SGD-INDUCED PARAMETER DRIFT IN THE SADDLE-TO-SADDLE DYNAMICS OF DEEP LINEAR NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

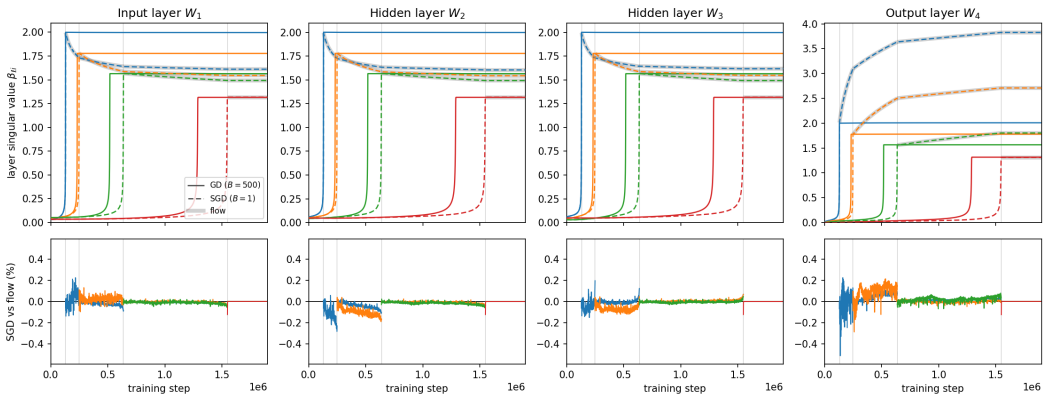
Stochastic gradient descent (SGD) and gradient descent (GD) can remain close in function space while selecting distinct parametrizations. We study this separation in deep linear teacher–student regression across initialization scales, in both online and offline training. Empirically, SGD and GD learn similar end-to-end maps throughout training, and SGD does not achieve lower population loss. However, for small initialization, where the network undergoes saddle-to-saddle dynamics, we find that SGD systematically redistributes scale across layers. In the absence of label noise, once a teacher mode is learned, its output-layer singular value grows while the corresponding singular values in earlier layers shrink, leaving the mode's end-to-end contribution approximately constant. This corresponds to motion along the rescaling symmetries of the network. Layer imbalances are not conserved by SGD; near rank-deficient saddles, their mean drift is a gradient flow that reduces the gradient noise. For Gaussian inputs with teacher-aligned covariances and decoupled learned modes, we show that this flow is driven by the residual error of the unlearned modes, and prove that SGD follows it while near a saddle. The flow predicts that the output-layer singular value grows as $\tau^{1/4}$, independent of depth, while the earlier-layer singular values decay as $\tau^{-1/[4(L-1)]}$, where L is the depth and $\tau = \eta^2 t$ is a learning-rate-scaled training time. With label noise, the layer singular values flow towards a unique residual-dependent fixed point until the next mode is learned. This fixed point shifts as additional modes are learned and, at full fit, recovers the noise equilibrium of Ziyin et al. (2024). These results characterize an implicit bias of SGD toward particular parametrizations that can be substantial even when the learned function remains close to that of GD.

1 INTRODUCTION

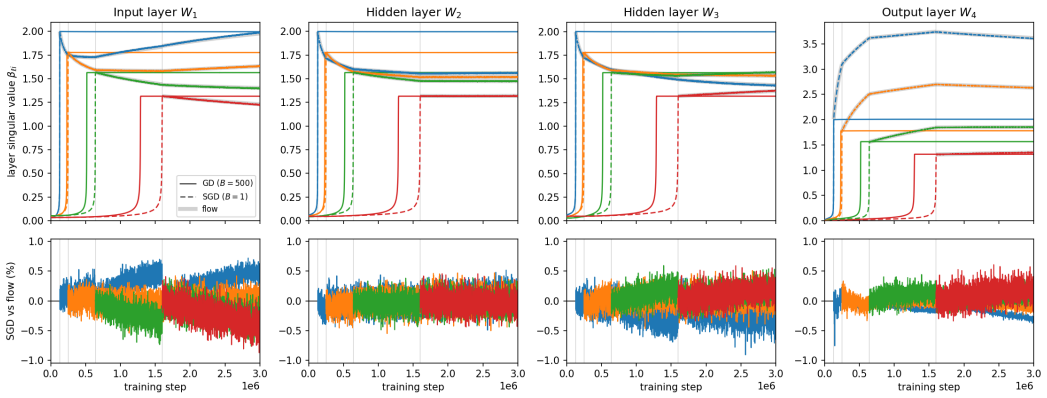
Characterizing the implicit biases of gradient noise is an important problem in the theory of deep learning (Simon et al., 2026): does the noise in stochastic gradient descent (SGD) change what a network learns? One line of work says it does. In diagonal linear networks, stochastic gradient flow converges to sparser interpolators than gradient flow (Pesme et al., 2021), and gradient noise attracts the dynamics toward simpler subnetworks (Chen et al., 2023). These results mostly concern offline, multi-epoch training near convergence. In the online regime of large-scale pre-training, which stops far from convergence, the evidence points the other way: small batches do not lower the population loss relative to large ones, and SGD stays close to the large-batch trajectory in both loss and function space, which Vyas et al. (2024) call the golden path. We call the claim that SGD stays on this path the golden path hypothesis. It has theoretical support for high-dimensional convex quadratics (Paquette et al., 2022) and for a Langevin model of SGD in deep linear networks (Corlouer et al., 2026).

The golden path hypothesis can hold in function space and still fail in parameter space. The end-to-end map of a deep linear network (DLN) is invariant under a general linear group acting on each hidden layer, so each function has a continuum of parametrizations. Gradient flow conserves the layer imbalances $W_\ell W_\ell^\top - W_{\ell+1}^\top W_{\ell+1}$ (Du et al., 2018; Arora et al., 2018; Marcotte et al., 2023), but finite step sizes and gradient noise break these conservation laws (Kunin et al., 2021). The parametrization matters even for a fixed function: it sets the neural tangent kernel and the lazy or rich character of later learning (Dominé et al., 2025; Kunin et al., 2024), the curvature of the loss (Dinh et al., 2017), and the layer-level quantities used by interpretability methods.

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.



(a) No label noise



(b) Isotropic label noise with $\sigma = 0.5$

Figure 1: **SGD moves the scale of learned modes into the output layer, following the flow** (5)**.** Layer singular values $\beta_{\ell i}$ of the four teacher modes (colors) of a depth-4, width-16 DLN trained online on a rank-4 teacher with singular values $(16, 10, 6, 3)$ and inputs $x \sim \mathcal{N}(0, I_5)$, (a) without and (b) with isotropic label noise ($\sigma = 0.5$). GD (batch size 500, solid) and SGD (batch size 1, dashed) use $\eta = 10^{-4}$ and start from the same small initialization (Appendix A). Grey lines: the flow, restarted from the SGD weights at each plateau (vertical lines). Bottom rows: relative difference between SGD and the flow. It stays within about 0.5% in (a), while the output-layer singular value of the first mode nearly doubles, and below 1% in (b), where it is mostly SGD fluctuation driven by the label noise.

Ziyin et al. (2024) show that gradient noise drives SGD along these symmetries toward a noise equilibrium. For DLNs, they derive a general condition for this equilibrium but solve it explicitly only at the global minimum. Without label noise, every per-sample gradient vanishes there, so the drift stops and the final parametrization is set by the drift during training. How gradient noise rescales the layers during training has not been studied before.

We study this question in DLNs trained as students of a teacher matrix. DLNs are a well-studied controlled setting for training dynamics (Nam et al., 2025). The initialization scale moves them from a lazy regime, in which the neural tangent kernel stays fixed (Jacot et al., 2018), through an intermediate mean-field regime, to a saddle-to-saddle regime, in which the modes of the teacher are learned one at a time in decreasing order of their singular values (Saxe et al., 2014; Jacot et al., 2021); saddle-to-saddle dynamics also occur in shallow ReLU networks and attention-only transformers (Zhang et al., 2026). We compare SGD with GD across initialization scales, in online and offline training. In offline training, GD is full-batch gradient descent on the training set. In online training, where every step uses fresh samples, we compare small-batch SGD with large-batch SGD; for brevity, we refer to large-batch SGD as GD throughout. We also consider singular input covari-

ances, for which the loss has many global minimizers in function space, so that SGD and GD could in principle converge to different functions.

Our contributions are as follows.

- **The golden path empirically holds in function space** (Section 3). Across initialization scales, in online and offline training, and for invertible as well as singular input covariances, SGD and GD learn nearly the same end-to-end map. In none of our experiments does minibatch noise lower the population loss relative to GD.
- **The golden path fails in parameter space** (Figure 1 for the detailed saddle-to-saddle dynamics, and Figure 5 for the comparison across regimes). In the saddle-to-saddle regime, once a teacher mode is learned, SGD increases its singular value in the output layer and decreases it in every earlier layer, keeping the end-to-end singular value of the mode fixed. The layer-scale drift is much weaker in the mean-field regime and essentially absent in the lazy regime.
- **At the saddles, the drift is a gradient flow driven by the unlearned modes** (Section 4). For Gaussian inputs, with input and label-noise covariances aligned with the teacher, we compute the drift of the layer imbalances in closed form. When the learned modes are decoupled, their layer singular values follow a gradient flow of the second moment of the stochastic gradient along the rescalings, and we prove that SGD follows this flow for up to $O(1/\eta^2)$ steps while it stays near a saddle. Without label noise, the error of the unlearned modes reaches only the output layer, so the flow has no fixed point: each output-layer singular value grows asymptotically as a fourth root of the slow time $\tau = \eta^2 t$, at a rate independent of depth, while the corresponding singular values in earlier layers shrink. These predictions match the measured layer singular values to within 2% (Figure 1a).
- **Label noise gives the drift a moving fixed point** (Section 4). With label noise, the flow has a unique fixed point, shifted toward the output layer by the remaining error, which SGD approaches during each plateau. As modes are learned, the fixed point moves and, at full fit, recovers the noise equilibrium of Ziyin et al. (2024) (Figure 1b).

2 Related work

SGD versus GD. Whether gradient noise helps generalization has long been debated. Small batches can generalize better in offline training (Smith et al., 2020), and full-batch training with an explicit regularizer that penalizes large gradients matches the performance of SGD (Geiping et al., 2022). In online training of convolutional networks and transformers, switching from high to low noise leads to the same solution as training with low noise throughout (Vyas et al., 2024). Paquette et al. (2022) prove that the golden path hypothesis holds for high-dimensional convex quadratic losses, using a continuous model of SGD (Paquette et al., 2024), and in the saddle-to-saddle regime of DLNs, an anisotropic Langevin model of SGD learns the same features, in the same order, as gradient flow (Corlouer et al., 2026). These works compare SGD and GD through the loss and the learned function; we also compare their parametrizations.

Symmetries, conservation laws, and noise equilibria. Continuous symmetries of the loss induce conservation laws under gradient flow (Kunin et al., 2021; Marcotte et al., 2023). In linear networks, these laws conserve the differences between the Gram matrices of adjacent layers, and in homogeneous networks the corresponding differences of layer norms, so layers that start balanced stay balanced (Du et al., 2018; Arora et al., 2018). Finite learning rates, momentum, weight decay, and gradient noise break them (Kunin et al., 2021). Along each rescaling, SGD drifts toward parametrizations with a smaller second moment of the stochastic gradient, and the conserved quantity moves toward a unique, initialization-independent noise equilibrium, at which the gradient noise on the two sides of the rescaling balances (Ziyin et al., 2024); for DLNs, these equilibria have been computed at the global minimum (Ziyin et al., 2024). Backward error analysis explains the drift: to second order in the learning rate, SGD follows the gradient flow of the loss plus a multiple of the mean squared norm of the minibatch gradients, which, unlike the loss, changes under the rescalings (Barrett & Dherin, 2021; Smith et al., 2021). Gradient noise also attracts the dynamics toward invariant sets corresponding to simpler subnetworks (Chen et al., 2023). At a global minimum with isotropic label noise, the second moment of the stochastic gradient is proportional to the trace of

the Hessian, so the drift is a flatness bias (Blanc et al., 2020; Damian et al., 2021; Li et al., 2022; Xu et al., 2026); since sharpness is not invariant under rescalings (Dinh et al., 2017), such biases act on the parametrization of a fixed function. We compute the drift at the saddles of DLNs, away from the global minimum. Without label noise, it has no finite equilibrium and sharpens the learned modes; with label noise, its equilibrium is shifted toward the output layer by the remaining error and recovers that of Ziyin et al. (2024) at full fit.

SGD and GD in linear networks. Linear networks have also revealed differences between SGD and GD beyond the loss. In diagonal linear networks, stochastic gradient flow selects sparser interpolators than gradient flow (Pesme et al., 2021). In linear networks, SGD yields weights of lower rank than GD (Varre et al., 2024), and with ℓ_2 regularization, SGD makes one-way jumps from high-rank to low-rank minima (Wang & Jacot, 2024); see also Lyu & Zhu (2023) for rank-one linear networks. The drift we describe keeps the end-to-end map fixed, rescales the learned modes across layers, and requires neither label noise nor regularization.

Training dynamics of deep linear networks. For aligned and balanced initializations, gradient flow in DLNs is exactly solvable (Saxe et al., 2014), and balanced factorizations give closed-form dynamics of the end-to-end singular values (Arora et al., 2019). From small initialization, gradient descent learns the modes of the target greedily, in a saddle-to-saddle sequence (Li et al., 2021; Jacot et al., 2021), as do diagonal linear networks (Pesme & Flammarion, 2023); from large initialization, it trains in the lazy regime of the neural tangent kernel (Jacot et al., 2018). The initialization scale and the imbalance between layers together determine where training falls between these regimes (Dominé et al., 2025; Kunin et al., 2024), and the dynamics from random initialization have also been characterized (Bordelon & Pehlevan, 2025). These analyses take the imbalance to be fixed by the initialization, as it is under gradient flow; we describe how gradient noise rescales the layers near the saddles.

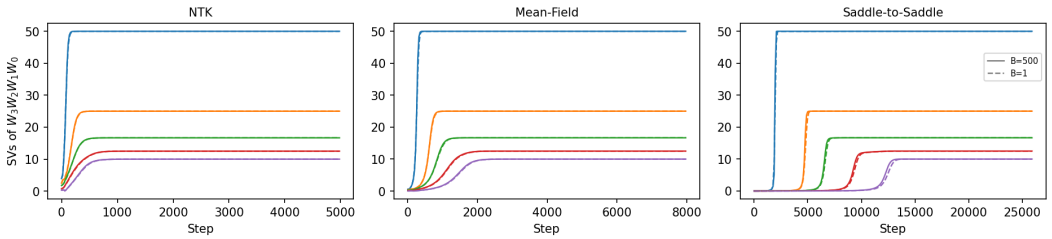
3 SGD AND GD LEARN NEARLY THE SAME FUNCTION

We first compare the functions that SGD and GD learn. We train DLNs of depth $L = 4$ and hidden width 50 on a rank-5 teacher with a square loss, from three initialization scales that produce the lazy, mean-field and saddle-to-saddle regimes, both online, with a fresh sample at every step, and offline, on a fixed training set of 500 samples. SGD uses batch size 1 and GD batch size 500, and both start from the same initialization (experimental details are in Appendix A). We compare the end-to-end maps W_{SGD} and W_{GD} directly, through their relative distance $\|W_{\mathrm{SGD}} - W_{\mathrm{GD}}\|_F / \|W_{\mathrm{GD}}\|_F$ and their singular values, and we compare their population losses, averaging SGD over independent runs.

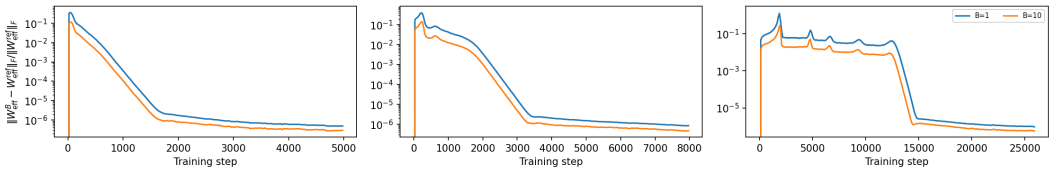
End-to-end maps. In all three regimes, the singular values of the end-to-end maps of SGD and GD overlap throughout training (Figure 2a). The relative distance between the two maps is largest when the function changes fastest (Figure 2b): early in training in the lazy and mean-field regimes, and at the saddle escapes in the saddle-to-saddle regime, where SGD lags slightly behind GD. On the plateaus it stays at a few percent, it is smaller for batch size 10 than for batch size 1, and once the teacher is fitted it drops by several orders of magnitude. SGD thus follows the path of GD in function space, up to a small delay at the saddle escapes. The singular values agree equally well in offline training and in the presence of label noise (Appendix A; Figure 4).

Population loss. In none of our runs does SGD reach a lower population loss than GD (Figure 3). GD's loss is significantly lower for most of training, and the ratio of the two losses peaks where the loss falls fastest, because there a small delay of SGD changes the loss by orders of magnitude.

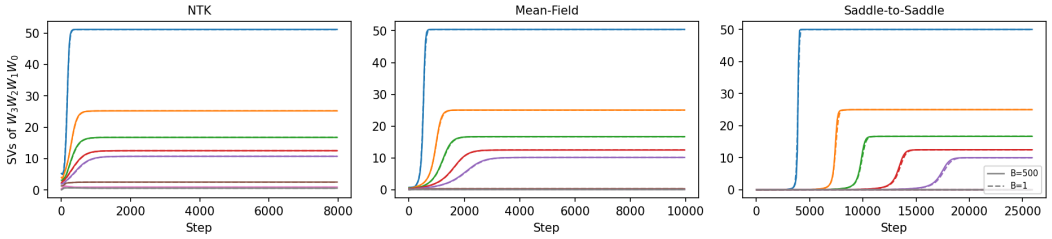
Singular input covariances. When the input covariance Σ_x is singular, the loss constrains the end-to-end map only on the range of Σ_x, so its global minimizers form a continuum in function space. Every input is orthogonal to a null vector v of Σ_x, and every sample gradient with respect to W_1 is an outer product with the input, so it vanishes on v. Hence $W_1 v$ keeps its initial value under both GD and SGD, and the learned map sends v to $W_L \cdots W_2 \, W_1(0) \, v$. SGD and GD can therefore differ on the null space only through $W_2, \ldots, W_L$, whose learned singular values are rescaled by



(a) Singular values of the end-to-end map, for GD (batch size 500, solid) and SGD (batch size 1, dashed).



(b) Relative distance between the end-to-end maps of SGD and GD, for SGD with batch size 1 and 10.



(c) As in (a), for an input covariance of rank 5 in dimension 8.

Figure 2: **SGD and GD learn nearly the same function.** Online training of a depth-4, width-50 DLN on a rank-5 teacher, in the lazy (NTK), mean-field and saddle-to-saddle regimes (columns).

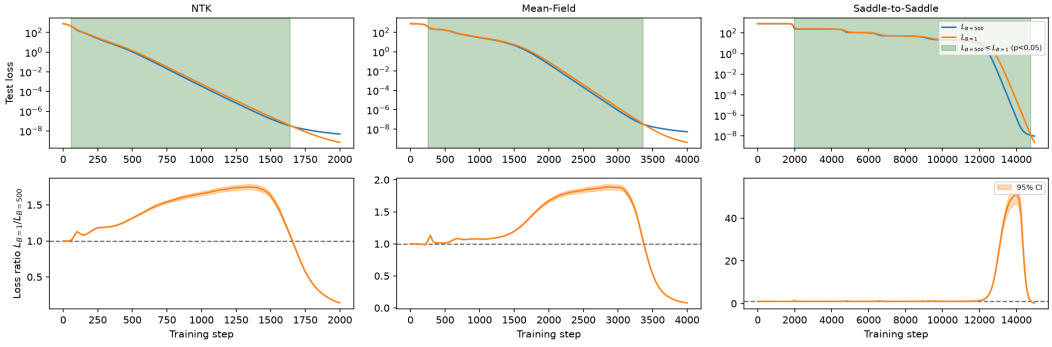


Figure 3: **Minibatch noise does not lower the population loss.** Online training of a depth-4, width-50 DLN in the three regimes. Top: population loss of GD (batch size 500) and of SGD (batch size 1, averaged over 10^4 runs); shading marks the steps at which GD's loss is significantly lower ($p < 0.05$). Bottom: ratio of SGD's loss to GD's, with a 95% confidence band. The apparent late-time dip of this ratio below one is a finite-precision artifact: once the loss reaches the numerical floor, the large-batch GD updates become too small to be resolved, rather than SGD attaining a lower population loss.

the drift of Section 4, and by an amount proportional to $W_1(0)v$. With an input covariance of rank 5 in dimension 8, SGD and GD reach nearly the same minimizer in all three regimes: their end-to-end singular values coincide, including those that the loss leaves unconstrained (Figure 2c).

4 GRADIENT NOISE RESCALES THE LAYERS AT THE SADDLES

We now explain the drift in Figure 1. The network function is invariant under the rescalings $W_\ell \mapsto GW_\ell$, $W_{\ell+1} \mapsto W_{\ell+1}G^{-1}$ for invertible G, so gradient flow conserves the imbalances $C_\ell = W_\ell W_\ell^\top - W_{\ell+1}^\top W_{\ell+1}$, $1 \le \ell < L$, which measure how scale is shared between adjacent layers. SGD does not conserve them. We study how they change near the low-rank saddles of saddle-to-saddle training, where the population gradient vanishes but the unlearned modes still produce prediction error, so the stochastic gradient is noisy even without label noise.

Setup. We train $M = W_L \cdots W_1$, with $W_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}$ and $L \ge 2$, on $x \sim \mathcal{N}(0, \Sigma_x)$ and $y = Ax + \varepsilon$, where the label noise ε is independent of x and has mean zero, covariance Σ_ε and finite fourth moment. We use online SGD with learning rate η on the sample loss $\frac{1}{2}\|Mx - y\|^2$, starting from a small random initialization. We write g_ℓ for the sample gradient with respect to W_ℓ, $\Phi = \mathbb{E}\sum_\ell \|g_\ell\|_F^2$ for its second moment, and $P_k = W_k \cdots W_1$, $S_k = W_L \cdots W_{k+1}$, with $P_0 = S_L = I$.

Imbalances under SGD. For symmetric K, let $T_{\ell,K}$ be the tangent at $s = 0$ of the rescaling $W_\ell \mapsto e^{sK}W_\ell$, $W_{\ell+1} \mapsto W_{\ell+1}e^{-sK}$; then $T_{\ell,K} = \nabla \frac{1}{2}\operatorname{tr}(KC_\ell)$. Since every sample loss is invariant under the rescalings, every sample gradient is orthogonal to every $T_{\ell,K}$. Hence, as C_ℓ is quadratic, an SGD step $W \mapsto W - \eta g$ changes it by exactly

$$C_\ell(W - \eta g) - C_\ell(W) = \eta^2\big(g_\ell g_\ell^\top - g_{\ell+1}^\top g_{\ell+1}\big). \tag{1}$$

Its conditional mean given the weights is $\eta^2 F_\ell(W)$, with the drift $F_\ell = \mathbb{E}[g_\ell g_\ell^\top - g_{\ell+1}^\top g_{\ell+1}]$, and differentiating Φ along the rescalings gives (Appendix B.2)

$$\operatorname{tr}(KF_\ell) = -\tfrac{1}{2}\, D\Phi[T_{\ell,K}] \qquad \text{for all symmetric } K. \tag{2}$$

On average, the imbalances thus change as if the weights moved along the rescalings in the direction that decreases Φ, and the drift vanishes exactly at the stationary points of Φ along the rescalings, which are the noise equilibria of Ziyin et al. (2024). This agrees with the backward-error picture of SGD as gradient flow on the loss with the explicit regularizer $\frac{\eta}{4}\Phi$ (Barrett & Dherin, 2021; Smith et al., 2021), but (1) and (2) are exact. Since $\mathbb{E}[g_\ell g_\ell^\top]$ is the outer product of the mean gradient plus the gradient covariance, F_ℓ is the drift that full-batch gradient descent would produce plus a part due to gradient noise. At the saddles considered below, the mean gradient vanishes, so only the noise part remains. A batch of m independent samples divides the gradient covariance by m, so with batch size m the drift at the saddles is F_ℓ/m.

The saddle manifold. Let $A = \sum_{j \le r} s_j u_j v_j^\top$ be the singular value decomposition of the teacher, and assume $\Sigma_x v_j = \lambda_j v_j$ and $\Sigma_\varepsilon u_j = \mu_j u_j$ with $\lambda_j, \mu_j > 0$, as holds when both the inputs and the label noise are isotropic. We index the modes in the order in which they are learned. Suppose that the first $q < r$ modes, $A_q = \sum_{i \le q} s_i u_i v_i^\top$, have been learned, and that every hidden layer has width at least q. We assume that the learned modes are nearly decoupled: each is carried by its own hidden coordinate in every layer, and the weights connecting different modes are small. Saxe et al. (2014) introduced this decoupled form to solve the learning dynamics exactly, and showed that training from small random initialization closely follows it (their Fig. 3). The loss and the SGD update are equivariant under orthogonal changes of coordinates in the input, hidden and output spaces, so we can fix coordinates in which $v_i = e_i$ and $u_i = e_i$ for $i \le r$, and in which mode i occupies coordinate i of each hidden layer. The exactly decoupled networks that compute A_q form the set

$$\mathcal{S}_q = \Big\{ W :\ W_\ell = \left(\begin{smallmatrix} D_\ell & 0 \\ 0 & 0 \end{smallmatrix}\right),\ D_\ell = \operatorname{diag}(\beta_{\ell 1}, \dots, \beta_{\ell q}),\ \beta_{\ell i} > 0,\ \textstyle\prod_\ell W_\ell = A_q \Big\}, \tag{3}$$

$\beta_{\ell i}$ is the singular value of W_ℓ associated with mode i. Every point of $\mathcal{S}_q$ is a critical point, and a saddle if every hidden layer is wider than q.

On $\mathcal{S}_q$, the unlearned modes produce error with covariance $\Gamma_q = (A - A_q)\Sigma_x(A - A_q)^\top$ in output directions that $W_L^\top$ annihilates, so the hidden layers receive only label noise. The remaining error is $\operatorname{tr}\Gamma_q = \sum_{j>q} \lambda_j s_j^2$. The drift and the second moment of the gradient are (Appendix B.3)

$$\begin{aligned} F_\ell &= X_{\ell-1}\, S_\ell^\top \Sigma_\varepsilon S_\ell - Y_{\ell+1}\, P_\ell \Sigma_x P_\ell^\top, \qquad \Phi = \sum_{\ell=1}^{L} X_{\ell-1} Y_\ell, \\ X_k &= \operatorname{tr}\big(P_k \Sigma_x P_k^\top\big), \qquad Y_k = \operatorname{tr}\big(S_k^\top (\Gamma_q + \Sigma_\varepsilon) S_k\big), \end{aligned} \tag{4}$$

where X_k and Y_k are the powers (traces of the covariances) of the activations of layer k and of the error reaching it. The first term of F_ℓ moves scale into layer ℓ in proportion to the label noise reaching it, and the second moves scale into layer $\ell+1$ in proportion to the activations of layer ℓ. The unlearned modes enter only through $Y_L = \operatorname{tr}(\Sigma_\varepsilon + \Gamma_q)$, which, like $X_0 = \operatorname{tr}\Sigma_x$, is constant on $\mathcal{S}_q$. Since the mean gradient vanishes on $\mathcal{S}_q$, Φ is there the total variance of the stochastic gradient, which we call the gradient noise.

A flow on $\mathcal{S}_q$. SGD does not stay exactly on $\mathcal{S}_q$: the weights outside the learned diagonal blocks are small but not zero, and the diagonal entries M_{ii} of the end-to-end map differ slightly from s_i. We therefore define a flow on $\mathcal{S}_q$ from the drift F_ℓ at its points, and show in Proposition 4.1 that SGD follows this flow to leading order. On $\mathcal{S}_q$, both C_ℓ and F_ℓ are diagonal, and a point of $\mathcal{S}_q$ is determined by its imbalances, because the products of its singular values along each mode are fixed. The equations

$$\frac{dC_\ell}{d\tau} = F_\ell, \qquad 1 \le \ell < L, \tag{5}$$

in the slow time $\tau = \eta^2 t$, where t counts SGD steps, therefore define this flow, along which each imbalance changes at the mean rate of an SGD step taken from the current point. Identifying each point of $\mathcal{S}_q$ with its singular-value coordinates $\beta = (\beta_1, \dots, \beta_q)$, the metric induced by the Euclidean metric on the weights is the Euclidean metric on β. By (2), the flow is therefore the gradient flow of Φ on $\mathcal{S}_q$:

$$\dot\beta = -\tfrac{1}{4}\,\operatorname{grad}_{\mathcal{S}_q}\Phi(\beta). \tag{6}$$

SGD follows the flow. To compare SGD with the flow, we use quantities defined for arbitrary weights, not only on $\mathcal{S}_q$. Let $D_\ell(W)$ be the matrix obtained from W_ℓ by retaining only the diagonal entries of its learned $q \times q$ block, and write $W_\ell = D_\ell(W) + E_\ell(W)$. For $\Lambda \ge 1$ and $\delta_W, \delta_M \ge 0$, let $\mathcal{N}$ be the set of weights satisfying

$$\|W_\ell\|_{\mathrm{op}} \le \Lambda, \qquad D_\ell(W) \succ 0, \qquad \|E_\ell(W)\|_F \le \delta_W, \qquad |M_{ii} - s_i| \le \delta_M \quad (i \le q),$$

so that δ_W and δ_M bound the deviation from $\mathcal{S}_q$, and let κ be the first step at which SGD leaves $\mathcal{N}$. We prove the following result in Appendix C.

Proposition 4.1. Let $W^{(0)} \in \mathcal{N}$, and let $\bar W(\tau)$ be the solution of (5) from the point of $\mathcal{S}_q$ with the same learned imbalances as $W^{(0)}$. For every $T > 0$, there exist constants $C_T < \infty$ and $\delta_0 > 0$ such that, if $\delta_W, \delta_M \le \delta_0$ and $0 < \eta \le 1$, then

$$\left(\mathbb{E}\max_{n \le T/\eta^2,\, n<\kappa} \sum_\ell \big\|D_\ell(W^{(n)}) - D_\ell(\bar W(\eta^2 n))\big\|_F^2\right)^{1/2} \le C_T\big(\eta + \delta_M + \delta_W^2\big).$$

Here C_T depends on T, Λ, L, the dimensions, the teacher, and the data moments, but not on η, δ_W, or δ_M.

On $\mathcal{S}_q$, the flow coordinates $\beta_{\ell i}$ are exactly the layer singular values associated with the learned modes. Proposition 4.1 controls instead the corresponding diagonal entries of the actual SGD weights, which need not themselves be singular values away from $\mathcal{S}_q$. If the singular value continuing from $\beta_{\ell i}$ is separated from the other singular values of W_ℓ and from zero by a gap $g > 0$, then standard singular-value perturbation gives

$$\big|\sigma_{\ell i}(W) - (D_\ell(W))_{ii}\big| = O\big(\delta_W^2/g\big)\,,$$

Thus, while these singular values remain separated, Proposition 4.1 implies that their dynamics are described to leading order by the singular-value flow $\beta_{\ell i}(\tau)$ on $\mathcal{S}_q$.

Without label noise. With $\Sigma_\varepsilon = 0$, (4) shows that only the last imbalance drifts, at the rate $-\operatorname{tr}\Gamma_q\, P_{L-1}\Sigma_x P_{L-1}^\top$. The modes evolve independently: the singular values of mode i in the earlier layers decrease, with their squared differences fixed by the conserved imbalances, while its output-layer singular value grows according to

$$\frac{d\beta_{Li}}{d\tau} = \frac{\lambda_i s_i^2 \operatorname{tr}\Gamma_q}{2\beta_{Li}^3}\,\frac{\sum_{\ell<L}(\beta_{Li}/\beta_{\ell i})^2}{1 + \sum_{\ell<L}(\beta_{Li}/\beta_{\ell i})^2}, \tag{7}$$

a closed equation, since the squared differences of the earlier-layer singular values are conserved and $\prod_{\ell<L} \beta_{\ell i} = s_i/\beta_{Li}$. As the output-layer singular value grows relative to the earlier-layer singular values, the fraction in (7) tends to one, giving

$$\beta_{Li}(\tau) \simeq \big(2\lambda_i s_i^2 \operatorname{tr}\Gamma_q\, \tau\big)^{1/4} \qquad (\tau \to \infty). \tag{8}$$

The exponent $1/4$ does not depend on the depth, the initial singular values or the number of learned modes. Since the earlier imbalances are frozen, the earlier layers remain approximately balanced from small initialization, so $\beta_{\ell i} \approx (s_i/\beta_{Li})^{1/(L-1)}$ for $\ell < L$, and they decay as

$$\beta_{\ell i}(\tau) \simeq s_i^{1/(L-1)} \big(2\lambda_i s_i^2 \operatorname{tr}\Gamma_q\, \tau\big)^{-1/[4(L-1)]} \qquad (\ell < L,\ \tau \to \infty). \tag{9}$$

Figure 1a compares this flow with SGD, plateau by plateau.

With label noise. Label noise makes the first term of F_ℓ in (4) nonzero: the label noise reaching each hidden layer moves scale back into it, which opposes the redistribution. If $\mu_i > 0$ for every learned mode, we prove in Appendix B.5 that Φ has a unique minimizer on $\mathcal{S}_q$, that it is the only fixed point of the flow, and that every trajectory converges to it. The fixed point is

$$\beta_{\ell i} = \rho := \Big(\frac{c_q^2}{X_0 Y_L}\Big)^{1/(2L)} \quad (1 < \ell < L), \qquad \frac{\beta_{Li}}{\beta_{1i}} = \Big(\frac{\lambda_i}{\mu_i}\,\frac{Y_L}{X_0}\Big)^{1/2}, \qquad c_q = \sum_{j\le q} s_j \sqrt{\lambda_j \mu_j}. \tag{10}$$

At this fixed point, the learned blocks of all intermediate layers equal ρI_q. Setting $\Gamma_q = 0$ recovers the equilibrium at the global minimum found by Ziyin et al. (2024). SGD approaches this fixed point until the next mode is learned; whether it reaches it depends on the length of the plateau. In Figure 1b, each plateau ends before SGD reaches it; after the last mode is learned, the input- and output-layer singular values move toward the equilibrium of Ziyin et al. (2024).

Plateaus and transitions. Each SGD step changes the imbalances by $O(\eta^2)$, so the parameters can only drift by $O(1)$ over $O(\eta^{-2})$ gradient steps. This requires long plateaus in gradient-flow time, which small initialization produces. When the next mode $q+1$ is learned, the network moves near $\mathcal{S}_{q+1}$ and the residual covariance drops from Γ_q to Γ_{q+1}. The transitions between plateaus, which we do not model, are short compared with the plateaus in gradient-flow time, so they should have little effect on the redistribution, as our experiments confirm (Figure 1).

Bottlenecks. If a hidden layer has width $b < r$, at most b modes can be fitted, and the remaining error stays positive. Without label noise, the drift therefore persists after the loss has converged.

Sharpening. On $\mathcal{S}_q$, the Hessian of the loss with respect to the layer singular values of mode i is $\lambda_i s_i^2\, n_i n_i^\top$, where n_i is normal to $\mathcal{S}_q$. Its nonzero eigenvalue $\lambda_i s_i^2 \|n_i\|^2$ is a lower bound on the largest eigenvalue of the full Hessian, and, by the inequality of arithmetic and geometric means, it is smallest on $\mathcal{S}_q$ when the singular values of mode i are equal. Without label noise, the drift moves a mode learned with equal layer singular values away from balance, so this curvature increases, at second order at first, and grows without bound as $\tau \to \infty$, since the smallest earlier-layer singular values vanish: the drift induced by SGD noise sharpens the network. This is the opposite of label noise at a global minimum, where $\Phi = \sigma^2 \operatorname{tr}\nabla^2 L$ for $\Sigma_\varepsilon = \sigma^2 I$, so that the drift flattens the network (Blanc et al., 2020; Li et al., 2022; Damian et al., 2021).

5 Discussion

In deep linear networks, gradient noise barely changes the learned function but changes how it is parametrized: near saddles, SGD follows a gradient flow of the second moment of the stochastic gradient along the rescalings. The modes not yet learned act on the learned ones like label noise, except that their error lies in output directions the network does not yet use, so it reaches only the output layer. The second moment then keeps falling as scale moves into the output layer: the noise equilibrium lies at infinity, and the learned modes sharpen. Label noise, which reaches every layer, restores a finite equilibrium, shifted toward the output layer by the remaining error (10), which becomes that of Ziyin et al. (2024) at full fit.

A golden path in function space. Our results support the golden path hypothesis in function space (Section 3), in line with Vyas et al. (2024), but not in parameter space. There, SGD rescales the learned modes systematically. At the saddles the drift comes entirely from gradient noise and scales as 1/m with batch size m, so large-batch SGD rescales the modes only slightly, and full-batch GD, whose gradient vanishes there, does not rescale them. Gradient flow can therefore model the function that SGD learns, but not how SGD distributes it across layers: analyses that assume balanced layers (Saxe et al., 2014; Arora et al., 2019; Corlouer et al., 2026) describe the function learned by SGD but not its layer scales.

Rescaling happens in the saddle-to-saddle regime. Without label noise, every per-sample gradient vanishes at full fit, so the imbalance accumulated during training is frozen. The parametrization reached by SGD is then a record of its path rather than an equilibrium, unlike the unique, initialization-independent equilibria of Ziyin et al. (2024). A mode learned early is exposed to the error of every later plateau, so earlier modes carry more imbalance. Since the imbalances move on the slow time $\tau = \eta^2 t$, the redistribution is appreciable only when training spends sufficiently long near the saddles. Consistent with this picture, the layer-scale drift is pronounced in the saddle-to-saddle regime, much weaker in the mean-field regime, and essentially absent in the lazy regime (Figure 5). When a hidden layer is narrower than the rank of the teacher, the error never vanishes and, without label noise, the redistribution continues after the loss has converged, a regime that may be closer to the online training of large models.

Why the parametrization matters. For a learned mode i, the quantity $\lambda_i s_i^2 \sum_\ell \beta_{\ell i}^{-2}$ is both the curvature of the loss along the mode and the tangent-kernel eigenvalue that sets how fast gradient descent re-fits the mode, for instance during fine-tuning. Without label noise, the drift makes it grow without bound, so SGD moves toward sharper parameterizations of the same function, which re-fit learned modes faster but tolerate only smaller learning rates. This is the opposite of label-noise SGD at a global minimum, which lowers the trace of the Hessian (Blanc et al., 2020; Damian et al., 2021; Li et al., 2022): lowering gradient fluctuations can instead sharpen a fixed function (Dinh et al., 2017; Xu et al., 2026). For interpretability, layer-local quantities such as weight norms, per-layer spectra, and magnitude-based pruning scores then depend on the learning rate and batch size, although the drift preserves the singular subspaces of the learned modes: it changes where the scale of each feature is stored, not which features are represented.

Limitations. Our results are restricted to deep linear networks, and our analysis to online SGD with Gaussian inputs and label noise whose covariances are aligned with the teacher. The analysis assumes that the learned modes are decoupled, holds while SGD stays near a saddle, and does not model the transitions between saddles. We also don't include weight decay, which opposes the drift, since an ℓ_2 penalty at a fixed end-to-end map is minimized by balanced layers.

Future work. Natural next settings are ReLU networks, whose hidden neurons have their own rescaling symmetry, as well as tree tensor networks (Furman et al., 2026) and deep quadratic networks (Kunin et al., 2025). It also remains open what happens once the sharpening brings training to the edge of stability (Cohen et al., 2021; Damian et al., 2023; Ghosh et al., 2025).

AI USE STATEMENT

We used ChatGPT and Claude as tools to assist in mathematical derivations and checking intermediate calculations. All proofs were verified by the first author. ChatGPT and Claude were used to assist with writing, editing, and LaTeX formatting.

REFERENCES

Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *International Conference on Machine Learning*, 2018.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

David G. T. Barrett and Benoit Dherin. Implicit gradient regularization. In *International Conference on Learning Representations*, 2021.

Guy Blanc, Neha Gupta, Gregory Valiant, and Paul Valiant. Implicit regularization for deep neural networks driven by an Ornstein–Uhlenbeck like process. In *Conference on Learning Theory*, 2020.

Blake Bordelon and Cengiz Pehlevan. Deep linear network training dynamics from random initialization: Data, width, depth, and hyperparameter transfer. *arXiv preprint arXiv:2502.02531*, 2025.

Feng Chen, Daniel Kunin, Atsushi Yamamura, and Surya Ganguli. Stochastic collapse: How gradient noise attracts SGD dynamics towards simpler subnetworks. In *Advances in Neural Information Processing Systems*, volume 36, pp. 35027–35063, 2023.

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. In *International Conference on Learning Representations*, 2021.

Guillaume Corlouer, Avi Semler, Alexander Strang, and Alexander Gietelink Oldenziel. Stochastic gradient descent in the saddle-to-saddle regime of deep linear networks. *arXiv preprint arXiv:2604.06366*, 2026.

Alex Damian, Tengyu Ma, and Jason D. Lee. Label noise SGD provably prefers flat global minimizers. In *Advances in Neural Information Processing Systems*, 2021.

Alex Damian, Eshaan Nichani, and Jason D. Lee. Self-stabilization: The implicit bias of gradient descent at the edge of stability. In *International Conference on Learning Representations*, 2023.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. In *International Conference on Machine Learning*, 2017.

Clémentine C. J. Dominé, Nicolas Anguita, Alexandra M. Proca, Lukas Braun, Daniel Kunin, Pedro A. M. Mediano, and Andrew M. Saxe. From lazy to rich: Exact learning dynamics in deep linear networks. In *International Conference on Learning Representations*, 2025.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems*, 2018.

Zach Furman, Stephan Wäldchen, Yangda Bei, and Liam Hodgkinson. Benign loss landscapes can coexist with worst-case hardness. *arXiv preprint arXiv:2609.13057*, 2026.

Jonas Geiping, Micah Goldblum, Phillip E. Pope, Michael Moeller, and Tom Goldstein. Stochastic training is not necessary for generalization. In *International Conference on Learning Representations*, 2022.

Avrajit Ghosh, Soo Min Kwon, Rongrong Wang, Saiprasad Ravishankar, and Qing Qu. Learning dynamics of deep matrix factorization beyond the edge of stability. In *International Conference on Learning Representations*, 2025.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*, 2021.

Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel L. K. Yamins, and Hidenori Tanaka. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. In *International Conference on Learning Representations*, 2021.

Daniel Kunin, Allan Raventós, Clémentine Dominé, Feng Chen, David Klindt, Andrew Saxe, and Surya Ganguli. Get rich quick: Exact solutions reveal how unbalanced initializations promote rapid feature learning. In *Advances in Neural Information Processing Systems*, 2024.

Daniel Kunin, Giovanni Luca Marchetti, Feng Chen, Dhruva Karkada, James B. Simon, Michael R. DeWeese, Surya Ganguli, and Nina Miolane. Alternating gradient flows: A theory of feature learning in two-layer neural networks. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2021.

Zhiyuan Li, Tianhao Wang, and Sanjeev Arora. What happens after SGD reaches zero loss? A mathematical framework. In *International Conference on Learning Representations*, 2022.

Bochen Lyu and Zhanxing Zhu. Implicit bias of (stochastic) gradient descent for rank-1 linear neural network. In *Advances in Neural Information Processing Systems*, volume 36, pp. 58166–58201, 2023.

Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Abide by the law and follow the flow: Conservation laws for gradient flows. In *Advances in Neural Information Processing Systems*, 2023.

Yoonsoo Nam, Seok Hyeong Lee, Clémentine C. J. Dominé, Yeachan Park, Charles London, Wonyl Choi, Niclas Göring, and Seungjai Lee. Position: Solve layerwise linear models first to understand neural dynamical phenomena (Neural collapse, emergence, Lazy/Rich regime, and grokking). In *International Conference on Machine Learning*, pp. 81897–81929, 2025.

Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Implicit regularization or implicit conditioning? Exact risk trajectories of SGD in high dimensions. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Homogenization of SGD in high-dimensions: Exact dynamics and generalization properties. *Mathematical Programming*, pp. 1–90, 2024.

Scott Pesme and Nicolas Flammarion. Saddle-to-saddle dynamics in diagonal linear networks. In *Advances in Neural Information Processing Systems*, 2023.

Scott Pesme, Loucas Pillaud-Vivien, and Nicolas Flammarion. Implicit bias of SGD for diagonal linear networks: a provable benefit of stochasticity. In *Advances in Neural Information Processing Systems*, volume 34, pp. 29218–29230, 2021.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

Jamie Simon, Daniel Kunin, Alexander Atanasov, Enric Boix-Adserà, Blake Bordelon, Jeremy Cohen, Nikhil Ghosh, Florentin Guth, Arthur Jacot, Mason Kamb, Dhruva Karkada, Eric J. Michaud, Berkan Ottlik, and Joseph Turnbull. There will be a scientific theory of deep learning. *arXiv preprint arXiv:2604.21691*, 2026.

Samuel L. Smith, Erich Elsen, and Soham De. On the generalization benefit of noise in stochastic gradient descent. In *International Conference on Machine Learning*, 2020.

Samuel L. Smith, Benoit Dherin, David G. T. Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. In *International Conference on Learning Representations*, 2021.

Aditya Vardhan Varre, Margarita Sagitova, and Nicolas Flammarion. SGD vs GD: Rank deficiency in linear networks. In *Advances in Neural Information Processing Systems*, volume 37, pp. 60133–60161, 2024.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Gal Kaplun, Sham M. Kakade, and Boaz Barak. Beyond implicit bias: The insignificance of SGD noise in online learning. In *International Conference on Machine Learning*, pp. 49698–49716, 2024.

Zihan Wang and Arthur Jacot. Implicit bias of SGD in L_2-regularized linear DNNs: One-way jumps from high to low rank. In *International Conference on Learning Representations*, 2024.

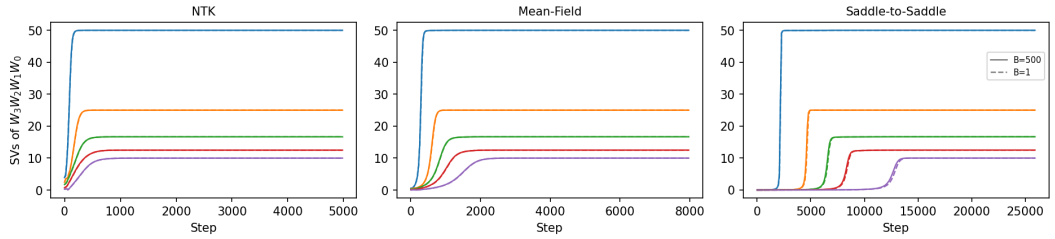
Yizhou Xu, Pierfrancesco Beneventano, Isaac Chuang, and Liu Ziyin. Does SGD seek flatness or sharpness? An exactly solvable model. *arXiv preprint arXiv:2602.05065*, 2026.

Yedi Zhang, Andrew Saxe, and Peter E. Latham. Saddle-to-saddle dynamics explains a simplicity bias across neural network architectures. In *International Conference on Learning Representations*, 2026.

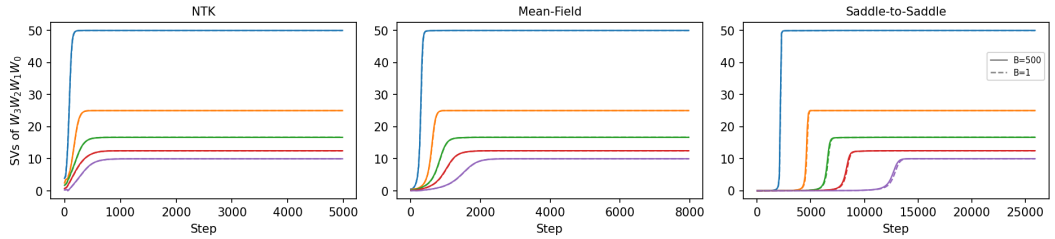
Liu Ziyin, Mingze Wang, Hongchao Li, and Lei Wu. Parameter symmetry and noise equilibrium of stochastic gradient descent. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

A EXPERIMENTAL DETAILS AND ADDITIONAL RESULTS

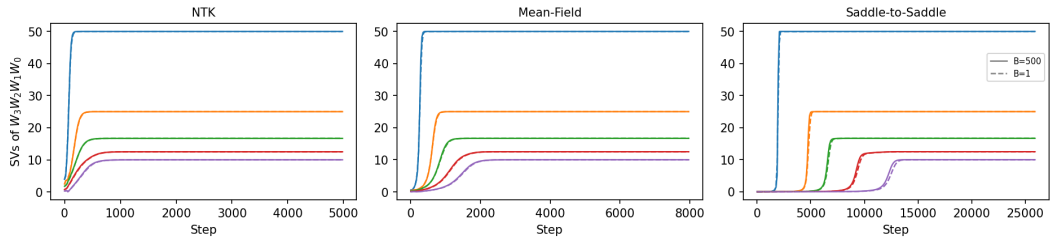
Experimental setup. The experiments in Section 3 use bias-free depth-4 DLNs with input and output dimension 5 and hidden width $h = 50$, trained with squared error and learning rate 10^{-4}. The initialization is $W_{\ell,ij}(0) \sim \mathcal{N}(0, h^{-\gamma})$, equivalently with standard deviation $h^{-\gamma/2}$; $\gamma = 0.75$, 1, and 1.5 give the lazy, mean-field, and saddle-to-saddle regimes, respectively. The comparisons use batch size 1 throughout; the offline panels compare against full-batch training, while the online panel compares against batch size 500. The noisy runs use label-noise standard deviation 0.2.



(a) Offline, no label noise: full batch versus batch size 1.



(b) Offline, label-noise standard deviation 0.2: full batch versus batch size 1.



(c) Online, label-noise standard deviation 0.2: batch size 500 versus batch size 1.

Figure 4: **The end-to-end singular-value agreement persists offline and with label noise.** Singular values of the end-to-end map for the same depth-4, width-50, rank-5 teacher–student setup as Figure 2a, across the lazy, mean-field, and saddle-to-saddle regimes. Figure 2a shows the remaining case, online training without label noise. In the offline panels, solid curves denote full-batch training; in the online panel, solid curves denote batch size 500. Dashed curves denote batch size 1.

Experiments of Figure 1. Figure 1 uses a bias-free depth-4 DLN with input and output dimension 5 and hidden width 16. The teacher is $A = \mathrm{diag}(16, 10, 6, 3, 0)$, the inputs are $x \sim \mathcal{N}(0, I_5)$, and the label noise is $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_5)$, with $\sigma = 0$ in Figure 1a and $\sigma = 0.5$ in Figure 1b. All weights are initialized i.i.d. from $\mathcal{N}(0, 0.0125^2)$, and GD and SGD start from the same initialization. Both are trained online, with fresh samples at every step, SGD with batch size 1 and GD with batch size 500, at learning rate $\eta = 10^{-4}$ in both Figure 1a and Figure 1b. For comparison, without label noise, SGD with batch size 1 diverges at $\eta = 2 \times 10^{-3}$ in this setup.

We measure $\beta_{\ell i}$ as follows. Set $h_{0,i} = v_i$. For $\ell < L$, compute $W_\ell h_{\ell-1,i}$ for all modes and orthonormalize these vectors by Gram–Schmidt, in the order in which the modes are learned, to obtain $h_{\ell,i}$. Then $\beta_{\ell i} = h_{\ell,i}^\top W_\ell h_{\ell-1,i}$ for $\ell < L$, and $\beta_{Li} = u_i^\top W_L h_{L-1,i}$. On $\mathcal{S}_q$ these are

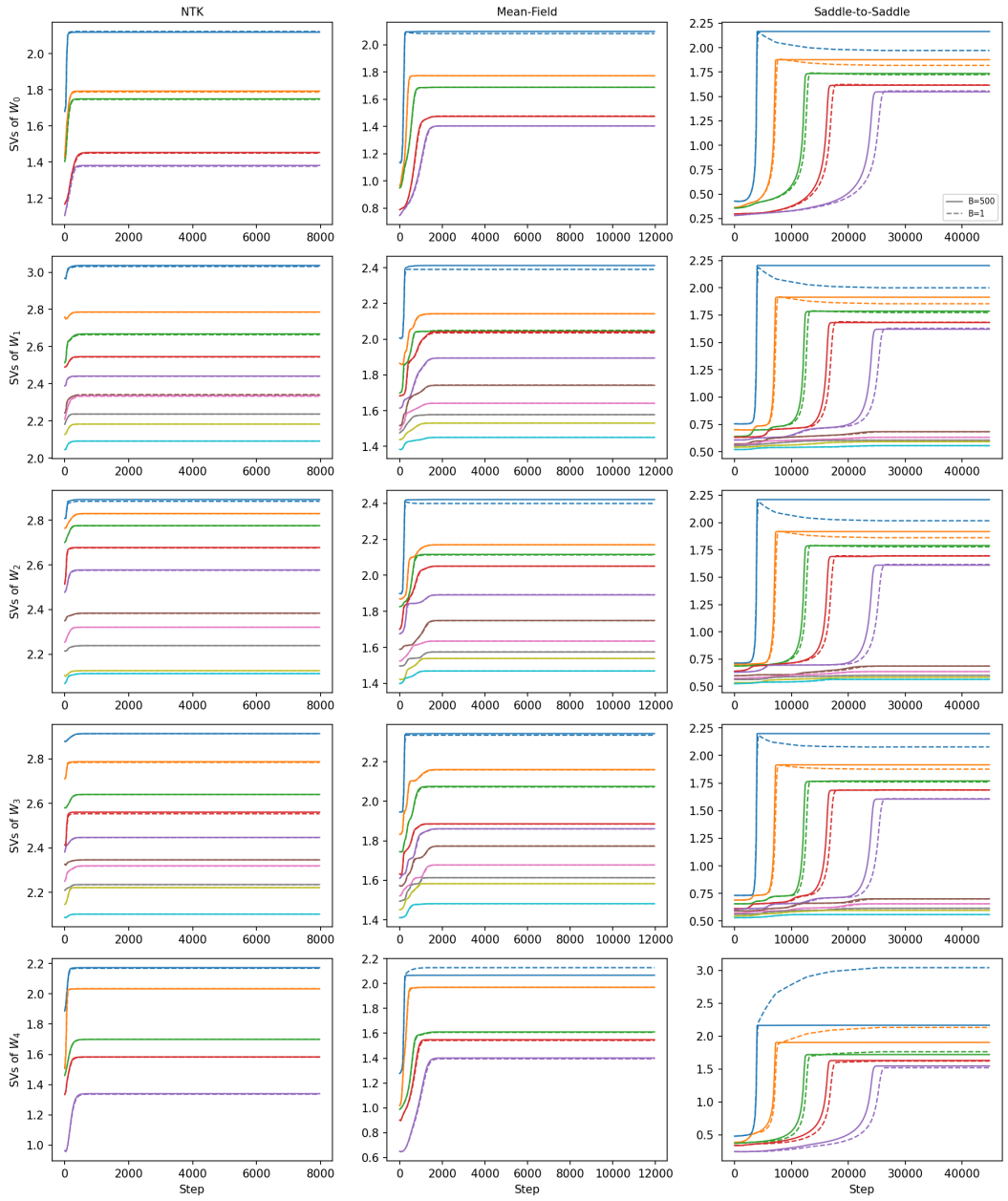


Figure 5: **Layer singular values across training regimes.** Layer singular values during online training of a depth-4, width-50 DLN on a rank-5 teacher in the lazy (NTK), mean-field, and saddle-to-saddle regimes, for GD (batch size 500, solid) and SGD (batch size 1, dashed), plotted against training step. The systematic separation between SGD and GD layer scales is pronounced in the saddle-to-saddle regime, much weaker in the mean-field regime, and essentially absent in the lazy regime.

the layer singular values of the learned modes, and near $\mathcal{S}_q$ they approximate the diagonal entries $D_\ell(W)$ of Proposition 4.1.

The flow is computed plateau by plateau. Mode q counts as learned at the first recorded step t_q (every 250 steps in Figure 1a and every 500 steps in Figure 1b) at which $u_q^\top M v_q \geq 0.99\, s_q$ for SGD. From t_q until t_{q+1}, or until the end of training after the last mode is learned, we integrate (5) with F_ℓ from (4), with $\lambda_i = 1$, $\mu_i = \sigma^2$, $X_0 = 5$ and $Y_L = \operatorname{tr}\Gamma_q + 5\sigma^2$. The integration starts from

the point of $\mathcal{S}_q$ with the same learned imbalances as the SGD weights at t_q, as in Proposition 4.1, and the flow is plotted against the step $t_q + \tau/\eta^2$.

B Derivations for Section 4

B.1 Notation and assumptions

We use the setup of Section 4. Thus $x \sim \mathcal{N}(0, \Sigma_x)$, $y = Ax + \varepsilon$, and the samples are independent across SGD steps. The noise is independent of x, has mean zero, covariance Σ_ε, and finite fourth moment. We use the aligned coordinates fixed in Section 4, so for the teacher coordinates $i \le r$,

$$Ae_i = s_i e_i, \qquad \Sigma_x e_i = \lambda_i e_i, \qquad \Sigma_\varepsilon e_i = \mu_i e_i,$$

with $s_i, \lambda_i > 0$. We fix $1 \le q < r = \operatorname{rank} A$ and assume every hidden layer has width at least q. Section B.4 takes $\Sigma_\varepsilon = 0$; Section B.5 assumes $\mu_i > 0$ for $i \le q$ and also applies when $q = r$.

We write

$$M = W_L \cdots W_1, \qquad P_k = W_k \cdots W_1, \qquad S_k = W_L \cdots W_{k+1}, \qquad P_0 = S_L = I,$$

and denote the sample gradient by g_ℓ. The imbalances are

$$C_\ell = W_\ell W_\ell^\top - W_{\ell+1}^\top W_{\ell+1}.$$

Let $A_q = \sum_{i \le q} s_i u_i v_i^\top$, as in Section 4; in the aligned input/output coordinates this is the diagonal map with first q diagonal entries s_i. Let $\Gamma_q = (A - A_q)\Sigma_x(A - A_q)^\top$. On $\mathcal{S}_q$ the first q diagonal entries of W_ℓ are $\beta_{\ell i} > 0$ and $\prod_\ell \beta_{\ell i} = s_i$. Dots denote derivatives in the slow time $\tau = \eta^2 t$. Unmarked matrix norms are Frobenius norms and $\|\cdot\|_{\mathrm{op}}$ is the operator norm.

B.2 Rescaling identities and equations (1)–(2)

Fix $1 \le \ell < L$ and a symmetric matrix K of the appropriate hidden-layer size. For one sample, the rescaling

$$W_\ell(s) = e^{sK} W_\ell, \qquad W_{\ell+1}(s) = W_{\ell+1} e^{-sK}$$

leaves the end-to-end map, and hence the sample loss, unchanged. Its tangent is $T_{\ell,K}$ from Section 4. Therefore

$$0 = \langle g, T_{\ell,K} \rangle = \operatorname{tr}\left[K\left(W_\ell g_\ell^\top - g_{\ell+1}^\top W_{\ell+1}\right)\right].$$

Since this holds for every symmetric K,

$$W_\ell g_\ell^\top + g_\ell W_\ell^\top = g_{\ell+1}^\top W_{\ell+1} + W_{\ell+1}^\top g_{\ell+1}.$$

Expanding C_ℓ after the step $W \mapsto W - \eta g$ therefore cancels all terms linear in η and gives exactly

$$C_\ell(W - \eta g) - C_\ell(W) = \eta^2\left(g_\ell g_\ell^\top - g_{\ell+1}^\top g_{\ell+1}\right),$$

which is equation (1).

The same rescaling gives the derivative of the gradient second moment. At the rescaled weights the only sample gradients that change are

$$g_\ell(s) = e^{-sK} g_\ell, \qquad g_{\ell+1}(s) = g_{\ell+1} e^{sK};$$

all other layer gradients are unchanged. Hence

$$\begin{aligned} D\Phi[T_{\ell,K}] &= -2\,\mathbb{E}\operatorname{tr}(K g_\ell g_\ell^\top) + 2\,\mathbb{E}\operatorname{tr}(K g_{\ell+1}^\top g_{\ell+1}) \\ &= -2\operatorname{tr}(K F_\ell), \end{aligned}$$

which is equation (2).

B.3 GRADIENT MOMENTS AND EQUATION (4)

Here we derive equation (4),

$$F_\ell = X_{\ell-1} S_\ell^\top \Sigma_\varepsilon S_\ell - Y_{\ell+1} P_\ell \Sigma_x P_\ell^\top, \qquad \Phi = \sum_{\ell=1}^{L} X_{\ell-1} Y_\ell.$$

By definition,

$$F_\ell = \mathbb{E}[g_\ell g_\ell^\top - g_{\ell+1}^\top g_{\ell+1}], \qquad \Phi = \sum_{\ell=1}^{L} \mathbb{E}\|g_\ell\|_F^2.$$

For arbitrary weights, write

$$R = M - A, \qquad Q = R\Sigma_x R^\top + \Sigma_\varepsilon, \qquad \bar{g}_\ell = S_\ell^\top R \Sigma_x P_{\ell-1}^\top.$$

Set $e = Rx - \varepsilon$. Since

$$g_\ell = S_\ell^\top e (P_{\ell-1} x)^\top,$$

and

$$\mathbb{E}[ee^\top \mid x] = Rxx^\top R^\top + \Sigma_\varepsilon,$$

we obtain

$$\mathbb{E}[g_\ell g_\ell^\top] = S_\ell^\top \mathbb{E}_x\big[\|P_{\ell-1}x\|^2 \big(Rxx^\top R^\top + \Sigma_\varepsilon\big)\big] S_\ell,$$

and

$$\mathbb{E}[g_\ell^\top g_\ell] = P_{\ell-1} \mathbb{E}_x\big[xx^\top \left(x^\top R^\top S_\ell S_\ell^\top R x + \operatorname{tr}(S_\ell^\top \Sigma_\varepsilon S_\ell)\right)\big] P_{\ell-1}^\top.$$

The Gaussian fourth-moment identity

$$\mathbb{E}[xx^\top (x^\top H x)] = \operatorname{tr}(H\Sigma_x)\Sigma_x + 2\Sigma_x H \Sigma_x$$

then gives

$$\mathbb{E}[g_\ell g_\ell^\top] = X_{\ell-1} S_\ell^\top Q S_\ell + 2\bar{g}_\ell \bar{g}_\ell^\top, \tag{B.1}$$

$$\mathbb{E}[g_\ell^\top g_\ell] = \operatorname{tr}(S_\ell^\top Q S_\ell) P_{\ell-1} \Sigma_x P_{\ell-1}^\top + 2\bar{g}_\ell^\top \bar{g}_\ell. \tag{B.2}$$

Hence, for arbitrary weights,

$$\begin{aligned} F_\ell &= X_{\ell-1} S_\ell^\top Q S_\ell - \operatorname{tr}(S_{\ell+1}^\top Q S_{\ell+1}) P_\ell \Sigma_x P_\ell^\top \\ &\quad + 2\big(\bar{g}_\ell \bar{g}_\ell^\top - \bar{g}_{\ell+1}^\top \bar{g}_{\ell+1}\big), \\ \Phi &= \sum_{\ell=1}^{L} \left[X_{\ell-1} \operatorname{tr}(S_\ell^\top Q S_\ell) + 2\|\bar{g}_\ell\|_F^2\right]. \end{aligned}$$

On $\mathcal{S}_q$ we have $R\Sigma_x R^\top = \Gamma_q$, $\bar{g}_\ell = 0$, and $S_\ell^\top \Gamma_q S_\ell = 0$ for $\ell < L$, so these reduce to (4).

B.4 WITHOUT LABEL NOISE

Here we derive the growth law (7), its asymptotic form (8), and the decay rates of the earlier-layer factors. We start from the flow already obtained in the main text,

$$\frac{d\beta}{d\tau} = -\frac{1}{4} \operatorname{grad}_{\mathcal{S}_q} \Phi(\beta). \tag{B.3}$$

Set $\Sigma_\varepsilon = 0$. On $\mathcal{S}_q$, (4) gives $Y_k = 0$ for $k < L$ and $Y_L = \operatorname{tr}\Gamma_q$, so

$$\Phi = \operatorname{tr}\Gamma_q \, X_{L-1} = \operatorname{tr}\Gamma_q \sum_{i \le q} \lambda_i \frac{s_i^2}{\beta_{Li}^2}.$$

Thus the modes decouple. Fix $i \le q$ and set

$$a_i = \lambda_i s_i^2 \operatorname{tr}\Gamma_q, \qquad n_i = (\beta_{1i}^{-1}, \dots, \beta_{Li}^{-1}).$$

Since n_i is normal to the constraint $\prod_\ell \beta_{\ell i} = s_i$, and

$$\frac{\partial \Phi}{\partial \beta_{\ell i}} = 0 \quad (\ell < L), \qquad \frac{\partial \Phi}{\partial \beta_{Li}} = -\frac{2a_i}{\beta_{Li}^3},$$

projecting the gradient onto the tangent space gives

$$\frac{d\beta_{Li}}{d\tau} = \frac{a_i}{2\beta_{Li}^3}\left(1 - \frac{\beta_{Li}^{-2}}{\|n_i\|^2}\right) = \frac{a_i}{2\beta_{Li}^3}\frac{\sum_{\ell<L}(\beta_{Li}/\beta_{\ell i})^2}{1+\sum_{\ell<L}(\beta_{Li}/\beta_{\ell i})^2},$$

which is (7). For $\ell < L$,

$$\frac{d\beta_{\ell i}}{d\tau} = -\frac{a_i}{2\beta_{Li}^4 \beta_{\ell i}\|n_i\|^2}, \qquad \frac{d}{d\tau}\beta_{\ell i}^2 = -\frac{a_i}{\beta_{Li}^4\|n_i\|^2},$$

independently of ℓ. Hence the squared differences between the earlier layers are constant. Moreover,

$$\frac{d}{d\tau}\beta_{Li}^4 = 2a_i \frac{\sum_{\ell<L}(\beta_{Li}/\beta_{\ell i})^2}{1+\sum_{\ell<L}(\beta_{Li}/\beta_{\ell i})^2}.$$

The earlier factors decrease while β_{Li} increases. If β_{Li} were bounded, then the fixed product and the conserved squared differences would keep all factors in a compact subset of $(0,\infty)$, on which the right-hand side of (7) is bounded below by a positive constant. Thus $\beta_{Li} \to \infty$. The fixed product then forces at least one earlier factor to zero, so the fraction above tends to one. Hence $d\beta_{Li}^4/d\tau \to 2a_i$, giving (8).

For the earlier layers, we now solve the balanced case exactly. Assume

$$\beta_{1i}(0) = \cdots = \beta_{L-1,i}(0).$$

Since their squared differences are constant, they remain balanced, so

$$\beta_{1i} = \cdots = \beta_{L-1,i} =: b_i.$$

Since the product along mode i is fixed,

$$b_i^{L-1}\beta_{Li} = s_i, \qquad \text{so} \qquad b_i = \left(\frac{s_i}{\beta_{Li}}\right)^{1/(L-1)}.$$

Using (8),

$$\beta_{\ell i} = b_i \sim s_i^{1/(L-1)}(2a_i\tau)^{-1/[4(L-1)]}, \qquad \ell < L,$$

which is (9) in the exactly balanced case. For small initialization, the earlier layers are approximately balanced when a mode is learned, so this balanced solution gives the approximation used in the main text.

Finally, the curvature quantity used in Section 4 is

$$\lambda_i s_i^2 \sum_{\ell=1}^{L} \beta_{\ell i}^{-2}.$$

From the projected flow,

$$\frac{d}{d\tau}\beta_{\ell i}^2 = -c_i \quad (\ell < L), \qquad c_i := \frac{a_i}{\beta_{Li}^4\|n_i\|^2} > 0,$$

and the fixed product gives $\frac{d}{d\tau}\beta_{Li}^2 = c_i\beta_{Li}^2\sum_{\ell<L}\beta_{\ell i}^{-2}$. Hence

$$\frac{d}{d\tau}\left(\lambda_i s_i^2 \sum_{\ell=1}^{L}\beta_{\ell i}^{-2}\right) = \lambda_i s_i^2 c_i \sum_{\ell<L}\beta_{\ell i}^{-2}\left(\beta_{\ell i}^{-2} - \beta_{Li}^{-2}\right).$$

For an equal-factor start this is zero initially and positive afterwards, because β_{Li} increases while the earlier factors decrease. It diverges as $\tau \to \infty$ because the earlier factors tend to zero.

B.5 WITH LABEL NOISE

Here we derive the fixed point (10). Since the flow is

$$\dot{\beta} = -\frac{1}{4}\,\mathrm{grad}_{\mathcal{S}_q}\,\Phi(\beta),$$

its fixed points are exactly the stationary points of Φ on $\mathcal{S}_q$. For $1 \le k < L$, write

$$p_{ki} = \prod_{a=1}^{k} \beta_{ai}^2, \qquad p_{0i} = 1, \qquad p_{Li} = s_i^2.$$

For $1 \le k < L$, the corresponding interior quantities on $\mathcal{S}_q$ are

$$X_k = \sum_{i \le q} \lambda_i p_{ki}, \qquad Y_k = \sum_{i \le q} \frac{\mu_i s_i^2}{p_{ki}}.$$

The boundary quantities remain those of the main text, $X_0 = \operatorname{tr}\Sigma_x$ and $Y_L = \operatorname{tr}(\Gamma_q + \Sigma_\varepsilon)$. At a fixed point, $F_\ell = 0$. Taking the ith diagonal entry of (4) gives

$$X_{\ell-1}\mu_i \frac{s_i^2}{p_{\ell i}} = Y_{\ell+1}\lambda_i p_{\ell i},$$

so

$$p_{\ell i} = s_i \sqrt{\frac{\mu_i}{\lambda_i}}\, t_\ell, \qquad t_\ell := \sqrt{\frac{X_{\ell-1}}{Y_{\ell+1}}}.$$

Therefore, with

$$c_q = \sum_{i \le q} s_i \sqrt{\lambda_i \mu_i},$$

we have $X_\ell = c_q t_\ell$ and $Y_\ell = c_q / t_\ell$. Also $t_0 = X_0/c_q$ and $t_L = c_q/Y_L$. Hence

$$t_\ell^2 = t_{\ell-1} t_{\ell+1},$$

so the ratios $t_\ell / t_{\ell-1}$ are constant. If this ratio is r, then

$$r^L = \frac{t_L}{t_0} = \frac{c_q^2}{X_0 Y_L}, \qquad r = \left(\frac{c_q^2}{X_0 Y_L}\right)^{1/L} = \rho^2.$$

Thus, for $1 < \ell < L$,

$$\beta_{\ell i}^2 = \frac{p_{\ell i}}{p_{\ell-1,i}} = r, \qquad \beta_{\ell i} = \rho,$$

and for the two end layers

$$\left(\frac{\beta_{Li}}{\beta_{1i}}\right)^2 = \frac{\lambda_i}{\mu_i}\frac{Y_L}{X_0}.$$

This is exactly (10): at the fixed point, the learned blocks of all intermediate layers are ρI_q. The calculation also shows that the fixed point is unique.

It remains only to justify convergence. From (4),

$$\Phi \ge \sum_{i \le q} s_i^2 \left(\frac{X_0 \mu_i}{\beta_{1i}^2} + \sum_{1 < \ell < L} \frac{\lambda_i \mu_i}{\beta_{\ell i}^2} + \frac{Y_L \lambda_i}{\beta_{Li}^2} \right).$$

Together with $\prod_\ell \beta_{\ell i} = s_i$, this makes the sublevel sets of Φ on $\mathcal{S}_q$ compact. Thus Φ attains a minimum; every minimizer is stationary, so the stationary point found above is also the unique minimizer. Since the flow is the gradient flow of Φ and has this unique stationary point, every trajectory converges to it. The same calculation applies when $q = r$, with $\Gamma_q = 0$.

C Proof of Proposition 4.1

We prove that, up to the exit time κ, the diagonal entries of SGD follow the flow on $\mathcal{S}_q$ with error $O(\eta + \delta_M + \delta_W^2)$ over $O(\eta^{-2})$ steps. Write $W_n = W^{(n)}$ for the SGD iterates.

For arbitrary weights, let D_ℓ be the matrix obtained from W_ℓ by retaining only the diagonal entries of its learned $q \times q$ block, and write $W_\ell = D_\ell + E_\ell$. Here the notation $D_\ell \succ 0$ in the main text means that these q retained diagonal entries are positive, including when W_ℓ is rectangular. Set

$$\widehat{\beta}_{\ell i} = (W_\ell)_{ii}, \qquad \delta = \delta_M + \delta_W^2,$$

and let $C(W)$ denote the array $((C_\ell(W))_{ii})_{\ell<L,\, i\le q}$. For each such array, let $\widetilde{W}(C) \in \mathcal{S}_q$ be the point with those imbalances and mode products s_i, and define

$$f(C) = \big((F_\ell(\widetilde{W}(C)))_{ii}\big)_{\ell<L,\, i\le q}.$$

Approximation near $\mathcal{S}_q$. We first show that the conditional drift of the SGD imbalances is $f(C(W)) + O(\delta)$.

A product contributing to the first q diagonal entries of M and containing exactly one E_ℓ vanishes. Hence, with $\widehat{m}_i = \prod_\ell \widehat{\beta}_{\ell i}$,

$$|\widehat{m}_i - s_i| \le \delta_M + C\delta_W^2.$$

Likewise,

$$\Big|(C_\ell(W))_{ii} - \widehat{\beta}_{\ell i}^2 + \widehat{\beta}_{\ell+1,i}^2\Big| \le C\delta_W^2.$$

Because $\|W_\ell\|_{\mathrm{op}} \le \Lambda$ and $\widehat{m}_i$ is close to $s_i > 0$, for sufficiently small δ_0,

$$\frac{s_i}{2\Lambda^{L-1}} \le \widehat{\beta}_{\ell i} \le \Lambda.$$

The positive factors on $\mathcal{S}_q$ depend smoothly on their imbalances and products, so the preceding two estimates give

$$\|\widehat{\beta}^2 - \widetilde{\beta}(C(W))^2\| \le C\delta. \tag{C.1}$$

Now let

$$d(W) = \big((F_\ell(W))_{ii}\big)_{\ell<L,\, i\le q}.$$

The first-order change of these entries in an off-block direction E vanishes: in (B.1)–(B.2), a single off-block factor cannot contribute to the first q diagonal entries or to the relevant traces. For the residual covariance, the linear perturbation of M has one learned–unlearned off-block index, whereas $A_q - A$ is supported on the remaining teacher coordinates. The resulting cross terms are therefore off-block and do not contribute to the learned diagonal entries or traces appearing in $d(W)$. Together with (C.1), this gives

$$\|d(W) - f(C(W))\| \le C\delta. \tag{C.2}$$

Comparison with the flow. Set $h = \eta^2$, $N = \lfloor T/h \rfloor$, and $C_n = C(W_n)$. For the sample used at step $n+1$, let

$$Z_{n+1,\ell i} = (g_\ell g_\ell^\top - g_{\ell+1}^\top g_{\ell+1})_{ii}.$$

Let $\mathcal{F}_n$ be the natural filtration generated by the samples through step n. Then $\mathbb{E}[Z_{n+1} \mid \mathcal{F}_n] = d(W_n)$, so

$$\mathcal{M}_n = h\sum_{t<n} \mathbf{1}_{\{t<\kappa\}}\big(Z_{t+1} - d(W_t)\big)$$

is a martingale. On $\mathcal{N}$ its increments have uniformly bounded second moments, and therefore Doob's inequality gives

$$\left(\mathbb{E}\max_{n\le N}\|\mathcal{M}_n\|^2\right)^{1/2} \le C_T\eta. \tag{C.3}$$

Let $\bar{C}(\tau) = C(\bar{W}(\tau))$. On the compact region reached before exit and by the flow on $[0, T]$, f is Lipschitz. Summing the exact imbalance update (1) and using (C.2) gives

$$C_n = C_0 + h \sum_{t<n} f(C_t) + \mathcal{M}_n + R_n, \qquad \|R_n\| \le C_T \delta.$$

The flow satisfies the corresponding Euler expansion

$$\bar{C}(nh) = C_0 + h \sum_{t<n} f(\bar{C}(th)) + \bar{R}_n, \qquad \|\bar{R}_n\| \le C_T h.$$

Discrete Gronwall and (C.3) therefore give

$$\left(\mathbb{E} \max_{\substack{n \le N \\ n < \kappa}} \|C_n - \bar{C}(nh)\|^2 \right)^{1/2} \le C_T(\eta + \delta). \tag{C.4}$$

Finally, smooth dependence of the positive factors on C and (C.1) imply

$$\|\widehat{\beta}(W_n)^2 - \bar{\beta}(nh)^2\| \le C_T\big(\delta + \|C_n - \bar{C}(nh)\|\big).$$

The uniform positive lower bound on the factors converts this to the same estimate for $\widehat{\beta} - \bar{\beta}$. Taking the maximal L^2 norm and using $h = \eta^2$ proves Proposition 4.1.