theta_{nn'}(t,t') \equiv \delta_{nn'} 1[t > t'] + 1[n > n'].$$

Intuitively, this tensor $\boldsymbol{\Theta}$, when applied to a vector $\boldsymbol{v}$, sums over all gradient updates for *earlier epochs or earlier steps within the same epoch*. The complete derivation is in App. B.

**Theorem 1** *The test error dynamics for each eigenmode $k$ is governed by transfer tensor $\boldsymbol{H}_k$ which encodes how the finite features $N$ or batch $B$ alter the dynamics. The error vector $v_k^0$ obeys*

$$v_{k,n}^0(t) = \sum_{n't'} H_{k,nn'}(t,t') \left( w_k^\star + \xi_{k,n'}^{\textit{Model}}(t') + \xi_{k,n'}^{\textit{Data}}(t') \right)$$

*where $\xi^{\textit{Model}}, \xi^{\textit{Data}}$ are Gaussian sources of noise related to (1) random model initialization and (2) random data + SGD fluctuations respectively. The noise and dynamics* ***decouple across eigenmodes*** *$k$. In a vectorized notation over time/epoch indices, the operator $\boldsymbol{H}_k \in \mathbb{R}^{T\mathcal{E}\times T\mathcal{E}}$ solves*

$$\boldsymbol{H}_k = (\boldsymbol{I} + \eta\lambda_k \boldsymbol{\Theta} \boldsymbol{R}_3 \boldsymbol{R}_1)^{-1} \ , \ \bar{\boldsymbol{H}} \equiv \sum_k \lambda_k \boldsymbol{H}_k$$

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

*and $\mathcal{D}$ casts an tensor to diagonal over times $(t,t')$ as $[\mathcal{D}(\boldsymbol{A})]_{nn'}(t,t') = A_{nn'}(t,t)\delta_{t,t'}$.*

While the **response functions** $\{\boldsymbol{H}_k, \boldsymbol{R}_1, \boldsymbol{R}_3\}$ encode the structural dynamics of the bias, computing the full loss requires solving for a system of **correlation functions** like $C_{nn'}^0(t,t') = \sum_k \lambda_k \left\langle v_{k,n}^0(t) v_{k,n'}^0(t') \right\rangle$. The noise statistics for $\xi$ fields, which contribute to the correlations, are themselves set by these correlation functions self-consistently. Solving this system (see Appendix B.3) can provide insight into the exact loss curve, which is generally non-monotonic. We show examples of our DMFT compared to the model dynamics in Figure 3.

**Ensembling** We can exactly characterize ensembling within our theory using an $E$-replicated DMFT which tracks joint dynamics of $E$ separate systems $\boldsymbol{v}_{e,n}^0(t)$, which each have separate model features $\boldsymbol{A}_e$ (App. C). We prove that the saddle point equations for the correlation functions are *replica symmetric* $C_{e,e'} = \delta_{e,e'} C_{\text{self}} + (1 - \delta_{e,e'}) C_{\text{cross}}$ with typical self-correlation and cross-replica correlation [6, 10, 26]. The correlation function for the ensembled model prediction $\bar{\boldsymbol{v}} = \frac{1}{E}\sum_e \boldsymbol{v}_e$ can be obtained from $\bar{C} = C_{\text{cross}} + \frac{1}{E}(C_{\text{self}} - C_{\text{cross}})$.We show that $C_{\text{self}}$ generates the non-monotonic blowup in the dynamics, while $C_{\text{cross}}$ is monotone.

**Theoretical Scaling Law under Source/Capacity** We now utilize our DMFT equations to describe the scaling law for models near the interpolation threshold. This setting is motivated since, under a compute budget $C \propto NPs$, the optimal strategy is to allocate $N \approx P$ [6, 14, 15].

**Theorem 2** *Under source/capacity conditions with model size approximately equal to total data $N \approx P$, an $E$ ensemble system at step $s = Tn + t$ obeys an approximate Chinchilla scaling law*

$$\mathcal{L} \sim s^{-(a-1)/b} + N^{-(a-1)} + P^{-(a-1)} + \frac{1}{E} P^{-(a-1+b/2)} s^{1/2}$$

*which implies optimal stopping time $s_\star \sim E^{\frac{2b}{2(a-1)+b}} P^b$.*

The overfitting term can be reduced through ensembling, increasing data, or through early stopping. The optimal epochs scale with total data as $\mathcal{E}_\star \sim P^{b-1}$, implying optimal epochs are constant for $b \approx 1$, consistent with Figure 1. We verify the $\sqrt{s}$ scaling law for the blowup in Figure 3.

**Conclusion** We studied a structured random feature model in multi-epoch SGD. This theory captures the impact of ensembling, feature structure, and multiple passes on the test loss dynamics. Future work could explore the influence of shuffling data across epochs on the SGD noise structure and ensemble statistics [19].

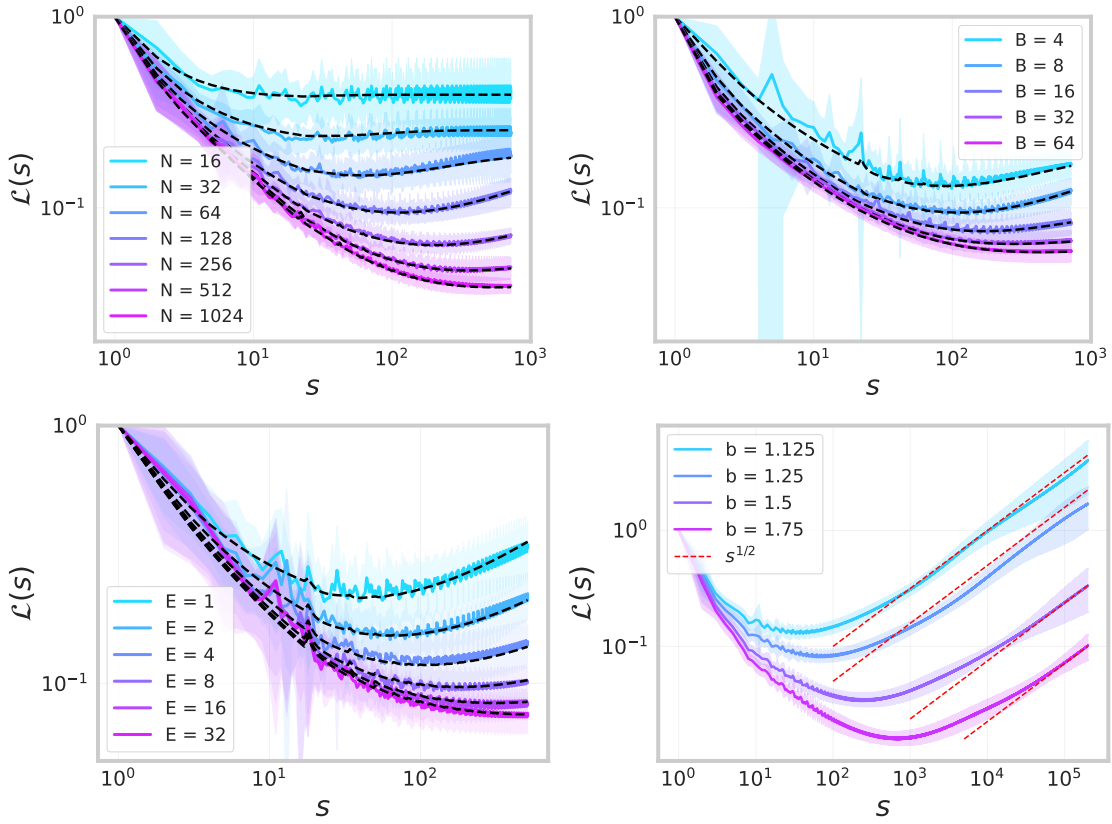


Figure 3: Loss dynamics as a functions of steps $s = nT + t$ in the multi-epoch power-law random feature model compared to DMFT (dashed black). We show the impact of varying $N, B, E$ and visualize the $s^{1/2}$ blowup across different capacity exponents $b$.

## References


[1] Alexander B Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*, 2024.

[2] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *arXiv preprint arXiv:2102.06701*, 2021.

[3] Blake Bordelon and Francesco Mori. Theory of optimal learning rate schedules and scaling laws for a random feature model. *arXiv preprint arXiv:2602.04774*, 2026.

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

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

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

[7] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. How feature learning can improve neural scaling laws, 2024. URL https://arxiv.org/abs/2409.17858.

[8] 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.

[9] Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data. *arXiv preprint arXiv:2112.07572*, 2021.

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

[11] Zhou Fan and Leda Wang. High-dimensional learning dynamics of multi-pass stochastic gradient descent in multi-index models. *arXiv preprint arXiv:2601.21093*, 2026.

[12] Damien Ferbach, Katie Everett, Gauthier Gidel, Elliot Paquette, and Courtney Paquette. Dimension-adapted momentum outscales sgd. *Advances in Neural Information Processing Systems*, 38:112780–112977, 2026.

[13] 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.

[14] 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.

[15] 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.

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

[17] 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.

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

[19] Francesca Mignacco and Pierfrancesco Urbani. The effective noise of stochastic gradient descent. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(8):083405, 2022.

[20] 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.

[21] 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.

[22] 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*, pages 3548–3626. PMLR, 2021.

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

[24] qlabs. Nanogpt slowrun: 100m tokens. infinite compute. lowest val loss wins. https://github.com/qlabs-eng/slowrun, 2026.

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

[26] 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*, pages 52198–52224. PMLR, 2025.

[27] 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.

[28] James B Simon, Dhruva Karkada, Nikhil Ghosh, and Mikhail Belkin. More is better in modern machine learning: when infinite overparameterization is optimal and overfitting is obligatory. *arXiv preprint arXiv:2311.14646*, 2023.

[29] 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.

[30] 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.

## Appendix A. Fitting Experimental Loss Curves to Scaling Forms

### A.1. Loss fit with respect to $s$ and $E$

Fixing $P$, we fit the following scaling law with respect to $s$ and $E$ and empirically

$$\mathcal{L}_{s,E} \;=\; \alpha_s^{(\text{strat})}\, E^{-\beta^{(\text{strat})}} + \mathcal{L}_{s,\infty},$$

where strat can be chosen from "init" and "init+shuffle". "init" means individuals forming an ensemble only differ in their model initializations; "init+shuffle" means they differ in both model initializations and the data order per epoch.

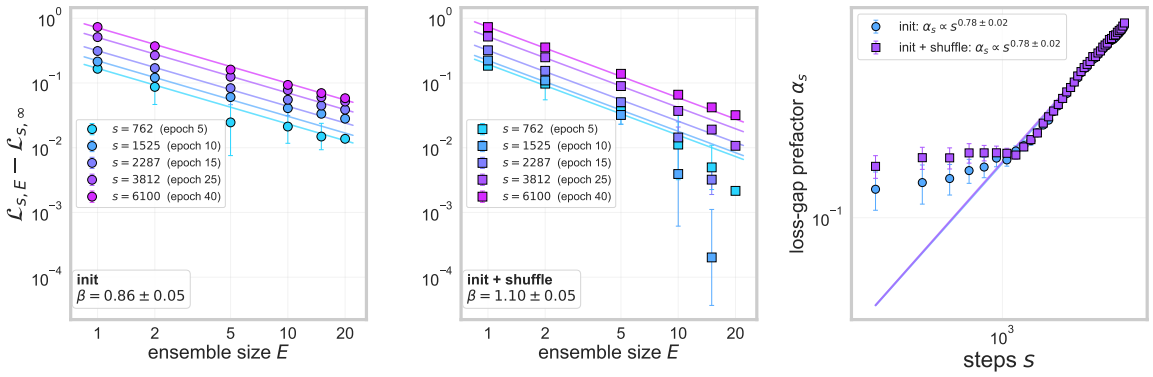


Figure 4: Ensembling reduces the validation loss by a power law in $E$ at every training step $s$, with strategy-specific exponents. Left and middle panels: $\mathcal{L}_{s,E} - \mathcal{L}_{s,\infty}$ vs $E$ on log-log axes, fitted as $\alpha_s E^{-\beta}$ with shared $\mathcal{L}_{s,\infty}$ across strategies and per-strategy $\beta$. Right: prefactor $\alpha_s$ as a function of steps $s$, with a power-law fit $\alpha_s = c\, s^p$ for each strategy.

### A.2. Global loss fit

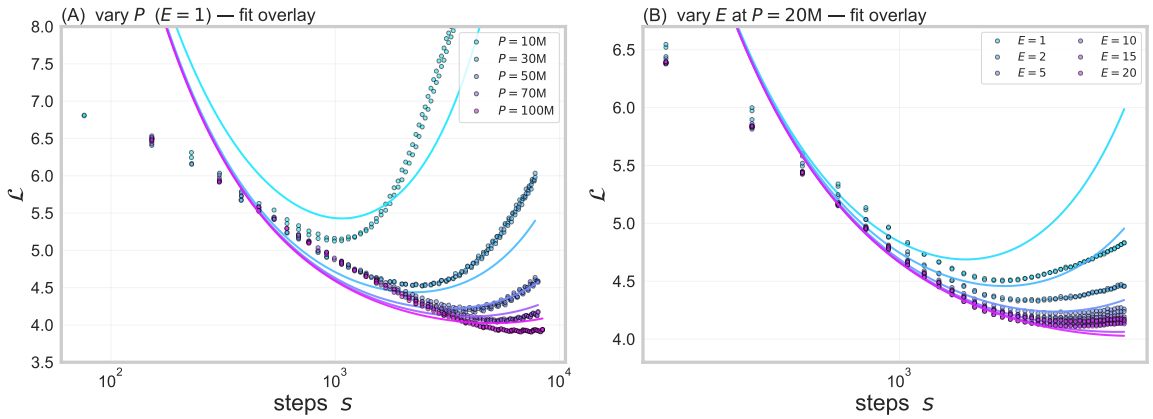


Figure 5: Global fit $\mathcal{L}(s, P, E, N) \;=\; c_1 s^{\delta_{s_1}} + c_2 N^{\delta_N} + c_3 P^{\delta_{P_1}} + (c_4/E)\, P^{\delta_{P_2}} s^{\delta_{s_2}} + \sigma^2$. (A) Validation loss vs steps for varying unique-data sizes $P$ at $E$=1; solid lines are the fit. (B) Same fit overlaid on ensemble-size sweep at $P$=20M. Both strategies pooled.

### A.3. Compute-matched comparison

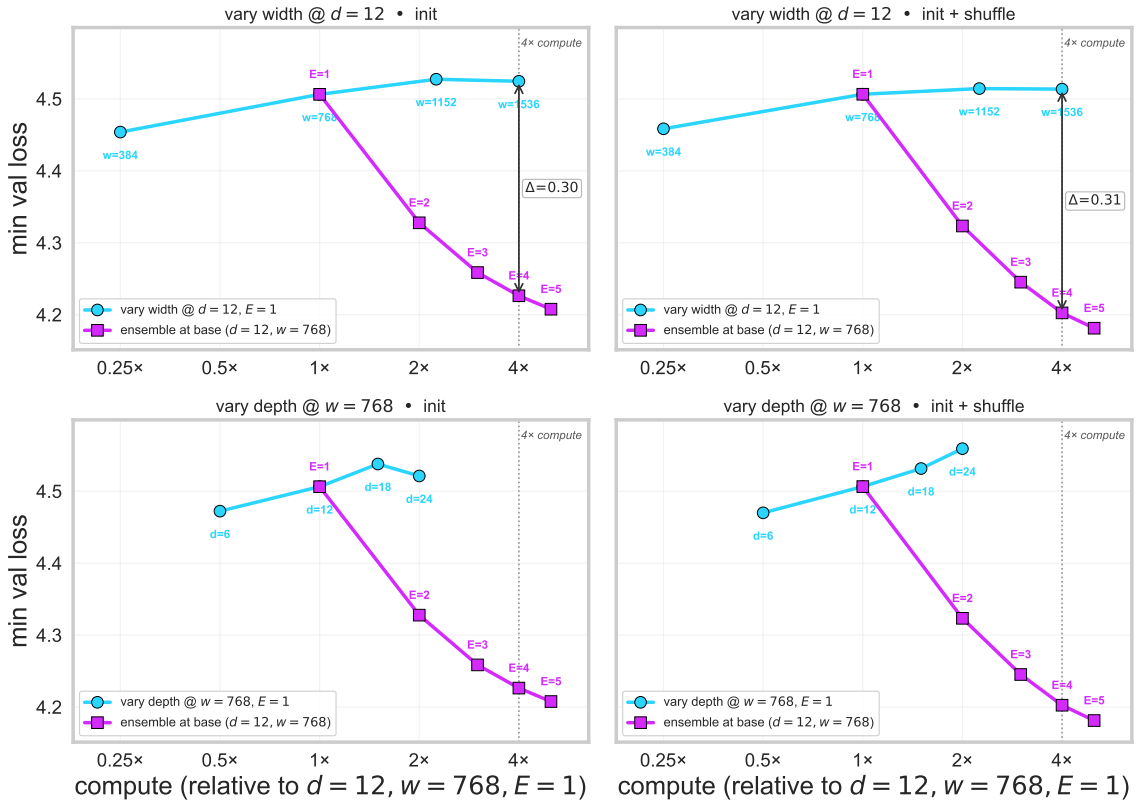


Figure 6: Compute-matched ensembling versus single-model scaling. Rows: vary width at fixed depth $d$=12 (top) / vary depth at fixed width $w$=768 (bottom). Columns: *init* / *init plus shuffle*. Cyan circles trace the single-model scaling line ($E = 1$, varying $N$ or $L$); magenta squares trace the ensemble-at-base line ($d$=12, $w$=768, varying $E$). Compute is taken proportional to non-embedding parameters $16\,LN^2$, multiplied by $E$ for ensembles, normalized to the $d$=12, $w$=768, $E$=1 baseline. Vertical dotted lines mark $4\times$ matched compute.

## Appendix B. Derivation of the Multi-Epoch DMFT Equations

In this appendix we derive the dynamical mean-field theory (DMFT) equations for the fixed-order multi-epoch random feature model. Throughout, all feature-space quantities are written in the eigenbasis of the population feature covariance,

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

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

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

We allow an arbitrary spectrum and target alignment in the DMFT equations. Power-law specializations are made only in Appendix E.

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 fixed-order multi-epoch causal operator is

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

Thus $\Theta_{\alpha\beta} = 1$ exactly when update $\beta$ occurs before state $\alpha$. We also define the fixed-minibatch projection operator $\mathcal{D}$. For any $S \times S$ matrix $\boldsymbol{A}$,

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

This operator keeps correlations between the same within-epoch minibatch label $t$ across different epochs, and removes correlations between different minibatch labels.

### B.1. 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 residual update is

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

with epoch boundary condition

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

When label noise is included, the corresponding minibatch noise vector is

$$\boldsymbol{\epsilon}(t) \in \mathbb{R}^B.$$

The noiseless case is recovered by setting $\sigma = 0$.

To derive the multi-epoch DMFT, we follow the standard random-feature construction [4, 6, 23] and introduce a sequence of auxiliary vector fields such that each stage of the update is linear in a single random matrix. This separates the contribution of the minibatch design $\boldsymbol{\Psi}(t)$ from that of

the random feature matrix $\boldsymbol{A}$, and provides the convenient starting point for the effective single-site stochastic process used below:

$$\begin{aligned}
\boldsymbol{v}_n^1(t) &= \boldsymbol{\Psi}(t)\boldsymbol{v}_n^0(t) + \sigma\boldsymbol{\epsilon}(t) \in \mathbb{R}^B, \\
\boldsymbol{v}_n^2(t) &= \frac{1}{B}\boldsymbol{\Psi}(t)^\top \boldsymbol{v}_n^1(t) \in \mathbb{R}^M, \\
\boldsymbol{v}_n^3(t) &= \boldsymbol{A}\boldsymbol{v}_n^2(t) \in \mathbb{R}^N, \\
\boldsymbol{v}_n^4(t) &= \frac{1}{N}\boldsymbol{A}^\top \boldsymbol{v}_n^3(t) \in \mathbb{R}^M.
\end{aligned}$$

Here $\boldsymbol{v}_n^1(t)$ is the minibatch residual vector, $\boldsymbol{v}_n^2(t)$ is the feature-space quantity obtained after passing through the data block, $\boldsymbol{v}_n^3(t)$ is the corresponding random-feature-space vector after multiplication by $\boldsymbol{A}$, and $\boldsymbol{v}_n^4(t)$ is the resulting feature-space gradient direction after multiplication by $\boldsymbol{A}^\top/N$.

With these definitions,

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

Equivalently, using the causal operator $\boldsymbol{\Theta}$,

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

Here $\boldsymbol{v}_\alpha^i$ denotes the vector field $\boldsymbol{v}_n^i(t)$ at $\alpha = (n,t)$. In stacked notation,

$$\boldsymbol{v}^0 = \mathbf{1}\otimes\boldsymbol{w}^\star - \eta\boldsymbol{\Theta}\boldsymbol{v}^4,$$

where $\boldsymbol{\Theta}$ acts only on the epoch-step indices.

### B.2. Single-site stochastic processes

We now pass from the high-dimensional random dynamics to the effective single-site stochastic process. There are three kinds of sites: a data site for $v^1$, a random-feature site for $v^3$, and a spectral mode $k$ for $v_k^0, v_k^2, v_k^4$. The fields remain coupled through the epoch-step indices.

The effective process is

$$v_{k,\alpha}^0 = w_k^\star - \eta\sum_\beta \Theta_{\alpha\beta}\, v_{k,\beta}^4, \tag{1}$$

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

$$v_{k,\alpha}^2 = u_{k,\alpha}^2 + \lambda_k\sum_\beta R_{1;\alpha\beta}\, v_{k,\beta}^0, \tag{3}$$

$$v_\alpha^3 = u_\alpha^3 + \frac{1}{N}\sum_\beta R_{2,4;\alpha\beta}\, v_\beta^3, \tag{4}$$

$$v_{k,\alpha}^4 = u_{k,\alpha}^4 + \sum_\beta R_{3;\alpha\beta}\, v_{k,\beta}^2. \tag{5}$$

Here $\epsilon_\alpha \equiv \epsilon_{(n,t)}$ denotes the effective single-site label-noise variable associated with the minibatch at epoch-step index $\alpha = (n,t)$, inherited from the minibatch noise vector $\boldsymbol{\epsilon}(t)$ in the original dynamics.

The zero-mean Gaussian fields have covariances

$$\begin{aligned}
\left\langle u^1_\alpha u^1_\beta \right\rangle &= [\boldsymbol{C}_0]_{\alpha\beta}, \\
\left\langle u^2_{k,\alpha} u^2_{\ell,\beta} \right\rangle &= \delta_{k\ell} \frac{\lambda_k}{B} [\mathcal{D}(\boldsymbol{C}_1)]_{\alpha\beta}, \\
\left\langle u^3_\alpha u^3_\beta \right\rangle &= [\boldsymbol{C}_2]_{\alpha\beta}, \\
\left\langle u^4_{k,\alpha} u^4_{\ell,\beta} \right\rangle &= \delta_{k\ell} \frac{1}{N} [\boldsymbol{C}_3]_{\alpha\beta}.
\end{aligned} \tag{6}$$

All omitted cross-covariances vanish. The label-noise field satisfies

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

Thus the label noise is correlated across epochs for the same within-epoch minibatch label, but uncorrelated across different within-epoch steps. The fixed-minibatch reuse structure enters through $\mathcal{D}$, which keeps only the same within-epoch minibatch label across different epochs.

The matrices $\boldsymbol{C}_i$ and $\boldsymbol{R}_{i,j}$ appearing in (1)–(6) are determined self-consistently. They are the DMFT order parameters. The correlation functions describe time-time covariances, while the response functions measure the sensitivity of the effective process to small dynamical perturbations.

### B.3. Correlation functions

We next define the correlation functions that enter the Gaussian covariances. They are

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

The test and train losses are read from the diagonal entries

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

From the single-site processes for $v^1$ and $v^3$, it is convenient to write

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

The operator $\mathcal{D}$ is applied to $\boldsymbol{C}_1$ when forming the covariance of $u^2$, reflecting the fact that only the same fixed minibatch label is reused across epochs.

### B.4. Response functions

We next define the response functions. We use the same convention as in the standard random-feature DMFT: $R_{0,2}$ measures the response of $v^0$ to a kick in the $v^2$ dynamics, $R_1$ the response of $v^1$ to a kick in its own dynamics, $R_{2,4}$ the response of $v^2$ to a kick in the $v^4$ dynamics, and $R_3$ the response of $v^3$ to a kick in its own dynamics.

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 (2) and (4), these responses obey

$$\boldsymbol{R}_1 = \left[ \boldsymbol{I} - \frac{1}{B} \mathcal{D}(\boldsymbol{R}_{0,2}) \right]^{-1}, \tag{7}$$
$$\boldsymbol{R}_3 = \left[ \boldsymbol{I} - \frac{1}{N} \boldsymbol{R}_{2,4} \right]^{-1}. \tag{8}$$

These equations express the feedback of a single added data site or random feature site through the rest of the system.

### B.5. Mode transfer functions

We now combine the equations for $v^0$, $v^2$, and $v^4$. From (1), (3), and (5), the residual for each mode satisfies

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

This identifies the deterministic signal term, the fluctuation terms, and the self-consistent feedback of the same mode. Therefore

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

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:

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

Substituting these into (7)–(8) gives the closed response equations

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

### B.6. Vectorized multi-epoch DMFT closure

We now collect the equations in the vectorized form used for numerical solution. All matrices below act on the $S = T\mathcal{E}$-dimensional epoch-step space. Define

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

This matrix is the effective propagation of the data-block Gaussian field $\boldsymbol{u}_k^2$ into $\boldsymbol{v}_k^2$.

The closed response equations are

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

The closed correlation equations are

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

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

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

The additive $\sigma^2$ is the irreducible label-noise contribution to the test mean-squared error. In noiseless experiments, this term is absent.

## Appendix C. Ensembling Theory: Replicated DMFT

We now extend the DMFT to ensembles. Consider $E$ independently initialized random-feature models indexed by $e = 1, \dots, E$. Each replica has an independent random feature matrix $\boldsymbol{A}^{(e)}$, but the replicas are trained on the same dataset. In the fixed-order setting studied here, they also share the same minibatch order.

For replica $e$, 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 replica average

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

Its covariance 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. Replica-symmetric decomposition

One can work out the full coupled set of DMFT equations for the $\sim E^2$ correlation and response functions. Under these equations, one finds that the response functions decouple $R_{e,e'} = \delta_{e,e'} R$. Further, one can easily show from the structure of the stochastic process that the saddle point equations for the correlation functions are replica-symmetric:

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

where s denotes the self-replica covariance and c the cross-replica covariance. Hence

$$\bar{\boldsymbol{C}}_0 = \frac{1}{E}\boldsymbol{C}_0^{\text{self}} + \frac{E-1}{E}\boldsymbol{C}_0^{\text{cross}} = \boldsymbol{C}_0^{\text{cross}} + \frac{1}{E}\left(\boldsymbol{C}_0^{\text{self}} - \boldsymbol{C}_0^{\text{cross}}\right). \tag{9}$$

Thus ensembling suppresses only the part of the covariance that is independent across replicas.

### C.2. Replicated effective process

Each replica obeys the same effective single-site process as in Appendix B.6:

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

The response operators $\boldsymbol{H}_k, \boldsymbol{R}_1, \boldsymbol{R}_3$ are single-replica objects and therefore do not depend on $E$.

The distinction between self and cross sectors comes entirely from the Gaussian forcing. The model-noise fields are independent across replicas because the feature matrices $\boldsymbol{A}^{(e)}$ are independent:

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

By contrast, the data-side Gaussian fields are correlated across replicas when the dataset and mini-batch 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 self and cross sectors:

$$\boldsymbol{C}_1^{\text{self}} = \boldsymbol{R}_1 \boldsymbol{C}_0^{\text{self}} \boldsymbol{R}_1^\top + \sigma^2 \boldsymbol{R}_1 \mathbf{1}\mathbf{1}^\top \boldsymbol{R}_1^\top, \qquad \boldsymbol{C}_1^{\text{cross}} = \boldsymbol{R}_1 \boldsymbol{C}_0^{\text{cross}} \boldsymbol{R}_1^\top + \sigma^2 \boldsymbol{R}_1 \mathbf{1}\mathbf{1}^\top \boldsymbol{R}_1^\top.$$

If label noise were independently resampled across replicas, the $\sigma^2$-term would appear only in the self sector.

### C.3. Self sector

The self-replica equations are exactly the single-replica DMFT equations with $\boldsymbol{C}_0, \boldsymbol{C}_1, \boldsymbol{C}_2$ replaced by $\boldsymbol{C}_0^{\text{self}}, \boldsymbol{C}_1^{\text{self}}, \boldsymbol{C}_2^{\text{self}}$:

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

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

### C.4. Cross sector

For $e \neq e'$, the model-noise covariance vanishes:

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

Therefore the $1/N$ model-noise term drops out of the cross-replica covariance. The cross covariance satisfies

$$\boldsymbol{C}_0^{\text{cross}} = \sum_k \lambda_k \boldsymbol{H}_k \left[ (w_k^\star)^2 \mathbf{1}\mathbf{1}^\top + \frac{\eta^2 \lambda_k}{B} \boldsymbol{\Theta} \boldsymbol{R}_3 \mathcal{D}(\boldsymbol{C}_1^{\text{cross}}) \boldsymbol{R}_3^\top \boldsymbol{\Theta}^\top \right] \boldsymbol{H}_k^\top. \tag{10}$$

The corresponding cross auxiliary covariance 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 $\boldsymbol{C}_0^{\text{cross}}$, since the $1/N$ model-noise channel is replica-independent.

### C.5. Self-minus-cross form

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 shared label-noise contribution cancels from $\Delta \boldsymbol{C}_0$. This form makes the ensemble mechanism explicit: independent model noise generates $\Delta \boldsymbol{C}_0$, while the data/minibatch feedback channel can amplify it across training time.

### C.6. Ensemble loss

Combining the replica-symmetric decomposition with (9), the ensemble covariance is

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

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

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

The limiting cases are immediate. For $E = 1$, $\bar{\boldsymbol{C}}_0 = \boldsymbol{C}_0^{\text{self}}$, recovering the single-model DMFT. As $E \to \infty$, $\bar{\boldsymbol{C}}_0 \to \boldsymbol{C}_0^{\text{cross}}$: independent model-initialization/random-feature variance is removed, while the covariance arising from shared data and shared labels remains.

**Independent shuffles.** The equations above correspond to fixed-order training with the same minibatch order in every replica. If each replica uses an independent data shuffle, the cross-replica operator $\mathcal{D}$ in (10) should be replaced by a cross-overlap operator $\mathcal{D}_{\text{cross}}$ measuring the probability that two replicas see the same example at the same update time. Heuristically, for independent random shuffles, this overlap is $O(1/T)$, so the cross data feedback is strongly reduced. This is why ensembles over both initialization and shuffle randomness can reduce the late-time growth more effectively than ensembles over initialization alone.

## Appendix D. Continuum Limits: Noisy SGD Theory vs Gradient Flow

We now discuss two small-learning-rate limits of the discrete multi-epoch DMFT. The two limits differ in the relative rate at which the learning rate is sent to zero compared to the within-epoch discretization. The first limit keeps the epoch index discrete and produces a noisy continuum-time SGD theory. The second limit averages over infinitely many infinitesimal passes through the data and recovers gradient flow on the empirical loss.

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

We first consider the formal scaling

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

Equivalently, we introduce the continuum within-epoch time

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

The scaling $B = P/T$ should be understood as a continuum idealization of fixed-data SGD rather than as a literal limit of integer-valued minibatch sizes. Each individual minibatch contains an infinitesimal fraction of the dataset, but one epoch still processes a total of $P$ examples. In this limit, the deterministic drift remains $O(1)$ over one epoch, while the effective data-noise covariance remains $O(P^{-1})$. The resulting continuum theory is therefore not gradient flow; it is a noisy SGD theory with a finite data-noise channel.

We use the combined index

$$\alpha = (n, \tau), \qquad n \in \mathbb{N}, \qquad \tau \in [0, 1].$$

The continuum causal operator is the integral operator

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

Equivalently, its kernel is

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

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

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

Thus $\mathcal{D}$ is local in the within-epoch time $\tau$, while still allowing correlations across different epochs $n, m$. This is the continuum analogue of reusing the same minibatch label at the same within-epoch time across epochs.

The corresponding single-site stochastic process is

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

Here all products denote integral-operator products over $(n, \tau)$. The Gaussian fields have covariances

$$\begin{aligned}
\left\langle u^1_n(\tau) u^1_m(\tau') \right\rangle &= C_{0;nm}(\tau, \tau'), \\
\left\langle u^2_{k,n}(\tau) u^2_{\ell,m}(\tau') \right\rangle &= \delta_{k\ell} \frac{\lambda_k}{P} [\mathcal{D}(\boldsymbol{C}_1)]_{nm}(\tau, \tau'), \\
\left\langle u^3_n(\tau) u^3_m(\tau') \right\rangle &= C_{2;nm}(\tau, \tau'), \\
\left\langle u^4_{k,n}(\tau) u^4_{\ell,m}(\tau') \right\rangle &= \delta_{k\ell} \frac{1}{N} C_{3;nm}(\tau, \tau').
\end{aligned}$$

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

$$\begin{aligned}
\boldsymbol{C}_1 &= \boldsymbol{R}_1 \left( \boldsymbol{C}_0 + \sigma^2 \mathbf{1}\mathbf{1}^\top \right) \boldsymbol{R}_1^\top, \qquad \boldsymbol{C}_3 = \boldsymbol{R}_3 \boldsymbol{C}_2 \boldsymbol{R}_3^\top, \\
\boldsymbol{C}_0 &= \sum_k \lambda_k \boldsymbol{H}_k \Bigg[ (w^\star_k)^2 \mathbf{1}\mathbf{1}^\top + \frac{1}{N} \boldsymbol{\Theta} \boldsymbol{R}_3 \boldsymbol{C}_2 \boldsymbol{R}_3^\top \boldsymbol{\Theta}^\top \\
&\qquad\qquad + \frac{\lambda_k}{P} \boldsymbol{\Theta} \boldsymbol{R}_3 \mathcal{D}(\boldsymbol{C}_1) \boldsymbol{R}_3^\top \boldsymbol{\Theta}^\top \Bigg] \boldsymbol{H}_k^\top, \\
\boldsymbol{C}_2 &= \sum_k \lambda_k^2 \boldsymbol{R}_1 \boldsymbol{H}_k \left[ (w^\star_k)^2 \mathbf{1}\mathbf{1}^\top + \frac{1}{N} \boldsymbol{\Theta} \boldsymbol{R}_3 \boldsymbol{C}_2 \boldsymbol{R}_3^\top \boldsymbol{\Theta}^\top \right] \boldsymbol{H}_k^\top \boldsymbol{R}_1^\top \\
&\quad + \frac{1}{P} \sum_k \lambda_k \boldsymbol{M}_k \mathcal{D}(\boldsymbol{C}_1) \boldsymbol{M}_k^\top.
\end{aligned}$$

This limit keeps the epoch structure explicit. Consequently, the response and correlation functions are not time-translation invariant in the global epoch-step index. The operator $\mathcal{D}$ is the signature of fixed-order multi-epoch SGD: it preserves the fact that the same infinitesimal minibatch label is reused at the same within-epoch time in later epochs.

### D.1.1. GRADIENT-FLOW LIMIT

We next consider a smaller-learning-rate limit in which many infinitesimal passes through the dataset occur on an $O(1)$ macroscopic time scale. Let

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

One epoch has duration

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

Thus, on an $O(1)$ interval of macroscopic time $u$, the algorithm makes $\mathcal{N}$ complete infinitesimal passes through the dataset. The minibatch-noise variance accumulated over $O(1)$ time is $O((P\mathcal{N})^{-1})$, so the stochastic minibatch fluctuations vanish as $\mathcal{N} \to \infty$. The limiting dynamics are therefore gradient flow on the empirical loss:

$$\partial_u \boldsymbol{v}^0(u) = -\left(\frac{1}{N}\boldsymbol{A}^\top \boldsymbol{A}\right)\left(\frac{1}{P}\boldsymbol{\Psi}^\top \boldsymbol{\Psi}\right)\boldsymbol{v}^0(u).$$

The auxiliary fields are now defined using the full dataset:

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

with

$$\partial_u \boldsymbol{v}^0(u) = -\boldsymbol{v}^4(u).$$

The corresponding single-site stochastic process is

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

The Gaussian fields have covariances

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

Compared to the noisy-SGD continuum limit, the projection operator $\mathcal{D}$ is absent. The full dataset is averaged at every infinitesimal time, so the data-noise covariance is the full two-time covariance $C_1(u, u')$, not its fixed-minibatch diagonal projection.

Let

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

In operator notation, the response equations are

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

The correlation equations are

$$\begin{aligned}
\boldsymbol{C}_1 &= \boldsymbol{R}_1 \left(\boldsymbol{C}_0 + \sigma^2 \mathbf{1}\mathbf{1}^\top\right) \boldsymbol{R}_1^\top, \qquad \boldsymbol{C}_3 = \boldsymbol{R}_3 \boldsymbol{C}_2 \boldsymbol{R}_3^\top, \\
\boldsymbol{C}_0 &= \sum_k \lambda_k \boldsymbol{H}_k \Bigg[ (w_k^\star)^2 \mathbf{1}\mathbf{1}^\top + \frac{1}{N} \boldsymbol{\Theta} \boldsymbol{R}_3 \boldsymbol{C}_2 \boldsymbol{R}_3^\top \boldsymbol{\Theta}^\top \\
&\qquad\qquad + \frac{\lambda_k}{P} \boldsymbol{\Theta} \boldsymbol{R}_3 \boldsymbol{C}_1 \boldsymbol{R}_3^\top \boldsymbol{\Theta}^\top \Bigg] \boldsymbol{H}_k^\top, \\
\boldsymbol{C}_2 &= \sum_k \lambda_k^2 \boldsymbol{R}_1 \boldsymbol{H}_k \left[ (w_k^\star)^2 \mathbf{1}\mathbf{1}^\top + \frac{1}{N} \boldsymbol{\Theta} \boldsymbol{R}_3 \boldsymbol{C}_2 \boldsymbol{R}_3^\top \boldsymbol{\Theta}^\top \right] \boldsymbol{H}_k^\top \boldsymbol{R}_1^\top \\
&\quad + \frac{1}{P} \sum_k \lambda_k \boldsymbol{M}_k \boldsymbol{C}_1 \boldsymbol{M}_k^\top,
\end{aligned} \tag{11}$$

where

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

In the stationary response closure, the response functions are time-translation invariant:

$$R(u, u') = R(u - u').$$

Using the Fourier convention

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

the response equations reduce to

$$\begin{aligned}
H_k(\omega) &= \frac{1}{i\omega + \lambda_k R_1(\omega) R_3(\omega)}, \\
R_1(\omega) &= 1 - \frac{1}{P} \sum_k \frac{\lambda_k R_1(\omega) R_3(\omega)}{i\omega + \lambda_k R_1(\omega) R_3(\omega)}, \\
R_3(\omega) &= 1 - \frac{1}{N} \sum_k \frac{\lambda_k R_1(\omega) R_3(\omega)}{i\omega + \lambda_k R_1(\omega) R_3(\omega)}.
\end{aligned}$$

Even under this stationary response closure, the correlation functions are not time-translation invariant because they continue to encode the signal contribution inherited from the initialization. We therefore use a two-frequency Fourier transform:

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

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

$$C_0(\omega,\omega') = \sum_k \lambda_k H_k(\omega) H_k(\omega') \left[ (w_k^\star)^2 + \frac{1}{N} C_3(\omega,\omega') + \frac{\lambda_k}{P} R_3(\omega) R_3(\omega') C_1(\omega,\omega') \right],$$
$$C_1(\omega,\omega') = R_1(\omega) R_1(\omega') C_0(\omega,\omega'),$$
$$C_2(\omega,\omega') = \sum_k \lambda_k H_k(\omega) H_k(\omega') \left[ \frac{1}{P}(i\omega)(i\omega') C_1(\omega,\omega') + \lambda_k R_1(\omega) R_1(\omega') \left( (w_k^\star)^2 + \frac{1}{N} C_3(\omega,\omega') \right) \right],$$

$$C_3(\omega,\omega') = R_3(\omega) R_3(\omega') C_2(\omega,\omega').$$

Label noise can be restored by replacing the $C_1$ source in the time-domain equations by $\boldsymbol{C}_0 + \sigma^2 \mathbf{1}\mathbf{1}^\top$, as in (11).

The test and train losses are obtained from the diagonal:

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

## Appendix E. Scaling form for interpolation-induced late-time growth

We now use the low-frequency form of the multi-epoch DMFT equations to study the late-time variance growth near the interpolation threshold. The goal is to identify the low-frequency denominator whose behavior controls the size of the correlation functions. We focus on the interpolation threshold $N = P$, where the model size and data size are balanced. At this threshold the finite-data and finite-width response functions become comparable. In the diagonal low-frequency approximation we 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

$$R(\omega) = 1 - \frac{1}{P} R(\omega)^2 \sum_k \lambda_k H_k(\omega). \tag{12}$$

This is the interpolation-threshold specialization of the low-frequency response equation in the random-feature DMFT.

**Fourier-domain correlation equations.** Recall that $C_0(\omega,\omega')$ and $C_2(\omega,\omega')$ are the double Fourier transforms of the correlation functions associated with $\boldsymbol{v}^0$ and $\boldsymbol{v}^2$, respectively. Here $\boldsymbol{v}^0$ is the discrepancy between the target weights and the model's effective weights, and $\boldsymbol{v}^2 = (1/B)\boldsymbol{\Psi}^\top \boldsymbol{v}^1$ is the auxiliary field after the mini-batch feature block.

Keeping the coupled $C_0$- and $C_2$-equations that control the late-time growth near the interpolation threshold, the equations read

$$C_0(\omega,\omega') = B_0(\omega,\omega') + \frac{1}{P}T_2(\omega,\omega')R(\omega)^2R(\omega')^2C_0(\omega,\omega') + \frac{1}{P}R(\omega)R(\omega')T_1(\omega,\omega')C_2(\omega,\omega'), \tag{13}$$

$$C_2(\omega,\omega') = B_2(\omega,\omega') + \frac{1}{P}(i\omega)(i\omega')T_1(\omega,\omega')R(\omega)R(\omega')C_0(\omega,\omega') + \frac{1}{P}T_2(\omega,\omega')R(\omega)^2R(\omega')^2C_2(\omega,\omega'). \tag{14}$$

Here

$$B_0(\omega,\omega') = \sum_k \lambda_k (w_k^\star)^2 H_k(\omega)H_k(\omega'),$$
$$B_2(\omega,\omega') = R(\omega)R(\omega')\sum_k \lambda_k^2 (w_k^\star)^2 H_k(\omega)H_k(\omega'),$$

and

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

The $T_q$ are spectral trace functions: $T_2$ controls the variance feedback to $C_0$ and $C_2$ through the random-feature projection, while $T_1$ controls the cross-coupling between them. The terms $B_0$ and $B_2$ are deterministic bias terms determined by the target coefficients $\{w_k^\star\}$ in the feature eigenbasis. The remaining terms describe the feedback of finite-data and finite-width fluctuations through the response functions.

Equations (13)–(14) can be written as

$$\boldsymbol{G}(\omega,\omega')\begin{pmatrix} C_0(\omega,\omega') \\ C_2(\omega,\omega') \end{pmatrix} = \begin{pmatrix} B_0(\omega,\omega') \\ B_2(\omega,\omega') \end{pmatrix}, \tag{15}$$

where

$$\boldsymbol{G}(\omega,\omega') = \begin{pmatrix} 1-\frac{1}{P}T_2R(\omega)^2R(\omega')^2 & -\frac{1}{P}R(\omega)R(\omega')T_1 \\ -\frac{1}{P}(i\omega)(i\omega')R(\omega)R(\omega')T_1 & 1-\frac{1}{P}T_2R(\omega)^2R(\omega')^2 \end{pmatrix}. \tag{16}$$

For compactness, $T_1, T_2$ in (16) denote $T_1(\omega,\omega')$ and $T_2(\omega,\omega')$. The asymmetry between the off-diagonal entries comes from the fact that the feedback from $C_0$ into $C_2$ is suppressed by $(i\omega)(i\omega')$, whereas the feedback from $C_2$ into $C_0$ is not. Equivalently, in the time-domain equations, the contribution to $C_2$ involves time derivatives of the discrepancy correlation, while the contribution to $C_0$ is obtained after propagating the auxiliary fluctuations through the response.

Solving (15) for $C_0$ introduces a denominator

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

which controls the multiplicative variance feedback. The interpolation-induced late-time growth is governed by the small-$\omega$ behavior of this denominator.

**Diagonal low-frequency reduction.** To extract the equal-time scaling, we examine the regime

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

For the scaling calculation we set $\omega' = \omega$ after deriving the two-frequency system. Then

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

and

$$\det \boldsymbol{G}(\omega) = \left(1 - \frac{1}{P} T_2(\omega) R(\omega)^4\right)^2 - \frac{1}{P^2} (i\omega)^2 R(\omega)^4 T_1(\omega)^2. \tag{17}$$

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 (12),

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

Hence

$$\frac{1}{P} R(\omega)^4 T_2(\omega) = 1 - R(\omega) - \frac{1}{P} i\omega R(\omega)^2 T_1(\omega). \tag{18}$$

Substituting (18) into (17) gives

$$\begin{aligned}1 - \Gamma(\omega, \omega) &= \left[R(\omega) + \frac{1}{P} i\omega R(\omega)^2 T_1(\omega)\right]^2 - \frac{1}{P^2} (i\omega)^2 R(\omega)^4 T_1(\omega)^2 \\ &= R(\omega)^2 + \frac{2}{P} i\omega R(\omega)^3 T_1(\omega).\end{aligned} \tag{19}$$

**Low-frequency scaling of the response.** At the interpolation threshold, the low-frequency solution has the scaling form

$$R(\omega)^2 = i\omega r^2 + o(\omega), \qquad \omega \to 0, \tag{20}$$

where $r$ is determined by

$$1 = \frac{1}{P} \sum_k \frac{\lambda_k r^2}{1 + \lambda_k r^2}. \tag{21}$$

Indeed, substituting $R(\omega)^2 = i\omega r^2$ into (12) gives (21) at leading order.

**Power-law assumption/specialization.** 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 [5, 6, 8, 27]:

$$\lambda_k \sim k^{-b}, \qquad \lambda_k (w_k^\star)^2 \sim k^{-a}, \qquad b > 1, \quad a > 1. \tag{22}$$

Here, $b$ is the spectral decay exponent and $a$ is the task-power exponent. The conditions $b > 1$ and $a > 1$ ensure that the spectral and target-energy tails appearing below are summable.

Let $k_\star$ be the cutoff satisfying

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

Using $\lambda_k \sim k^{-b}$ from (22), this gives

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

The sum in (21) is of order $k_\star$, so $k_\star \sim P$. Therefore

$$r^2 \sim P^b, \qquad r \sim P^{b/2}. \tag{23}$$

Next,

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

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

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

Thus

$$T_1(\omega) \sim (i\omega)^{-2} P^{1-b}. \tag{24}$$

Substituting (20), (23), and (24) into (19) gives

$$\begin{aligned} 1 - \Gamma(\omega, \omega) &= R(\omega)^2 + \frac{2}{P} i\omega R(\omega)^3 T_1(\omega) \\ &\sim P^b (i\omega) + \frac{2}{P} (i\omega)(i\omega)^{3/2} P^{3b/2} (i\omega)^{-2} P^{1-b} \\ &\sim P^b (i\omega) + 2P^{b/2} (i\omega)^{1/2}. \end{aligned} \tag{25}$$

The two terms in (25) dominate in different regimes. Their ratio 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 the relevant Fourier scale at training step $s$ is $|\omega| \sim s^{-1}$, the interpolation-induced growth governs the dynamics only at late times

$$s \gtrsim P^b.$$

For $s \ll P^b$, the analytic term $P^b(i\omega)$ dominates, so the dynamics remains in the standard pre-onset regime governed by mode-wise relaxation. In particular, one recovers the usual time-bottleneck decay. Therefore the square-root growth derived below should be understood as a genuinely late-time interpolation-threshold effect, rather than as the early-time behavior.

**Derivation of the standard time-bottleneck decay.** We briefly justify the scaling

$$\mathcal{L}_{\text{bias}}(s) \sim s^{-(a-1)/b}.$$

In the pre-onset regime $s \ll P^b$, the interpolation-induced feedback has not yet become dominant, and the dynamics is governed by the standard mode-wise relaxation mechanism. Since

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

the $k$-th mode relaxes on the timescale

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

Hence, by time $s$, only modes with

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

have been significantly learned, while the higher modes remain unresolved.

Using the power-law task assumption

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

the residual bias is dominated by the unresolved tail,

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

Approximating the sum by an integral gives

$$\mathcal{L}_{\text{bias}}(s) \sim \int_{s^{1/b}}^{\infty} k^{-a}\, dk \sim s^{-(a-1)/b},$$

which is the standard time-bottleneck exponent in power-law random feature models.

In the late-time regime $s \gtrsim P^b$, by contrast, the square-root term in (25) becomes dominant. In this regime,

$$1 - \Gamma(\omega, \omega) \sim P^{b/2} (i\omega)^{1/2}. \tag{26}$$

Thus the interpolation-induced contribution has a $s^{1/2}$ time dependence, with a prefactor $P^{-b/2}$ coming from the inverse denominator. This is the late-time growth mechanism that is absent in the standard one-pass theory and is generated here by repeated data reuse near the interpolation threshold.

**Bias scale.** It remains to identify the scale of the bias term. The leading bias term entering $C_0$ is

$$B_0(\omega,\omega) = \sum_k \lambda_k (w_k^\star)^2 H_k(\omega)^2.$$

Near the interpolation threshold, using $R(\omega)^2 = i\omega r^2$, we get

$$B_0(\omega,\omega) = \frac{1}{(i\omega)^2} \sum_k \frac{\lambda_k (w_k^\star)^2}{(1+\lambda_k r^2)^2}.$$

Using the task-power scaling $\lambda_k (w_k^\star)^2 \sim k^{-a}$ from (22), the cutoff is again $k_\star \sim P$. Hence the inner bias sum scales as

$$\mathcal{B}_0(P) := \sum_k \frac{\lambda_k (w_k^\star)^2}{(1+\lambda_k r^2)^2} \sim r^{-4} \sum_{k\le P} \frac{\lambda_k (w_k^\star)^2}{\lambda_k^2} + \sum_{k>P} \lambda_k (w_k^\star)^2.$$

Under the usual source-limited regime $a-1 < 2b$, the dominant contribution comes from the modes $k > P$, i.e., the part of the target spectrum not resolved at the effective cutoff $k_\star \sim P$:

$$\mathcal{B}_0(P) \sim \sum_{k>P} k^{-a} \sim P^{-(a-1)}. \tag{27}$$

This is the same finite-$P$ model/data bottleneck scale appearing in the standard random-feature learning curve. If $a-1 \ge 2b$, the bias scale is instead limited by the response cutoff and should be replaced by the corresponding $P^{-2b}$-type saturation scale.

**Resulting scaling form.** Combining (26) and (27), in the late-time regime $s \gtrsim P^b$, the corresponding time-domain contribution to the test loss scales as

$$\frac{1}{E}\, P^{-b/2}\, \mathcal{B}_0(P)\, s^{1/2} \;\sim\; \frac{1}{E}\, P^{-(a-1+b/2)}\, s^{1/2}. \tag{28}$$

The factor $E^{-1}$ applies only to fluctuations that are independent across ensemble members, such as independent random-projection fluctuations. Fluctuations coming from a shared dataset or a shared fixed mini-batch order are common across members and therefore are not reduced by ensembling.

A local scaling form near the interpolation threshold is therefore

$$\mathcal{L}(s,P,E) \approx \sigma^2 + c_1\, s^{-(a-1)/b} + c_2\, P^{-(a-1)} + \frac{c_3}{E}\, P^{-(a-1)}\, \Phi\!\left(\frac{s}{P^b}\right), \tag{29}$$

where $c_1, c_2, c_3$ are positive constants independent of $s, P, E$. Here $\sigma^2$ is the irreducible label-noise floor; in a well-specified noiseless setting, $\sigma = 0$. The crossover function $\Phi$ satisfies

$$\Phi(x) \to 0 \quad (x \to 0), \qquad \Phi(x) \sim x^{1/2} \quad (x \to \infty).$$

The last term in (29) separates the bias scale $P^{-(a-1)}$, coming from $\mathcal{B}_0(P)$ in (27), from the late-time denominator scale $P^{-b/2}$, coming from $1-\Gamma$ in (26). Indeed, for $s \gg P^b$,

$$\frac{c_3}{E}\, P^{-(a-1)}\, \Phi\!\left(\frac{s}{P^b}\right) \;\sim\; \frac{c_3}{E}\, P^{-(a-1+b/2)}\, s^{1/2},$$

recovering (28).

Equation (29) explains the observed non-monotone loss curves near the interpolation threshold. For $s \ll P^b$, the decreasing time-bottleneck term dominates. For $s \gtrsim P^b$, the interpolation-induced $s^{1/2}$ growth becomes visible and can drive the loss upward. The two central testable consequences are the onset scale $s \sim P^b$ across different values of $P$, and the $1/E$ reduction under ensembling over independent random projections. These predictions are examined empirically in the experiments, including Figure 1.

The scaling form is written in terms of the global step $s$ and the total number of distinct training samples $P$. The number of epochs $\mathcal{E}$ enters only through the maximal training time $s_{\max} = T\mathcal{E}$, where $T = P/B$ is the number of mini-batches per epoch. Thus the late-time regime is observable only if $T\mathcal{E} \gtrsim P^b$.