A SOLVABLE MODEL OF ADAPTIVE LEARNING RATE RESCALING: ACCELERATION, STABILITY & SCALING

Anonymous authors
Paper under double-blind review

ABSTRACT

A recurring design principle in modern optimizers is to decouple update magnitude from the raw gradient norm, yet its consequences for learning-curve and resource scaling remain unclear. We isolate this mechanism by studying normalized SGD in a random-feature model with power-law teacher and data covariance. Fixed-norm updates induce an effective learning rate that grows as gradients shrink. We derive a dynamical mean-field theory (DMFT) describing the joint dependence of the loss on training time, model width and batch size. Normalization initially accelerates SGD, mapping the power-law exponent $r_{\mathrm{SGD}} < 1$ to $2r_{\mathrm{SGD}}/(1 - r_{\mathrm{SGD}})$, with exponential convergence at $r_{\mathrm{SGD}} = 1$ and formal finite-time convergence for $r_{\mathrm{SGD}} > 1$. At finite step size, however, the same feedback ultimately breaks the acceleration and leads to marginal stability. The late-time theory yields width-limited, edge-of-stochastic-stability (EoSS), and deterministic edge-of-stability (EoS) regimes. These phases determine when larger batches or wider models reduce serial training time at comparable compute. We quantify in which of these phases increased batch size or width can compensate the excess compute use per step by fewer optimization steps to target loss. Linearized ResNet experiments on CIFAR-5M support the predicted acceleration, breakdown, and resource-scaling trends. Together, these results connect normalization-induced acceleration, EoS effects, and width–batch allocation within a solvable theory.

1 INTRODUCTION

Neural scaling laws describe how performance improves predictably with training time, model size and data, and provide a quantitative basis for allocating compute across these resources (Hoffmann et al., 2022; Kaplan et al., 2020). Several theoretical studies connect these laws to the spectral structure of data and tasks: learning proceeds progressively through the data spectrum, while finite training time, model size, and data impose different spectral cutoffs (Bahri et al., 2024b; Maloney et al., 2022; Bordelon et al., 2020; 2024a; Paquette et al., 2024). Only recently have studies begun to examine how these spectral learning dynamics interact with optimizer design (Paquette et al., 2026; Xiao et al., 2025), and many aspects of this interaction remain poorly understood.

Modern adaptive optimizers such as RMSProp, Adam, SOAP, and Muon induce trajectory-dependent effective learning rates through gradient-dependent preconditioning, even when the base learning rate is fixed (Tieleman & Hinton, 2012; Kingma & Ba, 2014; Vyas et al., 2025; Jordan et al., 2024). This rescaling changes both update direction and magnitude; grafting makes their distinct roles explicit by combining the direction of one optimizer with the magnitude of another (Agarwal et al., 2020). We isolate the feedback between update magnitude and learning dynamics by studying normalized SGD, which preserves the stochastic-gradient direction while fixing the update norm. With the normalization used below, this is equivalent to SGD with an endogenous scalar learning rate $\eta_{\mathrm{eff}}(t) = \eta\|\mathbf{g}(t)\|_2$, providing a tractable setting for studying how adaptive rescaling affects convergence and stability.

Gradient normalization creates a tension between acceleration and stability. As progressively weaker spectral modes are learned, the gradient norm decays and the effective learning rate grows, potentially accelerating progress through the spectral tail. At finite step size, however, the same growth eventually brings the leading spectral mode to the discrete-time stability threshold. The deterministic edge-of-stability (EoS) corresponds to marginal stability of the leading Hessian mode (Cohen et al.,

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.

2021), while the edge-of-stochastic-stability (EoSS) is associated with a stability limit imposed by minibatch fluctuations (Andreyev & Beneventano, 2024). The spectral tail therefore controls the transient acceleration, while the spectral edge controls its breakdown.

These stability limits also shape the computational value of increasing training resources. A larger batch can support a larger effective learning rate, but each update is more expensive (McCandlish et al., 2018; Shallue et al., 2019; Zhang et al., 2025; Meterez et al., 2026c); increasing width changes both approximation capacity and the spectrum that controls stability. Understanding these trade-offs requires distinguishing gains in effective step size from gains per unit compute. Our model makes this distinction explicit and connects the resulting predictions to experiments with linearized ResNets on CIFAR-5M (Nakkiran et al., 2021), using spectral exponents measured from the neural tangent kernel and target alignment.

We study a power-law random-feature model, which captures key features of empirically observed neural scaling laws while remaining analytically tractable (Maloney et al., 2022; Bordelon et al., 2024a). Within this setting, we ask:

How does adaptive learning rate shapes acceleration, and how do capacity and stability limits determine the benefits of increasing batch size and model width?

Our main contributions are:

- We derive a dynamical mean field theory (DMFT) for normalized SGD in a high-dimensional power-law random-feature model, characterizing the joint scaling of the loss with training time, model width, batch size, and finite step size.
- We characterize source- and capacity-dependent transient scaling of norm-SGD: normalization accelerates the SGD exponent r_{SGD} to $2r_{\mathrm{SGD}}/(1 - r_{\mathrm{SGD}})$ for $r_{\mathrm{SGD}} < 1$, while $r_{\mathrm{SGD}} = 1$ and $r_{\mathrm{SGD}} > 1$ yield exponential and finite-time convergence, respectively.
- DMFT yields a spectrum-dependent width-batch phase diagram: in overparameterized models, a width-dependent critical batch separates stochastic edge-of-stability (EoSS) from deterministic edge-of-stability (EoS), whereas underparameterized models are limited by finite-width error.
- We characterize compute–speed trade-offs across batch size and width: in the EoSS phase, a nearly linear batch-scaling regime reduces the steps to a target loss at almost unchanged compute. In the width-limited phase, we identify spectral regimes where the stationary effective learning rate grows superlinearly with width, giving rise to potential width speedups.
- We test the theory using linearized ResNets on CIFAR-5M. Measured kernel and target exponents yield predictions consistent with the accelerated loss decay and its crossover toward SGD-like scaling, while batch and width sweeps support the predicted efficiency trends in the explored regime.

We use norm-SGD to isolate how the feedback between gradient normalization and learning dynamics shapes convergence and stability. Our aim is to understand this mechanism rather than advocate norm-SGD as a practical optimizer.

1.1 Related Work

Theories of Neural Scaling Laws. Empirical scaling laws have been documented across several domains (Hestness et al., 2017), quantifying the dependence of loss on model size, data, and compute (Kaplan et al., 2020; Hoffmann et al., 2022). Theoretical explanations derive power-law exponents from kernel spectra and target alignment, or power-law random feature models (Spigler et al., 2020; Bordelon et al., 2020; Canatar et al., 2021; Bahri et al., 2024b; Maloney et al., 2022; Bahri et al., 2024a) or hierarchal data model (Cagnetta et al., 2026); with recent works extending the spectral framework to training dynamics, compute-optimal regimes and feature learning (Bordelon et al., 2024a; Paquette et al., 2024; Bordelon et al., 2024b).

Optimizers and Scaling. Whether optimizer gains persist with model size and compute has itself become a scaling question. Comparisons among AdamW, Shampoo, SOAP and Muon show that

reported gains depend on tuning, model scale, and training horizon (Loshchilov & Hutter, 2017; Gupta et al., 2018; Liu et al., 2025; Magakyan et al., 2026; Wen et al., 2026; Harari et al., 2026). Time dependent batch-size and learning rate schedules further shape finite-time scaling through update scale and gradient noise, motivating critical-batch scaling, and coupled schedules (Smith et al., 2017; McCandlish et al., 2018; Defazio et al., 2023; Zhang et al., 2025; Tissue et al., 2026; Meterez et al., 2026a;b). High-dimensional theory further resolves optimizer-dependent preconditioning and batch effects (Xiao et al., 2025; Paquette et al., 2026). Grafting and norm-based views separate update direction from magnitude (Agarwal et al., 2020; Bernstein & Newhouse, 2024). Following this decomposition, we study gradient normalization to isolate magnitude control while leaving the SGD direction unchanged.

Critical Batch Size and Compute Efficiency. Increasing batch size trades more computation per update for fewer serial updates, with diminishing returns beyond a problem-dependent "critical batch size" (McCandlish et al., 2018; Shallue et al., 2019). In LLM pretraining, prior work showed that critical batch size scales primarily with training data size rather than model size (Zhang et al., 2025). Batch-size schedules can also replace learning-rate decay while preserving similar learning curves (Smith et al., 2017); Seesaw develops this connection theoretically for noisy linear regression and extends it to normalized SGD under a variance-dominated approximation (Meterez et al., 2026a). Spectral analyses provide complementary perspective: high-dimensional theory identifies a spectrum-dependent batch threshold beyond which the convergence rate of SGD with momentum saturates (Lee et al., 2022). We study these thresholds and limits in normalized SGD, quantifying spectrum-dependent departures from linear effective-step scaling and separating the resulting efficiency consideration from the EoSS-EoS phase transition at which the effective learning rate freezes.

Training Stability and large step size effects. Large learning rates can lead to faster optimization, while driving training toward the boundary of discrete-time stability. On real objectives, GD often exhibits progressive sharpening, followed by edge-of-stability dynamics with persistent oscillations along sharp directions (Cohen et al., 2021; Arora et al., 2022; Cohen et al., 2022; 2024). In SGD, sampling fluctuations modify the stability criterion, producing an edge of stochastic stability governed by batch-dependent directional curvature and related critical step-size transitions (Xing et al., 2018; Mulayoff & Michaeli, 2024; Jastrzebski et al., 2020; Andreyev & Beneventano, 2024; Paquette et al., 2021; Agarwala & Pennington, 2025). Because curvature determines the stable learning-rate range, its scaling with model size is also central to optimal learning-rate transfer. The μP framework establishes such transfer across width and depth (Yang & Hu, 2021; Yang et al., 2021; Bordelon & Pehlevan, 2022; Yang et al., 2024; Bordelon et al., 2024c; Chizat, 2025; Dey et al., 2026), with recent work seeking mechanistic explanations for it through width-consistent sharpness and top spectral modes (Noci et al., 2024; Ghosh et al., 2026; Lauditi et al., 2026).

2 MODEL

We study online normalized SGD in a power-law random-feature model commonly used in theories of neural scaling laws (Maloney et al., 2022; Bahri et al., 2024b; Bordelon et al., 2024a; Paquette et al., 2024; Bodin & Macris, 2021). At each step t, a fresh batch of zero-mean inputs $\{\boldsymbol{\psi}_\mu^t\}_{\mu=1}^B \in \mathbb{R}^d$ is drawn independently with covariance $\mathbb{E}\left[\boldsymbol{\psi}_\mu^t \boldsymbol{\psi}_\rho^{s\top}\right] = \delta_{\mu\rho}\delta_{ts}\boldsymbol{\Lambda}$, where $\Lambda_{kk} = \lambda_k = k^{-b}$. A frozen random matrix $\boldsymbol{A} \in \mathbb{R}^{N\times d}$ projects the inputs onto N features, and $\boldsymbol{w}^t \in \mathbb{R}^N$ is trained on top of them. The target coefficients obey $\beta_k^2 = k^{-2a}$ in the covariance eigenbasis, with $2a + b > 1$ ensuring finite initial loss. Denoting the gradient by $\boldsymbol{g}$, the loss, predictor, and weight updates are

$$L^t := \frac{1}{2B}\sum_{\mu=1}^{B}\left(\boldsymbol{\beta}^\top\boldsymbol{\psi}_\mu^t - f^t\right)^2, \quad f_\mu^t := \frac{1}{\sqrt{N}}\boldsymbol{w}^{t\top}\boldsymbol{A}\boldsymbol{\psi}_\mu^t, \quad \boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta\sqrt{d}\left(\boldsymbol{g}^\top\boldsymbol{g}\right)^{-1/2}\boldsymbol{g}. \quad (1)$$

We take the proportional limit $d, N, B \to \infty$ at fixed width and batch ratios $\nu := \frac{N}{d}$ and $\alpha := \frac{B}{d}$.

In this limit, we use DMFT to reduce the high-dimensional dynamics to effective stochastic processes for individual spectral modes (Sompolinsky & Zippelius, 1981; De Dominicis, 1978; Crisanti & Sompolinsky, 2018; Mignacco et al., 2020; Helias & Dahmen, 2020). Their correlation and response functions, random at finite d, N, B, concentrate to deterministic limits. Equal-time correlations determine observables such as the loss and squared gradient norm C_3^{tt}, hence $\eta_{\text{eff}}^t = \eta/\sqrt{\nu C_3^{tt}}$,

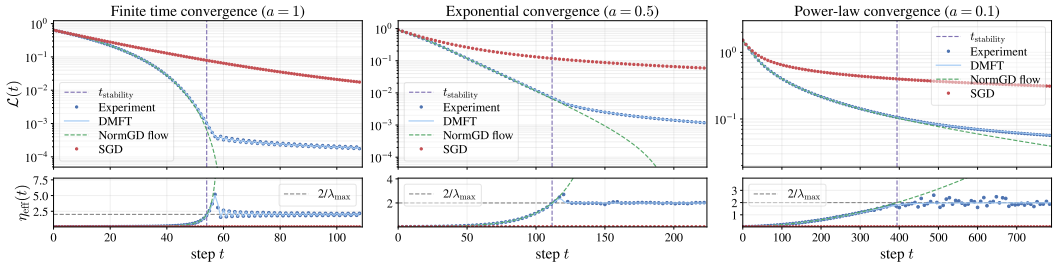


Figure 1: **Three transient convergence regimes.** Loss (**top**) and effective learning rate (**bottom**) at large α, ν. Before t_{stab}, simulations and DMFT follow the continuous-time prediction: formal finite-time convergence for $a > 1/2$ (**left**), exponential for $a = 1/2$ (**center**), and accelerated power-law convergence for $a < 1/2$ (**right**). SGD uses $\eta_{\text{SGD}} = \eta_{\text{eff}}(0)$ to match the initial update norm. At t_{stab}, η_{eff} reaches $2/\lambda_{\max}$ and the discrete dynamics depart from the flow prediction.

while responses encode how perturbations propagate across steps. The derivation and closed set of equations are given in App. B.

3 Continuous Time Dynamics

We first isolate the transient dynamics from the finite-width and minibatch effects that govern late-time behavior. As $\eta \to 0$, normalized SGD approaches $\partial_t \boldsymbol{w}(t) = -\sqrt{d}\left(\boldsymbol{g}^\top(t)\boldsymbol{g}(t)\right)^{-1/2}\boldsymbol{g}(t)$. These updates are parallel to ordinary gradient flow (GF) everywhere but have a different magnitude. Indeed, defining a new clock τ through $\frac{d\tau}{dt} := \left(\frac{g^\top(t)g(t)}{d}\right)^{-1/2}$ gives $\partial_\tau \boldsymbol{w}(\tau) = -\boldsymbol{g}(\tau)$. In this limit, normalization is therefore equivalent to reparameterizing time, with the effective learning rate increasing as the gradient norm decays. Consequently, it changes the temporal scaling of the loss while leaving the width- and data-limited exponents unchanged from GF (see e.g. (Bordelon et al., 2024a)). To extract this temporal scaling, we subsequently take $\alpha, \nu \to \infty$. In the τ-clock, the residual in mode k decays as $\exp(-\lambda_k \tau)$, so its relaxation time is $\tau_k \sim \lambda_k^{-1} \sim k^b$. At time τ, modes $k \lesssim k_\tau \sim \tau^{1/b}$ have been learned, leaving the loss dominated by the unresolved tail. Since each such mode contributes $\lambda_k \beta_k^2 \sim k^{-(2a+b)}$, summing this tail gives $L(\tau) \sim k_\tau^{-(2a+b-1)} \sim \tau^{-r_{\text{SGD}}}$, where $r_{\text{SGD}} = (2a + b - 1)/b$. Reexpressing the same trajectory in physical time gives

$$L(t) \sim \begin{cases} (T-t)^{2\frac{r_{\text{SGD}}}{r_{\text{SGD}}-1}} & a > 1/2 \\ e^{-2\sqrt{b}t} & a = 1/2 \\ t^{-2\frac{r_{\text{SGD}}}{1-r_{\text{SGD}}}} & a < 1/2 \end{cases} \tag{2}$$

(see Appendix E for details). Thus, normalization produces accelerated power-law convergence for $a < 1/2$, exponential convergence at $a = 1/2$, and formal finite-time convergence for $a > 1/2$. Figure 1 verifies these predictions at large but finite α and ν. Before t_{stab}, simulations and DMFT track the continuous-time prediction and outperform constant learning-rate SGD, whose learning rate is set to $\eta_{\text{eff}}(0)$ to match the initial update norm. The lower panels show $\eta_{\text{eff}}(t)$ growing toward $2/\lambda_{\max}$: once this stability threshold is reached, the discrete dynamics depart from the continuous-time prediction and enter the stability-controlled regime analyzed in the next section.

4 Discrete-Time Dynamics: Edge of Stability and Breakdown of the Continuous Description

At finite step size, the effective learning rate grows as the gradient decays until finite width, minibatch fluctuations, or deterministic instability cut off the acceleration. Figure 1 shows η_{eff}^t approaching a constant or stationary mean, denoted by η_{eff}^∞. Norm-SGD then has the same late-time update rule as SGD with constant step size η_{eff}^∞, with η_{eff}^∞ selected autonomously by the dynamics through three different mechanisms, as we show below. Because the feature map is untrained, the curvature is

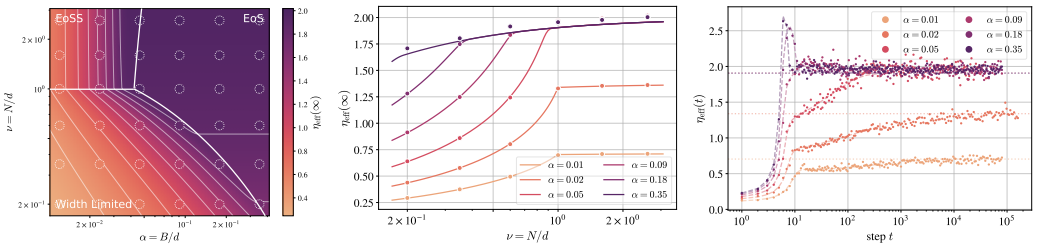


Figure 2: **Three late-time regimes: width limited, EoSS, and EoS. Left:** DMFT phase diagram in the (α, ν) plane. Color denotes $\eta_{\rm eff}^{\infty}$, solid curves phase boundaries, and open markers norm-SGD simulations. Increasing ν removes the width bottleneck, while increasing α suppresses minibatch noise and drives the EoSS–EoS transition; within EoS, the stability threshold depends weakly on ν. **Center:** $\eta_{\rm eff}^{\infty}$ versus ν, showing width-limited behavior and the crossover from batch-dependent EoSS branches to the common EoS branch as α increases. **Right:** $\eta_{\rm eff}^{t}$ at $\nu = 1$. Dots are simulations, dashed curves are DMFT, and horizontal dotted lines are stationary predictions.

time-independent; therefore the response to perturbations is stationary, and we seek time-translation-invariant (TTI) solutions.

Let R_3^{ts} denote the causal DMFT response associated with the gradient field whose equal-time correlation C_3^{tt} determines the squared gradient norm. In a stationary state, $R_3^{ts} = R_3(t-s)$. Its Z-transform (see App A for an introduction to the Z-transform and its use here) obeys

$$\hat{R}_3(z) = \left(1 + \frac{1}{\nu}\left\langle \frac{\lambda_k \eta_{\rm eff}^{\infty}}{z - 1 + \lambda_k \eta_{\rm eff}^{\infty} \hat{R}_3(z)} \right\rangle_k\right)^{-1}, \qquad \hat{R}_3(z) \to 1 \quad \text{as } z \to \infty \tag{3}$$

where $\langle \cdot \rangle_k = d^{-1} \sum_k (\cdot)$ and the large-z condition selects the causal branch. Equation 3 determines the response for a candidate value of $\eta_{\rm eff}^{\infty}$. We determine $\eta_{\rm eff}^{\infty}$ by imposing the gradient norm square C_3^{∞} and hence $\eta_{\rm eff}^{\infty}$ has a finite size as $t \to \infty$ and solve for it self-consistently. The resulting self-consistency condition admits three distinct stationary solutions, or phases, as derived in Appendix D and illustrated in Fig. 2. (1) In the under-parametrized regime $\nu < 1$, the random-features miss part of the teacher, producing irreducible error sustained by bias and noise. In the over-parameterized regime $\nu > 1$ or large batch α, the effective learning rate is instead stability-limited: (2) at small batch, collective minibatch fluctuations produce an edge of stochastic stability (EoSS), whereas at sufficiently large batch (3) the leading mode of the Hessian $\boldsymbol{H} = \boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}^{\top}/N$ becomes marginally stable, producing a deterministic edge of stability (EoS). The three phases are determined by

$$\underbrace{\nu = \left\langle \frac{\lambda_k \eta_{\rm eff}^{\infty} \rho}{1 + \lambda_k \eta_{\rm eff}^{\infty} \rho} \right\rangle_k \quad \& \quad \nu < 1,}_{\text{Phase 1: Width approximation error bias}} \quad \underbrace{\alpha = \left\langle \frac{\lambda_k \eta_{\rm eff}^{\infty} \hat{R}_3(-1)}{2 - \lambda_k \eta_{\rm eff}^{\infty} \hat{R}_3(-1)} \right\rangle_k,}_{\text{Phase 2: EoSS}} \quad \underbrace{\nu = \left\langle \left(\frac{\lambda_k \eta_{\rm eff}^{\infty} \hat{R}_3(-1)}{2 - \lambda_k \eta_{\rm eff}^{\infty} \hat{R}_3(-1)} \right)^2 \right\rangle_k.}_{\text{Phase 3: EoS}} \tag{4}$$

Each of the conditions in Eq. equation 4 together with equation 3, and additionally Eq. D.13 for Phase 1, determines $\eta_{\rm eff}^{\infty}$, $\hat{R}_3$, and ρ for Phase 1.

These conditions have direct spectral interpretation through random matrix theory. As we show in Appendix C, $\hat{R}_3$ is related to the Stieltjes transform of the random-feature Hessian $\boldsymbol{H}$, and the EoS phase condition can be shown to be equivalent to $\eta_{\rm eff}^{\infty} = \frac{2}{\lambda_{\max}(\nu)}$, where $\lambda_{\max}(\nu)$ is the deterministic equivalent of the upper edge of the spectrum of $\boldsymbol{H}$. Similarly, the EoSS condition can be compared with the usual mini-batch mean-square stability criterion for a given realization of the Hessian: $B = \mathrm{Tr}\left[\frac{\eta \boldsymbol{H}}{2 - \eta \boldsymbol{H}}\right]$ (Mulayoff & Michaeli, 2024). Remarkably, the crossing from EoSS to EoS happens at finite critical batch size α_c, rather than $\alpha \to \infty$, reflecting the fact that there is only a marginal $1/d$ benefit in stability by increasing α beyond α_c. α_c itself can be found analytically by comparing the conditions for phases 2 and 3 above at fixed ν and solving for α[1].

[1]This α_c is indeed critical in the statistical physics sense and marks a phase transition; but it is distinct from "critical batch size" introduced by (McCandlish et al., 2018).

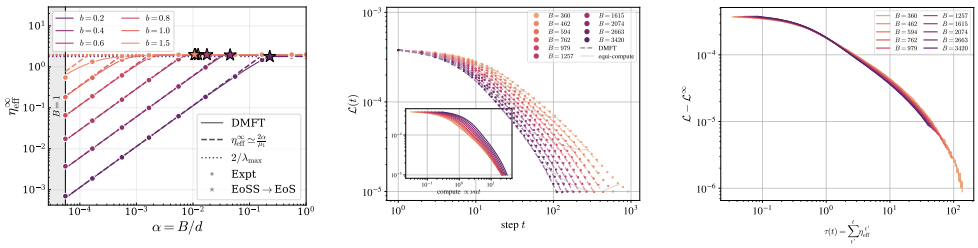


Figure 3: **Spectral control of batch-size scaling and saturation. Left:** Stationary effective learning rate $\eta_{\text{eff}}^{\infty}$ versus batch ratio $\alpha = B/d$, at fixed width ratio $\nu = N/d$. Solid curves show DMFT predictions and circles show simulations. Dashed lines indicate the small-batch linear approximation $2\alpha/\mu_1$, with $\mu_1 = \langle \lambda_k \rangle_k$, and dotted lines indicate the curvature ceiling $2/\lambda_{\max}(\nu)$. Stars mark the EoSS–EoS transition; the vertical line marks $B = 1$. Flatter spectra support a more extended increase in effective learning rate before saturation. **Center:** Training different batch sizes α on $a = 0.45$, $b = 0.2$ data. With equal compute, a larger batch size achieves the same level of performance faster (fewer steps). The inset shows the loss as a function of compute with a nearly collapsed curve across all batch sizes within a certain range. **Right:** The excess loss above the loss floor in equation D.14. Training curves for different batches *collapse* to one master curve when measures in time $\tau(t) = \sum_{t'}^{t} \eta_{\text{eff}}^{t'}$, implying that $\eta_{\text{eff}}^{t'}$, α together control compute vs. time tradeoff.

5 Batch Size and Width Speedups

While compute efficiency in power-law random-feature models has been studied extensively (for the case of SGD see (Bordelon et al., 2024a; Paquette et al., 2024)), another aspect is how to reduce the number of optimization steps needed to reach a target loss without substantially increasing compute, which we will refer to as *speed* or speedup. The three phases identified above yield distinct regimes:

- In Phase 1, increasing batch size speeds up optimization, but the reduction in steps does not offset the higher compute per step. Increasing width, however, can reduce the required steps enough to reach some target losses faster at roughly the equal total compute.
- In Phase 2, increasing batch size can result in speedup with roughly equal compute, but increasing width cannot.
- In Phase 3, neither larger batch size nor wider networks can lead to speedup with approximately equal compute if the model is over-parameterized ($\nu \geq 1$) or the total parameter count in the model is larger than the batch size ($\nu \geq \alpha$).

We quantify these results using $Q_\alpha(\alpha) := \eta_{\text{eff}}^{\infty}/\alpha$ (or $Q_\nu(\nu) := \eta_{\text{eff}}^{\infty}/\nu$) which measure whether the potential steps saving due to the increase $\eta_{\text{eff}}^{\infty}$ can compensate the increase in compute due to α (ν resp.). In all phases above these quantities would depend on the data and task through a, b, enabling speed up in some scenarios an forbidding it in others.

The absence of approximately equal-compute speedup follows from the scaling of the effective learning rate. In the small-step Phase-1 regime, increasing batch size yields only square-root gains; in Phase 2, increasing width yields gains slower than square-root. These gains cannot offset the linear increase in per-step compute. In Phase 3, increased batch size provides no effective-step gain, while the fractional gain from increasing is ν scales at most as $\sqrt{\nu}$ leading to a significant compute inefficiency. The derivations are given in Appendix F.4.

5.1 Batch Speedup at the Edge of Stochastic Stability

We study this trade-off on the stationary EoSS branch by varying $\alpha = B/d$ at fixed $\nu = N/d$ and input spectrum. Assuming a per-step cost proportional to α, we measure stationary effective time per unit compute by $Q_\alpha(\alpha) = \eta_{\text{eff}}^{\infty}(\alpha)/\alpha$. Linear growth of the effective learning rate would reduce the required number of updates proportionally, leaving the total compute unchanged. Throughout EoSS, however, $Q'_\alpha(\alpha) < 0$ for all b: larger batches yield only sublinear effective-step gains. Thus,

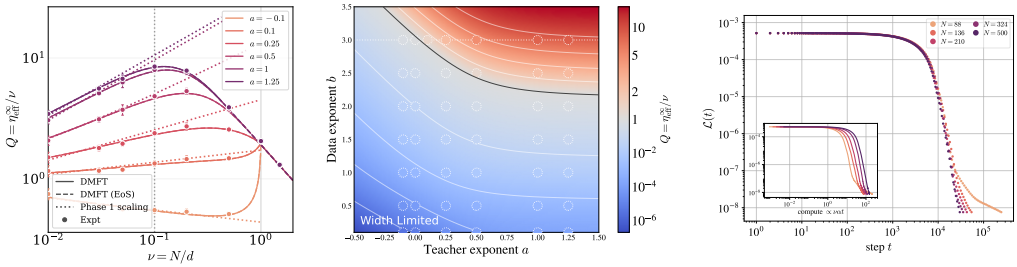


Figure 4: **Spectral dependence of width efficiency.** Stationary effective time per unit compute, quantified by $Q_\nu = \eta^\infty_{\text{eff}}/\nu$. **Left:** Width dependence of Q_ν at fixed $b = 3$, for different teacher exponents a. Solid curves show DMFT predictions before EoS, dashed curves indicate the EoS branch, and markers show simulations. Increasing Q_ν corresponds to superlinear growth of the effective learning rate with width; a maximum identifies an optimal width according to this stationary efficiency measure. **Center:** Dependence of Q_ν on the teacher and data exponents at fixed $\nu = 0.1$. The background shows DMFT predictions, while circles show simulation estimates using the same logarithmic color scale. Dotted reference lines indicate the fixed width and data exponent linking the two panels. Both panels use $d = 2000$, $B = 200$, and fixed bare learning rate $\eta = 10^{-5}$. **Right:** Loss as a function of steps (inset: compute) showing an example where wider networks train faster, while requiring approximately the same compute as narrower networks.

accumulating the same effective time requires fewer updates but more compute; nearly linear scaling keeps this overhead small.

Nevertheless, the departure from linear scaling can be arbitrarily small. Expanding the Phase 2 condition from Eq. 4 at small α gives

$$\eta^\infty_{\text{eff}} = \frac{2\alpha}{\mu_1}\left[1 - C\alpha + \mathcal{O}(\alpha^2)\right], \qquad C = \frac{\mu_2}{\mu_1^2} + \frac{1}{\nu} \tag{5}$$

where $\mu_j = \langle \lambda_k^j \rangle_k$. Both the relative efficiency penalty and the logarithmic slope deficit $1 - \partial \log \eta^\infty_{\text{eff}} / \partial \log \alpha$ therefore vanish as $C\alpha$. A small tolerated penalty δ gives the local estimate $\alpha_\delta \simeq \delta / C$ at fixed spectrum, thus defining a notion of approximately identical compute use, with $\delta = 0$ giving linear scaling – exactly the same compute and loss with less optimization steps.

This motivates α_δ as a local measure of the departure from linear batch scaling: it marks the batch size at which the stationary effective learning rate incurs a prescribed efficiency penalty δ. In this sense, α_δ provides a local analogue of empirical critical batch sizes defined by deviations from linear scaling in the number of steps required to reach a target loss (Zhang et al., 2025), provided that progress across batch sizes is governed by a common effective-time description in Fig. 3, right panel. It is distinct from α_c, the EoSS–EoS phase boundary, which instead marks the regime where the stationary effective learning rate saturates (Meterez et al., 2026c).

The spectrum controls how quickly this benefit deteriorates. For $\lambda_k = k^{-b}$, larger b concentrates curvature in fewer leading directions, increasing μ_2/μ_1^2 and the initial efficiency penalty. Flatter spectra spread curvature across more directions and support a more sustained benefit from increased batch size. Figure 3 (left) illustrates this distinction: steeper spectra allow larger effective steps at small α, but saturate sooner. Dimension determines whether this regime accommodates large physical batches. The leading linear approximation reveals the dependence on spectrum and dimension

$$\eta^\infty_{\text{lin}} = \frac{2\alpha}{\mu_1} \sim \begin{cases} 2(1-b)\alpha d^b, & 0 \le b < 1 \\ 2\alpha d / \zeta(b), & b \ge 1 \end{cases} \tag{6}$$

where $\zeta(b) = \sum_{k=1}^\infty k^{-b}$, $b > 1$. At fixed physical batch size B, increasing d reduces the leading effective-step gain for $b \le 1$, because the total curvature $\sum_k \lambda_k$ grows with dimension. For $b > 1$, this sum converges to $\zeta(b)$, making the linear prediction η^∞_{lin} independent of d.

Eventually, the effective learning rate reaches the curvature ceiling at the EoSS–EoS transition, marked by stars in Figure 3 (left). Beyond this point, further batching provides no effective-step gain and the efficiency deteriorates as $Q_\alpha \propto \alpha^{-1}$.

The right panel of Fig. 3 shows that the loss curves for different batch sizes approximately collapse when the excess loss $\mathcal{L} - \mathcal{L}^\infty$ is plotted against cumulative effective time, $\tau(t) := \sum_{t'}^{t} \eta_{\text{eff}}^{t'}$. Here, $\mathcal{L}^\infty$ is the corresponding loss floor from equation D.14. This collapse indicates that reaching a fixed target excess loss requires approximately the same effective time τ_* across batch sizes. The middle panel of Fig. 3 demonstrates the compute-speed tradeoff. The main plot shows the loss as a function of step size, showing that larger batches get to the same target loss faster, while the contour connects equal-compute points. The inset shows the same loss curves as a function of compute ($\propto BNt$) with very little spread in compute around $B = 1,000$ indicating significant savings in number of steps.

5.2 Width Speedup at the Width Limited Phase

Next, we vary the width ratio $\nu = N/d$ at fixed batch ratio α, base learning rate η, input b and teacher a spectra. Similarly, assuming a per-step cost proportional to ν, we measure stationary width efficiency by $Q_\nu(\nu) = \eta_{\text{eff}}^\infty(\nu)/\nu$. Whereas batch efficiency decreases throughout EoSS, width efficiency can increase on the width-limited branch: adding parameters can raise the stationary effective learning rate more than proportionally to their computational cost. This behavior is controlled by the residual target power $P_\nu := \frac{1}{d}\sum_{k=1}^{d} \frac{\lambda_k \beta_k^2}{1+\lambda_k \eta_{\text{eff}}^\infty \rho}$ left unrepresented at finite width, where the product $\eta_{\text{eff}}^\infty \rho$ is determined by the width constraint in Eq. 4 and it is independent of η. As derived in Appendix F.3, taking $\eta \ll 1$ at fixed $0 < \nu < 1$ gives

$$\eta_{\text{eff}}^\infty(\nu) \sim \eta \sqrt{\frac{\alpha}{\mu_1 P_\nu}}. \tag{7}$$

Since μ_1 is independent of width, reducing P_ν increases the stationary effective learning rate as $P_\nu^{-1/2}$. Efficiency improves only if this increase outweighs the additional cost per update. The data spectrum and teacher alignment jointly determine whether this condition holds. For $\lambda_k = k^{-b}$ and $\beta_k = k^{-a}$, with $b > 1$ and $2a + b > 1$, the width scalings in Appendix F.3 yield

$$Q_\nu(\nu) \sim \begin{cases} \nu^{(2a+b-3)/2}, & (1-b)/2 < a < 1/2 \\ \nu^{(b-2)/2}, & a \geq 1/2 \end{cases} \tag{8}$$

up to log factors of ν. As shown in Eq. 8, width speedup requires $Q_\nu(\nu)$ to grow super linearly. For flatter spectra, $0 < b \leq 1$, the corresponding asymptotic scalings instead predict decreasing efficiency within this width window for all admissible teacher exponents a. Figure 4 (left) shows stronger width-efficiency gains for larger a, which concentrates target power in the leading data modes. The curves reach an interior maximum at $\nu_* < 1$, while beyond this maximum, further widening no longer produces a proportional increase in the effective learning rate. Here, the decreasing efficiency on the EoS branch, where $Q'_\nu(\nu) < 0$ (see Appendix F.4), limits the width benefit. The middle panel shows the dependence of Q on both a, b when fixing $\nu = 0.1$, visualizing that higher a, b favor larger widths, while low a, b favor narrow widths while they are not capacity limited. Finally, the left panel of Fig. 4 shows a case where larger widths lead to faster training at approximately the same compute budget for the same target loss. In Fig. 9 in Appendix F.3, we show narrower widths that exhibit a capacity floor under the same parameters.

In this regime the loss floor L^∞ plays an important role, and narrow models can reach their capacity limit before the width speedup can be realized. In these cases, wider models trivially reach a target loss faster and using less compute, implying that some of the benefit is due to the lower loss floor.

6 Experiment with real data and networks

In this section, we experiment with training real ResNet in the lazy (kernel) regime on the CIFAR-5M binary classification task of classifying animate vs inanimate while training with normalized SGD. To match the setting to our theory, we extract a, b from the infinite-width deterministic NTK and treat the linearized neural networks eNTK as random features drawn from the NTK. In our experiments, we measured $b \simeq 1.8$ and $a \simeq -0.325$, close to the measurements of (Bordelon et al., 2024a) for the same architecture.

In Fig. 5, right panel, we show the results when varying the width in the range $w = 4$ ($N = 1,120$ random features) to $w = 256$ ($N = 4,135,936$) with fixed $B = 1,000$ for a single pass with no repetitions. On the left panel we show the results for the same setting when fixing $w = 128$ and

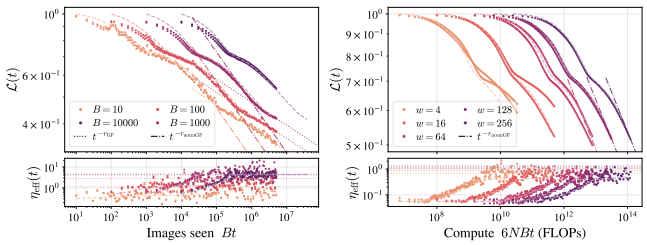


Figure 5: **Theory predicts training curves exponents for training with linearized networks.** Loss (**top**) and effective learning rate (**bottom**). Before $\eta_{\rm eff}$ hits its ceiling $\eta_{\rm eff}^{\infty}$, the loss decreases with an accelerated power law, after the ceiling is reached the power law degrades to the standard SGD power law. Throughout the narrowest width and smallest batch size are optimal, as predicted in the previous section for these a, b. (**left**) Loss as a function of images seen, Bt, for different batch sizes and fixed $w = 128$, $\eta = 0.1$ (proportional to compute). (**right**) Loss as a function of compute, calculated as $6NBt$, for different widths and fixed $B = 1,000$, $\eta = 0.01$

varying the batch size from $B = 10$ to $B = 10,000$. Since the NTK is infinite-dimensional, while any network has a finite parameter count and batch size, we adopt our theory to this dimension-free regime ($d \to \infty$ while N, B are finite) in Appendix H. This also means that $\nu < 1$ and phase 2 (EoSS) is inaccessible. As discussed in the previous section, this implies that the smallest batch size B is always compute optimal, which we confirm in the left panel. Additionally, based on results from the previous section, we expect that as long as the model is time-constrained rather than parameter-constrained, narrow networks will be more compute-efficient for our measured exponents a, b, which the experiments in the right panel confirm.

Additionally, we add to the figures rulers showing the normalized gradient flow accelerated exponent, and intersecting rulers showing the breakdown of acceleration and the change to the standard SGD exponent when the effective learning rate reaches marginal stability. Throughout, our deterministic DMFT dynamics track the stochastic dynamics well, even though the DMFT assumes Gaussian features and only uses the NTK spectrum, the NTK target coefficients, and N, B, with no information on the eNTK or other data, architecture, and task beyond those four quantities.

7 Outlook

Normalized SGD provides a solvable setting in which acceleration, stability, and resource allocation can be studied together. In continuous time, normalization preserves the gradient-flow trajectory while accelerating its clock, yielding accelerated power-law, exponential, or formal finite-time convergence depending on the teacher exponent. However, in discrete dynamics, the growing effective learning rate ultimately breaks this acceleration down through one of three mechanisms: width-limited noise and approximation error, a noise-driven EoSS, or deterministic EoS dynamics.

We relate these phases to compute efficiency and speedup. We find a critical batch ratio α_c that marks the phase transition between EoSS and EoS, and distinguish it from α_δ, the maximal batch ratio that allows approximately linear scaling of learning rate and batch size – which allows speedup with equal compute. We quantify the dependence of these scales on width and on data exponents.

Experiments with linearized ResNets on CIFAR-5M support the predicted acceleration and its breakdown, and connect the spectral description to observed batch- and width-dependent compute trends. Extending this connection beyond the kernel regime requires accounting for the evolution of features, target alignment, and curvature during training. Further directions include determining how learning-rate and batch-size schedules alter the trade-off, and relating effective-time efficiency to wall-clock performance under realistic hardware costs. Additionally, developing similar trade-off of compute–parameter count instead of compute-speedup could accelerate inference at the cost of increased compute cost in the training phase. Finally, extending the analysis to optimizers that also modify update direction would clarify which conclusions persist beyond global gradient normalization.

AI Use Statement. In this work, we used generative AI tools to help simplify the expression in App D.1 and which enabled the evaluation of the integrals, which provided "critical ingredients for proving mathematical claims". In this work, we have **not** used generative AI tools for:

- help develop theoretical models or conceptual frameworks
- formulate mathematical claims
- assist in the writing of proofs
- propose or refine hypotheses
- design or provide feedback on research methodology or experiments
- implement methods

The following are not applicable to this work:

- generate synthetic data sets
- assist with translation
- clean and reformat dataset
- support qualitative and thematic data analysis
- interpret results

Additionally, we used generative AI tools for create and edit figures, write code, and to edit and polish the manuscript. We take responsibility for the final content of this work, including text, claims and figures produced with the aid of generative AI.

Reproducibility Statement. We detial the experimental setting in Appendix G, code to reproduce all figure will be provided when the paper in made public.

REFERENCES

Naman Agarwal, Rohan Anil, Elad Hazan, Tomer Koren, and Cyril Zhang. Disentangling Adaptive Gradient Methods from Learning Rates, February 2020. URL http://arxiv.org/abs/2002.11803. arXiv:2002.11803 [cs.LG].

Atish Agarwala and Jeffrey Pennington. High dimensional analysis reveals conservative sharpening and a stochastic edge of stability, February 2025. URL http://arxiv.org/abs/2404.19261. arXiv:2404.19261 [cs.LG].

Arseniy Andreyev and Pierfrancesco Beneventano. Edge of stochastic stability: Revisiting the edge of stability for sgd. *arXiv preprint arXiv:2412.20553*, 2024.

Sanjeev Arora, Zhiyuan Li, and Abhishek Panigrahi. Understanding Gradient Descent on Edge of Stability in Deep Learning, October 2022. URL http://arxiv.org/abs/2205.09745. arXiv:2205.09745 [cs.LG].

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024a.

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining Neural Scaling Laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, July 2024b. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.2311878121. URL http://arxiv.org/abs/2102.06701. arXiv:2102.06701 [cs].

Jeremy Bernstein and Laker Newhouse. Old Optimizer, New Norm: An Anthology, December 2024. URL http://arxiv.org/abs/2409.20325. arXiv:2409.20325 [cs.LG].

Antoine Bodin and Nicolas Macris. Model, sample, and epoch-wise descents: exact solution of gradient flow in the random feature model. *Advances in neural information processing systems*, 34:21605–21617, 2021.

Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. *Advances in Neural Information Processing Systems*, 35:32240–32256, 2022.

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, April 2024a. URL http://arxiv.org/abs/2402.01092. arXiv:2402.01092 [cond-mat, stat].

Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. How Feature Learning Can Improve Neural Scaling Laws, September 2024b. URL http://arxiv.org/abs/2409.17858. arXiv:2409.17858 [cond-mat, stat].

Blake Bordelon, Lorenzo Noci, Mufan Li, Boris Hanin, and Cengiz Pehlevan. Depthwise hyperparameter transfer in residual networks: Dynamics and scaling limit. In *International Conference on Learning Representations*, volume 2024, pp. 22088–22127, 2024c.

Francesco Cagnetta, Allan Raventós, Surya Ganguli, and Matthieu Wyart. Deriving neural scaling laws from the statistics of natural language. *arXiv preprint arXiv:2602.07488*, 2026.

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.

Lénaïc Chizat. The hidden width of deep resnets: Tight error bounds and phase diagram. *arXiv preprint arXiv:2509.10167*, 2025.

Jeremy Cohen, Alex Damian, Ameet Talwalkar, J. Zico Kolter, and Jason D. Lee. Understanding Optimization in Deep Learning with Central Flows. October 2024. URL https://openreview.net/forum?id=sIE2rI3ZPs.

Jeremy M Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. *arXiv preprint arXiv:2103.00065*, 2021.

Jeremy M Cohen, Behrooz Ghorbani, Shankar Krishnan, Naman Agarwal, Sourabh Medapati, Michal Badura, Daniel Suo, David Cardoze, Zachary Nado, George E Dahl, et al. Adaptive gradient methods at the edge of stability. *arXiv preprint arXiv:2207.14484*, 2022.

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

C De Dominicis. Dynamics as a substitute for replicas in systems with quenched random impurities. *Physical Review B*, 18(9):4913, 1978.

Aaron Defazio, Ashok Cutkosky, Harsh Mehta, and Konstantin Mishchenko. When, why and how much? adaptive learning rate scheduling by refinement. 2023.

Nolan Dey, Bin Zhang, Lorenzo Noci, Mufan Li, Blake Bordelon, Shane Bergsma, Cengiz Pehlevan, Boris Hanin, and Joel Hestness. Don't be lazy: Completep enables compute-efficient deep transformers. *Advances in Neural Information Processing Systems*, 38:137707–137739, 2026.

Nikhil Ghosh, Denny Wu, and Alberto Bietti. Understanding the mechanisms of fast hyperparameter transfer. In *International Conference on Learning Representations*, volume 2026, pp. 65022–65063, 2026.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pp. 1842–1850. PMLR, 2018.

Gil Harari, Yoel Zimmermann, Ola Tangen Kulseng, Laura Zichi, Chuin Wei Tan, Marc L. Descoteaux, and Boris Kozinsky. Beyond Adam: SOAP and Muon for Faster, Label-Efficient Training of Machine Learning Interatomic Potentials. May 2026. URL https://openreview.net/forum?id=mYxGJcckHl#discussion.

Moritz Helias and David Dahmen. *Statistical Field Theory for Neural Networks*, volume 970 of *Lecture Notes in Physics*. Springer International Publishing, Cham, 2020. ISBN 978-3-030-46443-1 978-3-030-46444-8. doi: 10.1007/978-3-030-46444-8. URL http://link.springer.com/10.1007/978-3-030-46444-8.

Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

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.

Stanislaw Jastrzebski, Maciej Szymczak, Stanislav Fort, Devansh Arpit, Jacek Tabor, Kyunghyun Cho, and Krzysztof Geras. The break-even point on optimization trajectories of deep neural networks. *arXiv preprint arXiv:2002.09572*, 2020.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

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.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2014. URL http://arxiv.org/abs/1412.6980. arXiv:1412.6980 [cs.LG].

Clarissa Lauditi, Cengiz Pehlevan, and Blake Bordelon. Spectral dynamics in deep networks: Feature learning, outlier escape, and learning rate transfer. *arXiv preprint arXiv:2605.07870*, 2026.

Kiwon Lee, Andrew N. Cheng, Courtney Paquette, and Elliot Paquette. Trajectory of Mini-Batch Momentum: Batch Size Saturation and Convergence in High Dimensions, June 2022. URL http://arxiv.org/abs/2206.01029. arXiv:2206.01029 [math.OC].

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Gagik Magakyan, Pablo Parrilo, and Asuman Ozdaglar. Spectral scaling laws of muon. *arXiv preprint arXiv:2606.04058*, 2026.

Alexander Maloney, Daniel A. Roberts, and James Sully. A Solvable Model of Neural Scaling Laws, October 2022. URL http://arxiv.org/abs/2210.16859. arXiv:2210.16859 [cs].

Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training. *arXiv preprint arXiv:1812.06162*, 2018.

Alexandru Meterez, Depen Morwani, Jingfeng Wu, Costin-Andrei Oncescu, Cengiz Pehlevan, and Sham Kakade. Seesaw: Accelerating Training by Balancing Learning Rate and Batch Size Scheduling, July 2026a. URL http://arxiv.org/abs/2510.14717. arXiv:2510.14717 [cs.LG].

Alexandru Meterez, Pranav Ajit Nair, Depen Morwani, Cengiz Pehlevan, and Sham Kakade. Anytime pretraining: Horizon-free learning-rate schedules with weight averaging. *arXiv preprint arXiv:2602.03702*, 2026b.

Alexandru Meterez, Pranav Ajit Nair, Depen Morwani, Cengiz Pehlevan, Sham Kakade, and Alex Damian. A defense of the quadratic model. *arXiv preprint arXiv:2607.21716*, 2026c.

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.

Rotem Mulayoff and Tomer Michaeli. Exact Mean Square Linear Stability Analysis for SGD, June 2024. URL http://arxiv.org/abs/2306.07850. arXiv:2306.07850 [cs.LG].

Marc Mézard and Andrea Montanari. *Information, physics, and computation*. Oxford graduate texts. Oxford university press, Oxford, 2009. ISBN 978-0-19-857083-7.

Preetum Nakkiran, Behnam Neyshabur, and Hanie Sedghi. The Deep Bootstrap Framework: Good Online Learners are Good Offline Generalizers, February 2021. URL http://arxiv.org/abs/2010.08127. arXiv:2010.08127 [cs.LG].

Lorenzo Noci, Alexandru Meterez, Thomas Hofmann, and Antonio Orvieto. Super consistency of neural network landscapes and learning rate transfer. *Advances in Neural Information Processing Systems*, 37:102696–102743, 2024.

Courtney Paquette, Kiwon Lee, Fabian Pedregosa, and Elliot Paquette. SGD in the Large: Average-case Analysis, Asymptotics, and Stepsize Criticality. In *Proceedings of Thirty Fourth Conference on Learning Theory*, pp. 3548–3626. PMLR, July 2021. URL https://proceedings.mlr.press/v134/paquette21a.html.

Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+3 Phases of Compute-Optimal Neural Scaling Laws. November 2024. URL https://openreview.net/forum?id=aVSxwicpAk&noteId=4mjoq2kraU.

Elliot Paquette, Noah Marshall, Lucas Benigni, Guangyuan Wang, Atish Agarwala, and Courtney Paquette. Phases of Muon: When Muon Eclipses SignSGD, May 2026. URL http://arxiv.org/abs/2605.09552. arXiv:2605.09552 [math.OC].

Christopher J. Shallue, Jaehoon Lee, Joseph Antognini, Jascha Sohl-Dickstein, Roy Frostig, and George E. Dahl. Measuring the Effects of Data Parallelism on Neural Network Training, July 2019. URL http://arxiv.org/abs/1811.03600. arXiv:1811.03600 [cs.LG].

Samuel L Smith, Pieter-Jan Kindermans, Chris Ying, and Quoc V Le. Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*, 2017.

Haim Sompolinsky and Annette Zippelius. Dynamic theory of the spin-glass phase. *Physical Review Letters*, 47(5):359, 1981.

Stefano Spigler, Mario Geiger, and Matthieu Wyart. Asymptotic learning curves of kernel methods: empirical data versus teacher–student paradigm. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(12):124001, 2020.

T. Tieleman and G Hinton. Tieleman, T. and Hinton, G. (2012) Lecture 6.5-rmsprop Divide the Gradient by a Running Average of Its Recent Magnitude. COURSERA Neural Networks for Machine Learning, 4, 26-31. - References - Scientific Research Publishing, 2012. URL https://www.scirp.org/reference/referencespapers?referenceid=1911091.

Howe Tissue, Venus Wang, and Lu Wang. Scaling law with learning rate annealing. *Advances in Neural Information Processing Systems*, 38:91017–91053, 2026.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. SOAP: Improving and Stabilizing Shampoo using Adam, January 2025. URL http://arxiv.org/abs/2409.11321. arXiv:2409.11321 [cs.LG].

Kaiyue Wen, David Hall, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them. In *International Conference on Learning Representations*, volume 2026, pp. 144731–144838, 2026.

Ke Liang Xiao, Noah Marshall, Atish Agarwala, and Elliot Paquette. Exact risk curves of signSGD in high-dimensions: quantifying preconditioning and noise-compression effects. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *ICML'25*, pp. 68391–68439, Vancouver, Canada, July 2025. JMLR.org.

Chen Xing, Devansh Arpit, Christos Tsirigotis, and Yoshua Bengio. A walk with sgd. *arXiv preprint arXiv:1802.08770*, 2018.

Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pp. 11727–11737. PMLR, 2021.

Greg Yang, Edward Hu, Igor Babuschkin, Szymon Sidor, Xiaodong Liu, David Farhi, Nick Ryder, Jakub Pachocki, Weizhu Chen, and Jianfeng Gao. Tuning large neural networks via zero-shot hyperparameter transfer. *Advances in Neural Information Processing Systems*, 34:17084–17097, 2021.

Greg Yang, Dingli Yu, Chen Zhu, and Soufiane Hayou. Tensor programs vi: Feature learning in infinite depth neural networks. In *International Conference on Learning Representations*, volume 2024, pp. 55099–55150, 2024.

Hanlin Zhang, Depen Morwani, Nikhil Vyas, Jingfeng Wu, Difan Zou, Udaya Ghai, Dean Foster, and Sham Kakade. How does critical batch size scale in pre-training? In *International Conference on Learning Representations*, volume 2025, pp. 66756–66782, 2025.

A A SHORT INTRODUCTION TO THE Z-TRANSFORM

The Z-transform packages the sequence $f^0, f^1, f^2, \ldots, f^t, \ldots$ into a function of a complex variable, with each power of z^{-1} labeling one time step; the result is a generating function for a sequence, which can also be seen as the discrete-time analogue of the Laplace transform under the substitution $z = e^s$. This transformation diagonalizes operators that have a shift symmetry, or as they are more commonly called time translation invariant (TTI) operators. Practically, the transform turns shifts and convolutions into algebraic manipulations.

A.1 FROM A SEQUENCE TO A FUNCTION

For a causal sequence f^t, defined for $t \geq 0$, the Z-transform is

$$\hat{f}(z) := \sum_{t=0}^{\infty} f^t z^{-t} = f^0 + f^1 z^{-1} + f^2 z^{-2} + \cdots, \tag{A.1}$$

where first equality is the definition and the second makes its bookkeeping role explicit. For example, the transform of the sequence $1, a, a^2, \ldots$ is the geometric series

$$\sum_{t=0}^{\infty} a^t z^{-t} = \frac{z}{z-a}, \qquad |z| > |a|. \tag{A.2}$$

The condition $|z| > |a|$ is the *region of convergence* of this series. For a causal sequence, the region of convergence is the exterior of a circle whose radius is set by the growth rate of the sequence. Keeping track of this region tells us where algebraic manipulations and contour integrals are valid.

In particular, $z \to 1$ corresponds to the small-s limit $s \to 0$ in Laplace transform or the small frequency limit in Fourier transform $\omega \to 0$, which is why the behavior near $z = 1$ contains information about long times.

More generally, the sequence can be recovered from its transform by extracting the coefficient of z^{-t}

$$f^t = \oint_\gamma \frac{z}{2\pi i z}\, z^t \hat{f}(z), \tag{A.3}$$

where the contour γ is a circle lying in the region of convergence. This is the familiar coefficient-extraction formula for a power series. In practice, one rarely use this method to invert the series directly; it is more useful as a way to understand how the singularities of $\hat{f}$ determine the behavior of f^t.

A.2 SHIFTS AND RECURRENCES

A change of summation index gives the transform of a forward shift,

$$\sum_{t=0}^{\infty} f^{t+1} z^{-t} = z\big(\hat{f}(z) - f^0\big). \tag{A.4}$$

It follows that the forward difference $(Df)^t = f^{t+1} - f^t$ satisfies

$$\widehat{Df}(z) = (z-1)\hat{f}(z) - z f^0. \tag{A.5}$$

Thus $z-1$ plays the role of the perhaps more familiar Laplace variable s for a time derivative, while the remaining term carries the initial condition.

This shift rule turns linear recurrences into algebraic equations. For example, if

$$x^{t+1} = a x^t + b^t, \tag{A.6}$$

with prescribed input b^t, transforming both sides and using Eq. equation A.4 gives

$$\hat{x}(z) = \frac{z x^0 + \hat{b}(z)}{z-a}. \tag{A.7}$$

The denominator already reveals the natural mode a^t.

A.3 TIME-TRANSLATION INVARIANCE AND DIAGONALIZATION

Consider a causal matrix K^{ts} acting on time indices. It is time-translation invariant, or TTI, when it depends only on the lag between its indices,

$$K^{ts} = K(t-s), \qquad K(u) = 0 \quad \text{for } u < 0. \tag{A.8}$$

Its action on a sequence is therefore a convolution,

$$(Kf)^t := \sum_{s=0}^{t} K(t-s) f^s, \qquad \hat{K}(z) := \sum_{u=0}^{\infty} K(u) z^{-u}. \tag{A.9}$$

Ignoring the initial boundary for a moment, a geometric sequence $f^s = z^s$ is an eigenvector of every TTI matrix:

$$\sum_{s} K(t-s) z^s = z^t \sum_{u} K(u) z^{-u} = \hat{K}(z) z^t. \tag{A.10}$$

The Z-transform is therefore a change to the common eigenbasis of TTI matrices, and $\hat{K}(z)$ is the corresponding eigenvalue. This is the precise sense in which it diagonalizes the time dynamics. Matrix products become products of their Z-transforms; equivalently, convolution with a sequence obeys

$$\widehat{Kf}(z) = \hat{K}(z)\hat{f}(z). \tag{A.11}$$

For example, the delayed running-sum matrix has $K(u) = \mathbf{1}_{u>0}$, so

$$\hat{K}(z) = \sum_{u=1}^{\infty} z^{-u} = \frac{1}{z-1}, \qquad |z| > 1. \tag{A.12}$$

Together with Eq. equation A.5, this shows that $z-1$ and $1/(z-1)$ represent discrete differentiation and integration, just as s and $1/s$ do in the Laplace transform.

A.4 READING DYNAMICS FROM SINGULARITIES

Equation equation A.2 gives the basic dictionary: a pole at $z = a$ produces a mode proportional to a^t. Such a mode decays when $|a| < 1$ and grows when $|a| > 1$. On the unit circle it is marginal: $a = 1$ gives a constant, $a = -1$ a period-two oscillation, and $a = e^{i\omega}$ an undamped oscillation at frequency ω. The unit circle is therefore the discrete-time stability boundary. More generally, a pole of order m adds a factor t^{m-1} to the exponential mode.

If a sequence approaches a finite limit as $t \to \infty$, its asymptotic value can be obtained using the final-value theorem

$$\lim_{t\to\infty} f^t = \lim_{z\to 1} (z-1)\hat{f}(z). \tag{A.13}$$

Thus a nonzero constant at late times appears as a simple pole at $z = 1$, with coefficient isolated by multiplying by $z - 1$. A growing mode or an undamped oscillation violates the pole condition and prevents the limit from existing.

The same picture can be phrased in frequency language. Evaluating a stable transform at $z = e^{\epsilon + i\omega}$ and taking $\epsilon \downarrow 0$ gives its response to a discrete-time frequency ω. The point $z = 1$ is zero frequency, while $z = -1$ is the highest discrete-time frequency and corresponds to sign alternation from one step to the next.

A.5 BEYOND TTI MATRICES

An ordinary Z-transform suffices for a TTI matrix because $K^{ts} = K(t-s)$ depends on only one time lag. A general two-time matrix $F^{t_1 t_2}$, such as a correlation matrix, is not necessarily TTI and retains two independent time indices. We therefore introduce one transform variable for each index:

$$\tilde{F}(z_1, z_2) := \sum_{t_1, t_2 = 0}^{\infty} F^{t_1 t_2} z_1^{-t_1} z_2^{-t_2}. \tag{A.14}$$

The double transform preserves the diagonalization of TTI multiplication on either side. If K and L are TTI matrices, then

$$\widetilde{KFL^\top}(z_1, z_2) = \hat{K}(z_1)\hat{L}(z_2)\tilde{F}(z_1, z_2). \tag{A.15}$$

For our use, the equal-time diagonal is especially important. If F is naturally diagonal in time, $F^{t_1 t_2} = \delta^{t_1 t_2} f^{t_1}$, then

$$\tilde{F}(z_1, z_2) = \hat{f}(z_1 z_2). \tag{A.16}$$

because the two time indices are equal. For a general F, define

$$f^t := F^{tt}, \qquad \hat{f}(\zeta) := \sum_{t=0}^{\infty} F^{tt}\zeta^{-t}. \tag{A.17}$$

We keep the product $z_1 z_2 = \zeta$ fixed by writing

$$z_1 = \sqrt{\zeta}\, w, \qquad z_2 = \frac{\sqrt{\zeta}}{w}. \tag{A.18}$$

The double transform thus also have the degree of freedom $w^{t_2 - t_1}$. Integrating along the unit circle removes every nonzero power of w, leaving only $t_1 = t_2$:

$$\begin{aligned}\hat{f}(\zeta) &= \oint_{|w|=1} \frac{w}{2\pi i w}\, \tilde{F}\left(\sqrt{\zeta}\, w, \frac{\sqrt{\zeta}}{w}\right) \\ &= \int_0^{2\pi} \frac{\theta}{2\pi}\, \tilde{F}\left(\sqrt{\zeta} e^{i\theta}, \sqrt{\zeta} e^{-i\theta}\right).\end{aligned} \tag{A.19}$$

Here ζ follows the behavior averaged across time, while $w = e^{i\theta}$ resolves relative-time difference and we use the angular integration to project onto zero relative time and hence onto the diagonal.

A.6 Use in the steady-state analysis

We now specialize these facts to the calculation in this paper. Once η_{eff}^t approaches a constant, the matrices Θ and R_3 become TTI. The Z-transforms diagonalize the time indices, turning matrix products and inverses into scalar algebra. In particular, $\hat{\Theta}(z) = \eta_{\text{eff}}/(z-1)$, while the physical branch of the self-consistent equation for $\hat{R}_3$ is selected by $\hat{R}_3(z) \to R_3(0) = 1$ as $z \to \infty$.

The correlation functions remain genuine two-time objects. We therefore solve for their double transforms and apply Eq. equation A.19 to obtain the equal-time loss and gradient correlations $\hat{c}_1(\zeta)$ and $\hat{c}_3(\zeta)$, and the final-value theorem gives

$$c_3(\infty) = \lim_{\zeta \to 1}(\zeta - 1)\hat{c}_3(\zeta), \qquad \eta_{\text{eff}}^{\infty} = \frac{\eta}{\sqrt{\nu c_3(\infty)}}. \tag{A.20}$$

The transformed DMFT equations contain several possible sources for this simple pole at $\zeta = 1$. Which one becomes singular distinguishes the width-limited, edge-of-stability, and stochastic edge-of-stability phases.

B Deriving the Dynamical Mean-Field Theory Equations

Starting from the loss

$$L = \frac{1}{2B}\sum_{\mu=1}^{B}\left(\beta^\top \psi_\mu - \frac{1}{\sqrt{N}} w^\top A \psi_\mu\right)^2$$

The gradient is

$$g = \frac{\partial L}{\partial w_k} = -\frac{1}{\sqrt{N}} A \frac{1}{B}\psi^T \psi \left(\beta - \frac{1}{\sqrt{N}} A^T w\right)$$

so the weight updates are

$$w_{t+1} = w_t - \sqrt{d}\eta \left(g^T g\right)^{-1/2} g$$

multiplying by $\frac{1}{\sqrt{N}}A^T$ from the left

$$\frac{1}{\sqrt{N}}A^T w_{t+1} = \frac{1}{\sqrt{N}}A^T w_t - \sqrt{d}\eta\frac{1}{\sqrt{N}}A^T\left(g^T g\right)^{-1/2} g$$

and subtracting our equation form β

$$\beta - \frac{1}{\sqrt{N}}A^T w_{t+1} = \beta - \frac{1}{\sqrt{N}}A^T w_t + \sqrt{d}\eta\frac{1}{\sqrt{N}}A^T\left(g^T g\right)^{-1/2} g.$$

uUsing the following fields

$$\begin{aligned}
&\mathbb{R}^d \ni h_0^t := \beta - \frac{1}{\sqrt{N}}A^\top w^t \qquad \mathbb{R}^B \ni h_1^t := \frac{1}{\sqrt{\alpha d}}\psi_t h_0^t \qquad \mathbb{R}^d \ni h_2^t := \frac{1}{\sqrt{B}}\psi_t^\top h_1^t \\
&\mathbb{R}^N \ni h_3^t := \frac{1}{\sqrt{\nu d}}A h_2 \qquad \mathbb{R}^d \ni h_4^t := \frac{1}{\sqrt{N}}A^\top h_3^t
\end{aligned} \tag{B.1}$$

and the dynamics close the loop by

$$h_0^{t+1} = h_0^t - \sqrt{d}\eta\left(h_3^{t\top} h_3^t\right)^{-1/2} h_4^t.$$

B.1 Field Averaging Using the Cavity Method

Below, we use the cavity method (Mézard & Montanari, 2009) to find a mean-field description for the fields $\{h_i\}_{i=0}^4$ by adding a small perturbation to the rows and columns of ψ and A and reading off their collective average effect on a fresh coordinate in the same field. Along the way, we introduce average response functions $R_{i\to j}$ that denote the effect of a perturbation in a field h_i on the field h_j; these response functions are unknown and have to be solved for self-consistently.

Adding a row to ψ Throughout, we use $\tilde{\cdot}$ to denote a perturbed quantity and $\bar{\cdot}$ to denote a fresh independent entry in the field. We ignore $1/d$ corrections and use $\to$ to denote the large d limit after employing the law of large numbers (LLN) and the central limit theorem (CLT) when summing independent quantities.

$$\begin{aligned}
h_1^t &= \frac{1}{\sqrt{\alpha d}}\psi_t h_0^t + \frac{1}{\sqrt{\alpha d}}\psi_t \tilde{h}_0^t \\
\bar{h}_1^t &= \frac{1}{\sqrt{\alpha d}}\tilde{\psi}_t h_0^t + \frac{1}{\sqrt{\alpha d}}\tilde{\psi}_t \tilde{h}_0^t \\
h_2^t &= \frac{1}{\sqrt{B}}\psi_t^T h_1^t + \frac{1}{\sqrt{B}}\psi_t^T \delta h_1^t + \tilde{\psi}_t^T \bar{h}_1^t
\end{aligned}$$

Reading the effect of the perturbation on the cavity

$$\bar{h}_1^t = \frac{1}{\sqrt{\alpha d}}\tilde{\psi}_t^T h_0^t + \frac{1}{\sqrt{\alpha d}}\frac{1}{\sqrt{B}}\int \tilde{\psi}_t^T \frac{\partial h_0^t}{\partial h_2^{t'}}\tilde{\psi}_{t'}\bar{h}_1^{t'}\, dt'.$$

Using the fact that the new row $\tilde{\psi}$ is independent of the old fields, we can use CLT on the first expression and LLN on the second, keeping in mind that batches are independent

$$\bar{h}_1^t \to u_1 + \frac{1}{\alpha}\int^t \delta\left(t-t'\right) R_{2\to 0}\left(t,t'\right)\bar{h}_1^{t'}\left(t'\right)\, dt' = u_1\left(t\right) + \frac{1}{\alpha}R_{2\to 0}\left(t,t\right)\bar{h}_1\left(t\right)$$

with $u_1 \sim \mathcal{N}\left(0, \delta^{tt'}\frac{1}{\alpha}C_0^{tt'}\right)$

$$\left\langle u_1^{t\top} u_1^{t'}\right\rangle = \frac{1}{\alpha d}\left\langle\left(\psi_t^T h_0^t\right)^\top \psi_t^T h_0^t\right\rangle = \mathrm{Tr}\left[\left\langle\psi_t\psi_t^\top\right\rangle\left\langle h_{0k}^t h_{0k}^{t\top}\right\rangle\right] = \frac{1}{\alpha d}\sum_{k=1}^d \delta^{tt'}\lambda_k\left\langle h_{0k}^t h_{0k}^{t'}\right\rangle =: \delta^{tt'}\frac{1}{\alpha}C_0^{tt'},$$

and the response

$$R_{2\to 0} = \frac{1}{d}\sum_k \lambda_k\left\langle\frac{\partial h_{0k}^t}{\partial h_{2k}^{t'}}\right\rangle = \frac{1}{d}\sum_k \lambda_k R_{2\to 0k}.$$

However, since the causal chain runs $h_0^t \to h_1^t \to h_2^t \to h_3^t \to h_4^t \to h_0^{t+1}$ we see that h_2^t only effects h_0^{t+1} which forces $R_{2\to 0}^{tt} = 0$

Adding a column to ψ Similarly, adding a column to ψ we find the following perturbed fields

$$h_1^t = \frac{1}{\sqrt{\alpha d}}\psi_t h_0^t + \frac{1}{\sqrt{\alpha d}}\psi_t \tilde{h}_0^t + \tilde{\psi}_t \bar{h}_0^t$$
$$h_2^t = \frac{1}{\sqrt{B}}\psi_t^T h_1^t + \frac{1}{\sqrt{B}}\psi_t^T \delta h_1^t$$
$$\bar{h}_2^t = \frac{1}{\sqrt{B}}\tilde{\psi}_t^T h_1^t + \frac{1}{\sqrt{B}}\tilde{\psi}_t^T \delta h_1^t.$$

Again, reading off the effect of the perturbation on the cavity, we find

$$\bar{h}_{2k}^t = \frac{1}{\sqrt{B}}\psi_{tk}^T h_1^t + \frac{1}{\sqrt{B}}\frac{1}{\sqrt{\alpha d}}\int \tilde{\psi}_{tk}^T \frac{\partial h_1^t}{\partial h_1^{t'}} \tilde{\psi}_{t'k} \bar{h}_0^{t'}\, dt'$$

and in the high-dimensional limit

$$\bar{h}_{2k}^t \to u_{2k} + \lambda_k \int^t \delta\left(t-t'\right) R_1\left(t,t'\right) \bar{h}_0\left(t'\right)\, dt' = u_2\left(t\right) + \lambda_k R_{1k}\left(t,t\right) \bar{h}_0\left(t\right)$$

with

$$\left\langle u_2^t u_2^{t'} \right\rangle = \delta^{tt'} \lambda_k \left\langle h_{1k}^t h_{1k}^{t'} \right\rangle =: \delta^{tt'} \lambda_k C_1^{tt'}$$
$$R_{1k} = \lambda_k \left\langle \frac{\partial h_{1k}^t}{\partial h_{1k}^{t'}} \right\rangle$$

We can again use causality to simplify the response R_1^{tt}. As we have seen above h_1 is just noise uncorrelated in time, so the only contribution is the immediate one, thus $R_1 = 1$

Cavity results Repeating the same process for a column and a row of A and collecting all the results together, we find

$$h_0^{t+1} = h_0^t - \eta \left(C_3^{tt}\right)^{-1/2} h_4^t$$
$$h_1^t = u_1^t \qquad u_1^t := \mathcal{N}\left(0, \delta^{tt'}\frac{1}{\alpha}C_0^{tt'}\right)$$
$$h_{2k}^t = u_{2k}^t + \lambda_k h_{0k}^t \qquad u_{2k}^t := \mathcal{N}\left(0, \delta^{tt'}\lambda_k C_1^{tt'}\right)$$
$$h_3^t = u_3 + \frac{1}{\nu}\sum_{t'=0}^t R_{4\to 2}^{tt'} h_3^{t'} \qquad u_3 := \mathcal{N}\left(0, \frac{1}{\nu}C_2\right)$$
$$h_{4k}^t = u_{4k} + \sum_{t'=0}^t R_3^{tt'} h_{2k}^{t'} \qquad u_4 := \mathcal{N}\left(0, C_3\right)$$

$$C_0 := \frac{1}{d}\sum_k \lambda_k \left\langle h_{0k} h_{0k} \right\rangle \qquad R_{2\to 0}^D := \frac{1}{d}\sum_k \lambda_k \left\langle \frac{\delta h_{0k}}{\delta u_{2k}} \right\rangle$$
$$C_1 := \left\langle h_1 h_1 \right\rangle \qquad R_1 := \left\langle \frac{\delta h_1}{\delta u_1} \right\rangle$$
$$C_2 := \frac{1}{d}\sum_k \left\langle h_{2k} h_{2k} \right\rangle \qquad R_{4\to 2} := \frac{1}{d}\sum_k \left\langle \frac{\delta h_{2k}}{\delta u_{4k}} \right\rangle$$
$$C_3 := \left\langle h_3 h_3 \right\rangle \qquad R_3 := \left\langle \frac{\delta h_3}{\delta u_3} \right\rangle.$$

Therefore, the dynamics can be described as

$$h_0^{t+1} = h_0^t - \eta \left(C_3^{tt}\right)^{-1/2} \left(u_{4k} + \sum_{t'=0}^t R_3^{tt'} \left(u_{2k}^{t'} + \lambda_k h_{0k}^{t'} \right) \right)$$

The test $\mathcal{L}$ and train L loss can be expressed as

$$\mathcal{L}(t) = \frac{1}{d}\sum_k \lambda_k h_{0k}^2(t) = C_0(t,t); \quad L(t) = C_1(t,t)$$

and since it is online, we only have to account for the difference in normalization (normalized by P vs by d)

$$C_0(t,t) = \alpha C_1(t,t)$$

Which can also be seen more directly by considering $h_1^t = u_1^t \Rightarrow \left\langle h_1^t h_1^{t'} \right\rangle = \left\langle u_1^t u_1^{t'} \right\rangle \Rightarrow C_1^{tt'} = \delta^{tt'} \frac{1}{\alpha} C_0^{tt'}$.

However, these dynamics still describe the random field h_1 and injected colored noises $u_{2,k},\ u_{4,k}$. We would like to have a fully deterministic description in terms of the concentration response and correlation functions.

B.2 Closing the DMFT equations

Here, we derive a closed set of equations in terms of correlations and responses. Taking the derivative of the h_3 equation by u_3

$$\frac{\delta}{\delta u_3^s} h_3^t = \frac{\delta}{\delta u_3}\left(u_3 + \frac{1}{\nu}\sum_{t'=0}^{t} R_{4\to 2}^{tt'} h_3^{t'}\right)$$

and using $R_3 = \frac{\delta h_3^t}{\delta u_3^s}$

$$R_3^{ts} = \delta^{ts} + \frac{1}{\nu}\sum_{t'=0}^{t} R_{4\to 2}^{tt'} R_3^{ts}.$$

Similarly taking a derivative h_2 w.r.t. u_4

$$R_{4\to 2k}^{ts} = \lambda_k R_{4\to 0k}^{ts}$$

repeating the same process for $R_{4\to 0}$ and R_4

$$\left(\delta^{t+1,t'} - \delta^{t,t'}\right) R_{4\to 0k}^{t',s} = -\eta\left(\nu C_3^{tt}\right)^{-1/2} R_{4k}^{ts}$$

$$R_{4k}^{ts} = \delta^{ts} + \sum_{t'=0}^{t} R_3^{tt'} R_{4\to 2k}^{t's}.$$

Using matrix notation in time domain, and introduction the integration operator $\Theta^{ts} = \eta 1_{s<t}(\nu c_3^s)^{-1/2}$

$$\Theta = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ \frac{\eta}{\sqrt{\nu C_3^{00}}} & 0 & 0 & 0 & 0 \\ \frac{\eta}{\sqrt{\nu C_3^{00}}} & \frac{\eta}{\sqrt{\nu C_3^{11}}} & \ddots & \vdots & \vdots \\ \frac{\eta}{\sqrt{\nu C_3^{00}}} & \frac{\eta}{\sqrt{\nu C_3^{11}}} & \ddots & 0 & \vdots \\ \frac{\eta}{\sqrt{\nu C_3^{00}}} & \frac{\eta}{\sqrt{\nu C_3^{11}}} & \ddots & \frac{\eta}{\sqrt{\nu C_3^{tt}}} & \ddots \end{pmatrix}$$

we can solve the above equations for R_3 to find a self-consistent equation

$$R_3 = I - \frac{1}{\nu d}\sum_{k=1}^{d} \lambda_k \left(\Theta\left[I + \lambda_k R_3 \Theta\right]^{-1}\right) R_3. \tag{B.2}$$

We can also close the equation of h_0

$$h_{0k} = h_{0k}^0 j - \Theta\left(u_{4k} + R_3\left(u_{2k} + \lambda_k h_{0k}\right)\right)$$

with j the all ones time vector. We can formally solve this to find a fully explicit equation for h_0

$$h_{0k} = \left[I + \lambda_k \Theta R_3\right]^{-1}\left(h_{0k}^0 j - \Theta\left(u_{4k} + R_3 u_{2k}\right)\right).$$

Similarly for h_2

$$h_{2k}^t = u_{2k}^t + \lambda_k\left[I + \lambda_k \Theta R_3\right]^{-1}\left(h_{0k}^0 j - \Theta\left(u_{4k} + R_3 u_{2k}\right)\right)$$

B.2.1 CORRELATIONS

Starting with the h_3 equation we see

$$h_3^t - \frac{1}{\nu}\sum_{t'=0}^{t} R_{4\to 2}^{tt'} h_3^{t'} = u_3$$

multiplying by the same equation at time s and taking expectation

$$\left\langle \left(h_3^t - \frac{1}{\nu}\sum_{t'=0}^{t} R_{4\to 2}^{tt'} h_3^{t'}\right)\left(h_3^s - \frac{1}{\nu}\sum_{s'=0}^{t} R_{4\to 2}^{ss'} h_3^{s'}\right)\right\rangle = \left\langle u_3^t u_3^s \right\rangle$$

$$\nu C_3^{ts} - \sum_{s'=0}^{t} R_{4\to 2}^{ss'} C_3^{s't} - \sum_{t'=0}^{t} R_{4\to 2}^{tt'} C_3^{t's} + \frac{1}{\nu}\sum_{t'=0}^{t}\sum_{s'=0}^{t} R_{4\to 2}^{tt'} R_{4\to 2}^{ss'} C_3^{t's'} = C_2$$

using $R_3 = I + \frac{1}{\nu} R_{4\to 2} R_3$ and simplifying we can rewrite this as

$$C_3 = \frac{1}{\nu} R_3 C_2 R_3^\top .$$

Similarly calculating C_2 directly

$$h_{2k}^t = u_{2k}^t + \lambda_k \left[I + \lambda_k \Theta R_3\right]^{-1} \left(h_{0k}^0 j - \Theta\left(u_{4k} + R_3 u_{2k}\right)\right)$$

$$\begin{aligned}\left\langle h_{2k} h_{2k}^\top \right\rangle = \Big\langle &\Big(u_{2k} + \lambda_k \left[I + \lambda_k \Theta R_3\right]^{-1} \left(h_{0k}^0 j - \Theta\left(u_{4k} + R_3 u_{2k}\right)\right)\Big) \\ &\Big(u_{2k}^t + \lambda_k \left[I + \lambda_k \Theta R_3\right]^{-1} \left(h_{0k}^0 j - \Theta\left(u_{4k} + R_3 u_{2k}\right)\right)\Big)^\top \Big\rangle\end{aligned}$$

we can simplify it to

$$\begin{aligned}C_{2k} = \Big(&\lambda_k c_1 - \lambda_k^2 c_1 \left(\left[I + \lambda_k \Theta R_3\right]^{-1} \Theta R_3\right)^\top - \lambda_k^2 \left[I + \lambda_k \Theta R_3\right]^{-1} \Theta R_3 c_1 \\ &+\lambda_k^2 \left[I + \lambda_k \Theta R_3\right]^{-1} \left(\left(\left\langle \left(h_{0k}^0\right)^2 \right\rangle jj^\top + \Theta\left(C_3 + \lambda_k R_3 c_1 R_3^\top\right)\Theta^\top\right)\right)\left[I + \lambda_k \Theta R_3\right]^{-\top}\Big).\end{aligned}$$

We also have that

$$C_1^{tt'} = \delta^{tt'} c_1^t = \delta^{tt'} \frac{1}{\alpha} C_0^{tt'}.$$

Finally a direct calculation for $C_{0k} = \left\langle h_{0k} h_{0k}^\top \right\rangle$ gives

$$\begin{aligned}C_{0k} &= \left\langle \left(\left[I + \lambda_k \Theta R_3\right]^{-1} \left(h_{0k}^0 j - \Theta\left(u_{4k} + R_3 u_{2k}\right)\right)\right)\left(\left[I + \lambda_k \Theta R_3\right]^{-1} \left(h_{0k}^0 j - \Theta\left(u_{4k} + R_3 u_{2k}\right)\right)\right)^\top \right\rangle \\ &= \left[I + \lambda_k \Theta R_3\right]^{-1} \left(\left(\left\langle \left(h_{0k}^0\right)^2 \right\rangle jj^\top + \Theta\left(C_3 + \lambda_k R_3 c_1 R_3^\top\right)\Theta^\top\right)\right)\left[I + \lambda_k \Theta R_3\right]^{-\top}\end{aligned}$$

which concludes the all required correlation functions.

B.3 RESULTS

To get a more intutive understanding we can define $H_k := \left[I + \lambda_k \Theta R_3\right]^{-1}$ and $\langle f_k \rangle_k := \frac{1}{d}\sum_k f_k$, $\left\langle \left(h_{0k}^0\right)^2 \right\rangle_{h^0} := \beta_k^2$ and re-write the C_0 equation

$$\begin{aligned}C_{0k} &= \beta_k^2 H_k j \left(jH_k\right)^\top + H_k \Theta\left(C_3 + \lambda_k R_3 c_1 R_3^\top\right)\left(\Theta H_k\right)^\top \\ C_0 &= \left\langle \lambda_k \left(\beta_k^2 H_k j \left(jH_k\right)^\top + H_k \Theta\left(C_3 + \lambda_k R_3 c_1 R_3^\top\right)\left(\Theta H_k\right)^\top\right)\right\rangle_k\end{aligned}$$

we now see that the first term propagates forward the initial conditions, and the second propagates forward the current state.

For C_2 we may notice that $I - \lambda_k H_k \Theta R_3 = H_k \left(H_k^{-1} - \lambda_k \Theta R_3\right) = H_k \left(I + \lambda_k \Theta R_3 - \lambda_k \Theta R_3\right) = H_k$ so

$$\begin{aligned} C_{2k} &= \left(\lambda_k \left(I - \lambda_k \left(H_k \Theta R_3\right)\right) c_1 \left(I - \lambda_k \left(H_k \Theta R_3\right)^\top\right) + \lambda_k^2 H_k \left(\beta_k^2 j j^\top + \Theta C_3 \Theta^\top\right) H_k^\top\right) \\ &= \lambda_k^2 \beta_k^2 H_k j \left(j H_k\right)^\top + \lambda_k H_k c_1 H_k^\top + \lambda_k^2 H_k \Theta C_3 \left(\Theta H_k\right)^\top \\ C_2 &= \left\langle \lambda_k^2 \beta_k^2 H_k j \left(j H_k\right)^\top + \lambda_k H_k \left(c_1 + \lambda_k \Theta C_3 \Theta^\top\right) H_k^\top \right\rangle_k \end{aligned}$$

we can see the first term again propagates the initial conditions and the second propagates forward the current state.

To complete these we have the simpler relations

$$C_1^{tt'} = \delta^{tt'} c_1^t = \delta^{tt'} \frac{1}{\alpha} C_0^{tt'} \tag{B.3}$$

$$C_3 = \frac{1}{\nu} R_3 C_2 R_3^\top .$$

The diagonal of C_2 takes an especially simple form pretty simple

$$C_{2k}^{tt} = \left\langle \left(u_{2k}^t + \lambda_k h_{0k}^t\right)\left(u_{2k}^t + \lambda_k h_{0k}^t\right)\right\rangle = \left\langle u_{2k}^t u_{2k}^t \right\rangle + 2\lambda_k \left\langle u_{2k}^t h_{0k}^t \right\rangle + \lambda_k^2 \left\langle h_{0k}^t h_{0k}^t \right\rangle$$

since u_{2k}^t, h_{0k}^t are uncorrelated by causality we have (defining $C_{0k}^{tt'} := \langle h_{0k}^t h_{0k}^t \rangle$)

$$\left\langle u_{2k}^t u_{2k}^t \right\rangle + \lambda_k^2 \left\langle h_{0k}^t h_{0k}^t \right\rangle = \lambda_k C_1^{tt} + \lambda_k^2 \left\langle h_{0k}^t h_{0k}^t \right\rangle = \lambda_k C_1^{tt} + \lambda_k^2 C_{0k}^{tt}$$

now averaging over modes to get C_2^{tt}

$$c_2^t = C_2^{tt} = C_1^{tt} \frac{1}{d} \sum_k \lambda_k + \frac{1}{d} \sum_k \lambda_k^2 C_{0k}^{tt} = \bar{\lambda} C_1^{tt} + \frac{1}{d} \sum_k \lambda_k^2 C_{0k}^{tt}$$

using the equal time identity of C_0, C_1

$$c_2^t = C_2^{tt} = \frac{\langle \lambda_k \rangle_k}{\alpha} C_0^{tt} + \left\langle \lambda_k^2 C_{0k}^{tt} \right\rangle_k$$

We can also see the first term is SGD noise as it vanishes for $\alpha \to 1$

As the operators in these equations are causal (and therefore lower triangular) we can solve them numerically by stepping through time.

C Random Matrix Theory Interpretation

In this section we show that the Z-transform of the response function can be interpreted as the companion Stieltjes transform to $\Lambda^{1/2} A^T A \Lambda^{1/2}$, giving another interpretation to its role, and showing that the EoS condition is equivalent to the familiar sharpness stability condition $\eta_{\text{eff}} = 2/\lambda_{\max}$; where $\lambda_{\max}$ is the largest eigenvalue fo the random matrix $\Lambda^{1/2} A^T A \Lambda^{1/2}$.

We start by showing the equivalence to the Stieltjes transform; by definition

$$h_{0k} = h_{0k}^0 j - \Theta h_4.$$

now applying the cavity argument only to the ψ randomness

$$\begin{aligned} h_3^t &= \frac{1}{\sqrt{\nu d}} A h_2 \\ &= \frac{1}{\sqrt{\nu d}} A \left(u_2(t) + \Lambda \bar{h}_0(t)\right) \\ &= \frac{1}{\sqrt{\nu d}} A \left(u_2(t) + \Lambda \left(h_{0k}^0 j - \Theta h_4\right)\right) \\ &= \frac{1}{\sqrt{\nu d}} A \left(u_2(t) + \Lambda \left(h_{0k}^0 j - \Theta \frac{1}{\sqrt{N}} A^T h_3^t\right)\right). \end{aligned}$$

To find the response operator, we introduce a small source u_3 to h_3 such that $h_3^t = \frac{1}{\sqrt{\nu d}} A h_2 + u_3$ and take a derivative w.r.t. to it

$$\frac{\delta h_3}{\delta u_3} = I - \Lambda \frac{1}{\sqrt{\nu d}} A \Theta \frac{1}{\sqrt{N}} A^T \frac{\delta h_3}{\delta u_3}$$

where $\frac{\delta u_2}{\delta u_3} = 0$ since they are two unrelated noise sources. We can now solve for R_3

$$\left(I + \frac{1}{\sqrt{\nu d}} A \Lambda \Theta \frac{1}{\sqrt{N}} A^T \right) \frac{\delta h_3}{\delta u_3} = I$$

$$\frac{\delta h_3}{\delta u_3} = \left[I + A \Lambda \Theta \frac{1}{N} A^T \right]^{-1}$$

Noting that Θ is an operator in time domain, whereas A, Λ are operators in feature and input space we can commute them to find

$$\frac{\delta h_3}{\delta u_3} = \left[I + \Theta \frac{A \Lambda A^T}{N} \right]^{-1}$$

Z-transforming the equations and then multiplying by $\hat{\Theta}$

$$\hat{\Theta} \frac{\delta h_3}{\delta u_3} = \left[\hat{\Theta}^{-1} + \frac{A \Lambda A^T}{N} \right]^{-1}$$

finally, averaging in feature space we see that $\underline{m} := \hat{\Theta} \hat{R}_3 = \hat{\Theta} \frac{1}{N} \operatorname{Tr} \left[\frac{\delta h_3}{\delta u_3} \right]$ is the Stieltjes transform of $\frac{A \Lambda A}{N}$ with an argument $w = -\hat{\Theta}^{-1} = -\frac{z-1}{\eta_{\text{eff}}}$

$$\underline{m} = \operatorname{Tr} \left[\frac{A \Lambda A^T}{N} - w \right]^{-1}.$$

We can now reinterpret the self consistent equation for $\hat{R}_3$ in terms of $\underline{m}, w$

$$\frac{1}{\hat{R}_3(z)} = 1 + \frac{1}{\nu} \left\langle \frac{\lambda_k \eta_{\text{eff}}}{z - 1 + \lambda_k \eta_{\text{eff}} \hat{R}_3(z)} \right\rangle_k \iff w = -\frac{1}{\underline{m}} + \frac{1}{\nu} \left\langle \frac{\lambda_k}{1 + \lambda_k \underline{m}} \right\rangle_k$$

which is the Silverstein equation for the companion Stieltjes transform of $\Lambda^{1/2} A^T A \Lambda^{1/2} / N$, i.e. the Silverstein equation of the matrix $\frac{A \Lambda A}{N}$ written in input space (d); noting that the two matrices share the same non-zero spectrum.

This perspective gives a particularly sharp interpretation for the edge of stability phase condition, and justify its name. The largest eigenvalue of a Wishart is given by $\underline{m}(-1)$ that solves

$$1 = \frac{1}{\nu} \left\langle \frac{\lambda_k^2 \underline{m}^2}{(1 + \lambda_k \underline{m})^2} \right\rangle$$

and given this solution for $\underline{m}$, the largest eigenvalue λ_{max} is explicit

$$\lambda_{\text{max}} = -\frac{1}{\underline{m}} + \frac{1}{\nu} \left\langle \frac{\lambda_k}{1 + \lambda_k \underline{m}} \right\rangle.$$

Now rewriting our conditions in terms of $w, \underline{m}$

$$\nu = \left\langle \left(\frac{\lambda_k \frac{\eta_{\text{eff}}}{2} \hat{R}_3(-1)}{-1 + \lambda_k \frac{\eta_{\text{eff}}}{2} \hat{R}_3(-1)} \right)^2 \right\rangle \iff 1 = \frac{1}{\nu} \left\langle \left(\frac{\lambda_k \underline{m}(z=-1)}{1 + \lambda_k \underline{m}(z=-1)} \right)^2 \right\rangle$$

$$r_-^{-1} = 1 + \frac{1}{\nu} \left\langle \frac{\lambda_k \eta_{\text{eff}}}{-2 + \lambda_k \eta_{\text{eff}} r_-} \right\rangle_k \iff \frac{2}{\eta_{\text{eff}}} = -\frac{1}{\underline{m}} + \frac{1}{\nu} \left\langle \frac{\lambda_k}{1 + \lambda_k \underline{m}} \right\rangle_k$$

so we find that the effective learning is exactly $\eta_{\text{eff}} = \frac{2}{\lambda_{\text{max}}}$ - the canonical EoS, expressed self-consistently for a random matrix.

D LONG TIME BEHAVIOR AND EFFECTIVE LEARNING RATE STEADY STATE

As discussed in the main text, after a certain amount of time, the solution to the DMFT equations in section B.3 settles on a steady state for the effective learning rate. Our results uncover three possible phases for this steady state: (1) edge of stability, (2) edge of stochastic stability, and (3) width-limited regime.

To find these steady states, we assume $\eta_{\text{eff}}^t \equiv \eta_{\text{eff}}^\infty$ and solve for η_{eff}^∞ self-consistently. Note that this assumption is *not* limited to a fixed point; for example, in a limit cycle, the η_{eff}^∞ we find would be its zero-frequency component. For the rest of this section, we drop the ∞ from η_{eff}^∞ for notational brevity.

To solve these equations, we use the z-transform, which diagonalizes time operators that are TTI, i.e. $A(t,s) = A(t-s)$. We start by z-transforming the time integration operator Θ. Writing the operator Θ^{ts} as a function function of the time lag $u = t - s$

$$\Theta^u = \eta_{\text{eff}} \sum_{v>0} \delta^{u,v}$$

where only positive time lags contribute and using the definition of the Z-transform we find

$$\hat{\Theta} = \eta_{\text{eff}} \sum_u \sum_{v>0} \delta^{u,v} z^{-u} = \eta_{\text{eff}} \sum_{u>0} z^{-u} = \eta_{\text{eff}} \frac{1}{z-1}.$$

Inspecting the self-consistent equation for R_3 we see that if Θ is TTI then R_3 must be TTI too, transforming Eq. B.2 we find

$$\hat{R}_3(z) = 1 - \frac{1}{\nu d} \sum_{k=1}^{d} \frac{\lambda_k \eta_{\text{eff}} \frac{1}{z-1} \hat{R}_3(z)}{1 + \lambda_k \eta_{\text{eff}} \frac{1}{z-1} \hat{R}_3(z)}$$

where the physical solution lives on the branch connected to $\lim_{z\to\infty} \hat{R}_3 = 1 \iff R_3\left(t - s = 0\right) = 1$.

Similarly, we can use the double z-transform to describe the correlation functions, which are not TTI.[2] We use the notation

$$\tilde{C}_k(z_1, z_2) := \sum_{t_1,t_2} C_j^{t_1,t_2} z_1^{-t_1} z_2^{-t_2} \tag{D.1}$$

$$\hat{c}_i(\zeta) := \sum_t C_i^{t,t} \zeta^{-t} \tag{D.2}$$

and we have the important relation between the two

$$\hat{c}_j = \int_0^{2\pi} \frac{d\theta}{2\pi} \tilde{C}_j \left(\sqrt{\zeta} e^{i\theta}, \sqrt{\zeta} e^{-i\theta}\right),$$

as shown in App. A.

[2]To see that, consider a system with the fact that $C_0^{tt} = \mathcal{L}^t$; if it were TTI $C_0^{tt} \equiv \text{const}$, so the system could never move.

Using the variables ζ, w with $z_1 = \sqrt{\zeta} w$ and $z_2 = \sqrt{\zeta}/w$, we then have the Z-transforms of C_0, C_3

$$\begin{aligned}
\tilde{C}_0\left(\sqrt{\zeta}w, \sqrt{\zeta}/w\right) &= \left\langle \lambda_k \beta_k^2 \widehat{H_k j}\left(\sqrt{\zeta}w\right) \widehat{H_k j}\left(\sqrt{\zeta}/w\right)\right\rangle_k \\
&+ \frac{\eta_{\text{eff}}^2 \left\langle \lambda_k \hat{H}_k\left(\sqrt{\zeta}w\right) \hat{H}_k\left(\sqrt{\zeta}/w\right)\right\rangle}{\left(\sqrt{\zeta}w - 1\right)\left(\sqrt{\zeta}/w - 1\right)} \tilde{C}_3\left(\sqrt{\zeta}w, \sqrt{\zeta}/w\right) \\
&+ \frac{\eta_{\text{eff}}^2 \hat{R}_3\left(\sqrt{\zeta}w\right) \hat{R}_3\left(\sqrt{\zeta}/w\right) \left\langle \lambda_k^2 \hat{H}_k\left(\sqrt{\zeta}w\right) \hat{H}_k\left(\sqrt{\zeta}/w\right)\right\rangle}{\left(\sqrt{\zeta}w - 1\right)\left(\sqrt{\zeta}/w - 1\right)} \hat{c}_1(\zeta) \\
\tilde{C}_3\left(\sqrt{\zeta}w, \sqrt{\zeta}/w\right) &= \frac{\hat{R}_3\left(\sqrt{\zeta}w\right) \hat{R}_3\left(\sqrt{\zeta}/w\right)}{\nu - \eta_{\text{eff}}^2 \frac{\hat{R}_3\left(\sqrt{\zeta}w\right)}{\sqrt{\zeta}w - 1} \frac{\hat{R}_3\left(\sqrt{\zeta}/w\right)}{\sqrt{\zeta}/w - 1} \left\langle \lambda_k^2 \hat{H}_k\left(\sqrt{\zeta}w\right) \hat{H}_k\left(\sqrt{\zeta}/w\right)\right\rangle_k} \\
&\times \left(\left\langle \lambda_k^2 \widehat{H_k j}\left(\sqrt{\zeta}w\right) \widehat{H_k j}\left(\sqrt{\zeta}/w\right)\right\rangle_k + \left\langle \lambda_k \hat{H}_k\left(\sqrt{\zeta}w\right) \hat{H}_k\left(\sqrt{\zeta}/w\right)\right\rangle_k \hat{c}_1(\zeta)\right)
\end{aligned} \tag{D.3}$$

Using the relation $c_1^t = \frac{1}{\alpha} C_0^{tt}$ in Eq. equation B.3 we find the expression for the loss $\hat{c}_1$ and the gradient norm squared $\hat{c}_3$

$$\begin{aligned}
\hat{c}_1(\zeta) &= \frac{\hat{\mathcal{B}}(\zeta)}{1 - \hat{\mathcal{K}}(\zeta)} \\
\hat{c}_3(\zeta) &= \hat{\mathcal{S}}(\zeta) + \hat{\mathcal{T}}(\zeta) \frac{\hat{\mathcal{B}}(\zeta)}{1 - \hat{\mathcal{K}}(\zeta)}
\end{aligned} \tag{D.4}$$

with the following definitions:

$$\begin{aligned}
\hat{\mathcal{B}}(\zeta) &:= \frac{1}{\alpha} \oint \frac{dw}{2\pi i w} \left(B_1 + \frac{M_1}{\nu - M_2} B_2\right) \\
\hat{\mathcal{K}}(\zeta) &:= \frac{1}{\alpha} \oint \frac{dw}{2\pi i w} \left(\frac{\left(\sqrt{\zeta}w - 1\right)\left(\sqrt{\zeta}/w - 1\right)}{\eta_{\text{eff}}^2 \hat{R}\left(\sqrt{\zeta}w\right) \hat{R}_3\left(\sqrt{\zeta}/w\right)} \frac{M_1^2\left(\sqrt{\zeta}w, \sqrt{\zeta}/w\right)}{\nu - M_2\left(\sqrt{\zeta}w, \sqrt{\zeta}/w\right)} + M_2\left(\sqrt{\zeta}w, \sqrt{\zeta}/w\right)\right) \\
\hat{\mathcal{S}}(\zeta) &:= \oint \frac{dw}{2\pi i w} \frac{\hat{R}\left(\sqrt{\zeta}w\right) \hat{R}_3\left(\sqrt{\zeta}/w\right) B_2\left(\sqrt{\zeta}w, \sqrt{\zeta}/w\right)}{\nu - M_2\left(\sqrt{\zeta}w, \sqrt{\zeta}/w\right)} \\
\hat{\mathcal{T}}(\zeta) &:= \oint \frac{dw}{2\pi i w} \frac{\left(\sqrt{\zeta}w - 1\right)\left(\sqrt{\zeta}/w - 1\right)}{\eta_{\text{eff}}^2} \frac{M_1\left(\sqrt{\zeta}w, \sqrt{\zeta}/w\right)}{\nu - M_2\left(\sqrt{\zeta}w, \sqrt{\zeta}/w\right)} \\
B_a(z_1, z_2) &:= \left\langle \lambda_k^a \beta_k^2 \frac{z_1}{z_1 - 1 + \lambda_k \eta_{\text{eff}} \hat{R}_3(z_1)} \frac{z_2}{z_2 - 1 + \lambda_k \eta_{\text{eff}} \hat{R}_3(z_2)} \right\rangle \\
M_a(z_1, z_2) &:= \left\langle \lambda_k^a \frac{\eta_{\text{eff}} \hat{R}_3(z_1)}{z_1 - 1 + \lambda_k \eta_{\text{eff}} \hat{R}_3(z_1)} \frac{\eta_{\text{eff}} \hat{R}_3(z_2)}{z_2 - 1 + \lambda_k \eta_{\text{eff}} \hat{R}_3(z_2)} \right\rangle.
\end{aligned}$$

As explained in Appendix A, we can find the long-time behavior using the final value theorem

$$\eta_{\text{eff}}^{\infty} = \eta / \sqrt{c_3(\infty)} = \eta / \sqrt{\lim_{\zeta \to 1} (\zeta - 1) \hat{c}_3(\zeta)}$$

giving us a self-consistency condition and allowing us to determine $\eta_{\text{eff}}^{\infty}$ (where we kept the ∞ superscript explicit to emphasize the steady-state condition).

Thus, self-consistency requires a simple pole at $\zeta = 1$. Inspecting the terms we find three candidates

$$\begin{aligned}
&(1) \quad \lim_{\zeta \to 1} \left| e^{i\theta} - 1 + \lambda_k \eta_{\text{eff}} \hat{R}_3 \right|^2 = 0; \\
&(2) \quad \lim_{\zeta \to 1} \nu - \left\langle \frac{\lambda_k^2}{\left| e^{i\theta} - 1 + \lambda_k \eta_{\text{eff}} \hat{R}_3 \right|^2} \right\rangle \eta_{\text{eff}}^2 \left| \hat{R}_3 \right|^2 = 0; \\
&(3) \quad \lim_{\zeta \to 1} \hat{\mathcal{K}}(\zeta) - 1 = 0.
\end{aligned}$$

As we show below, these three conditions correspond to the three phases from the main text. Before detailing the phases, we evaluate the contour integral in $\hat{\mathcal{B}}, \hat{\mathcal{K}}, \hat{\mathcal{S}}, \hat{\mathcal{T}}$ as they would be used throughout.

D.1 Evaluating the contour integral in $\hat{\mathcal{K}}, \hat{\mathcal{T}}$

A central quantity we will use in this section is

$$g_k(z) := \frac{\lambda_k}{q - \lambda_k}; \quad q := \frac{1 - z}{\eta_{\text{eff}} \hat{R}_3(z)}$$

first we show five identities

(Identity 1) Multiplying the fraction in the definition ofg_k by the conjugate of the denominator

$$g_k(z) = \frac{\lambda_k(\bar{q} - \lambda_k)}{|q - \lambda_k|^2} = \frac{\lambda_k \bar{q}}{|q - \lambda_k|^2} - \frac{\lambda_k^2}{|q - \lambda_k|^2} = \frac{\bar{q}}{\lambda_k} |g_k|^2 - |g_k|^2 \tag{D.5}$$

(Identity II) Taking the conjugate of D.5 and subtracting it from D.5 while noting that $\overline{g_k(z)} = g_k(\bar{z})$, we find

$$g_k(z) - g_k(\bar{z}) = (\bar{q} - q) \frac{|g_k|^2}{\lambda_k} \tag{D.6}$$

(Identity III) similarly multiplying D.5 by q and the conjugate by $\bar{q}$ and subtracting

$$q g_k(z) - \bar{q} g_k(\bar{z}) = \frac{|q|^2}{\lambda_k} |g_k|^2 - q |g_k|^2 - \left(\frac{|q|^2}{\lambda_k} |g_k|^2 - \bar{q} |g_k|^2 \right) = (\bar{q} - q) |g_k|^2 . \tag{D.7}$$

(Identity IV) another useful identify is found by writing the $\hat{R}_3$ self-consistent equation in terms of g we find for $\hat{R}_3 \neq 0$

$$\hat{R}_3(z) = \left(1 - \frac{1}{\nu} \frac{1}{\hat{R}_3} \langle g_k \rangle_k \right)^{-1} \Rightarrow q \langle g_k \rangle_k = q\nu \left(\hat{R}_3(z) - 1 \right) = \nu \left(\frac{1 - z}{\eta_{\text{eff}}} - q \right) \tag{D.8}$$

(Identity V) finally, using D.7 and then identity D.8 we find

$$\nu - \left\langle |g_k|^2 \right\rangle = \nu - \left\langle \frac{q g_k(z) - \bar{q} g_k(\bar{z})}{(\bar{q} - q)} \right\rangle = \nu - \frac{\nu \left(\frac{1 - z}{\eta_{\text{eff}}} - q \right) - \nu \left(\frac{1 - \bar{z}}{\eta_{\text{eff}}} - \bar{q} \right)}{(\bar{q} - q)}$$

which can be simplified to

$$\nu - \left\langle |g_k|^2 \right\rangle = -\frac{\nu(\bar{z} - z)}{\eta_{\text{eff}}(\bar{q} - q)} \tag{D.9}$$

D.1.1 Simplifying $\hat{\mathcal{T}}$

we can now simplify $\hat{\mathcal{T}}$ on the contour integral where $z_1 = \bar{z}_2$. Replacing the denominator by the identity D.9 and then rewriting using g_k

$$\hat{\mathcal{T}} = \oint \frac{dz}{2\pi i z} \frac{|z - 1|^2}{-\frac{\nu(\bar{z} - z)}{\eta_{\text{eff}}(\bar{q} - q)}} \left\langle \frac{\lambda_k \left| \hat{R}_3 \right|^2}{\left| z - 1 + \lambda_k \eta_{\text{eff}} \hat{R}_3 \right|^2} \right\rangle = \oint \frac{dz}{2\pi i z} \frac{|z - 1|^2}{-\frac{\nu(\bar{z} - z)}{(\bar{q} - q)}} \frac{1}{\eta_{\text{eff}}} \left\langle \frac{|g_k|^2}{\lambda_k} \right\rangle$$

using D.6 and then D.8

$$\hat{\mathcal{T}} = \oint \frac{dz}{2\pi i z} \frac{|z - 1|^2}{-\nu(\bar{z} - z)} \frac{1}{\eta_{\text{eff}}} \langle g_k(z) - g_k(\bar{z}) \rangle = \oint \frac{dz}{2\pi i z} \frac{|z - 1|^2}{-\nu(\bar{z} - z)} \frac{1}{\eta_{\text{eff}}} \left(\nu \left(\hat{R}_3(z) - 1 \right) - \nu \left(\overline{\hat{R}_3(z)} - 1 \right) \right)$$

simplifying

$$\hat{\mathcal{T}} = \oint \frac{dz}{2\pi i z} \frac{|z - 1|^2}{\eta_{\text{eff}}} \left(\frac{\hat{R}_3(\bar{z}) - \hat{R}_3(z)}{\bar{z} - z} \right)$$

D.1.2 SIMPLIFYING $\hat{\mathcal{K}}$

Where we also used the fact $\overline{\hat{R}_3(z)} = \hat{R}_3(\overline{z})$ which can be seen since the Laurent coefficients of $\hat{R}_3$ are real (as there are time steps in the z-transform and R_3 is real!).

Similarly for $\hat{\mathcal{K}}$

$$\hat{\mathcal{K}}(\zeta=1) = \frac{1}{\alpha}\oint \frac{dz}{2\pi i z}\left(\frac{\left|e^{i\theta}-1\right|^2\left\langle \frac{\lambda_k}{\left|e^{i\theta}-1+\lambda_k\eta_{\text{eff}}\hat{R}_3\right|^2}\right\rangle^2 \eta_{\text{eff}}^2\left|\hat{R}_3\right|^2}{\nu-\left\langle \frac{\lambda_k^2}{\left|e^{i\theta}-1+\lambda_k\eta_{\text{eff}}\hat{R}_3\right|^2}\right\rangle \eta_{\text{eff}}^2\left|\hat{R}_3\right|^2} + \left\langle \frac{\lambda_k^2}{\left|e^{i\theta}-1+\lambda_k\eta_{\text{eff}}\hat{R}_3\right|^2}\right\rangle \eta_{\text{eff}}^2\left|\hat{R}_3\right|^2\right).$$

$$\hat{\mathcal{K}}(\zeta=1) = \frac{1}{\alpha}\oint \frac{dz}{2\pi i z}\frac{|z-1|^2\left\langle \frac{|g_k|^2}{\lambda_k}\right\rangle^2}{\nu-\left\langle |g_k|^2\right\rangle} + \left\langle |g_k|^2\right\rangle = \frac{1}{\alpha}\oint \frac{dz}{2\pi i z}\frac{|q|^2\left\langle \frac{|g_k|^2}{\lambda_k}\right\rangle^2 + \nu\left\langle |g_k|^2\right\rangle - \left\langle |g_k|^2\right\rangle^2}{\nu-\left\langle |g_k|^2\right\rangle}$$

Going further requires a sixth, more involved identity

(Identity VI) We would need an additional identity

$$(\overline{q}-q)\left(|q|^2\left\langle\frac{|g_k|^2}{\lambda_k}\right\rangle^2 - \left\langle |g_k|^2\right\rangle^2\right) = q\left\langle g_k(z)\right\rangle\left\langle g_k(z)\right\rangle - \bar{q}\left\langle g_k(\overline{z})\right\rangle\left\langle g_k(\bar{z})\right\rangle \tag{D.10}$$

we prove it below:taking D.6 multiplying by q and averaging, and multiplying it's conjugate by $\bar{q}$ and averaging and then taking the product

$$q\left(\left\langle g_k(z)\right\rangle - \left\langle g_k(\overline{z})\right\rangle\right)\bar{q}\left(\left\langle g_k(\bar{z})\right\rangle - \left\langle g_k(z)\right\rangle\right) = q(\overline{q}-q)\left\langle\frac{|g_k|^2}{\lambda_k}\right\rangle\bar{q}(q-\bar{q})\left\langle\frac{|g_k|^2}{\lambda_k}\right\rangle$$

$$|q|^2\left(\left\langle g_k(z)\right\rangle - \left\langle g_k(\overline{z})\right\rangle\right)\left(\left\langle g_k(\bar{z})\right\rangle - \left\langle g_k(z)\right\rangle\right) = |q|^2\left(2|q|^2-\bar{q}^2-q^2\right)(q-\bar{q})\left\langle\frac{|g_k|^2}{\lambda_k}\right\rangle^2$$

$$|q|^2\left(2\left\langle g_k(z)\right\rangle\left\langle g_k(\bar{z})\right\rangle - \left\langle g_k(z)\right\rangle\left\langle g_k(z)\right\rangle - \left\langle g_k(\overline{z})\right\rangle\left\langle g_k(\bar{z})\right\rangle\right) = |q|^2\left(2|q|^2-\bar{q}^2-q^2\right)\left\langle\frac{|g_k|^2}{\lambda_k}\right\rangle^2$$

now averaging D.7 and taking its product with its conjugate

$$\left(q\left\langle g_k(z)\right\rangle - \bar{q}\left\langle g_k(\overline{z})\right\rangle\right)\left(\bar{q}\left\langle g_k(\bar{z})\right\rangle - q\left\langle g_k(z)\right\rangle\right) = (\overline{q}-q)(q-\bar{q})\left\langle|g_k|^2\right\rangle^2$$

$$|q|^2\left\langle g_k(z)\right\rangle\left\langle g_k(\bar{z})\right\rangle - q^2\left\langle g_k(z)\right\rangle\left\langle g_k(z)\right\rangle - \bar{q}^2\left\langle g_k(\overline{z})\right\rangle\left\langle g_k(\bar{z})\right\rangle + |q|^2\left\langle g_k(\overline{z})\right\rangle\left\langle g_k(z)\right\rangle = \left(2|q|^2-\bar{q}^2-q^2\right)\left\langle|g_k|^2\right\rangle^2$$

$$2|q|^2\left\langle g_k(z)\right\rangle\left\langle g_k(\bar{z})\right\rangle - q^2\left\langle g_k(z)\right\rangle\left\langle g_k(z)\right\rangle - \bar{q}^2\left\langle g_k(\overline{z})\right\rangle\left\langle g_k(\bar{z})\right\rangle = \left(2|q|^2-\bar{q}^2-q^2\right)\left\langle|g_k|^2\right\rangle^2.$$

Subtracting the two results and simplifying yields the desired result

$$(\overline{q}-q)\left(|q|^2\left\langle\frac{|g_k|^2}{\lambda_k}\right\rangle^2 - \left\langle|g_k|^2\right\rangle^2\right) = q\left\langle g_k(z)\right\rangle\left\langle g_k(z)\right\rangle - \bar{q}\left\langle g_k(\overline{z})\right\rangle\left\langle g_k(\bar{z})\right\rangle$$

Simplifying $\hat{\mathcal{K}}$ substituting into $\hat{\mathcal{K}}$ identities D.10, D.9,D.7

$$\frac{1}{\alpha}\oint\frac{dz}{2\pi i z}\frac{\frac{q\langle g_k(z)\rangle\langle g_k(z)\rangle-\bar{q}\langle g_k(\overline{z})\rangle\langle g_k(\bar{z})\rangle}{(\overline{q}-q)} + \nu\frac{q\langle g_k(z)\rangle-\bar{q}\langle g_k(\overline{z})\rangle}{(\bar{q}-q)}}{-\frac{\nu(\bar{z}-z)}{\eta_{\text{eff}}(\bar{q}-q)}}$$

we can now simplify

$$\frac{1}{\alpha}\oint\frac{dz}{2\pi i z}\frac{q\left\langle g_k(z)\right\rangle\left(\left\langle g_k(z)\right\rangle+\nu\right) - \bar{q}\left\langle g_k(\overline{z})\right\rangle\left(\left\langle g_k(\overline{z})\right\rangle+\nu\right)}{-\frac{\nu(\bar{z}-z)}{\eta_{\text{eff}}}}$$

using D.8 in the form $q\left(\langle g_k\rangle_k+\nu\right)=\nu\left(\frac{1-z}{\eta_{\text{eff}}}\right)$

$$\frac{1}{\alpha}\oint\frac{dz}{2\pi iz}\frac{\langle g_k\left(z\right)\rangle\,\nu\left(\frac{1-z}{\eta_{\text{eff}}}\right)-\langle g_k\left(\overline{z}\right)\rangle\,\nu\left(\frac{1-\bar{z}}{\eta_{\text{eff}}}\right)}{-\frac{\nu(\bar{z}-z)}{\eta_{\text{eff}}}}$$

which can be simplified to

$$\frac{1}{\alpha}\oint\frac{dz}{2\pi iz}\frac{\langle g_k\left(\overline{z}\right)\rangle\left(1-\bar{z}\right)-\langle g_k\left(z\right)\rangle\left(1-z\right)}{\bar{z}-z}$$

using D.8 again in the form $\langle g_k\rangle_k=\nu\left(\hat{R}_3\left(z\right)-1\right)$

$$\hat{\mathcal{K}}=\frac{\nu}{\alpha}\oint\frac{dz}{2\pi iz}\frac{\left(\hat{R}_3\left(\bar{z}\right)-1\right)\left(1-\bar{z}\right)-\left(\hat{R}_3\left(z\right)-1\right)\left(1-z\right)}{\bar{z}-z}.$$

D.1.3 Evaluating the contour integral for $\hat{\mathcal{K}}$ at $\zeta=1$

We can expand $\hat{R}_3$ in a Laurent series back to time domain $\hat{R}_3=\sum_{t=0}^{1}R_tz^{-t}=1+\sum_{t=1}^{1}R_tz^{-t}$, additionally at $\zeta=1$ we have $\bar{z}=z^{-1}$, so after some simplification the expression becomes

$$\hat{\mathcal{K}}\left(\zeta=1\right)=-\frac{\nu}{\alpha}\oint\frac{dz}{2\pi iz}\frac{\sum_{t=1}^{\infty}R_tz^t+\sum_{t=1}^{1}R_tz^{-(t-1)}}{1+z}$$

pulling out a shared $\sum_{t=1}^{1}R_tz^{-(t-1)}$ and expanding out the polynomial

$$-\frac{\nu}{\alpha}\sum_{t=1}^{\infty}R_tz^{-(t-1)}\oint\frac{dz}{2\pi iz}\frac{z^{2t-1}+1}{1+z}=-\frac{\nu}{\alpha}\sum_{t=1}^{\infty}R_t\sum_{j=0}^{2t-2}\left(-1\right)^j\oint\frac{dz}{2\pi iz}z^{j-t+1}$$

which we can finally directly integrate using $\oint\frac{d\theta}{2\pi}e^{i\theta(j-t+1)}=\delta_{j-t+1,0}$ with δ the Kronecker delta which gives after the j summation and pushing in the minus from outside

$$\frac{\nu}{\alpha}\sum_{t=1}^{\infty}R_t\left(-1\right)^t=\frac{\nu}{\alpha}\left(\hat{R}_3\left(-1\right)-1\right)$$

which we recognize at $\hat{R}_3-1$ evaluated at at $z=-1$. Using identity D.8 once more we reach the final result

$$\hat{\mathcal{K}}\left(\zeta=1\right)=\frac{\langle g_k\left(z=-1\right)\rangle_k}{\alpha}.$$

D.1.4 Evaluating the contour integral for $\hat{\mathcal{T}}$ at $\zeta=1$

Expanding in a Laurent series again

$$\hat{\mathcal{T}}=\oint\frac{dz}{2\pi iz}\frac{\left|z-1\right|^2}{\eta_{\text{eff}}}\left(\frac{\sum_{t=1}^{\infty}R_tz^t-\sum_{t=1}^{\infty}R_tz^{-t}}{\bar{z}-z}\right)=\frac{1}{\eta_{\text{eff}}}\sum_{t=1}^{\infty}R_t\oint\frac{dz}{2\pi iz}\frac{\left(z-1\right)\left(z^{-1}-1\right)}{\left(z^{-1}-z\right)}\left(z^t-z^{-t}\right)$$

calculating the integral explicitly again and resumming we arrive at

$$\hat{\mathcal{T}}\left(\zeta=1\right)=\frac{2\left(\hat{R}_3\left(-1\right)-1\right)}{\eta_{\text{eff}}}=\frac{2\left\langle g_k\right\rangle_k}{\nu\eta_{\text{eff}}}$$

D.2 Width Limited Pole $\lim_{\zeta\to1}\left|e^{i\theta}-1+\lambda_k\eta_{\text{eff}}\hat{R}_3\right|^2=0$

This term can vanish in two separate ways, either for a specific k, or for all k. The former admits no consistent solution and here we focus here on the latter. For this expression to vanish for all k, we require $e^{i\theta}-1$ and $\eta_{\text{eff}}\hat{R}_3$ vanish simultaneously.

Requiring $e^{i\theta}-1$ vanishes immediately implies $\theta=0$ or equivalently $z=1$. Using the notation $r_{:}=R_3\left(1\right)$ we find two cases. For $r>0$ substituting z=1 into

$$\hat{R}_3\left(z\right)=\left(1+\frac{1}{\nu}\left\langle\frac{\lambda_k\eta_{\text{eff}}}{z-1+\lambda_k\eta_{\text{eff}}\hat{R}_3\left(z\right)}\right\rangle_k\right)^{-1}$$

we immediately find

$$r=1-\frac{1}{\nu}$$

so a solution exists only for $\nu>1$. Another solution can be found for $r=\rho\left(z-1\right)$ in this case

$$1=\lim_{z\to1}\rho\left(z-1\right)+\frac{1}{\nu}\left\langle\frac{\lambda_k\eta_{\text{eff}}\rho\left(z-1\right)}{z-1+\lambda_k\eta_{\text{eff}}\rho\left(z-1\right)}\right\rangle_k$$

which gives the condition

$$\left\langle\frac{\lambda_k\eta_{\text{eff}}\rho}{1+\lambda_k\eta_{\text{eff}}\rho}\right\rangle_k=\nu.$$

The physical branch is the one where $\rho>0$ so this condition can only be satisfied for $\nu<1$ as the denominator is greater than the numerator and ν is a positive real quantity (with $\nu=1$ a borderline case giving a square root divergence).

The above conditions determine the product $\eta_{\text{eff}}\rho$, but to isolate each one individually, we must determine η_{eff} from the residue of $\hat{c}_3$ at $\zeta=1$. Inspecting $\hat{\mathcal{S}}$, $\hat{\mathcal{T}}$, and $\hat{\mathcal{K}}$ the singularity cancels in all of them and only $\hat{\mathcal{B}}$ maintains a singular part, so

$$\lim_{\zeta\to1}(\zeta-1)\hat{c}_3(\zeta)=\frac{\hat{\mathcal{T}}\left(1\right)}{1-\hat{\mathcal{K}}\left(1\right)}\lim_{\zeta\to1}\left(\zeta-1\right)\hat{\mathcal{B}}\left(\zeta\right).$$

Evaluating it at the limit

$$\lim_{\zeta\to1}\left(\zeta-1\right)\hat{\mathcal{B}}\left(\zeta\right)=\lim_{\zeta\to1}\left(\zeta-1\right)\frac{1}{\alpha}\int_0^{2\pi}\frac{d\theta}{2\pi}\left(\left\langle\frac{\lambda_k\beta_k^2}{\left|1+\lambda_k\eta_{\text{eff}}\rho\right|^2}\right\rangle\frac{1}{\left|\sqrt{\zeta}e^{i\theta}-1\right|^2}\right.\tag{D.11}$$
$$\left.+\frac{\left\langle\frac{\lambda_k}{\left|1+\lambda_k\eta_{\text{eff}}\rho\right|^2}\right\rangle\eta_{\text{eff}}^2\rho^2}{\nu-\left\langle\frac{\lambda_k^2}{\left|1+\lambda_k\eta_{\text{eff}}\rho\right|^2}\right\rangle\eta_{\text{eff}}^2\rho^2}\left\langle\frac{\lambda_k^2\beta_k^2}{\left|1+\lambda_k\eta_{\text{eff}}\rho\right|^2}\right\rangle\frac{1}{\left|\sqrt{\zeta}e^{i\theta}-1\right|^2}\right)$$

and after integration the limit yields

$$\lim_{\zeta\to1}\left(\zeta-1\right)\hat{\mathcal{B}}\left(\zeta\right)=\frac{1}{\alpha}\left(\left\langle\frac{\lambda_k\beta_k^2}{\left(1+\lambda_k\eta_{\text{eff}}\rho\right)^2}\right\rangle+\frac{\left\langle\frac{\lambda_k}{\left(1+\lambda_k\eta_{\text{eff}}\rho\right)^2}\right\rangle\eta_{\text{eff}}^2\rho^2}{\nu-\left\langle\frac{\lambda_k^2}{\left(1+\lambda_k\eta_{\text{eff}}\rho\right)^2}\right\rangle\eta_{\text{eff}}^2\rho^2}\left\langle\frac{\lambda_k^2\beta_k^2}{\left(1+\lambda_k\eta_{\text{eff}}\rho\right)^2}\right\rangle\right)$$
$$=\frac{1}{\alpha}\left\langle\frac{\lambda_k\beta_k^2}{1+\lambda_k\eta_{\text{eff}}\rho}\right\rangle.\tag{D.12}$$

The final result is therefore

$$\lim_{\zeta\to1}(\zeta-1)\hat{c}_3(\zeta)=\frac{2\left(\hat{R}_3(-1)-1\right)}{\alpha\eta_{\text{eff}}\left[1-\frac{\nu}{\alpha}\left(\hat{R}_3(-1)-1\right)\right]}\left\langle\frac{\lambda_k\beta_k^2}{1+\lambda_k\eta_{\text{eff}}\,\rho}\right\rangle_k$$
$$\nu=\left\langle\frac{\lambda_k\eta_{\text{eff}}\rho}{1+\lambda_k\eta_{\text{eff}}\rho}\right\rangle_k$$
$$\hat{R}_3\left(-1\right)=\left(1+\frac{1}{\nu}\left\langle\frac{\lambda_k\eta_{\text{eff}}}{\lambda_k\eta_{\text{eff}}\hat{R}_3\left(-1\right)-2}\right\rangle_k\right)^{-1}.\tag{D.13}$$

D.3 Edge of Stability Pole $\lim_{\zeta\to 1}\nu - \left\langle \frac{\lambda_k^2}{\left|e^{i\theta}-1+\lambda_k\eta_{\text{eff}}\hat{R}_3\right|^2}\right\rangle \eta_{\text{eff}}^2\left|\hat{R}_3\right|^2 = 0$

In this limit, we have

$$\left\langle \frac{\lambda_k^2}{\left|e^{i\theta}-1+\lambda_k\eta_{\text{eff}}\hat{R}_3\right|^2}\right\rangle \eta_{\text{eff}}^2\left|\hat{R}_3\right|^2 = \nu$$

Using the identity D.8

$$\frac{z-1}{\eta_{\text{eff}}} = \frac{z-1}{\eta_{\text{eff}}\hat{R}_3(z)}\left(1+\frac{1}{\nu}\left\langle \frac{\lambda_k}{\frac{1-z}{\eta_{\text{eff}}\hat{R}_3(z)}-\lambda_k}\right\rangle_k\right)$$

and taking the imaginary part

$$\Im\left[\frac{1-z}{\eta_{\text{eff}}}\right] = \Im\left[\frac{1-z}{\eta_{\text{eff}}\hat{R}_3(z)}\right]\left(\underbrace{1-\frac{1}{\nu}\eta_{\text{eff}}^2\left|\hat{R}_3\right|^2\left\langle \frac{\lambda_k^2}{\left|z-1+\lambda_k\eta_{\text{eff}}^2\left|\hat{R}_3\right|^2\right|^2}\right\rangle_k}_{=0}\right) = 0,$$

where the last term is zero by assumption, we find that $\Im\left[\frac{1-z}{\eta_{\text{eff}}}\right]=0$, therefore $z=\pm 1$, but only the case $z=-1$ gives a physically consistent solution.

Letting $r_- := \hat{R}_3(z=-1)$ we find the condition

$$\left\langle \frac{\lambda_k^2\eta_{\text{eff}}^2|r_-|^2}{|-2+\lambda_k\eta_{\text{eff}}r_-|^2}\right\rangle = \nu$$

which is the condition for Phase 1 - EoS. This condition has to be solved together with the self-consistent equation for $\hat{R}_3$ in Eq. 3

$$r_-^{-1} = 1+\frac{1}{\nu}\left\langle \frac{\lambda_k\eta_{\text{eff}}}{-2+\lambda_k\eta_{\text{eff}}r_-}\right\rangle_k.$$

Practically, we can first solve for the product $\eta_{\text{eff}}r_-$, and then solve the second one for the ratio, which determines the physical η_{eff}^∞. In Appendix C we show that the result can be interpreted as $\eta_{\text{eff}}^\infty = 2/\lambda_{\max}$ with $\lambda_{\max}$ the largest eigenvalue of the random matrix $A\frac{\Lambda}{N}A^T$.

D.4 Edge of Stochastic Stability Pole $\lim_{\zeta\to 1}\hat{\mathcal{K}}(\zeta)-1=0$

We already evaluated $\hat{\mathcal{K}}(\zeta=1)$ in Appendix D.1 and found

$$\hat{\mathcal{K}}(\zeta=1) = \frac{\langle g_k(z=-1)\rangle_k}{\alpha}.$$

Our EoSS pole assumption thus directly implies

$$\left\langle \frac{\lambda_k\eta_{\text{eff}}r_-}{2-\lambda_k\eta_{\text{eff}}r_-}\right\rangle_k \overset{!}{=} \alpha$$

which has to be solved together with the r_- equation

$$r_- = \left(1+\frac{1}{\nu}\left\langle \frac{\lambda_k\eta_{\text{eff}}}{-2+\lambda_k\eta_{\text{eff}}r_-}\right\rangle_k\right)^{-1}.$$

$$\nu(1-r_-) = \left\langle \frac{\lambda_k\eta_{\text{eff}}r_-}{-2+\lambda_k\eta_{\text{eff}}r_-}\right\rangle_k.$$

Given the relation

$$\lim_{\zeta\to 1}(\zeta-1)c_3 = \hat{\mathcal{T}}(1)\lim_{\zeta\to 1}(\zeta-1)c_1$$

We can also deduce the final loss using

$$\lim_{\zeta\to 1}(\zeta-1)c_1 = \frac{1}{\hat{\mathcal{T}}(1)}\frac{\eta^2}{\eta_{\text{eff}}^2} \tag{D.14}$$

with the η_{eff} that solves the condition above.

E Short Derivation of the GF Scaling Law

Here we give a simple heuristic derivation for the scaling law $L(\tau) \sim \tau^{r_{\text{SGD}}} = \tau^{-\frac{2a+b-1}{b}}$. Writing the loss as a sum over spectral modes $e_k(t) := w_k^* - w_k(t)$ and using the solution $e_k(t) = e_k(0)\exp(-2\lambda_k t)$

$$L = \frac{1}{2}\sum_k \lambda_k e_k^2(t) = \sum_k \lambda_k e_k^2(0)\exp\left(-2\lambda_k\tau\right)$$

and plugging in $\lambda_k = k^{-b}$, $e_k(0) = k^{-a}$ we find

$$\sum_k k^{-2a-b}\exp\left(-2k^{-b}\tau\right).$$

From this we can conclude that at time τ approximately down to $\frac{1}{\tau}$ are learned. Plugging the eigenvalue scaling we find that modes up to k_* are learned

$$\tau_k \sim \frac{1}{2\lambda_k} \sim k^b \Rightarrow k_* = \tau^{1/b}.$$

Summing only modes higher than k_* we finally get an explicit scaling with time

$$L\left(\tau\right) = \frac{1}{2}\sum_{k=k_*(\tau)}^{\infty} k^{-2a-b} \approx \frac{1}{2}\int_{\tau^{1/b}}^{\infty} k^{-(2a+b)}\,dk = \frac{1}{2}\frac{1}{1-(2a+b)}\left[k^{-(2a+b-1)}\right]_{\tau^{1/b}}^{\infty} \tag{E.1}$$

$$\sim \tau^{-\frac{2a+b-1}{b}} = \tau^{-r_{\text{SGD}}}. \tag{E.2}$$

In the same way, the gradient norm in the τ time is

$$C_3(\tau) = \sum_k \lambda_k^2 e_k^2(\tau) = \sum_k k^{-2(a+b)}\exp(-2k^{-b}\tau). \tag{E.3}$$

Using the same cutoff $\kappa_\star \sim \tau^{1/b}$ leads to

$$C_3(\tau) \approx \int_{k_\star}^{\infty} k^{-2(a+b)}dk \sim \tau^{-\frac{2a+2b-1}{b}} = \tau^{-(r_{\text{SGD}}+1)}. \tag{E.4}$$

We now return to the physical time t of normalized gradient flow. Since by construction $\frac{dt}{d\tau} = \sqrt{C_3(\tau)}$, therefore

$$\frac{dt}{d\tau} = \tau^{-\frac{r_{\text{SGD}}+1}{2}}. \tag{E.5}$$

The resulting relation between τ and t depends on whether r_{SGD} is smaller than, equal to, or larger than one. Since

$$r_{\text{SGD}} - 1 = \frac{2a-1}{b},$$

these cases correspond respectively to $a < 1/2$, $a = 1/2$, and $a > 1/2$. For $a < 1/2$, we have $r_{\text{SGD}} < 1$, and hence

$$t(\tau) \sim \int^{\tau} du\, u^{-\frac{r_{\text{SGD}}+1}{2}} \sim \tau^{\frac{1-r_{\text{SGD}}}{2}}.$$

Therefore,

$$\tau(t) \sim t^{\frac{2}{1-r_{\text{SGD}}}},$$

and substituting into $L(\tau) \sim \tau^{-r_{\mathrm{SGD}}}$ gives

$$L(t) \sim t^{-\frac{2r_{\mathrm{SGD}}}{1-r_{\mathrm{SGD}}}}.$$

At $a = 1/2$, we have $r_{\mathrm{SGD}} = 1$, so

$$\frac{dt}{d\tau} \sim \frac{1}{\tau}, \qquad t \sim \log \tau.$$

Thus τ grows exponentially with physical time and the loss decays exponentially. Retaining the prefactor in the continuum integral gives

$$C_3(\tau) \sim \frac{1}{4b\tau^2}, \qquad \tau(t) \sim e^{2\sqrt{b}\,t}, \qquad L(t) \sim e^{-2\sqrt{b}\,t}.$$

Finally, for $a > 1/2$, we have $r_{\mathrm{SGD}} > 1$, and the clock integral converges to a finite value T. Its remaining tail satisfies

$$T - t(\tau) \sim \int_{\tau}^{\infty} du\, u^{-\frac{r_{\mathrm{SGD}}+1}{2}} \sim \tau^{-\frac{r_{\mathrm{SGD}}-1}{2}}.$$

Consequently

$$\tau(t) \sim (T-t)^{-\frac{2}{r_{\mathrm{SGD}}-1}},$$

and

$$L(t) \sim (T-t)^{\frac{2r_{\mathrm{SGD}}}{r_{\mathrm{SGD}}-1}}.$$

Collecting the three regimes

$$L(t) \sim \begin{cases} (T-t)^{\frac{2r_{\mathrm{SGD}}}{r_{\mathrm{SGD}}-1}}, & a > \frac{1}{2}, \\ e^{-2\sqrt{b}\,t}, & a = \frac{1}{2}, \\ t^{-\frac{2r_{\mathrm{SGD}}}{1-r_{\mathrm{SGD}}}}, & a < \frac{1}{2}. \end{cases} \tag{E.6}$$

F Effective Learning Rate Scaling with Batch Size and Width

F.1 Effective learning rate and batch scaling in Phase 2

We derive the stationary effective learning rate $\eta_{\mathrm{eff}}^{\infty}$ on the Phase 2 (EoSS) branch, and quantify its departure from linear scaling with batch size. That is, we quantify when increasing batch size can lead to the same loss with fewer optimization steps and nearly the same compute budget – close to linear scaling of $\eta_{\mathrm{eff}}^{\infty}$ with α. We also show that super-linear scaling is never possible in our setting, meaning the best possible outcome is to break even in terms of compute.

Concretely, we study how $\eta_{\mathrm{eff}}^{\infty}$ varies with $\alpha = B/d$ at fixed $\nu = N/d$ and fixed input spectrum $\{\lambda_k\}_k$.

First, let $r_- = \hat{R}_3(-1)$. The response equation and the Phase 2 condition together read

$$\frac{1}{r_-} = 1 - \frac{1}{\nu}\Big\langle \frac{\lambda_k \eta_{\mathrm{eff}}}{2 - \lambda_k \eta_{\mathrm{eff}} r_-} \Big\rangle_k, \quad \alpha = \Big\langle \frac{\lambda_k \eta_{\mathrm{eff}}}{2 - \lambda_k \eta_{\mathrm{eff}} r_-} \Big\rangle_k r_-. \tag{F.1}$$

Since r_- is independent of k, this just leads to

$$r_- = 1 + \frac{\alpha}{\nu}. \tag{F.2}$$

Given that $\eta_{\mathrm{eff}}^{\infty} > 0$ and $\alpha \geq 0$, this implies that

$$\alpha \geq \frac{\mu_1}{2}\eta_{\mathrm{eff}}^{\infty}(1 + \alpha/\nu), \quad 0 \leq \eta_{\mathrm{eff}}^{\infty} \leq \frac{2\alpha}{\mu_1(1+\alpha/\nu)} \tag{F.3}$$

with $\mu_1 = \langle \lambda_k \rangle_k$. Thus, at fixed spectrum, the spectral products tend to zero and the denominators can be expanded geometrically. Keeping every term through order α^2 gives

$$\alpha = \frac{\mu_1}{2}\eta_{\text{eff}}^{\infty} + \frac{\mu_1}{2\nu}\alpha\eta_{\text{eff}}^{\infty} + \frac{\mu_2}{4}{\eta_{\text{eff}}^{\infty}}^2 + \mathcal{O}(\alpha^3). \tag{F.4}$$

Matching the coefficients to obtain $\eta_{\text{eff}}^{\infty}(\alpha)$ yields to

$$\eta_{\text{eff}}^{\infty}(\alpha) = \frac{2\alpha}{\mu_1}\left[1 - \alpha\left(\frac{\mu_2}{\mu_1^2} + \frac{1}{\nu}\right)\right] + \mathcal{O}(\alpha^3). \tag{F.5}$$

Then, the leading term in α sets the linear scaling while the quadratic term determines its first relative correction

$$\frac{\eta_{\text{eff}}^{\infty}}{\eta_{\text{lin}}^{\infty}} = 1 - C\alpha + \mathcal{O}(\alpha^2) \tag{F.6}$$

with $C = \frac{\mu_2}{\mu_1^2} + 1/\nu$. Notice the sign of this correction is always negative, so the growth of $\eta_{\text{eff}}^{\infty}$ is always sublinear, and larger Cs give smaller tipping points or smaller batch size in the sense of (McCandlish et al., 2018).

The logarithmic slope measures the fractional increase in effective learning rate produced by a fractional increase in batch size

$$p(\alpha) = \frac{d\log\eta_{\text{eff}}^{\infty}}{d\log\alpha}. \tag{F.7}$$

The equality of the two logarithmic derivatives follows from keeping d fixed. Linear scaling corresponds to $p = 1$, whereas $0 < p < 1$ corresponds to sublinear growth. Differentiating the expansion above gives

$$\frac{d\eta_{\text{eff}}^{\infty}}{d\alpha} = \frac{2}{\mu_1}\left[1 - 2C\alpha + \mathcal{O}(\alpha^2)\right] \tag{F.8}$$

and therefore

$$p(\alpha) = 1 - C\alpha + \mathcal{O}(\alpha^2). \tag{F.9}$$

Thus, the coefficient C controls the initial departure from linear scaling. We can trade off some compute efficiency for fewer optimization steps by introducing a small tolerance δ; the local condition $p \geq 1 - \delta$ gives

$$\alpha_\delta = \frac{\delta}{\mu_2/\mu_1^2 + 1/\nu} + \mathcal{O}(\delta^2) \tag{F.10}$$

where the expansion is taken as $\delta \to 0$ at fixed spectrum and ν; yielding a mathematically concrete notion of the tradeoff mentioned above.

For the power law spectrum $\lambda_k = k^{-b}$, define

$$S_j(d,b) = \sum_{k=1}^{d} k^{-jb}, \qquad \mu_1 = \frac{S_1}{d}, \qquad \frac{\mu_2}{\mu_1^2} = \frac{dS_2}{S_1^2}. \tag{F.11}$$

Using the asymptotics of power-law sums, we obtain

$$\frac{\mu_2}{\mu_1^2} \asymp \begin{cases} 1, & 0 \leq b < \frac{1}{2} \\ \log d, & b = \frac{1}{2} \\ d^{\,2b-1}, & \frac{1}{2} < b < 1 \\ d/(\log d)^2, & b = 1 \\ d, & b > 1. \end{cases} \tag{F.12}$$

Consequently, at fixed ν, the coefficient of the initial slope deficit $1 - p$ remains finite for $b < 1/2$, grows logarithmically for $b = 1/2$ and grows with dimension for $b > 1/2$. The linear scaling window size is inversely proportional to that, i.e., macroscopic for $b < 1/2$ and vanishing for $b > 1$.

For completeness, the leading effective learning rate is

$$\eta_{\text{lin}}^{\infty} = \frac{2\alpha}{\mu_1} \sim \begin{cases} 2(1-b)\alpha d^b, & 0 \leq b < 1 \\ 2\alpha d/\log d, & b = 1 \\ 2\alpha d/\zeta(b), & b > 1. \end{cases} \tag{F.13}$$

Here $\zeta(b) = \sum_{k=1}^{\infty} k^{-b}$ denotes the Riemann zeta function, whose value is finite for $\mathbb{R} \ni b > 1$. Together, these expressions specify the leading effective learning rate and the dimension dependence of its first relative correction in every spectral regime.

F.2 CRITICAL BATCH RATIO SEPARATING EOSS AND EOS

We determine the critical batch ratio $\alpha_c = B_c/d$ by imposing the Phase 2 and Phase 3 conditions simultaneously. We consider $\lambda_k = k^{-b}$ and derive large d asymptotics at fixed $\nu > 0$ and fixed b.

At the transition, define

$$s_c = \frac{\eta_{\text{eff}}^{\infty}(\alpha_c)}{2}\left(1 + \frac{\alpha_c}{\nu}\right). \tag{F.14}$$

The two phase conditions become

$$\nu d = \sum_{k=1}^{d}\left(\frac{s_c}{k^b - s_c}\right)^2, \qquad \alpha_c = \frac{1}{d}\sum_{k=1}^{d}\frac{s_c}{k^b - s_c}. \tag{F.15}$$

with $0 < s_c < 1$. These equations determine the transition within the stationary stability-limited regime. For fixed $b > 0$, we first separate the contribution of the largest eigenvalue, corresponding to $k = 1$. The squared phase condition reads

$$\nu d = \left(\frac{s_c}{1 - s_c}\right)^2 + \sum_{k=2}^{d}\left(\frac{s_c}{k^b - s_c}\right)^2. \tag{F.16}$$

The remaining squared terms grow more slowly than d. Thus, at large d, the first term supplies the leading contribution, giving

$$\frac{s_c}{1 - s_c} \simeq \sqrt{\nu d}, \qquad s_c \to 1. \tag{F.17}$$

We now use this result in the unsquared phase condition. Keeping the top-mode contribution and setting $s_c \simeq 1$ in the remaining terms gives

$$B_c \simeq \sqrt{\nu d} + \sum_{k=2}^{d}\frac{1}{k^b - 1}. \tag{F.18}$$

The first term is the contribution of the largest eigenvalue, while the second collects the remaining spectrum.

For $0 < b < 1$, the spectral sum grows as $d^{1-b}/(1-b)$. It grows logarithmically at $b = 1$ and approaches a finite constant for $b > 1$. Comparing this contribution with $\sqrt{\nu d}$ gives the leading critical batch ratio:

$$\alpha_c(\nu; b, d) \simeq \begin{cases} \sqrt{\nu}, & b = 0 \\ \dfrac{d^{-b}}{1-b}, & 0 < b < \frac{1}{2} \\ \dfrac{2 + \sqrt{\nu}}{\sqrt{d}}, & b = \frac{1}{2} \\ \sqrt{\dfrac{\nu}{d}}, & b > \frac{1}{2}. \end{cases} \tag{F.19}$$

The flat-spectrum result $b = 0$ is exact. The remaining expressions are leading approximations as $d \to \infty$ at fixed $b > 0$ and $\nu > 0$.

For $b < 1/2$, the remaining spectrum dominates the critical batch, and the dependence on ν enters at smaller order. At $b = 1/2$, both contributions have the same order and must be added. For $b > 1/2$, the largest eigenvalue determines the leading result $B_c \simeq \sqrt{\nu d} = \sqrt{N}$, while the dependence on b appears in subleading corrections.

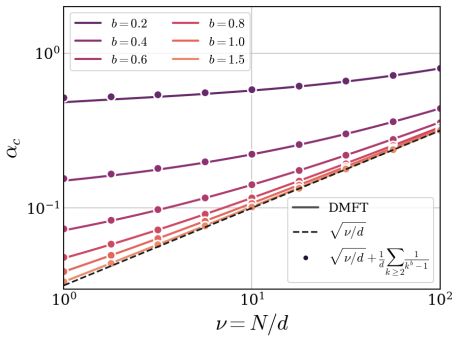


Figure 6: Critical batch ratio α_c versus width ratio ν. Solid curves show DMFT predictions, the dashed line shows the leading top-mode estimate $\sqrt{\nu/d}$, and markers show the large-dimension approximation $\sqrt{\nu/d} + \frac{1}{d}\sum_{k=2}^{d} \frac{1}{k^b-1}$. The remaining spectrum raises the saturation threshold substantially for smaller b.

F.3 Scaling with Width at Phase 1

For the width-limited (Phase 1) branch, define

$$\mu_1 = \frac{1}{d}\sum_k \lambda_k \tag{F.20}$$

$$\Phi(\eta_{\text{eff}}^{\infty} r_-) = \frac{1}{d}\sum_k \frac{\lambda_k \eta_{\text{eff}}^{\infty} r_-}{2 - \lambda_k \eta_{\text{eff}}^{\infty} r_-} \tag{F.21}$$

$$P_\nu = \frac{1}{d}\sum_k \frac{\lambda_k \beta_k^2}{1 + \lambda_k \eta_{\text{eff}}^{\infty} \rho}. \tag{F.22}$$

The three conditions to be satisfied can be then read as

$$\nu = \frac{1}{d}\sum_k \frac{\lambda_k \eta_{\text{eff}}^{\infty} \rho}{1 + \lambda_k \eta_{\text{eff}}^{\infty} \rho} \tag{F.23}$$

$$r_- = \left(1 - \frac{1}{\nu}\frac{\Phi(\eta_{\text{eff}}^{\infty} r_-)}{r_-}\right)^{-1} \tag{F.24}$$

$$\frac{\eta^2}{\nu \eta_{\text{eff}}^{\infty}} = \frac{2(r_- - 1)}{\alpha\left(1 - \frac{\nu}{\alpha}(r_- - 1)\right)} P_\nu. \tag{F.25}$$

From the response function equation, one can then solve for r_- and get

$$r_- = 1 + \frac{\Phi(\eta_{\text{eff}}^{\infty} r_-)}{\nu} \tag{F.26}$$

from which the third equation simplify as

$$\alpha = \Phi(\eta_{\text{eff}}^{\infty} r_-)\left(1 + \frac{2\eta_{\text{eff}}^{\infty} P_\nu}{\eta^2}\right). \tag{F.27}$$

We first specify the physical branch on which the expansion is performed. Writing $y = \eta_{\text{eff}}^{\infty} r_-$, the EoS condition derived above is

$$G(y_c) = \nu, \qquad G(y) = \frac{1}{d}\sum_k \left(\frac{\lambda_k y}{2 - \lambda_k y}\right)^2. \tag{F.28}$$

On $0 \leq y < 2/\lambda_{\max}$, $G(y)$ increases continuously from 0 to ∞. Hence its unique root in this interval satisfies $0 < y_c < 2/\lambda_{\max}$. The stable Phase-1 branch lies below this EoS endpoint, so $0 < y < y_c$ and therefore

$$0 < \lambda_k y < 2 \qquad \forall k. \tag{F.29}$$

Each denominator in $\Phi(y)$ is thus positive and no larger than 2, giving

$$\Phi(y) = \frac{1}{d}\sum_k \frac{\lambda_k y}{2-\lambda_k y} \geq \frac{1}{d}\sum_k \frac{\lambda_k y}{2} = \frac{\mu_1 y}{2}. \tag{F.30}$$

Moreover, $\Phi(y) > 0$ implies $r_- = 1 + \Phi(y)/\nu \geq 1$. Substituting $y = \eta_{\text{eff}}^{\infty} r_-$ we obtain

$$\Phi(\eta_{\text{eff}}^{\infty} r_-) \geq \frac{\mu_1 \eta_{\text{eff}}^{\infty} r_-}{2} \geq \frac{\mu_1 \eta_{\text{eff}}^{\infty}}{2}. \tag{F.31}$$

This bound follows from the stability branch selection. Using this bound in the stationary closure gives

$$\alpha \geq \frac{\mu_1 P_\nu}{\eta^2}(\eta_{\text{eff}}^{\infty})^2. \tag{F.32}$$

Rearranging

$$0 < \eta_{\text{eff}}^{\infty} \leq \eta\sqrt{\frac{\alpha}{\mu_1 P_\nu}}. \tag{F.33}$$

Since α, $\mu_1 > 0$ and $P_\nu > 0$ remain fixed, $\eta \to 0$ therefore implies $\eta_{\text{eff}}^{\infty} \to 0$. The stationary closure gives $0 \leq \Phi(y) \leq \alpha$, hence $1 \leq r_- \leq 1 + \alpha/\nu$. Therefore $y = \eta_{\text{eff}}^{\infty} r_- \to 0$, implying $\Phi(y) \to 0$ and $r_- \to 1$. Expanding Φ near zero then gives

$$\frac{\Phi(y)}{\eta_{\text{eff}}^{\infty}} = \frac{\mu_1}{2} r_- + o(1) \longrightarrow \frac{\mu_1}{2}. \tag{F.34}$$

Rearranging the exact stationary closure

$$\frac{(\eta_{\text{eff}}^{\infty})^2}{\eta^2} = \frac{\alpha - \Phi(y)}{2P_\nu\left[\Phi(y)/\eta_{\text{eff}}^{\infty}\right]} \longrightarrow \frac{\alpha}{\mu_1 P_\nu}, \tag{F.35}$$

and taking the positive square root yields

$$\eta_{\text{eff}}^{\infty}(\nu) \sim \eta\sqrt{\frac{\alpha}{\mu_1 P_\nu}}. \tag{F.36}$$

Consequently, the leading width efficiency is

$$Q_\nu(\nu) = \frac{\eta_{\text{eff}}^{\infty}(\nu)}{\nu} \sim \frac{\eta}{\nu}\sqrt{\frac{\alpha}{\mu_1 P_\nu}}. \tag{F.37}$$

Since at fixed input spectrum, μ_1 is independent of ν, the remaining width dependence is determined by P_ν.

For the power-law spectrum

$$\mu_1 = \begin{cases} \frac{d^{-b}}{1-b}, & 0 \leq b < 1 \\ \frac{\log d}{d}, & b = 1 \\ \frac{\zeta(b)}{d}, & b > 1. \end{cases} \tag{F.38}$$

This is the same μ_1 that appears in the small-α EoSS branch calculation. To extract the width dependence, we now specialize to $\lambda_k = k^{-b}$ and $\beta_k = k^{-a}$. Writing $m_\nu = \eta_{\text{eff}}^{\infty}\rho$, the width equation and the residual target power become

$$\nu = \frac{1}{d}\sum_{k=1}^{d} \frac{m_\nu}{k^b + m_\nu} \tag{F.39}$$

$$P_\nu = \frac{1}{d}\sum_{k=1}^{d} \frac{k^{-2a}}{k^b + m_\nu}. \tag{F.40}$$

The first equation determines m_ν independently of the teacher exponent a. Once m_ν is known, the second equation determines the dependence on both a and b.

For $\lambda_k = k^{-b}$, the denominator $k^b + m_\nu$ changes behavior around $k \sim m_\nu^{1/b}$: below this index it is dominated by m_ν, and above it by k^b. To obtain asymptotic width scalings, we focus on an intermediate-width regime where this crossover is well separated from both spectral endpoints, i.e.

$$1 \ll m_\nu^{1/b} \ll d. \tag{F.41}$$

Both ranges then contain many modes, allowing their contributions to the spectral sums to be estimated separately. This condition specifies a window of ν given d, it is not required by the full stationary equations.

For modes with $k \ll m_\nu^{1/b}$, the denominator is dominated by m_ν, whereas for $k \gg m_\nu^{1/b}$ it is dominated by k^b. Splitting the sums into these two ranges gives the scaling estimates

$$d\nu \asymp m_\nu^{1/b} + m_\nu \sum_{k \gtrsim m_\nu^{1/b}}^{d} k^{-b}, \tag{F.42}$$

$$dP_\nu \asymp \frac{1}{m_\nu} \sum_{k \lesssim m_\nu^{1/b}} k^{-2a} + \sum_{k \gtrsim m_\nu^{1/b}}^{d} k^{-(2a+b)}. \tag{F.43}$$

Here $\asymp$ denotes equality of leading scaling, without specifying the multiplicative constant.

Now, to determine m_ν, we approximate the width equation by

$$d\nu \asymp m_\nu^{1/b} + m_\nu \int_{m_\nu^{1/b}}^{d} k^{-b}\, dk. \tag{F.44}$$

For $0 < b < 1$, the integral is dominated by its upper endpoint and gives $m_\nu d^{1-b}$, which dominates the first term.

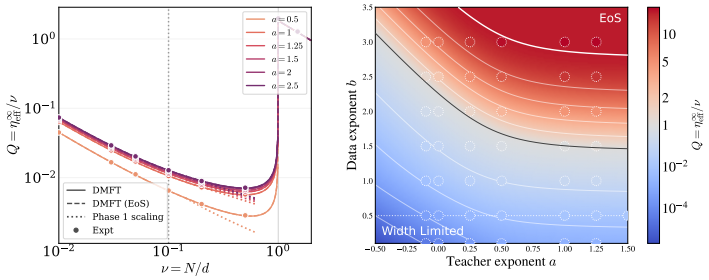


Figure 7: **Width efficiency and spectral dependence. Left:** $Q_\nu = \eta_{\text{eff}}^{\infty}/\nu$ versus width for $b = 0.5$ and different teacher exponents a. DMFT predictions (solid; dashed in EoS) are compared with simulations (symbols) and anchored Phase-1 scalings (dotted), which predict $Q_\nu \propto \nu^{-1/2}$ for $a > 1/2$, with a logarithmic correction at $a = 1/2$. **Right:** Q_ν across (a, b) at $\nu = 0.1$, with theory and simulation markers sharing the color scale. The thick white curve marks the EoS boundary; the dark contour marks $Q_\nu = 1$. For these parameters, no interior efficiency maximum is observed for $\nu < 1$: $Q_\nu(\nu)$ decreases to a minimum before rising toward a peak near the interpolation threshold $\nu = 1$.

For $b = 1$, it gives $m_\nu \log(d/m_\nu)$. For $b > 1$, the lower endpoint dominates, giving a contribution of order $m_\nu^{1/b}$, comparable to the first term. Therefore,

$$\nu \asymp \begin{cases} m_\nu d^{-b}, & 0 < b < 1 \\ \dfrac{m_\nu}{d} \log\left(\dfrac{d}{m_\nu}\right), & b = 1 \\ \dfrac{m_\nu^{1/b}}{d}, & b > 1. \end{cases} \tag{F.45}$$

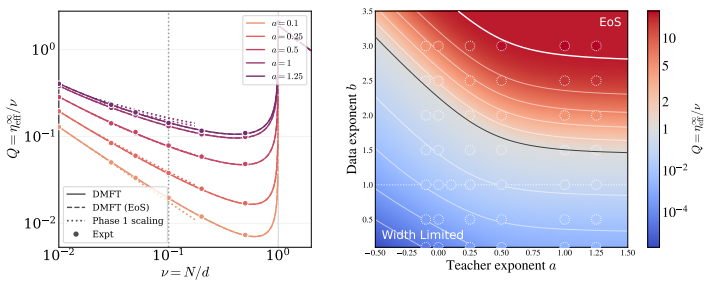


Figure 8: **Width efficiency at $b = 1$. Left:** $Q_\nu = \eta_{\rm eff}^{\infty}/\nu$ versus width for different teacher exponents a. DMFT predictions (solid; dashed in EoS) agree with simulations (symbols). Dotted curves show anchored Phase-1 scalings, including the logarithmic corrections at $b = 1$; for $a > 1/2$, $Q_\nu \propto [\nu \log(1/\nu)]^{-1/2}$. Efficiency initially decreases, reaches a minimum, and rises toward a peak near $\nu = 1$. **Right:** Q_ν across (a, b) at $\nu = 0.1$, with theory and simulation markers sharing the color scale. The thick white curve marks the EoS boundary, and the dark contour marks $Q_\nu = 1$. The vertical and horizontal dotted reference lines indicate $\nu = 0.1$ and $b = 1$, respectively.

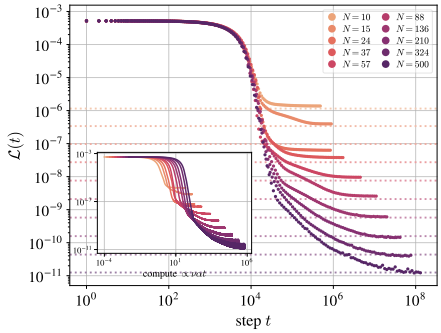


Figure 9: Complimentary figure to 4 right panel. Here we train a larger range of widths to convergence, showing that the width speedup can sometimes be subsumed by the loss floor, making larger widths for compute and speed efficient.

Inverting these relations yields

$$m_\nu \asymp \begin{cases} \nu d^b, & 0 < b < 1 \\ \dfrac{\nu d}{\log(1/\nu)}, & b = 1 \\ (\nu d)^b, & b > 1. \end{cases} \tag{F.46}$$

We next evaluate the two contributions to P_ν. The contribution from modes below the crossover satisfies

$$\frac{1}{m_\nu} \sum_{k \lesssim m_\nu^{1/b}} k^{-2a} \asymp \begin{cases} m_\nu^{(1-2a-b)/b}, & a < 1/2 \\ m_\nu^{-1} \log m_\nu, & a = 1/2 \\ m_\nu^{-1}, & a > 1/2. \end{cases} \tag{F.47}$$

Indeed, the sum grows as its upper cutoff to the power $1 - 2a$ for $a < 1/2$, grows logarithmically for $a = 1/2$, and converges for $a > 1/2$.

We assume $2a+b > 1$, ensuring that $\sum_k \lambda_k \beta_k^2$ converges as $d \to \infty$. Consequently, the contribution from modes above the crossover is

$$\sum_{k \gtrsim m_\nu^{1/b}}^{d} k^{-(2a+b)} \asymp m_\nu^{-(2a+b-1)/b}. \tag{F.48}$$

Comparing the two contributions, we obtain

$$P_\nu \asymp \frac{1}{d}\begin{cases} m_\nu^{-(2a+b-1)/b}, & \frac{1-b}{2} < a < \frac{1}{2} \\ m_\nu^{-1}\log m_\nu, & a = \frac{1}{2} \\ m_\nu^{-1}, & a > \frac{1}{2}. \end{cases} \tag{F.49}$$

Substituting the residual-power scaling into $Q_\nu(\nu)$ gives

$$Q_\nu(\nu) \asymp \frac{\eta}{\nu}\sqrt{\frac{\alpha d}{\mu_1}}\begin{cases} m_\nu^{(2a+b-1)/(2b)}, & (1-b)/2 < a < 1/2 \\ \sqrt{m_\nu/\log m_\nu}, & a = 1/2 \\ \sqrt{m_\nu}, & a > 1/2. \end{cases} \tag{F.50}$$

For $b > 1$, substituting the corresponding m_ν form Eq. F.45 yields, at fixed d, α, η

$$Q_\nu(\nu) \asymp \begin{cases} \nu^{(2a+b-3)/2}, & (1-b)/2 < a < 1/2 \\ \nu^{(b-2)/2}/\sqrt{\log(\nu d)}, & a = 1/2 \\ \nu^{(b-2)/2}, & a > 1/2 \end{cases} \tag{F.51}$$

up to factors independent of ν. Thus efficiency grows with width when $2a + b > 3$ for $a < 1/2$, or $b > 2$ for $a > 1/2$, with a logarithmic correction at $a = 1/2$. This scaling alone does not determine the optimal width. The gain is eventually reversed by stability: on the EoS branch, $Q'_\nu(\nu) < 0$. When this branch is reached below $\nu = 1$, the increasing and decreasing regimes imply an efficiency maximum at an underparameterized width. Its location is determined by the full stationary equations.

For $0 < b < 1$, substituting m_ν gives

$$Q_\nu(\nu) \asymp \begin{cases} \nu^{(2a-b-1)/(2b)}, & (1-b)/2 < a < 1/2 \\ \dfrac{\nu^{-1/2}}{\sqrt{\log\big((1-b)\nu d^b\big)}}, & a = 1/2 \\ \nu^{-1/2}, & a > 1/2. \end{cases} \tag{F.52}$$

Thus all three regimes give decreasing $Q_\nu(\nu)$. The marginal spectrum $b = 1$ requires a separate treatment. The width constraint becomes

$$\nu d \sim m_\nu \log\left(\frac{d}{m_\nu}\right). \tag{F.53}$$

Writing $\ell_\nu = \log(1/\nu)$, its leading logarithmic solution is

$$m_\nu \sim \frac{\nu d}{\ell_\nu}, \qquad \nu \ll 1. \tag{F.54}$$

Substitution then gives

$$Q_\nu(\nu) \asymp \begin{cases} \dfrac{\nu^{a-1}}{\ell_\nu^a}, & 0 < a < 1/2 \\ \dfrac{1}{\sqrt{\nu\ell_\nu \log\big(\nu d/\ell_\nu\big)}}, & a = 1/2 \\ \dfrac{1}{\sqrt{\nu\ell_\nu}}, & a > 1/2. \end{cases} \tag{F.55}$$

Here too, efficiency decreases with width within the asymptotic regime and the negative powers of ν dominate the logarithmic corrections.

F.4 Absence of approximately equal-compute speedup.

Approximately equal-compute speedup requires the stationary effective learning rate to grow approximately linearly with the resource being increased. We examine the complementary resource directions to those permitting this behavior.

Batch size in Phase 1. At fixed width and spectrum, the width constraint fixes $\eta_{\text{eff}}^{\infty}\rho$, so P_ν is independent of batch size. The small-step asymptotics in equation F.36 therefore gives

$$\eta_{\text{eff}}^{\infty} \sim \eta\sqrt{\frac{\alpha}{\mu_1 P_\nu}}, \qquad \frac{\partial \log \eta_{\text{eff}}^{\infty}}{\partial \log \alpha} \longrightarrow \frac{1}{2}.$$

Thus the gain is square-root rather than approximately linear. For example, doubling the batch reduces the updates needed to accumulate the same effective time by only a factor $\sqrt{2}$, increasing total compute by a factor $\sqrt{2}$. Increasing batch size therefore does not provide approximately equal-compute speedup in this regime.

Width in Phase 2. At fixed batch size and spectrum, the EoSS condition $\Phi(y) = \alpha$ fixes y independently of width. The response equation gives

$$\eta_{\text{eff}}^{\infty} = \frac{\nu y}{\nu + \alpha}, \qquad \frac{\partial \log \eta_{\text{eff}}^{\infty}}{\partial \log \nu} = \frac{\alpha}{\nu + \alpha}.$$

The physical Phase-2 branch satisfies $\nu > 1$ and $G(y) < \nu$. Since $\Phi(y)^2 \leq G(y)$, we obtain $\alpha < \sqrt{\nu} < \nu$, and hence

$$\frac{\partial \log \eta_{\text{eff}}^{\infty}}{\partial \log \nu} < \frac{1}{2}.$$

Width gains are therefore slower than square-root throughout Phase 2. In particular, doubling width while remaining in this phase increases the compute required for the same effective time by more than a factor $\sqrt{2}$.

Batch size and width in Phase 3. At fixed width, the deterministic stability threshold $\eta_{\text{eff}}^{\infty} = 2/\lambda_{\max}(\nu)$ is independent of batch size. Additional batching therefore provides no stationary speedup.

For width, the response equation and the EoS condition give

$$\eta_{\text{eff}}^{\infty} = \frac{\nu y}{\nu + \Phi(y)}, \qquad G(y) = \nu.$$

Using $y\Phi'(y) = \Phi(y) + G(y)$ and differentiating along the EoS branch yields

$$\frac{\partial \log \eta_{\text{eff}}^{\infty}}{\partial \log \nu} = \frac{\Phi(y)}{\nu + \Phi(y)}.$$

For $\nu \geq 1$, the inequality $\Phi(y)^2 \leq G(y) = \nu$ immediately bounds this slope by $1/(1+\sqrt{\nu}) \leq 1/2$. For $\nu < 1$, Phase 3 is reached from the width-limited branch. At the transition to EoS, the Phase-1 closure gives

$$\alpha = \Phi(y)\left(1 + \frac{2\eta_{\text{eff}}^{\infty} P_\nu}{\eta^2}\right) \geq \Phi(y).$$

Hence throughout Phase 3, $\Phi(y) \leq \alpha$. Assuming $N > B$, equivalently $\nu > \alpha$, we therefore obtain

$$\frac{\partial \log \eta_{\text{eff}}^{\infty}}{\partial \log \nu} = \frac{\Phi(y)}{\nu + \Phi(y)} \leq \frac{\alpha}{\nu + \alpha} < \frac{1}{2}.$$

Thus, also for $\nu < 1$, increasing width yields at most square-root gains in the stationary effective learning rate when $N > B$. We do not treat here the case of an under-parameterized network where the batch size α is larger than the total parameter count ν, and we do not exclude speedups in this case; instead we note that this case is far removed from modern practice.

G EXPERIMENTAL DETAILS

Task and data. We train on CIFAR-5M with the binary animate/inanimate target, $y = +1$ for classes 2–7 and $y = -1$ otherwise, so that $\mathbb{E}[y^2] = 1$. We rescaled each coordinate's color scale to $[-1, 1]$ and used online training with no repetition: each of the 5×10^6 training images is used exactly once, so every step draws a fresh batch such that the total number of steps satisfies $T\,BT = 5 \times 10^6$ in every run.

Architecture. We use the residual CNN of (Bordelon et al., 2024a): a 3×3 stem convolution with w channels, three residual groups of one block each with the third at stride 2, average pooling, and a scalar readout. A block is $\mathrm{Relu} \to \mathrm{Conv}_{3\times3} \to \mathrm{Relu} \to \mathrm{Conv}_{3\times3}$ summed with its shortcut. The width w is constant across groups and there is no batch normalization. The number of trainable parameters is

$$N(k) = 63w^2 + 28w. \tag{G.1}$$

Linearized training. We train the linearization of the network around its initialization, $f(\delta\boldsymbol{\theta}) = \boldsymbol{J}(\boldsymbol{\theta}_0)\,\delta\boldsymbol{\theta}$, on the half mean-squared error, with the normalized update $\delta\boldsymbol{\theta} \leftarrow \delta\boldsymbol{\theta} - \eta\,\boldsymbol{g}/\|\boldsymbol{g}\|_2$. Accordingly we use $\eta^t_{\mathrm{eff}} = \eta/\|\boldsymbol{g}^t\|_2$. We note that this follows our dimension-free update described in App. H and differs from our proportional regime where we used $\sqrt{d}\eta$ for the update.

Spectral exponents. We take λ_k and β_k from the analytic infinite-width NTK, evaluated on $M = 16{,}384$ images and diagonalized in numerically; we find $\mathrm{tr}\,\boldsymbol{\Lambda} = 0.690$ and by the defintion of the labels $\sum_k \lambda_k\beta_k^2 = \mathbb{E}[y^2] = 1$. We fit b from the decay of λ_k and $2a + b - 1$ from the decay of the residual target power

$$1 - C(k) := \frac{\sum_{k'=k}^{M} \lambda_{k'}\beta_{k'}^2}{\sum_{k'=1}^{M} \lambda_{k'}\beta_{k'}^2},$$

both over the window $k \in [20, 512]$ where we find a good power fit with parameters

$$b = 1.81, \qquad 2a + b - 1 = 0.17, \qquad a = -0.32, \qquad \lambda_{\max} = 0.44, \tag{G.2}$$

close to the values reported in (Bordelon et al., 2024a) for the same architecture. Hence $r_{\mathrm{SGD}} = 0.095$, with predicted transient exponent $2r_{\mathrm{SGD}}/(1 - r_{\mathrm{SGD}}) = 0.21$.

The DMFT takes the measured spectrum for $k \le 1024$ and the fitted power law beyond it, continued to $d = 10^{10}$. We found that taking larger d made a negligible difference.

Sweep. We vary width and batch size over

w	4	16	64	128	256
N	1,120	16,576	259,840	1,035,776	4,135,936
B	2	10	100	1,000	10,000
T	$2.5 \cdot 10^6$	$5 \cdot 10^5$	$5 \cdot 10^4$	$5 \cdot 10^3$	500
η	0.01	0.1	0.2	0.5	

with two seeds per cell, for 241 runs in total.

Evaluation. We report the mean-squared error on a fixed held-out set of 4,096 images, evaluated at 150 logarithmically spaced steps, and record $\|\boldsymbol{g}^t\|_2$ and η^t_{eff} at every step. The DMFT is integrated to up to $T = 3800$ steps due to memory constraints.

H THE DIMENSION-FREE LIMIT

The main text focuses on the proportional limit $d, N, B \to \infty$ at fixed $\nu = N/d$ and $\alpha = B/d$, here we extend the results to the case of taking the limit $d \to \infty$ first, and only then taking $N, B \gg 1$ – a dimension free setting. This limit corresponds to the kernel settings and our NTK and linearized networks experiments where the representations live in an infinite-dimensional kernel space. We show that at fixed N, B the theory has a regular limit there in which the dimension drops out of every equation, leaving N, B, η and the spectrum. We then translate the steady state of App. D

into this setting, where phase 1 (width-limited) and phase 3 (EoS) survive but phase 2 (EoSS) does not. The main change is that we no longer average over k modes, but take the trace (sun), so with $\lambda_k = k^{-b}$ and $\beta_k = k^{-a}$,

$$\operatorname{Tr}\Lambda = \zeta(b), \qquad \mathcal{L}^0 = \sum_k \lambda_k \beta_k^2 = \zeta(2a+b), \tag{H.1}$$

and the limit exists for a trace-class kernel, $b > 1$, alongside the usual $2a + b > 1$.

To define this limit, we also need to modify the update and remove the $\sqrt{d}$ scaling,

$$\boldsymbol{w}^{t+1} = \boldsymbol{w}^t - \eta \left(\boldsymbol{g}^\top \boldsymbol{g}\right)^{-1/2} \boldsymbol{g}, \qquad \eta_{\text{eff}}^t = \frac{\eta}{\|\boldsymbol{g}^t\|_2}. \tag{H.2}$$

which should be contrasted with equation 1.

H.1 The Dimension-Free DMFT Equations

In App. B.3 the dimension enters only through $\langle\cdot\rangle_k = d^{-1}\sum_k$ and the ratios ν, α, and these appear only in the combinations $\nu^{-1}\langle\cdot\rangle_k$ and $\alpha^{-1}\langle\cdot\rangle_k$. Multiplying every relation by d therefore removes the dimension and trades the intensive quantities for convergent ones (under this appendix's additional assumption $b > 1$),

$$\langle\cdot\rangle_k \mapsto \sum_k, \quad \nu \mapsto N, \quad \alpha \mapsto B, \quad C_1 \mapsto C_0/B, \quad \nu C_3 \mapsto N C_3 \quad d C_2 \mapsto \Sigma_2. \tag{H.3}$$

Keeping $H_k = [I + \lambda_k \Theta R_3]^{-1}$ and $\Theta^{ts} = \eta\, 1_{s<t} \left(N C_3^{ss}\right)^{-1/2} = \eta_{\text{eff}}^s 1_{s<t}$, the closed set of App. B.3 becomes

$$\begin{aligned} R_3 &= I - \frac{1}{N}\sum_k \lambda_k \left(\Theta\left[I + \lambda_k R_3 \Theta\right]^{-1}\right) R_3 \\ {}_0 &= \sum_k \lambda_k \left[\beta_k^2 H_k j \left(H_k j\right)^\top + H_k \Theta \left(C_3 + \lambda_k R_3 c_1 R_3^\top\right)\left(H_k \Theta\right)^\top\right] \\ \Sigma_2 &= \sum_k \left[\lambda_k^2 \beta_k^2 H_k j \left(H_k j\right)^\top + \lambda_k H_k \left(c_1 + \lambda_k \Theta C_3 \Theta^\top\right) H_k^\top\right], \end{aligned} \tag{H.4}$$

with j the all-ones time vector. These are the $\nu, \alpha \to 0$ limit of the main-text closure at fixed N and B, and are exact for $1 \ll N, B \ll d$. The same substitution in equation 3 gives the stationary response

$$\hat{R}_3(z) = \left(1 + \frac{1}{N}\sum_k \frac{\lambda_k \eta_{\text{eff}}^\infty}{z - 1 + \lambda_k \eta_{\text{eff}}^\infty \hat{R}_3(z)}\right)^{-1}. \tag{H.5}$$

H.2 Steady State and the Absence of the EoSS Phase

Multiplying Eq. equation 4 by d, the three conditions read

$$\underbrace{\sum_k \frac{\lambda_k \eta_{\text{eff}}^\infty \rho}{1 + \lambda_k \eta_{\text{eff}}^\infty \rho} = N}_{\text{Phase 1: width}}, \qquad \underbrace{\operatorname{Tr}\left[\frac{\Lambda \eta_{\text{eff}}^\infty \hat{R}_3(-1)}{2 - \Lambda \eta_{\text{eff}}^\infty \hat{R}_3(-1)}\right] = B}_{\text{Phase 2: EoSS}}, \qquad \underbrace{\operatorname{Tr}\left[\left(\frac{\Lambda \eta_{\text{eff}}^\infty \hat{R}_3(-1)}{2 - \Lambda \eta_{\text{eff}}^\infty \hat{R}_3(-1)}\right)^2\right] = N}_{\text{Phase 3: EoS}}, \tag{H.6}$$

Taking the $z = -1$ response equation to the same limit gives

$$\operatorname{Tr}\left[\frac{\Lambda \eta_{\text{eff}}^\infty \hat{R}_3(-1)}{2 - \Lambda \eta_{\text{eff}}^\infty \hat{R}_3(-1)}\right] = N\left(\hat{R}_3(-1) - 1\right), \tag{H.7}$$

so phase 2 reads $\hat{R}_3(-1) = 1 + B/N$, while multiplying equation D.13 by d and using $\eta_{\text{eff}}^\infty = \eta/\sqrt{N C_3^\infty}$ puts phase 1 in the form

$$\eta^2 = \frac{2\eta_{\text{eff}}^\infty N\left(\hat{R}_3(-1) - 1\right)}{B - N\left(\hat{R}_3(-1) - 1\right)} \sum_k \frac{\lambda_k \beta_k^2}{1 + \lambda_k \eta_{\text{eff}}^\infty \rho}, \tag{H.8}$$

which always has a solution: its mode sum now runs over infinitely many modes and is no longer bounded by d, which is the origin of the $\nu < 1$ condition in the main text. The right-hand side of equation H.8 grows continuously from zero and diverges precisely at $\hat{R}_3(-1) = 1 + B/N$, so it matches any η at a smaller effective learning rate than phase 2 requires. The width-limited state is therefore reached first for every N, B and η, and phase 2 is excluded. The only transition left is width-limited to EoS, at a critical batch size obtained by evaluating equation H.8 on the EoS branch,

$$B_{1\to 3} = N\left(\hat{R}_3(-1) - 1\right)\left(1 + \frac{2\eta_{\text{eff}}^{\infty}}{\eta^2}\sum_k \frac{\lambda_k \beta_k^2}{1 + \lambda_k \eta_{\text{eff}}^{\infty}\rho}\right). \tag{H.9}$$